Enhanced Computer Vision Methods for Cancer Detection and Precision Guidance in Medical Imaging

Enhanced Computer Vision Methods for Cancer Detection and Precision Guidance in Medical Imaging

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr. S.K. Lenaerts, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op woensdag 18 december 2024 om 16:00 uur.

door

Christiaan Günter Alwyn Viviers

geboren te Pretoria, Zuid-Afrika

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. J.K.A.H.P.J. Kok
1e promotor:	prof.dr.ir. P.H.N. de With
copromotor(en):	dr.ir. F. van der Sommen
leden:	prof.dr. D.V. Stoyanov (University College London, UK)
	prof.dr. W.M. Prokop (Radboud Universiteit Nijmegen)
	prof.dr.ir. D. Ruijters
	prof.dr. M. Luyer (Catharina Ziekenhuis Eindhoven)
adviseur:	dr.ir. J.J. Hermans, MD (Radboud UMC Nijmegen)
	dr. L. Filatova (Philips Healthcare)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Enhanced Computer Vision Methods for Cancer Detection and Precision Guidance in Medical Imaging

Christiaan Günter Alwyn Viviers

Cover photo: ”In a fantastical sky, an ancient engineering zeppelin soars through fluffy, cotton-candy-like clouds that billow with soft, ethereal light. As it maneuvers precisely and skillfully to evade the thick, swirling clouds, the airship searches for hidden cancer cells in their crystalline forms shimmering against the tapestry of the cosmos.” by Christiaan Viviers with Flux AI, Adobe Firefly and Adobe Photoshop.

Cover design: Christiaan Viviers & Kathryn Moy
Printed by: ADC Nederland.

ISBN 978-90-386-6234-3
NUR-code 959

Summary

Enhanced Computer Vision Methods for Cancer Detection and Precision Guidance in Medical Imaging

Medical imaging is a cornerstone of modern healthcare, facilitating diagnosis, treatment planning, and patient management through visualization of the anatomy. Imaging modalities such as X-ray, CT, MRI, ultrasound and especially natural (RGB) images have transformed the field, offering detailed views that are essential for understanding a wide range of medical conditions. However, the interpretation of these images often relies heavily on the skill and expertise of clinicians, and diagnostic errors can arise from subtle anatomical variations or complex pathological conditions that are difficult to detect. The advent of advanced computer vision techniques, particularly deep learning-based image analysis methods, offers a promising solution to these challenges. The objective of this research is to develop and enhance computer vision techniques for various applications in medical imaging. Specifically, this work explores methods to support early cancer detection, improve segmentation accuracy and uncertainty quantification, detect abnormalities/anomalies across multiple imaging modalities, and provide reliable guidance during minimally invasive surgeries.

The detection and treatment of pancreatic ductal adenocarcinoma (PDAC), a highly aggressive and lethal cancer, have been significantly enhanced by the application of computer-aided detection (CADe) systems. Chapter 3 focuses on the development of a CADe system for PDAC, reviewing prior work in the field and introducing a new method that integrates secondary features into the detection process. Early detection of pancreatic cancer is difficult, since the pancreas has a complicated 3D shape and the early tumor does not always appear hypodense and clearly visible in CT. Clinical experts refer to secondary tumor-indicative features to enhance their understanding and detection performance. These secondary tumor-indicative features improve the CADe system's ability to detect pancreatic tumors when integrated along with the CT scan. The developed CADe system yields a high detection performance of 0.99 Area Under the Receiver Operating Characteristic (AUROC), thereby suggesting potential improvement in patient outcomes. The chapter also covers advanced segmentation techniques, enabling accurate identification of the pancreas and surrounding anatomical structures, forming a comprehensive framework for PDAC detection and localization.

Chapter 4 addresses the concept of uncertainty in medical image segmenta-

tion. In image analysis, accounting for uncertainty in segmentation is essential for improving the reliability of model predictions, especially given the significant interobserver and intraobserver variability, commonly observed among clinicians. This variability underscores the inherent uncertainty in interpreting medical images, making robust uncertainty quantification a critical factor. The research develops methods to quantify aleatoric uncertainty in both 2D and 3D medical image segmentation tasks. A significant contribution in this chapter is the incorporation of Normalizing Flows (NFs) into the Probabilistic U-Net (PU-Net), which allows for more flexible posterior distributions. This enhancement improves the model's ability to handle uncertainty in ambiguous images (14% improvement in the GED metric).

Chapter 5 explores a concrete application of the PU-Net, which further builds on both the preceding chapters by developing a method for predicting the resectability of PDAC, a critical determinant in formulating appropriate treatment strategies. The research uses segmentation models as a proxy for involvement prediction in two steps. First, the models accurately delineate both the tumor (0.66 DSC) and vasculature (veins 0.88 DSC, arteries 0.86 DSC). Second, any involvement of the tumor from the computed segmentation maps is computed (high sensitivity of 0.88 and specificity of 0.86) and then a degree of involvement is calculated (0.89 accuracy with 3D PU-Net). By integrating uncertainty into these models, the framework provides clinicians with more reliable resectability assessments, offering critical support in surgical planning. These advancements have significant implications for improving surgical outcomes in pancreatic cancer patients, particularly those eligible for resection.

Chapter 6 investigates the use of generative models to detect both semantic and covariate Out-of-Distribution (OOD) data. OOD involves identifying inputs that deviate from the model's training data, signaling unfamiliar or novel scenarios. A novel semantic OOD detection methodology utilizing wavelet-based Normalizing Flows is introduced. It is shown that the modeled Haar-wavelet components are effective at distinguishing between the semantic difference in the abundant benign melanoma and previously unseen malignant melanoma (0.79 AUROC), while utilizing a small 1.25-Million parameter model. Additionally, the chapter further explores the detection of OOD covariate shifts, offering a comprehensive strategy for identifying distributional anomalies for avoiding erroneous predictions and analyses. CovariateFlow is introduced as a novel method that models the heteroscedastic high-frequency image components, thereby improving the ability to detect covariate shift. A detection score of 0.75 AUROC on CIFAR10(-C), 0.72 AUROC on ImageNet200-(C) and 0.93 AUROC in the X-ray setting is obtained. These methods are instrumental in enhancing the safety and reliability of data-driven diagnostic tools by detecting corrupted or unfamiliar data samples, particularly in clinical environments where distribution shifts can lead to diagnostic errors.

Finally, Chapter 7 introduces the development of a pose estimation technique for use in image-guided surgeries, specifically those involving X-ray imaging. The

research presents a general-purpose 6-degrees-of-freedom (6-DoF) pose estimation model that can account for the variability in X-ray imaging geometries, improving the accuracy and real-time performance of surgical guidance systems. The effectiveness of the proposed YOLOv5-6D model as a general-purpose approach for 6-DoF object/instrument pose estimation is tested on the public LINEMOD benchmark, obtaining 96.84% ADD(-S) at 42 FPS. In the X-ray domain, the same model shows efficacy in three settings: (1) test Cube at 99.27% ADD(-S), (2) a cannulated cancellous spinal screw at 96.87% ADD(-S) and (3) the screws in a human-spine phantom at 92.41% ADD(-S). This work addresses a critical challenge in minimally invasive surgeries, such as precise instrument positioning during spinal surgeries. This work enhances the state-of-the-art precision and safety of these procedures.

The research presented in this thesis makes substantial contributions to the advancement of computer-aided detection, image-guided surgery, and medical image segmentation, offering novel solutions to critical challenges in medical imaging. A major outcome is the development of enhanced CADe systems for pancreatic cancer detection, incorporating secondary tumor-indicative features that improve diagnostic accuracy and enable earlier detection, which is pivotal for better patient outcomes. This thesis also introduces innovative improvements for uncertainty quantification, addressing the inherent variability and ambiguity in medical images, for the purpose of improving the robustness of segmentation models used in clinical practice. Additionally, key breakthroughs in Out-of-Distribution (OOD) detection are presented, offering a powerful approach to ensure the reliability of medical image analysis, particularly in the presence of data that falls outside the model's training distribution, a frequent issue in clinical environments. The work on 6-DoF pose estimation further strengthens the thesis by advancing real-time guidance systems in X-ray towards higher accuracy, offering practical applications in minimally invasive surgeries where precision and safety are paramount. These contributions are poised to accelerate the adoption of data-driven tools in healthcare, leading to more precise, reliable, and efficient medical diagnostics and interventions.

Samenvatting

Verbeterde computervisiemethodes voor kankerdetectie en precisiegeleide instrumenten in medische beeldvorming

Medische beeldvorming is een hoeksteen van de moderne gezondheidszorg en ondersteunt de diagnose, het behandelplan en het patiëntmanagement door middel van visualisatie van de anatomie. Beeldvormingstechnieken zoals röntgen, CT, MRI, echografie en vooral natuurlijke (RGB) beelden hebben het werkgebied veranderd door gedetailleerde visuele representaties te bieden die essentieel zijn voor het begrijpen van een breed scala aan medische aandoeningen. De interpretatie van de gebruikte beelden hangt echter sterk af van de vaardigheden en expertise van clinici, waarbij diagnostische fouten kunnen voortkomen uit subtiele anatomische variaties of complexe pathologische condities die moeilijk te detecteren zijn. De opkomst van geavanceerde technieken in computervisie, met name diep lerende (*deep learning*) beeldanalyse, biedt een veelbelovende oplossing voor deze uitdagingen. Het doel van dit onderzoek is om computervisietechnieken te ontwikkelen en te verbeteren voor diverse toepassingen bij medische beeldvorming. Het onderzoek in deze dissertatie richt zich specifiek op methoden ter ondersteuning van vroege kankerdetectie, betere nauwkeurigheid van segmentatie en kwantificering van onzekerheid, detectie van afwijkingen/anomalieën in verschillende beeldvormingstechnieken, en betrouwbare instrumentbegeleiding tijdens minimaal invasieve operaties.

De detectie en behandeling van een ductaal adenocarcinoom in/bij de alvleesklier (PDAC), een zeer agressieve en dodelijke vorm van kanker, zijn aanzienlijk verbeterd door de toepassing van computerondersteunde detectiesystemen (CADe). Hoofdstuk 3 behandelt de ontwikkeling van een CADe-systeem voor PDAC, waarin eerder werk op dit gebied wordt besproken en een nieuwe methode wordt geïntroduceerd die secundaire kenmerken integreert in het detectieproces. Vroege detectie van alvleesklierkanker is uitdagend, aangezien de alvleesklier een complexe driedimensionale structuur heeft en vroege tumoren niet altijd hypodens en duidelijk zichtbaar zijn in CT-beelden. Daarom gebruiken klinische experts secundaire indicatieve tumorkenmerken om hun begrip en detectieprestaties te verbeteren. Deze secundaire tumorkenmerken versterken het vermogen van het CADe-systeem om alvleeskliertumoren te detecteren wanneer ze worden geïntegreerd met de CT-scan. Het ontwikkelde CADe-systeem behaalt een hoge detectieprestatie van 0.99 *Area Under the Receiver Operating Characteristic* (AUROC), wat suggereert dat het de patiëntuitkomsten kan verbeteren. Het

hoofdstuk behandelt ook geavanceerde segmentatietechnieken, die een nauwkeurige identificatie van de alvleesklier en omliggende anatomische structuren mogelijk maken. Dit biedt een goed en uitgebreid raamwerk voor succesvolle detectie en lokalisatie van PDAC.

Hoofdstuk 4 beschrijft het concept van onzekerheid in medische beeldsegmentatie. Bij beeldanalyse is het van cruciaal belang om rekening te houden met onzekerheid in segmentatie voor het verbeteren van de betrouwbaarheid van modelvoorspellingen, vooral gezien de significante variabiliteit bij individuele waarnemers en tussen waarnemers onderling, hetgeen vaak gebeurt bij clinici. Deze variabiliteit benadrukt de inherente veranderlijkheid en onzekerheid bij het interpreteren van medische beelden, wat robuuste kwantificering van onzekerheid tot een kritieke factor maakt. Het onderzoek ontwikkelt methoden om aleatorische onzekerheid te kwantificeren voor zowel 2D- als 3D-segmentatietaken. Een belangrijke bijdrage in dit hoofdstuk is de integratie van zogenaamde *Normalizing Flows* (NFs) ofwel een statistische modellering in het probabilistische U-Net model (PU-Net), waarmee meer flexibiliteit wordt geboden bij gebruik van *a-posteriori* kansverdelingen. Deze verbetering verhoogt het vermogen van het model om met onzekerheid om te gaan in ambigue beelden (14% verbetering in de GED-metriek).

Hoofdstuk 5 verkent een concrete toepassing van het PU-Net model, die voortbouwt op de voorgaande hoofdstukken door een methode te ontwikkelen voor het voorspellen van een mogelijke operatieve verwijdering (reseceerbaarheid) van het PDAC, een kritieke factor bij het formuleren van geschikte behandelstrategieën. Het onderzoek maakt gebruik van segmentatiemodellen als autorisatiemethode voor de betrokkenheidsvoorspelling in twee stappen. Eerst vindt een nauwkeurige modelsegmentatie plaats van zowel de tumor (0,66 DSC) als de bloedvaten (aders 0,88 DSC, slagaders 0,86 DSC). Vervolgens wordt elke betrokkenheid van de tumor op basis van de berekende segmentatiegebieden bepaald (hoge sensitiviteit van 0,88 en specificiteit van 0,86), en wordt de mate van betrokkenheid berekend (0,89 nauwkeurigheid met het 3D PU-Net model). Door onzekerheid in deze modellen te integreren, biedt het gehele raamwerk aan clinici een betere betrouwbaarheid bij de beoordeling van reseceerbaarheid, wat essentiële ondersteuning biedt bij chirurgische planning. Deze vooruitgang heeft positieve implicaties voor alvleesklierkankerpatiënten die in aanmerking komen voor een operatieve alvleesklierverwijdering.

Hoofdstuk 6 onderzoekt het gebruik van generatieve modellen om zowel semantische als covariate *Out-of-Distribution* (OOD)-data te detecteren. OOD omvat het identificeren van inputbeelden die afwijken van de trainingsdata van het model, wat duidt op onbekende of nieuwe scenario's. Een nieuwe semantische OOD-detectiemethode wordt geïntroduceerd, die gebruik maakt van *wavelet*-gebaseerde *Normalizing Flows* distributies. Het wordt aangetoond dat de gemodelleerde *Haar-wavelet* componenten effectief zijn in het onderscheiden van het semantische verschil tussen de overvloedig aanwezige goedaardige melanomen en voorheen onbekende sporadische kwaadaardige melanomen (0,79 AUROC), terwijl bovendien een klein model met slechts 1,25 miljoen parameters wordt

gebruikt. In het hoofdstuk gaat het onderzoek verder in op de detectie van OOD *covariate shifts* en biedt een uitgebreide strategie voor het identificeren van afwijkingen in distributies om foutieve voorspellingen en analyses te voorkomen. Een CovariateFlow model wordt geïntroduceerd als een nieuwe methode die de heteroscedastische hoogfrequente beeldcomponenten modelleert en zo het vermogen verbetert om covariate distributieverschuivingen te detecteren. Een detectiescore van 0,75 AUROC op CIFAR10(-C), 0,72 AUROC op ImageNet200(-C) en 0,93 AUROC met röntgenbeelddata wordt behaald. Deze methoden zijn van essentieel belang voor het verbeteren van de veiligheid en betrouwbaarheid van datageleide diagnostische hulpmiddelen door gecorrumpeerde of onbekende datasamples te detecteren, vooral in klinische omgevingen waar datadistributieverschuivingen kunnen leiden tot diagnostische fouten.

Tot slot introduceert hoofdstuk 7 de ontwikkeling van een pose-schattingstechniek voor instrumentgebruik bij beeldgeleide operaties, met name geschikt voor röntgenbeelden. Het onderzoek presenteert een algemeen toepasbaar model voor pose-schatting met zes vrijheidsgraden (6-DoF), dat rekening houdt met de variabiliteit in geometrie bij röntgenbeelden en dat de nauwkeurigheid en realtime prestaties van chirurgische begeleidingssystemen verbetert. De effectiviteit van het voorgestelde YOLOv5-6D-model als algemene aanpak voor 6-DoF-object-/instrument-poseschatting wordt getest op de openbare LINEMOD-dataset, dat resulteert in 96,84% ADD(-S) bij 42 FPS (frames/seconde). In de röntgenomgeving toont hetzelfde model zijn effectiviteit onder drie omstandigheden: (1) testset Cube met 99,27% ADD(-S), (2) een gecannuleerde spinale schroef met 96,87% ADD(-S) en (3) de schroeven in een menselijk ruggenmergfantoom met 92,41% ADD(-S). Dit werk richt zich op een kritieke uitdaging optredend bij minimaal invasieve operaties, zoals de precieze plaatsing van instrumenten tijdens rugoperaties. Hiermee wordt de algemeen haalbare precisie en veiligheid van dergelijke procedures verbeterd.

Het onderzoek in dit proefschrift levert aanzienlijke bijdragen aan de vooruitgang in computerondersteunde detectie, beeldgeleide chirurgie en medische beeldsegmentatie, en biedt nieuwe oplossingen voor kritieke uitdagingen in de medische beeldvorming. Een belangrijk resultaat is de ontwikkeling van een verbeterd CADe-systeem voor de detectie van alvleesklierkanker, met gebruikmaking van secundaire tumorkenmerken die de diagnostische nauwkeurigheid verbeteren, vroege detectie mogelijk maken wat cruciaal is voor betere patiëntuitkomsten. Dit proefschrift introduceert ook innovatieve verbeteringen voor kwantificering van onzekerheid waarmee de inherente variabiliteit en ambiguïteit in medische beelden worden aangepakt om daarmee de robuustheid van segmentatiemodellen in de klinische praktijk te verbeteren. Bovendien worden belangrijke doorbraken in *Out-of-Distribution* (OOD)-detectie gepresenteerd, met krachtige methoden om de betrouwbaarheid van medische beeldanalyse te waarborgen, vooral bij datagebruik met statistiek buiten de trainingsdistributie van het model, een veelvoorkomend probleem in klinische omgevingen. Het werk aan 6-DoF poseschatting versterkt het proefschrift verder door realtime begeleidingssystemen

voor instrumenten in röntgenbeelden naar een hoger nauwkeurigheidsniveau te brengen, wat praktische toepassingen biedt in minimaal invasieve operaties waar precisie en veiligheid essentieel zijn. Deze bijdragen versnellen de acceptatie van datageleide hulpmiddelen in de gezondheidszorg, wat leidt tot meer precieze, betrouwbare en efficiënte medische diagnostiek en interventies.

Contents

Chapter 1 Introduction

1.1 Background

1.1.1 Deep learning in medical image analysis

The rapid advancement of machine learning (ML), particularly Deep Learning (DL), has catalyzed transformative changes across numerous fields, with medical imaging standing at the forefront of these innovations. Deep learning, especially through architectures such as convolutional neural networks (CNNs), has unlocked unprecedented possibilities for automated image analysis, dramatically improving the accuracy, speed, and consistency of medical diagnostic processes. These developments holds an immense promise for medical imaging applications, where modalities like X-ray, computed tomography (CT), Ultrasound and especially applications utilizing natural images (RGB) are poised to benefit from the enhanced precision and efficiency enabled by DL-driven solutions.

Deep learning models excel in capturing complex patterns, making them highly effective for the intricate and often subtle features present in medical images. While medical imaging datasets are relatively limited compared to those in other domains, advancements in data augmentation, transfer learning, and model optimization have enabled DL models to perform exceptionally well, even with smaller, highly specialized datasets. These models have been successfully applied to critical tasks, such as detecting abnormalities, segmenting anatomical structures, and predicting clinical outcomes from medical scans. In radiology, for instance, CNNs trained on curated and annotated datasets have shown strong capabilities in detecting pulmonary nodules on chest X-rays, identifying tumors in MRI scans, and assessing the severity of conditions on Computed Tomography (CT) scans with accuracy levels comparable to expert radiologists [1]. Recent DL-advancements in RGB-based endoscopy video analysis for early Barrett's neoplasia detection, the Computer-Aided Detection (CADe) system showcased superior performance to endoscopists in detecting neoplasia [2].

In computed tomography (CT) and magnetic resonance imaging (MRI), deep learning models have played a crucial role in advancing image analysis beyond the capabilities of conventional image analysis methods. These models have been instrumental in enhancing tasks such as image segmentation, tumor detection, and the accurate identification of pathological structures. By automating these complex processes, deep learning algorithms significantly reduce the time required

for image interpretation, thus improving the efficiency of clinical workflows and can even bring diagnostic expertise to locations previously not available. Moreover, their ability to provide detailed anatomical insights potentially facilitating more precise planning of surgeries and therapies, contributing to better patient outcomes [3].

X-ray imaging remains a fundamental tool in clinical diagnostics due to its speed, accessibility, and relatively low cost, making it indispensable for detecting a wide range of conditions, from bone fractures and lung infections to cardiovascular abnormalities. Despite its ubiquity, interpreting X-ray images is inherently complex, requiring considerable expertise to distinguish subtle anomalies from normal variations. Deep learning techniques have revolutionized this process, enabling the automation of key diagnostic tasks. Beyond diagnostics, X-ray imaging is increasingly employed in minimally invasive surgeries, where real-time imaging is critical for procedures such as catheter, cardiac valve and stent placement. Deep learning further enhances these applications by automating the identification of anatomical landmarks and providing precise guidance during interventions, reducing both the need for more invasive approaches and recovery times [4]. The continued integration of deep learning in X-ray imaging, from diagnostics to interventional radiology, is poised to further elevate the quality and efficiency of patient care.

1.1.2 Advancements of deep learning for enhanced image analysis

The field of deep learning for image analysis has seen remarkable advancements since the introduction of AlexNet [5] in 2012, which was pivotal in demonstrating the power of deep convolutional networks by winning the ImageNet [6] challenge. AlexNet's success marked a shift from traditional handcrafted features to data-driven feature learning, combined with advancements in parallelizing computations with modern[1] Graphics Processing Unit (GPU) for training, accelerated the growth of the computing technology. Subsequent architectures such as the ResNet [7] model, with innovations like skip connections, introduced scaling laws which indicate that deeper and larger models equally deliver improved performance. The evolution continued with vision transformers (ViT) [8], culminating in models like DINOv2[9], which leverage attention mechanisms for enhanced feature learning over global image contexts, offering superior performance in self-supervised learning [10]. Recently, architectures such as SAMv2 [11] have emerged, focusing on open-set segmentation tasks, blending the strengths of transformers with convolutional backbones, enabling more robust handling of diverse and unseen image data. These developments have not only pushed the boundaries of accuracy and efficiency but also expanded the applicability of deep learning to more complex and open-ended tasks in image analysis.

Initially, deep learning models were developed primarily for natural image analysis tasks, such as object recognition and classification. However, these archi-

[1] Nvidia Corp., CA, USA

tectures have been successfully adapted for more specialized domains, including medical imaging. For instance, U-Net [12], originally designed for biomedical image segmentation, has been widely adopted across both medical and non-medical fields, becoming a staple in tasks such as tumor segmentation [13] and general semantic segmentation. This adaptability is further enhanced through techniques like transfer learning [14], where models pretrained on large-scale datasets like ImageNet [6] are fine-tuned to detect specific patterns in medical images. Notably, innovations tailored to medical applications often flow back into general image analysis, driving new capabilities forward, with examples such as the earlier mentioned U-Net being the backbone in state-of-the-art Diffusion generative models [15]. This cross-domain influence highlights the reciprocal relationship between medical and general image analysis, where advancements in one domain catalyze progress in the other.

Despite the rapid advancements of DL in medical imaging, several critical challenges should be addressed for these technologies to be fully integrated into clinical practice. For CADx and CADe (CAD) systems to gain widespread adoption, they need to align more closely with clinical workflows or even even be integrated into existing medical equipment. This requires not only enhancing the accuracy of these models, but more critically, ensuring their robustness across diverse clinical environments and varying user expertise. Medical images often present ambiguities due to low-contrast changes between anatomical structures, variable/poor image quality, or rare pathologies, making accurate interpretation difficult. DL models should be able to handle (part of) these complex cases, particularly in high-stakes applications like tumor detection or trauma assessment, where errors can have significant consequences. Therefore, continued advancements in model architectures, training techniques, and more comprehensive datasets are essential to ensure deep learning systems that can meet or exceed the performance of human experts in all clinical scenarios.

In addition to diagnostic accuracy, there is a need for novel deep learning solutions that extend beyond diagnosis. Predicting clinical outcomes, such as patient prognosis or treatment response, remains an underexplored area with substantial potential for positive impact. Similarly, DL could play a crucial role in real-time decision-making support during surgeries, providing guidance to surgeons by identifying critical structures and aiding in instrument maneuvering. However, the reliability of DL models remains a major hurdle, especially when confronted with Out-of-Distribution (OOD) data, such as rare diseases or unexpected variations that differ from the data used in training. To ensure trustworthiness in clinical settings, models should be developed with robust mechanisms to detect and handle OOD scenarios, and safeguard against adversarial errors. Addressing these challenges will be crucial for the broader adoption of deep learning in clinical practice, paving the way for more reliable medical applications.

1.2 Computer-aided detection in medical imaging

Computer-aided detection (CADe) systems in medical imaging represent an evolving field that integrates advanced computational techniques with image analysis to enhance diagnostic processes across a variety of medical specialties. CAD systems are designed to assist healthcare professionals by increasing the accuracy and efficiency of diagnoses through automated detection and characterization of visual/anatomical abnormalities and diseases in medical images.

One foundational study by Giger and Suzuki [16] outlines the significant role CAD systems play in providing a "second opinion" in diagnostic imaging, enhancing the ability of medical professionals to detect subtle or overlooked features in images. Roth *et al*. [17] further discuss improvements in CAD through convolutional neural networks (CNNs), which provide high sensitivity in detecting anomalies (e.g. cancer) across various imaging modalities, such as CT scans and MRIs. Their study highlights the application of deep learning techniques to enhance detection accuracy and reduce false positives. Moreover, a comprehensive review by Chen *et al*. [18] and Li *et al*. [19] detail various Computer-Aided Diagnosis (CADx) methods that extend beyond detection to include the probabilistic assessment and classification of detected lesions, thereby aiding in therapeutic decisions. Their study underscores the integration of pattern recognition and ML as pivotal in evolving CAD systems.

In clinical settings, CAD systems have shown substantial benefits in areas such as mammography[20], lung nodule detection [21], early Barrett's esophagus [2] and the identification of pancreatic cancer [22]. The potential of CAD systems continues to expand with advancements in machine learning, specifically deep learning. Nagaraj *et al*. [23] predict a shift towards fully automated diagnostic systems that could operate with minimal human oversight. The integration of such deep learning techniques suggests enhanced capabilities of CAD systems, making them more robust in handling a broad range of medical imaging tasks. The evolution of CAD systems has also been marked by the increasing use of hybrid models that combine multiple diagnostic parameters and imaging modalities, leading to more comprehensive and accurate systems. The development of such systems is detailed in the works of Shin *et al*. [24], who have explored the integration of various deep learning models to optimize the performance of CAD systems across different stages of disease detection and evaluation.

1.2.1 Pancreatic cancer

1.2.1.A Background on the pancreas

The pancreas is a vital organ located in the abdomen, playing a crucial role in the digestion of foods and regulation of blood sugar levels. It is approximately 15 cm long and is located across the back of the abdomen, behind the stomach. The pancreas is divided into three sections: the head, the body, and the tail. It produces digestive enzymes that help break down fats, proteins, and carbohydrates, and it secretes hormones such as insulin and glucagon into the bloodstream to help the

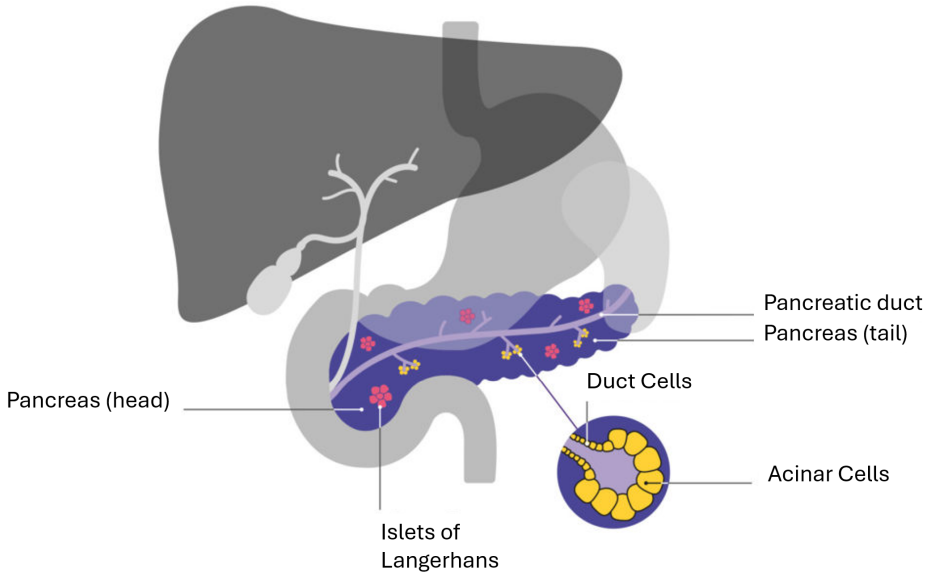


Figure 1.1 *Diagram showing the anatomy of the pancreas. Diagram obtained from https://www.pancreaticcancer.org.uk/*

body use or store the glucose it derives from food.

Pancreatic cancer is a challenging and aggressive disease and arises when abnormal cells within the pancreas grow uncontrollably and form tumors. Unfortunately, this type of cancer is often diagnosed at later stages, since early-stage pancreatic cancer rarely causes symptoms. This late diagnosis confirms its reputation as one of the most lethal types of cancer. Among the various forms of pancreatic cancer, Pancreatic Ductal Adenocarcinoma (PDAC) is the most common type, accounting for about 90% of the cases. This malignancy originates in the lining of the pancreatic ducts through which digestive enzymes flow.

The risk factors for developing pancreatic cancer include age (most patients are over 45), smoking, chronic pancreatitis, diabetes, family history of the disease, and certain genetic disorders. Treatment options vary depending on the stage of the disease and may include surgery, chemotherapy, radiation therapy, or a combination of these treatments aimed at managing symptoms and prolonging life.

1.2.1.B Pancreatic ductal adenocarcinoma (PDAC)

Advanced machine learning methodologies, specifically deep CNNs, have demonstrated exceptional capabilities in processing vast datasets of medical images. These networks are adept at detecting small details and even sometimes indiscernible alterations in pancreatic tissues. This enhanced detection capability is pivotal in the early identification and management of PDAC.

Pancreatic ductal adenocarcinoma (PDAC) is one of the leading causes of

cancer related deaths worldwide, with a dismal prognosis and an overall 5-year survival rate of only 9% [25], [26]. Although pancreatic cancer treatment has improved over the past years through centralization and optimization of treatment strategies, overall survival has not significantly improved [27], [28]. Pancreatic cancer often causes only a few non-specific symptoms before it develops into advanced stages of disease. The most commonly presented symptoms in patients with PDAC are pain, jaundice, steatorrhea, and weight loss [29]. As a result, more than 75% of the patients present with irresectable or metastatic disease [30], [31]. Therefore, early detection of pancreatic tumors holds significant promise by enabling potential curative treatment [32]. Subsequently, characterization of pancreatic tumors is important in order to tailor specific treatments, determine surgical resectability, and identify each patient for curative treatment as well as possible.

Radiological imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) are key in providing information on the presence of disease and the relation to vessels surrounding the pancreas, which determines the resectability [33]. Standardized resectability criteria are used to tailor the need for neoadjuvant therapy and to select patients for a minimally invasive approach [34]–[36]. However, determining resectability, especially after neoadjuvant therapy, is extremely difficult and mostly inaccurate at this time [37], [38]. Tumor regression after neoadjuvant treatment is rarely visible on CT, and the extent of vascular involvement tends to be overestimated [39]. Artificial intelligence offers a unique opportunity to improve the early detection and characterization of pancreatic cancer. Over the past decades, deep learning-based algorithms have been developed that can provide pixel-level segmentation maps of relevant anatomy [40], [41].

1.2.2 Guidance during minimally invasive surgeries

Minimally invasive surgeries (MIS) have transformed the landscape of surgical procedures by reducing patient recovery time, minimizing surgical trauma, and improving overall outcomes. However, these procedures demand a high level of precision, often requiring real-time guidance to navigate complex anatomical structures and avoid critical areas. Image-guided intervention and image-guided surgery can both particularly from image-based deep learning methods, where the latter has emerged as a powerful tool in enhancing the accuracy and safety of these surgeries. This section explores the recent advancements in image-based deep learning methods that support clinicians during minimally invasive procedures in radiology.

1.2.2.A Real-time surgical navigation

In the context of minimally invasive surgeries, image-based deep learning models are increasingly being used to enhance surgical navigation[42]. These models can process intraoperative imaging data, such as ultrasound, endoscopic or fluoroscopy images in real-time, providing surgeons with precise guidance of the instrument utilized during the surgery.

Initially, CNNs have been effectively utilized to segment organs, tumors, and

other critical structures in real-time, aiding surgeons in navigating through these regions with precision. Shvets *et al*. [43] demonstrated the use of CNNs for real-time instrument segmentation in endoscopic images, thereby improving the surgeon's ability to identify and track surgical tools during the procedure. This capability reduces the risk of accidental damage to tissues and enhances the surgeon's situational awareness, thereby improving surgical outcomes.

1.2.2.B Augmented reality and deep learning: enhancing surgical precision

One of the most promising applications of deep learning in MIS is the integration of Augmented Reality (AR) with real-time imaging data [42]. AR, powered by deep learning algorithms, overlays critical information, such as anatomical landmarks, blood vessels and nerves, potential risk zones or surgical tool guidance directly onto the surgeon's view [44]. This technology provides an intuitive, visual guide that enhances the surgeon's ability to perform complex tasks with higher precision.

For instance, deep learning models can predict the 3D structure of organs from 2D imaging data and project this information onto the surgical field. A recent study by Ramalhinho *et al*. [45] explored the use of deep learning for AR in liver surgeries, where the models provided real-time guidance by superimposing the liver's vascular structure onto the surgeon's view. This approach not only enhances the accuracy of the surgery, but also reduces the cognitive load on the surgeon by providing clear and continuous visual guidance.

A recent study by Malhotra *et al*. [46] reviews the application of augmented reality (AR) in surgical navigation, emphasizing its integration with deep learning technologies across various surgical fields such as neurosurgery, orthopedic surgery, and laparoscopic surgery. The findings highlight that AR-based systems enhance the precision and safety of surgical procedures by providing real-time 3D visualization and guidance, which are crucial for minimally invasive surgeries. However, the study notes that despite significant progress, the widespread adoption of AR in surgery encounters challenges, particularly in accuracy and validation, necessitating further research and development to fully realize its potential.

1.3 Modeling image data distributions

The deep learning-based methods utilized in CAD application, have proven highly effective at extracting complex patterns from image data by learning representations that either implicitly or explicitly capture the underlying distribution of the data they are trained on. Through training on large datasets, these models are optimized to generalize to new data and unseen test data. Implicitly, models like CNNs encode the statistical properties of the data in their learned weights, whereas explicit approaches, such as generative models, aim to model the full distribution of the input space. This capability to encode the intricacies of the data distribution is a key reason for the success of deep learning in fields like medical

image analysis, where understanding subtle patterns in images is crucial for tasks such as diagnosis or segmentation.

One of the significant advantages of learning data distributions is the ability to leverage these representations for Out-of-Distribution (OOD) detection and uncertainty modeling. By understanding what constitutes the normal data distribution, a model can be more effective at identifying inputs that deviate from this distribution, signaling potential anomalies or unexpected inputs, which is critical in medical imaging scenarios. Furthermore, uncertainty modeling can quantify the model's confidence in its predictions, providing an additional layer of reliability, especially when dealing with edge cases or rare conditions. This becomes invaluable in healthcare, where accurate and reliable decision-making can impact patient outcomes.

Deep generative models such as Variational Autoencoders (VAEs) [47], Generative Adversarial Networks (GANs) [48], Diffusion Models [15] and Normalizing Flows (NFs) [49] have revolutionized the ability to model image data distributions. These methods can capture the underlying distribution of complex image data by learning to generate images that are indistinguishable from real data. VAEs, for instance, encode images into a latent space that represents the data distribution and can be used to generate new, realistic images by sampling from this space. GANs, on the other hand, utilize a generator and a discriminator to learn and model data distributions through adversarial training.

In medical imaging, these models have been used to enhance tasks such as image reconstruction, denoising, and even generating synthetic data for training purposes. A well-modeled distribution can capture the nuances of medical imaging data, which is often characterized by high variability and subtle features that are crucial for accurate diagnosis. For instance, Chen *et al*. [50] demonstrated how GANs could model the distribution of MRI images for tasks such as image enhancement and artifact removal, thereby improving the quality and accuracy of subsequent analyses.

Beyond modeling the data distribution, these models can be integrated with task-specific models, such as segmentation networks, to express uncertainty. In radiological applications, understanding the uncertainty of a model's predictions is critical, especially in scenarios where the consequences of incorrect predictions can be severe. By leveraging probabilistic deep learning approaches, such as Bayesian neural networks or ensemble methods, it is possible to quantify uncertainty in segmentation tasks. For instance, a model trained to segment tumors in radiological images can use the underlying data distribution to identify regions where the model's predictions are uncertain. This is particularly useful in identifying borderline cases where the model is unsure if a region is cancerous or not, allowing radiologists to review these areas with greater scrutiny. Early work by Kendall and Gal [51] explored how Bayesian deep learning models could provide uncertainty estimates in segmentation tasks, leading to more robust and interpretable models in medical imaging.

Modeling the complete data distribution also offers a powerful mechanism

for out-of-distribution (OOD) detection. In radiology, OOD detection is crucial for identifying cases where the input data does not conform to the distribution the model was trained on, which may indicate an unusual or previously unseen pathology. By evaluating whether a new sample comes from the learned distribution, it is possible to flag potentially anomalous cases for further investigation. For example, in a model trained on a specific type of imaging modality, an OOD detector can identify when an image is from a different modality or contains features that are not well-represented in the training data. This capability is particularly important in clinical settings, where reliance on a model's output without recognizing its limitations could lead to diagnostic errors. The work by Hendrycks and Gimpel [52] on simple OOD detection methods using softmax scores, and more advanced methods like deep generative models, illustrates how these techniques can be effectively employed in medical imaging to enhance the safety and reliability of Artificial Intelligence (AI) systems in analysis methods.

1.4 Challenges

In the previous sections, we briefly explored the significant advancements that deep learning has brought to medical image analysis, particularly in enhancing the capabilities of computer-aided diagnosis (CAD) systems. These innovations have led to more accurate, efficient, and automated diagnostic tools, improving patient outcomes and streamlining clinical workflows. However, despite these achievements, several challenges persist in fully realizing the potential of deep learning in medical imaging, but also enable novel treatment strategies featuring enhanced background information and insights. This section will discuss some of these key challenges, outlining the technical aspects to be addressed in this thesis.

1.4.1 Pancreatic cancer treatment

1.4.1.A Accurate PDAC detection

Deep learning has emerged as a powerful tool in the detection and treatment of pancreatic cancer, particularly in identifying Pancreatic Ductal Adenocarcinoma (PDAC). However, the effectiveness of deep learning models heavily depends on the availability of large, annotated datasets. These models excel in detecting fine-grained details within medical images, but their performance can be significantly hampered in data-scarce environments.

Pancreatic cancer exhibits secondary, tumor-indicative features that can indicate tumor well before it is clearly visible in CT volumes. In this thesis, the specific challenge is to incorporate such domain-specific knowledge of PDAC into these models to enhance their detection capabilities by allowing the algorithms to focus on these clinically relevant features that may not be apparent in a generic training dataset. This feature integration can lead to more accurate and earlier detection, ultimately improving patient outcomes.

1.4.1.B Segmentation uncertainty quantification

A critical challenge that remains unresolved is the inherent ambiguity in segmenting tumors, i.e. defining tumor borders precisely. This ambiguity originates from the inter- and intra-observer variability of the clinicians' uncertainty when segmenting the tumor, which is partly introduced by the limited resolution of the CT imaging process and the low contrast between the tumor and surrounding anatomy.

To describe the ambiguity in the data analysis process by means of statistical tools and models is a key challenge. This challenge occurs, since the current uncertainty models are limited in capacity and do not match well with the medical imaging problems considered. To address this, it is crucial to develop models capable of effectively describing uncertainty and providing a means of quantification. This would allow the models to express their confidence levels in predictions, especially when faced with ambiguous tumor boundaries. If the ambiguous boundaries of a structure can be accurately modelled, the overall segmentation of structure will also become more accurate.

1.4.2 Out-of-Distribution detection

Out-of-distribution (OOD) detection in medical image analysis has been predominantly used in two manners. (1) Semantic OOD refers to detecting anomalous samples (e.g. rare tumor samples) when trained on a healthy domain. (2) Covariate OOD detection ensures the safe deployment of secondary CAD systems in dynamic environments to filter OOD samples or to detect faulty imaging equipment. Figure 1.2 depicts the high-level concepts of modeling a distribution from observed samples. New samples are then evaluated under the modeled distributions to determine the likelihood they originate from the same underlying distribution.

1.4.2.A Semantic OOD detection

Detecting anomalous samples as semantically OOD is valuable. However, existing approaches often focus on the wrong aspects of the data, such as high-frequency features, which may not be relevant for distinguishing between In-Distribution (ID) and OOD samples.

The key challenge is to identify and model the components of the training distribution (In-Distribution set) that provide most information for determining if a new sample is OOD. Generative models, an effective approach to OOD detection, are limited in modeling capacity. To compensate for this limitation, key-informative components to the OOD detection problem can be identified and captured instead. For instance, in the context of detecting melanoma, emphasizing semantic context and low-frequency image components can enhance the detection of semantic OOD samples.

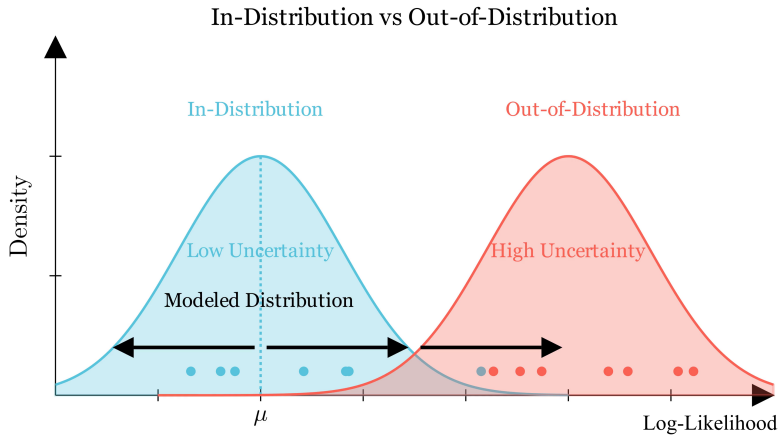


Figure 1.2 *Conceptual diagram of modeling a distribution and utilizing the modeled distribution to detect OOD samples or estimate uncertainty. The arrows represent a notion of increasing uncertainty.*

1.4.2.B Covariate OOD detection

Existing OOD detection methods largely overlook the issue of covariate shift, where the distribution of high-level features changes over time or between different environments, while still offering consistent low-level features. This is particularly troublesome in medical imaging, because slight variations in imaging protocols or patient demographics can lead to significant shifts in data distribution. The key challenge is to explicitly define covariate shift, particularly through novel approaches, which can contribute to the robustness and reliability of imaging systems in clinical settings.

1.4.3 Accurate and real-time pose estimation

Accurate and real-time pose estimation is a well-established task in RGB images, with current methods achieving high levels of accuracy across various applications. However, these advancements have not yet been effectively translated to X-ray imaging, which presents specific challenges resulting from the complexity and variability of the imaging environment. One of the key challenges in X-ray pose estimation is to account for variable imaging geometries, such as changing source-image distances, field-of-view adjustments, and detector variations, which all have a significant impact on the accuracy of pose predictions.

1.4.3.A General purpose 6-DoF pose estimation

The key challenge is to develop pose estimation techniques to accommodate the aforementioned variable geometries inherent in X-ray imaging. The aim is to develop a method that ensures reliable and accurate pose estimation, regardless of changes in the acquisition setup. Such advancement would enhance the precision and safety of image-guided surgical procedures, particularly in complex

interventions such as spinal surgeries where accurate instrument placement is critical.

1.5 Problem statement and research questions

This section defines a problem statement based on the observations from this chapter and formulates specific research questions following from this problem definition.

1.5.1 Problem statement

This thesis addresses four problems. The first is the exploitation of incorporating domain-specific knowledge in PDAC detection models. The second involves modeling of ambiguity (uncertainty) in the data to improve the accuracy in delineating cancerous lesions. The third problem involves the accurate detection of semantic and covariate Out-of-Distribution data in dynamic clinical environments. Finally, the fourth problem entails developing an advanced instrument pose estimation technique capable of real-time and accurate pose predictions, particularly capable of handling variable imaging geometries.

1.5.2 Research questions

From the above problem statement, a number of specific research questions are derived, which are formulated below.

RQ1: Incorporating domain-specific knowledge in PDAC detection

The accurate detection of pancreatic ductal adenocarcinoma (PDAC) remains a significant challenge in the field of medical imaging due to its subtle early-stage presentation and complex anatomical surroundings. Recent advances in computer-aided detection (CADe) systems, leveraging deep learning techniques, offer the potential to significantly enhance the accuracy of PDAC detection. Based on these observations, this research aims to explore state-of-the-art techniques for pancreatic CADe systems, focusing on the integration of secondary, PDAC-indicative features (the domain knowledge) that can improve detection performance. This investigation further continues by addressing the components critical for an end-to-end PDAC CADe system, from the acquisition of CT scans to the accurate assessment of tumor presence, aiming to provide a comprehensive detection system. Bearing these aspects in mind, this research raises the following questions.

- RQ1a: *Can we effectively include PDAC-indicative features into a PDAC CADe system to enhance the detection performance?*
- RQ1b: *What is a possible setup for a complete end-to-end pancreatic CADe system?*

RQ2: Accurate ambiguity modeling for improved segmentation

Accurately segmenting anatomical structures in medical imaging is crucial for diagnosis and treatment planning, especially in ambiguous cases where ground-

truth data are uncertain. It is expected that the quantification of aleatoric uncertainty in medical image segmentation can be improved by examining the use of advanced probabilistic models, such as Normalizing Flows integrated with the Probabilistic U-Net (PU-Net). Furthermore, by extending these uncertainty models to 3D processing, the consistency, efficiency, and accuracy of uncertainty quantification can be enhanced. This provides more reliable segmentation maps that can be used confidently in clinical practice. These views raise the following research questions.

- RQ2a: *How can we model ambiguous ground-truths (aleatoric uncertainty) to improve the accuracy of segmentation maps?*
- RQ2b: *Does aleatoric uncertainty modeling in 3D improve the accuracy, consistency and execution speed?*

RQ3: Exploring uncertainty modeling in PDAC resectability prediction

The prediction of the resectability of pancreatic ductal adenocarcinoma (PDAC) is crucial for determining appropriate treatment strategies, yet remains a complex task due to the intricate relationship between the tumor and surrounding vasculature. This research aims to develop a segmentation-based approach to accurately delineate PDAC and relevant vascular structures, facilitating the automated determination of tumor resectability. Furthermore, the study focuses on exploring the role of model uncertainty in making clinically relevant resectability predictions, ultimately contributing to more informed and precise surgical planning. Through this investigation, the following research questions are raised.

- RQ3a: *How accurate should a PDAC and relevant vasculature segmentation algorithm be to obtain a feasible automated prediction of resectability?*
- RQ3b: *Is the integration of model uncertainty into prediction models applicable and sufficiently useful for resectability prediction?*

RQ4: Density modeling for Out-of-Distribution detection

Out-of-distribution (OOD) detection is a critical aspect of ensuring the reliability of CAD systems in medical imaging, or detecting rare occurrences of complex diseases such as malignant melanoma. Density modeling techniques, such as wavelet-based Normalizing Flows, can more accurately model the wavelet details of the In-Distribution (ID) melanoma to improve the detection of OOD samples in skin melanoma images. Covariate shifts is known to adversely affect image quality and as a result, the image analysis methods using these images. To detect these covariate shifts, unsupervised generative models are strong candidates with their ability for OOD detection. However, modeling the complete data distribution may be difficult and it is expected that when explicitly modeling high-frequency heteroscedastic components of the data distribution, these generative model could be more effective in detecting and quantifying covariate shifts. Bearing these aspects in mind, the following research questions are formulated.

- RQ4a: *Can generative models effectively detect and quantify semantic and/or covariate shifts in natural and X-ray images?*
- RQ4b: *How can high-frequency heteroscedastic image components be explicitly modeled and does this lead to improved OOD covariate shift detection performance?*
- RQ4c: *Do the decomposed frequency components of an image contain sufficient information to improve OOD detection performance?*

RQ5: Instrument Pose Estimation in X-ray

Accurate and efficient pose estimation of surgical instruments during X-ray image-guided interventions is critical for the success of the intervention. This research seeks to establish a general-purpose method for 6-degrees of freedom (6-DoF) pose estimation, addressing the challenges of variable imaging geometries inherent in X-ray procedures. By incorporating X-ray imaging geometry into the pose estimation process, this study aims to develop a model that enhances the accuracy and performance of pose predictions. Integrating these mentioned components raises the following research questions.

- RQ5a: *How to develop a general-purpose method that is both accurate and fast for 6-DoF pose estimation?*
- RQ5b: *How can X-ray data and the imaging geometry be effectively incorporated into the 6-DoF pose estimation process, and what impact does this have on the model performance?*

1.6 Contributions

This section provides an overview of the scientific contributions presented in this thesis. These contributions can be linked to four categories, which are elaborated below.

1.6.1 Contributions to pancreatic cancer treatment

This research contributes to the early detection and treatment via resectability prediction of pancreatic ductal adenocarcinoma (PDAC), one of the most lethal forms of cancer. A deep residual 3D U-Net that integrates secondary, tumor-indicative features is developed to segment the PDAC from contrast-enhanced CT scans. These secondary features, which are identified and utilized by expert radiologists from the CT data, are crucial in identifying PDAC at an earlier stage when surgical resection is still viable. This feature integration enhances the detection capabilities beyond the level that is achieved with primary tumor detection alone. A complete end-to-end PDAC detection processing chain is established, which realizes a high PDAC detection performance of 0.99 Area Under Curve of the Receiver Operating Characteristic (AUROC) on an internal test set and a perfect sensitivity on a public benchmark.

In addition to detection, this research is one of the first investigations into to the critical task of predicting resectability of PDAC. Accurate resectability prediction is essential for determining whether a patient can undergo surgery with a curative intent. The study introduces a novel segmentation-based approach, utilizing multiple nnU-Nets, 3D U-Nets and Probabilistic 3D U-Nets (9 models in total). This overall approach does not only identify the tumor, but also delineates relevant vascular structures that are critical for surgical planning. From the generated segmentation maps of the tumor and vasculature, an automated "angle of involvement" is derived that estimates the extent of the tumor growth. This angle aligns with the established clinical practice on pancreatic cancer treatment, making the model output suitable for clinical use. By integrating model uncertainty into the segmentation predictions, the research further enhances the reliability of the resectability assessments, providing clinicians with valuable information that can guide well-funded treatment decisions.

This multi-stage approach encompasses both early detection and resectability prediction. It offers a comprehensive solution that can lead to more personalized and effective treatment strategies for pancreatic cancer patients, ultimately improving survival rates and quality of life.

1.6.2 Contributions to improving quantification of uncertainty

In the realm of medical image segmentation, the conducted research advances the state-of-the-art methods in uncertainty quantification. This is an important component for accurate segmentation under ambiguous ground truths and to ensure the reliability of automated diagnostic systems. Early methods for quantifying uncertainty often rely on mapping the uncertainty through Gaussian distributions, which may not fully capture the complexity and variability inherently occurring in medical imaging data. This research introduces the integration of Normalizing Flows (NFs) into the Probabilistic U-Net (PU-Net), allowing for more flexible and expressive posterior distributions. This combination enables a more accurate representation of aleatoric uncertainty (14% improvement in GED and 13% in Hungarian IoU), which is the inherent uncertainty in the data itself, particularly when dealing with ambiguous ground truths.

Moreover, the research extends these advancements by applying the aforementioned enhanced uncertainty quantification methods to three-dimensional (3D) medical imaging data. This extension is crucial as many clinical applications, such as volumetric tumor segmentation and 3D anatomical modeling, require the analysis of 3D datasets. By developing a 3D probabilistic U-Net with NFs that support 3D processing, the research addresses the challenges of maintaining consistency, efficiency, and accuracy in multi-dimensional segmentation tasks. This 3D extension is validated on 3D lung nodule segmentation and aleatoric uncertainty quantification. Using both quantitative (accuracy and efficiency) and qualitative (consistency) evaluations the method efficacy is proved. By advancing these techniques, the research lays a foundation for robust uncertainty quantification in medical imaging that can better adapt to the real demands of clinical practice.

1.6.3 Contributions to methodologies for Out-of-Distribution detection

Out-of-Distribution (OOD) detection is a crucial aspect of computer vision, as it can detect novel semantic objects, such as tumors in a healthy cohort. In addition, it can be employed as a safe-guard to detect statistical deviations in images that could adversely affect the image quality and image analysis models using them. This research makes significant strides in this field by utilizing generative models and an image decomposition process that reveal the valuable components for either (1) semantic or (2) covariate OOD detection.

Specifically, (1) wavelet-based NFs (Wavelet Flow) are employed to model the wavelet detail coefficients of benign melanoma, showcasing clear deviations in the low-frequency components when testing for malignant melanoma. This approach realizes a 5-% increase in AUROC compared to the baselines, whilst reducing model size to only 1.25 M parameters. This work demonstrates that by focusing on the most relevant image components, a novel OOD detection model is created that significantly improves semantic OOD detection performance.

Similarly, to detect (2) covariate shift, this research introduces CovariateFlow, a novel methodology that applies conditional Normalizing Flows to explicitly model the high-frequency heteroscedastic components of images. The decomposition is obtained through a Gaussian filter and the conditional relation between the high-frequency and low-frequency components are modeled with a series of coupling flow steps. Additionally, this research introduces the Normalized Score Distance (NSD) as a unified metric that combines typicality and log-likelihood for more effective OOD detection in NFs. The culmination of these improvements enable accurate covariate shift detection across a wide range of covariate factors and different datasets. The method is also applied to covariate shift detection in X-ray images, in which it exhibits a strong performance compared to the baseline models (implemented generative models).

This study is one of the first to explore covariate shift to this extent, thoroughly comparing various generative models on both natural and X-ray images. Besides this novelty, these contributions are particularly valuable in clinical environments, where the ability to detect distribution shifts—whether due to changes in imaging protocols, patient populations, or even equipment malfunctions—is critical for maintaining the accuracy and reliability of diagnostic systems.

1.6.4 Contributions to methodologies for object 6-DoF pose estimation

State-of-the-art minimally invasive surgeries can utilize pose estimation systems in guiding surgical instruments during procedures. This research contributes to this field by developing a general-purpose method, called YOLOv5-6D, for 6-degrees-of-freedom (6-DoF) pose estimation, that fully supports the challenges of X-ray imaging. The proposed method predicts 2D keypoints of the projected 3D bounding box of the object. This sets the method apart from direct pose estimation approaches, since the camera model can freely change at execution time and the pose can still accurately be computed with standard Perspective-n-Point (PnP) methods that incorporate the per-frame camera geometry.

Moreover, the research has practical implications for enhancing the safety and efficacy of image-guided procedures, particularly in complex surgeries such as spinal operations, where precise instrument placement is critical. The methods is extensively validated on three benchmarks: (1) a public RGB benchmark (LINEMOD) consisting of every-day objects, (2) an X-ray test cube confirming the effective transition to the X-ray domain and (3) a spinal screw trained on clear images and tested on more complex X-ray images with a human spine phantom. The obtained accuracy (96.84% ADD(-S)), superior inference speed (41.88 FPS) and generalization ability of the proposed YOLOv5-6D models across these diverse datasets proves its efficacy.

By developing a methodology that can adapt to different and varying acquisition geometries, the research ensures that the pose estimation model remains accurate across a wide range of clinical scenarios. This adaptability is key to the generalization of the method and its potential for widespread application in various medical imaging contexts.

1.7 Thesis outline and scientific background

This chapter sets the stage for the thesis by providing a comprehensive overview of the problem domain, its significance, and the specific challenges addressed. The thesis layout is depicted in Figure 1.3. After the generic introduction, the thesis starts with the background section, which introduces the reader to the context of the research on enhancing computer vision methods for cancer detection and precision guidance in medical imaging. The subsequent chapters delve into the prominent role of deep learning in medical image analysis and Computer-aided detection (CADe) in medical imaging, particularly focusing on Pancreatic ductal adenocarcinoma (PDAC). The research is then dedicated to a more generic problem on modeling image data distributions, and explores the statistical and computational methods used to model medical images for further analysis e.g. OOD detection. The last technical chapter discusses computer-aided guidance during minimally invasive surgeries and the integration of learning-based methods in surgical procedures to enhance precision and outcomes. The last chapter concludes the research. The following paragraphs summarize the results and scientific background of the individual chapters.

Chapter 2 introduces on the recent advancements in deep learning technologies applied to medical image analysis. It starts with an introduction to Convolutional Neural Networks (CNNs) in medical imaging, outlining their role in enhancing image interpretation. Key Convolutional Neural Network (CNN) architectures are briefly outlined, including those for image classification, image segmentation, object detection, and pose estimation, highlighting their applications and performances in medical imaging. The chapter then discusses generative models, starting with Variational AutoEncoders (VAEs), covering their theoretical foundations and practical implementations for non-linear latent variable models. Normalizing Flows are introduced as a method for data modeling through prob-

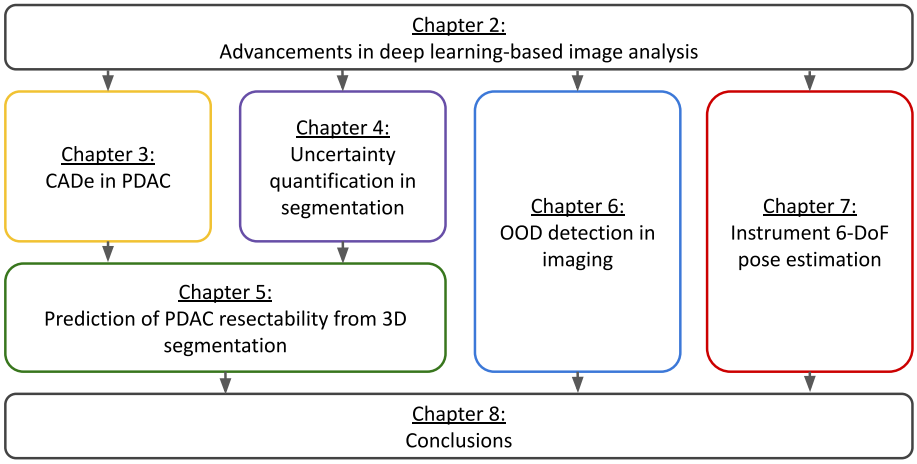


Figure 1.3 *Schematic layout of the thesis.*

ability content preservation and enabling image generation and OOD detection capabilities. The final sections address uncertainty in deep learning, differentiating between deterministic and distributional methods, and exploring uncertainty disentanglement to improve model reliability and interpretation in medical applications.

Chapter 3 is dedicated to the application of computer-aided detection (CAD) in diagnosing PDAC. It begins with detecting pancreatic cancer, explaining the clinical significance and the impact of deep-learning methods on early detection. Previous work in PDAC detection is summarized to overview existing literature. The chapter then introduces a novel approach for PDAC detection that utilizes clinically-relevant secondary features. Further sections discuss the automated segmentation of the external and clinically-relevant features, focusing on multi-stage segmentation approaches for pancreas and bile duct segmentation. The development and evaluation of a comprehensive framework for pancreatic head cancer detection and localization on CT is then presented as a novel end-to-end approach (yielding a score of 99% AUROC in PDAC detection). The chapter concludes with a discussion of challenges and future directions in PDAC Detection.

The review on PDAC detection literature was published in the Journal on Clinical Medicine [22]. The proposed approach to PDAC detecting incorporating clinical features was presented at the MICCAI 2022 conference [53] and the method for obtaining these clinical features at SPIE Medical Imaging 2023 conference [54]. Finally, the complete end-to-end PDAC detection was presented in the Cancers journal [55].

Chapter 4 introduces uncertainty quantification in medical image segmentation and presents a review of existing methods for quantifying uncertainties in image segmentation, introducing Probabilistic U-Net (PU-Net) and its limitations. The chapter then builds on this model to introduce a novel approach for improving aleatoric uncertainty quantification using Normalizing Flows. This

improved PU-Net realizes an increase of 14% increase in a generalized energy distance and 13% in Hungarian IoU over the baseline. Further sections extend the framework to probabilistic 3D segmentation for aleatoric uncertainty quantification, that improves the accuracy, efficiency and consistency of uncertainty presentation over 2D approaches. The 3D approach is also a step closer to the practical clinical application of these uncertainty models.

The improved PU-Net was published at the MICCAI 2021 UNSURE workshop [56] and its extension to 3D at the SPIE Medical Imaging 2023 conference [57].

Chapter 5 focuses on predicting the resectability of pancreatic tumors, defined by the degree of involvement with surrounding anatomy. Building on Chapter 3 and Chapter 4, tumor detection in the CT scan is covered and a localized segmentation model capable of capturing the uncertainty in the relevant anatomy and tumor is developed. In this first of its kind exploration, a deep learning-based framework is utilized to automatically assess tumor-vessel involvement, which is essential for determining tumor resectability. It is shown that the best performing model detects involvement with a 88% sensitivity and a 86% specificity.

This research was presented at the International Conference of Computer Vision (ICCV) in 2023 at the CVAMD workshop [58].

Chapter 6 addresses the challenge of detecting Out-of-Distribution (OOD) data. Efficient OOD Detection with wavelet-based Normalizing Flows is introduced as a novel approach, followed by a validation of its application to melanoma detection yielding a 5% improvement in AUROC over the baseline. Subsequent sections delve into generative models for OOD covariate Shift detection, introducing concepts like covariate shift and typicality. The chapter presents CovariateFlow, an approach to OOD covariate shift detection, including definitions, model design, and experimental validation. The model achieves an average AUROC of 75% in detecting covariate shift of all degradation types in natural images and 93% in the X-ray setting.

The semantic OOD detection for early melanoma detection was published at the MICCAI 2022 conference [59] and the CovariateFlow model at the European Conference on Computer Vision (ECCV) 2024, Uncertainty in Computer Vision workshop [60].

Chapter 7 discusses the development of a general-purpose technique for 6-degrees of freedom pose estimation of instruments in medical imaging. The proposed YOLOv5-6D pose architecture achieves accurate and fast 6-DoF pose estimation, while generalizing across different and varying acquisition geometries and image complexities. This approach is tested for object pose estimation in RGB images (96.84% average distance difference (ADD) symmetric (-S)), test objects in X-ray imaging (99.27% ADD(-S)) and finally bone-screw pose estimation for spinal surgeries in X-ray image-guided interventions (92.41% ADD(-S)).

The initial concept of the chapter was published at the SPIE Medical Imaging 2022 conference [61] and the final YOLOv5-6D model in the IEEE Transactions on Image Processing (TIP) in 2024 [62].

Chapter 8, summarizes the key findings of the thesis, discussing the results of individual chapters and the corresponding research questions. The chapter concludes with a brief future outlook and research challenges.

Chapter 2 Recent advancements in DL-based image analysis

This chapter is added for completeness and provides the necessary technical background information and a brief introduction to the deep learning-based image analysis algorithms applied in this thesis. Depending on the background of the reader, some sections of the chapter may be skipped for reading.

A simple guide is as follows. Chapter 3 and Chapter 5 on the localization and analysis of PDAC, utilize the information on segmentation architectures in Section 2.1. Chapter 4 on segmentation uncertainty largely builds on the information on generative models discussed in Section 2.2 and the uncertainty quantification techniques in Section 2.3. Chapter 6 on OOD detection extends the generative modeling techniques introduced in Section 2.2. Finally, Chapter 7 presents a 6-DoF pose estimation model and is based on the details from Section 2.1.4.

2.1 Convolutional neural networks in medical imaging

Convolutional Neural Networks (CNNs) have become a cornerstone in the field of medical image analysis. These networks provide tools for image classification, segmentation, and object detection, each playing a crucial role in enhancing diagnostic processes. The following image analysis tasks and CNN-based architectures have significantly advanced the field of automated medical image analysis.

2.1.1 Image classification

Image classification is a fundamental task in computer vision that aims to categorize images into predefined classes. Over the past decade, deep convolutional neural networks (CNNs) have dramatically improved the state-of-the-art in image classification performance. Early breakthrough architectures like AlexNet [5] demonstrated the power of deep CNNs trained on large datasets. VGGNet [63] showed how increased depth through stacking many 3×3 convolutional layers can further enhance accuracy. GoogLeNet [64] introduced the inception module to efficiently expand the network width and capture multi-scale features.

More recently, ResNet [7] addressed the challenges of training very deep networks by introducing skip connections that enable residual learning. The ResNet architecture allows for effective training of extremely deep networks, significantly outperforming previous architectures. As described by He *et al*. [7], ResNet achieves state-of-the-art accuracy on the ImageNet dataset while having lower

complexity than the VGGNet. The residual learning framework enables the training of networks that are substantially deeper (more layers) than those used previously, leading to significant accuracy gains from increased depth. Figure 2.1 depicts the skip connection and introduces a residual learning building block, where the drawing represents one ResNet layer.

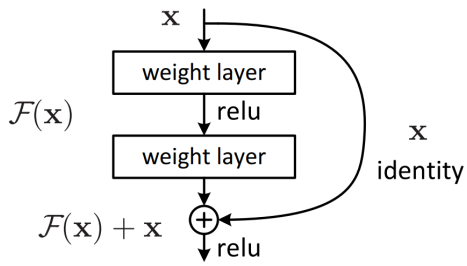


Figure 2.1 *Diagram of a residual learning building block. Image taken from ResNet paper [7]*

In radiology, CNNs are used to detect and classify various pathologies. For example, a Rajpurkar *et al*. [65] introduced CheXNet, a deep CNN trained on a large dataset of chest X-rays (ChestX-ray14 [66]) to detect pneumonia. CheXNet outperformed radiologists in identifying pneumonia, showcasing the potential of CNNs in improving the diagnostic accuracy in radiology. Further advancements in a study by Ardila *et al*. [67] present an end-to-end deep learning model that predicts the risk of lung cancer by analyzing CT images. Beyond radiology, CNNs have shown remarkable success in other medical imaging fields. In histopathology, CNNs have been utilized to classify tissue images, aiding in the detection of cancers. Coudray *et al*. [68] used deep learning to distinguish lung adenocarcinoma from squamous cell carcinoma with performance comparable to expert pathologists. Additionally, CNNs have advanced dermatology by classifying skin lesions from dermoscopic images, as demonstrated by Esteva *et al*. [69], whose model achieved dermatologist-level performance in identifying skin cancer. These applications illustrate the considerable impact of CNNs on medical image classification, enabling earlier and more accurate diagnoses across a broad spectrum of diseases.

2.1.2 Image segmentation

Image segmentation is another important application of CNNs in medical imaging. Unlike classification, segmentation aims to partition an image into multiple segments (sets of pixels) to simplify or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation can be approached through several methods. Conventionally, techniques like thresholding, edge detection, and region-growing were used. However, these methods often struggle with complex images and variability in lighting, texture, and shapes. Deep learning approaches, particularly CNNs, have demonstrated

superior performance by learning features directly from the data with higher robustness. Several key architectures have been developed for deep learning-based image segmentation, which are summarized below.

- Fully Convolutional Network (FCN) for semantic segmentation [70]: FCNs replace fully-connected layers with convolutional layers to output spatial maps, enabling pixel-wise prediction.
- SegNet [71]: Badrinarayanan *et al*. have proposed SegNet, which employs an encoder-decoder architecture to capture context and reconstruct segmentation maps.
- U-Net [12]: Developed by Ronneberger *et al*., the U-Net is particularly designed for biomedical image segmentation, utilizing a symmetric encoder-decoder structure with skip connections to combine low-level and high-level features.

2.1.2.A U-Net: convolutional networks for biomedical image segmentation

The U-Net [12] architecture (Figure 2.2) is a pioneering and the most established model in the domain of biomedical image segmentation. The U-Net consists of two main parts: the encoding path and the decoding path. The encoding path (left side of Figure 2.2) is a typical convolutional network with repeated application of two 3×3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) [72] and a 2×2 max-pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. This path captures context through increasingly abstract representations. The decoding path (right side of Figure 2.2) consists of an upsampling of the feature map followed by a 2×2 convolution ("up-convolution") that halves the number of feature channels. This is concatenated with the correspondingly cropped feature map from the encoding path, followed by two 3×3 convolutions and ReLU. This structure enables precise localization by combining high-resolution features from the encoding path with the upsampled output.

Training objective: The U-Net training strategy includes many standard data augmentation techniques, with the addition of elastic deformations, to increase the robustness and generalization ability of the model. This is particularly important in biomedical applications, where annotated data are scarce. The architecture also incorporates a weighted loss function to emphasize the separation of touching objects, which is a common challenge in cell segmentation tasks. The model can be trained with the standard loss functions such as a pixel-wise Cross-Entropy (CE) Loss, Mean-Squared Error (MSE) losses (commonly used for regression tasks), or using a Dice-Sørensen coefficient (DSC) loss which is particularly useful for segmentation tasks in medical imaging. The DSC loss quantifies the similarity between predicted and GT segmentation maps.

Performance: The U-Net has achieved significant success in various biomedical segmentation tasks. It won the ISBI cell tracking challenge 2015 by a large margin

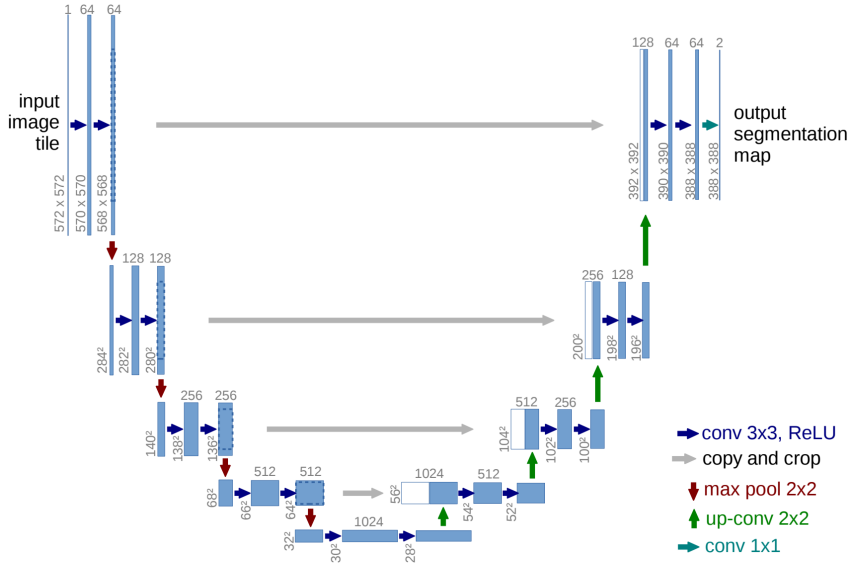


Figure 2.2 *Schematic of the U-net architecture. Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The spatial sizes are provided at the lower-left edge of each box. White boxes represent concatenated feature maps. The arrows denote the different operations. Figure obtained from the original paper [12]*

in several categories, demonstrating its efficiency and accuracy. The U-Net can be trained end-to-end using very few images only, making it a practical solution for many image analysis tasks where the data are limited.

2.1.3 Object detection

Object detection is a fundamental task in computer vision that involves identifying and localizing objects within an image. Unlike image classification, which assigns a single label to an entire image, object detection aims to detect and classify multiple objects within the image and provide their precise locations in the form of delineated bounding boxes. This technology has numerous applications, including autonomous driving, security surveillance, and medical image analysis. Object detection algorithms typically generate bounding boxes around objects of interest in an image. The process involves several key steps, mentioned below.

- *Feature Extraction:* Extracting object features from the input image using convolutional neural networks (CNNs) to obtain feature maps.
- *Region Proposal:* Generating a set of candidate regions or bounding boxes that may contain objects.
- *Classification and refining localization:* Classifying the proposed regions into object categories and refining the bounding boxes to better fit the objects.

Several significant architectures have been developed for object detection, each contributing to improving the accuracy and efficiency.

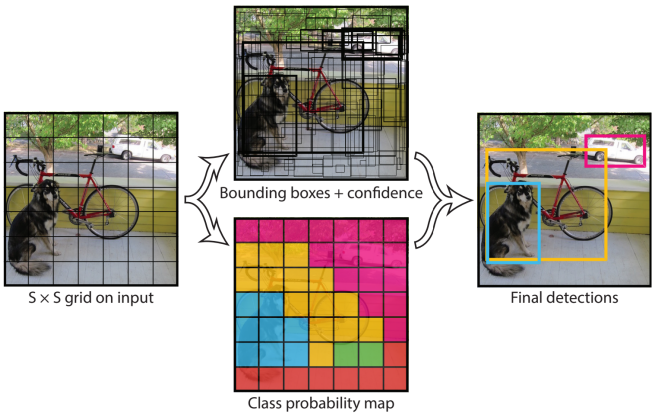


Figure 2.3 *Diagram of the YOLO model performing detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, the confidence values related to those boxes, and C class probabilities. These predictions are encoded as a tensor of size $S \times S \times (B*5 + C)$. The image is taken from the first YOLO paper [76].*

- R-CNN (Region-based Convolutional Neural Networks) [73]: R-CNN uses selective search to generate region proposals and applies a CNN to classify and refine each proposal. While accurate, R-CNN is computationally intensive due to its multi-stage pipeline.
- Fast R-CNN [74]: Being an improvement over R-CNN, Fast R-CNN combines region proposal and classification into a single network, significantly reducing computation time by sharing convolutional features.
- Faster R-CNN [75]: Faster R-CNN introduces a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, enabling nearly cost-free region proposals and improving detection speed.
- YOLO [76]: The You Only Look Once (YOLO) architecture is a first-of-its-kind architecture that predicts object location and class probabilities on a grid from the full image in one forward pass, significantly improving speed and accuracy over prior approaches.

2.1.3.A YOLO

The You Only Look Once (YOLO) architecture [76], proposed by Redmon *et al.*, represents a paradigm shift in object detection by framing it as a single regression problem, directly predicting bounding boxes and class probabilities from full images in one forward pass through the network. There are many versions of YOLO, currently reaching up to YOLOv11 [77]. The following section discusses the core principle of the architecture.

Architecture: As depicted in Figure 2.3, the YOLO model divides the input image into an $S \times S$ grid, where each grid cell is responsible for predicting B bounding boxes and confidence scores for those boxes. Each bounding box consists of five predictions: the image coordinates of the box center (x, y), its width

and height (w, h), and a confidence score indicating the presence of an object and the accuracy of the box. Each grid cell also predicts C conditional class probabilities, assuming that an object is present. These probabilities are multiplied by the individual box confidence predictions to produce the final scores for the bounding boxes. The YOLO network consists of a single CNN that processes the entire image at once, making it extremely fast. The original YOLO network uses 24 convolutional layers, followed by 2 fully-connected layers. The network architecture enables real-time object detection, enabling a balance between speed and accuracy.

Training objective: YOLO's loss function combines multiple components: (1) Localization loss, which measures errors in the predicted bounding-box coordinates, (2) Confidence loss, that measures errors in the predicted confidence scores for an object to be in a particular grid cell, and (3) Classification loss, which accounts for errors in the predicted class probabilities for each grid cell.

Performance: YOLO is known for its fast inference speed, capable of processing images at 45 frames per second. The model trades-off accuracy against speed, when compared to models like Faster R-CNN. Its ability to perform real-time object detection makes it suitable for real-time applications and areas where inference speed is crucial.

2.1.3.B Object detection in medical image analysis

Object detection is increasingly applied in medical image analysis, particularly in radiology, to identify and localize abnormalities such as tumors, lesions, and other pathological findings. DeepLesion [78], proposed by Yan *et al.*, is a large-scale dataset and deep learning framework for universal lesion detection in CT scans. The framework uses a variant of Faster R-CNN to detect a wide variety of lesions. RetinaNet [79] combines the speed of single-shot detectors like YOLO with the accuracy of two-stage detectors like Faster R-CNN. The model utilizes a focal loss to handle class imbalance. This architecture has been applied to detect microcalcifications in mammograms.

Deep learning-based object detection has significantly advanced the capability to detect and localize objects within images accurately and efficiently. Architectures like R-CNN, Faster R-CNN, and YOLO have paved the way for real-time and high-accuracy detection solutions. The continuous development and application of these techniques promise further advancements in the field of medical image analysis.

2.1.4 Object pose estimation

Object pose estimation is a valuable task in computer vision that involves determining the position and orientation of objects in a 3D space from 2D images. Unlike object detection, which focuses on identifying and localizing objects within an image, pose estimation aims to provide more detailed information about the object's spatial configuration. This information is essential for applications such as robotic manipulation, augmented reality, and in specialized medical image analysis applications. Object pose estimation typically involves predicting six

Degrees-of-Freedom (6-DoF) of an object: three for translation (x, y, z-coordinates) and three for rotation (roll, pitch, yaw).

The prediction process can be divided into three main steps. (1) The model extracts key object features from the image using convolutional neural networks (CNNs) or other feature extraction methods. (2) Pose prediction is performed using these features to estimate the object's position and orientation in the 3D space. (3) Optimization techniques are applied to refine the predicted pose for improving accuracy.

Pose estimation can be approached through various methods, including direct regression, where the network directly predicts the pose parameters, or by generating intermediate representations, like 2D keypoints or heatmaps, which are then used to infer the pose.

2.1.4.A YOLO 6D

The YOLO 6D model depicted in Figure A.1 in Appendix A is an extension of the YOLO architecture specifically designed for real-time 6-DoF object pose estimation. This model combines the speed and efficiency of the YOLO framework with the capability to predict 6D poses.

Architecture: YOLO 6D extends the original YOLO framework by incorporating 6D pose estimation capabilities. The architecture processes an input image and predicts key points corresponding to the projection of the 3D bounding box of the object. The 2D-3D correspondences from the prediction can then be used to solve the object pose. The network consists of several convolutional layers, followed by fully-connected layers. The key components of YOLO 6D in the order of operation include the following steps.

- *Grid Division:* The image is divided into an $S \times S$ grid. Each grid cell predicts several bounding boxes and their associated confidence scores, class probabilities, and 2D projections of the 3D corners of an object.
- *2D-to-3D Correspondences:* The YOLO 6D model predicts the 2D image coordinates of the projected corners of the 3D bounding box of the object. These 2D projections are used to establish correspondences between the 2D image and the 3D model.
- *Pose Estimation:* Using the 2D-3D correspondences, the 6-DoF pose of the object is estimated through a Perspective-n-Point (PnP) algorithm [80], which solves for the object's position and orientation in the camera coordinate system.

Training objective: YOLO 6D is trained using heavy augmentations of real-world datasets. The loss function is a combination of localization loss, confidence loss, classification loss, and key point projection loss. The key point loss ensures accurate prediction of the 2D projections of the 3D corners, which is crucial for precise pose estimation. Various transformations are applied to the training images, such

as scaling, rotation, and translation, to make the model robust to different viewpoints and lighting conditions.

Performance: YOLO 6D is known for its real-time performance, capable of processing images at high throughput rates while providing accurate 6D pose estimates.

2.1.4.B Object pose estimation in medical image analysis

While object pose estimation in medical imaging has hardly been attempted, landmark localization plays a role in medical image analysis, especially in radiology, where accurate spatial information about anatomical structures is essential for diagnosis and treatment planning. Payer [81] proposed a method for detecting anatomical landmarks in X-ray images using a heatmap-based approach. This method is used for tasks such as cephalometric landmark detection in dental radiographs and vertebrae localization in images of the spine.

Deep learning-based landmark prediction and pose estimation have significantly advanced the ability to determine the position and orientation of objects in 3D space. Architectures like YOLO 6D have demonstrated the feasibility of real-time 6D pose estimation, providing both speed and accuracy. In medical image analysis, prediction models are transforming radiological procedures by enabling precise localization and orientation of anatomical structures, which are crucial for diagnosis, treatment planning, and surgical interventions.

These architectures and applications of CNNs in medical image analysis not only help to automate and refine diagnostic procedures, but also significantly contribute to early detection and improved prognosis, thereby enhancing the overall efficacy of medical interventions.

2.2 Generative models

2.2.1 Variational Autoencoders

A Variational Autoencoder (VAE) [47] is a type of latent variable generative model. It consists of an autoencoding architecture that generates a probabilistic latent representation, $\mathbf{z} \sim p(\mathbf{z})$ of the input data ($\mathbf{x}$), from which, after training, can be sampled to generate new data points, $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$. We introduce the latent variables $\mathbf{z}$ and the joint distribution factorized as $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. This modeling approach expresses the generative process. During training, we only have access to the input data $\mathbf{x}$. As such, through probabilistic inference we can compute the marginal distribution (integrate over the unknown variable, $\mathbf{z}$). The (marginal) likelihood function is then calculated as

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \tag{2.1}$$

To compute the integral (marginal likelihood) involves integrating over all possible configurations of the latent variables $\mathbf{z}$, weighted by their likelihood and

prior probability. In general, this integral is intractable because the dimensionality of $\mathbf{z}$ can be very high, making the integral over such a space computationally prohibitive (dimensionality). Thus, VAEs employ a technique called *variational inference* to approximate $p(\mathbf{z}|\mathbf{x})$ with a simpler, parameterized distribution $q_\phi(\mathbf{z}|\mathbf{x})$.

2.2.1.A Variational inference for non-linear latent variable models

Since the integral cannot be calculated exactly, a simple approach is to use the Monte Carlo approximation, leading to

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[p(\mathbf{x}|\mathbf{z})\right] \tag{2.2a}$$

$$\approx \frac{1}{K}\sum_k p(\mathbf{x}|\mathbf{z}_k). \tag{2.2b}$$

In line 2.2b, a sample from the prior over the latents is drawn, $\mathbf{z}_k \sim p(\mathbf{z})$. This method is straightforward and feasible, due to the rapid growth of computational capabilities, allowing to sample a substantial number of points. Nonetheless, from a statistical perspective, when the latent variable $\mathbf{z}$ is multi-dimensional and the dimensionality K is large, we encounter the curse of dimensionality. Consequently, to adequately explore the space, the required number of samples escalates exponentially with K. If an insufficient number of samples are taken, the resulting approximation will be notably inaccurate.

Advanced Monte Carlo methods [82] can be employed to tackle this problem, but these techniques are still vulnerable to the same curse of dimensionality. An effective alternative is to use variational inference [83]. To this end, we consider a set of variational distributions parameterized by ϕ, denoted as $\{q_\phi(\mathbf{z})\}_\phi$. For example, Gaussian distributions parameterized by their means and variances, where $\phi = \{\mu, \sigma^2\}$, could be used. These distributions are well-defined, and it is assumed they assign non-zero probability mass across all possible values of $\mathbf{z}$ within the space $\mathcal{Z}^M$. Consequently, the logarithm of the marginal distribution can be approximated using a log-likelihood framework. Taking this perspective,

$$\ln p(\mathbf{x}) = \ln \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \tag{2.3a}$$

$$= \ln \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \tag{2.3b}$$

$$= \ln \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}\left[\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z})}\right] \tag{2.3c}$$

$$\geq \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})} \ln\left[\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z})}\right] \tag{2.3d}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}\left[\ln p(\mathbf{x}|\mathbf{z}) + \ln p(\mathbf{z}) - \ln q_\phi(\mathbf{z})\right] \tag{2.3e}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}\left[\ln p(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}\left[\ln q_\phi(\mathbf{z}) - \ln p(\mathbf{z})\right]. \tag{2.3f}$$

We have used Jensen's inequality [84], [85] in Equation (2.3d), which suggests the

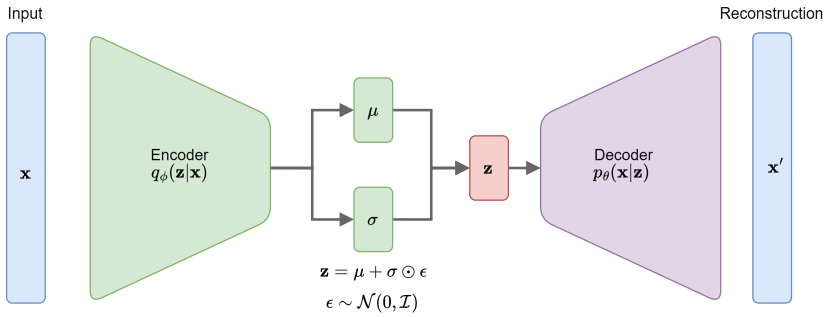


Figure 2.4 *Schematic of the Variational Autoencoder. The input data and its reconstruction is represented by the blue blocks. The green block depicts the encoder that is trained to predict a mean and standard deviation of a low-dimensional multivariate Gaussian distribution. The decoder is depicted in purple.*

lower bound on the likelihood. Using a neural network to estimate the parameters (amortization) of the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ instead of $q_\phi(\mathbf{z})$ for each $\mathbf{x}$, we obtain:

$$\ln p(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln q_\phi(\mathbf{z}|\mathbf{x}) - \ln p(\mathbf{z})\right]. \tag{2.4}$$

This approach results in a similar autoencoder architecture, however, characterized by a stochastic encoder, $q_\phi(\mathbf{z}|\mathbf{x})$, and a stochastic decoder, $p(\mathbf{x}|\mathbf{z})$. The term 'stochastic' underscores that both the encoder and decoder operate as probability distributions, distinguishing this model from deterministic autoencoders. This model (Figure 2.4) is the Variational Autoencoder [47], [86] and utilizes the amortized variational posterior. The lower bound of the log-likelihood function in this model is termed the Evidence Lower Bound (ELBO).

As depicted in Equation (2.4), the ELBO comprises of two parts: The first component $\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z})\right]$, acts as the negative reconstruction error, illustrating the process where $\mathbf{x}$ is encoded to $\mathbf{z}$ and then decoded back. The second component,
$\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln q_\phi(\mathbf{z}|\mathbf{x}) - \ln p(\mathbf{z})\right]$, serves as a regularizer and is equivalent to the Kullback-Leibler divergence (KL) between the two distributions.

2.2.1.B Practical variational autoencoder

In practice, there are a few additional components that make VAEs work well for high-dimensional data such as images, which are listed below.

1. *Encoder:* This part of the network takes the input data and transforms it into a distribution over the latent space. It is typically performed with a series of convolutional, non-linear activations and downsampling layers [47]. However, recent work [87] extended the approach to utilize transformers [8] and the attention mechanism for obtaining $q_\phi(\mathbf{z}|\mathbf{x})$ as well.

2. *Reparameterization trick:* The latent variable $\mathbf{z}$ is typically sampled from a distribution $q_\phi(\mathbf{z}|\mathbf{x})$ parameterized by ϕ, like a Gaussian distribution with mean μ and variance σ^2. Sampling $\mathbf{z}$ directly from $q_\phi(z|x)$ introduces a level of stochasticity that hinders the direct computation of gradients with respect to ϕ. To overcome this, the reparameterization trick reformulates $\mathbf{z}$ as a deterministic function of ϕ and an independent random variable ϵ, which is the source of randomness. Variable $\mathbf{z}$ can be computed as $\mathbf{z} = \mu + \sigma \odot \epsilon$, where ϵ is a noise variable sampled from a standard Gaussian distribution, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ is the element-wise multiplication.
3. *Decoder:* This component samples points from the latent distribution defined by the encoder and attempts to construct the data from these samples. The low-dimensional sample $\mathbf{z}$, is progressively transformed with convolutional layers and upsampled using either an interpolation method or a learnable transposed convolution to obtain images.

VAEs have significantly advanced the field of generative modeling, providing a robust framework for learning complex distributions and generating new data instances that are similar to the given dataset. The flexibility of VAEs allows them to be applied across a broad spectrum of applications, from image generation and enhancement to complex data imputation and anomaly detection. Moreover, the ongoing research into VAEs continues to yield a wide range of improvements, including enhancing model stability, increasing sample diversity, and reducing the gap between the variational approximation and the true posterior. Researchers are also exploring hybrid models that combine VAEs with other neural architectures, aiming to leverage their generative capabilities while overcoming some of their inherent limitations. As the field progresses, these enhancements promise to unlock even more sophisticated capabilities, further cementing the role of VAEs in the toolkit of advanced machine learning methodologies.

2.2.2 Normalizing Flows

Normalizing Flows (NFs) (Figure 2.5) [88] represent a class of generative models that provide a robust framework for *exact* likelihood estimation, a feature that sets them apart in the landscape of statistical modeling. At the core of NFs is the utilization of a sequence of bijective (one-to-one) transformations, which systematically map between complex data distributions and simpler and tractable latent distributions. This bijectivity is crucial, as it ensures the exact computation of the likelihood for any sample, by enabling the reverse transformation process. This effectively circumvents the often intractable integrations typically occurring in generative modeling.

The fundamental mechanism of NFs—transforming the target data distribution $p_{\mathbf{x}}(\mathrm{X})$ into a known, simple distribution (typically a Gaussian) $p_{\mathbf{z}}(\mathrm{Z})$ via invertible mappings (forward normalizing direction), enables precise density estimation. Sampling from the base distribution and applying the inverse trans-

formation (reverse, generative direction) enables sample generation. Specifically, if $\mathbf{z}_0$ is a random variable with a simple, known density $p_0(\mathbf{z}_0)$, a normalizing flow transforms z_0 into another random variable z_K using a series of invertible transformations $f_1, f_2, \ldots, f_K$:

$$z_k = f_k^{-1}(z_{k-1}), \quad k = 1, 2, \ldots, K, \tag{2.5}$$

where z_K is the output random variable whose distribution we aim to model, and f_k are the transformation functions. Each f_k is required to be differentiable and its inverse f_k^{-1} must also be differentiable. The inherent tractability of the likelihood function across these transformations not only facilitates efficient inference, but also enhances the interpretability and applicability of the model in diverse domains. Consequently, NFs are increasingly favored in the field of generative modeling, offering a compelling blend of flexibility, precision, and computational feasibility. By constraining these transformations to be bijective, inverse transformations exist that transform the base distribution $p_0(\mathbf{z}_0)$ to the complex distribution $p_k(\mathbf{z}_k)$ (generative direction) and, thereby enabling sample generation. An NF is optimized with a maximum likelihood objective [88], specified by

$$\log p(\mathbf{x}) = \log p_0\left(\mathbf{z}_0\right) - \sum_{i=1}^{K} \log \left| \det \frac{df_i}{d\mathbf{z}_{i-1}} \right|, \tag{2.6}$$

which is further explained in the next subsection.

2.2.2.A Probability content preservation

The change of variables theorem allows to compute the probability density function of the transformed variable $\mathbf{z}_K$ in terms of the base variable $\mathbf{z}_0$ and the transformation applied. The derivation is as follows for a single transformation f that maps a variable $\mathbf{x}$ to a variable $\mathbf{z}$ (i.e., $\mathbf{z} = f(\mathbf{x})$). The function is invertible, so $\mathbf{x} = f^{-1}(\mathbf{z})$. Given this invertible and differentiable function f where $\mathbf{z} = f(\mathbf{x})$, the probability density functions of $\mathbf{z}$ and $\mathbf{x}$, denoted as $p_z(\mathbf{z})$ and $p_x(\mathbf{x})$, respectively,

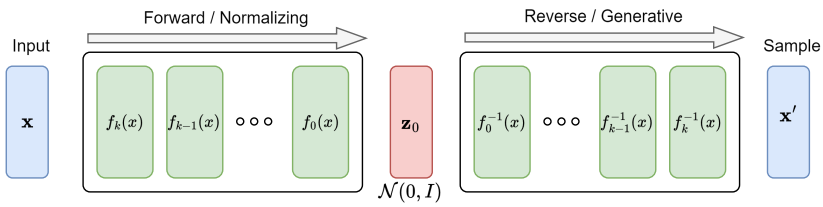


Figure 2.5 *Illustration concept of the Normalizing Flow architecture. The model (green) is trained to transform the input data (blue) to a Normal distribution (red) through a series of invertible transformations in the forward direction. Sampling from the Normal distribution and applying the inverse flow steps in the reverse or generative direction enables creating new sample data points.*

are related by the following equality

$$\int p_x(\mathbf{x})\, dx = \int p_z(\mathbf{z})\, dz = 1. \tag{2.7}$$

Under the transformation $\mathbf{z} = f(\mathbf{x})$, the differential dz is given by:

$$dz = f'(\mathbf{x})\, dx \tag{2.8}$$

where $f'(\mathbf{x})$ is the derivative of f with respect to $\mathbf{x}$. In the multi-dimensional case, $f'(\mathbf{x})$ is replaced by the Jacobian matrix $J_f(\mathbf{x})$ of the transformation, and dz becomes the determinant of the Jacobian, $|\text{det}J_f(\mathbf{x})|dx$. Substituting the transformed differential, the integral over $\mathbf{z}$ can be rewritten into

$$\int p_z(\mathbf{z})\, dz = \int p_z(f(\mathbf{x})) \cdot |\text{det}J_f(\mathbf{x})|dx. \tag{2.9}$$

Given that both integrals must equal 1 (as per Equation 2.7), we establish that

$$\int p_x(\mathbf{x})\, dx = \int p_z(f(\mathbf{x})) \cdot |\text{det}J_f(\mathbf{x})|dx. \tag{2.10}$$

By the property of integral equality, we infer that

$$p_x(\mathbf{x}) = p_z(f(\mathbf{x})) \cdot |\text{det}J_f(\mathbf{x})|. \tag{2.11}$$

With $\mathbf{z} = f(\mathbf{x})$ and its inverse transformation $\mathbf{x} = f^{-1}(\mathbf{z})$, we can express $p_z(\mathbf{z})$ in terms of $p_x(\mathbf{x})$ starting with Equation (2.11) as

$$p_z(\mathbf{z}) = p_x(f^{-1}(\mathbf{z})) \cdot |\text{det}J_{f^{-1}}(\mathbf{z})|. \tag{2.12}$$

Here, $J_{f^{-1}}(\mathbf{z})$ is the Jacobian of the inverse transformation f^{-1}. For multiple transformations $\mathbf{z}_0 \to \mathbf{z}_1 \to \ldots \to \mathbf{z}_K$, the density function at each step incorporates the determinant of the Jacobian of the transformation from the previous step

$$p(\mathbf{x}) = p_{z_K}(\mathbf{z}_K) = p_{z_0}(\mathbf{z}_0) \prod_{k=1}^{K} \left| \det \frac{df_k^{-1}}{dz_k}(\mathbf{z}_k) \right|, \tag{2.13}$$

where $\mathbf{z}_k = f_k(\mathbf{z}_{k-1})$ and $\mathbf{z}_0$ has a known distribution p_{z_0}. This framework allows the modeling of a complex distribution p_{z_K} through simpler, sequential transformations.

2.2.2.B Types of Flows

Various types of transformations can be employed at each flow step to develop models that effectively capture the complexities of intricate probability distributions. The transformation step in a flow model should adhere to several crucial criteria. Such important criteria are that the transformations should be: (1) in-

vertible to ensure no loss of information, (2) differentiable to facilitate the computation of the Jacobian determinant, (3) the Jacobian determinant itself should be tractable for efficient probability density updates during training, and finally (4) the transformations should be flexible to model complex, multi-modal distributions, enabling the system to adapt to diverse data characteristics. This thesis employs three specific types of flows: planar flows, radial flows, and coupling flows. Each has its own formulation and characteristics, making them suitable for various applications in density estimation and variational inference.

1. **Planar Flow:** Planar flows [89] apply a non-linear transformation to data distributions via a planar deformation. The transformation can expand and contract distributions along a specific direction. The function for planar flows is defined as
$$f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^T\mathbf{z} + b), \tag{2.14}$$
where parameter $\mathbf{z}$ is the original variable, $\mathbf{u}$ and $\mathbf{w}$ are learnable parameters (vectors $\in \mathbb{R}^D$), b is a learnable bias parameter $\in \mathbb{R}$, h is a non-linear activation function, typically a hyperbolic tangent (tanh). The absolute Jacobian of this transformation is:
$$\left|\det\frac{\partial f}{\partial z}\right| = \left|1 + \mathbf{u}^T h'(\mathbf{w}^T\mathbf{z} + b)\mathbf{w}\right|, \tag{2.15}$$
where h' is the derivative of h with respect to $\mathbf{x}$, which can be any smooth element-wise non-linearity function.

2. **Radial Flow:** Radial flows [89] introduce a radial transformation, affecting the distribution around a specific point. The radial flow is defined by
$$f(\mathbf{z}) = \mathbf{z} + \beta\, h(\alpha, r)(\mathbf{z} - \mathbf{z}_0), \tag{2.16}$$
where $r = \|\mathbf{z} - \mathbf{z_0}\|$ is the Euclidean distance from $\mathbf{z}$ to $\mathbf{z}_0$, β and α are learnable parameters controlling the strength and smoothness of the transformation and $h(\alpha, r) = 1/(\alpha + r)$ models the radial effect. The absolute Jacobian is given by
$$\left|\det\frac{\partial f}{\partial z}\right| = (1 + \beta h(\alpha, r))^{d-1}\,(1 + \beta h(\alpha, r) + \beta h'(\alpha, r)r)\,. \tag{2.17}$$

3. **Coupling Flow:** Coupling flows, introduced by Dinh *et al*. [90], [91], allow for expressive transformations, while keeping the Jacobian determinant tractable. Coupling flows split the input vector ($\mathbf{z}$) into two parts, $\mathbf{z}_{1:d}$ and $\mathbf{z}_{d+1:D}$. The transformation uses one part to transform the other part,
$$\begin{aligned}\mathbf{y}_{1:d} &= \mathbf{z}_{1:d}\\ \mathbf{y}_{d+1:D} &= g(\mathbf{z}_{d+1:D}; \theta(\mathbf{z}_{1:d})),\end{aligned}$$

and its inverse becomes

$$\begin{aligned}\mathbf{z}_{1:d} &= \mathbf{y}_{1:d} \\ \mathbf{z}_{d+1:D} &= g^{-1}(\mathbf{y}_{d+1:D}; \theta(\mathbf{y}_{1:d})),\end{aligned}$$

where g is a differentiable transformation, parameterized by θ, which are functions of the unchanged part $\mathbf{z}_{1:d}$. The Jacobian of this transformation is triangular,

$$\begin{bmatrix} \mathbf{I}_d & \mathbf{0} \\ \frac{\partial \mathbf{y}_{d+1:D}}{\partial \mathbf{z}_{1:d}} & \frac{\partial g'(\mathbf{z}_{d+1:D}; \theta(\mathbf{z}_{1:d}))}{\partial \mathbf{z}_{1:d}} \end{bmatrix}, \tag{2.20}$$

making the determinant

$$\det\left(\frac{\partial z}{\partial y}\right) = \prod_{i=d+1}^{D} \frac{\partial z_i}{\partial y_i}. \tag{2.21}$$

It is evident that θ can be any arbitrary function, such as a neural network, which underscores the flexibility and power of coupling layers with tractable Jacobian determinants. However, these layers process only half of the input at a time, necessitating an additional transformation to manipulate the entire input vector. A permutation layer is a straightforward, yet effective transformation that can be integrated within a coupling layer. Since a permutation is volume preserving, its Jacobian determinant is invariably equal to unity, it can be strategically applied after each coupling layer to alter the order of variables, such as reversing them in order. This approach ensures thorough processing of the input while maintaining the efficiency and simplicity of the model structure.

NFs are inherently designed to model continuous probability density functions (PDFs), however, in practice, data is often discrete. As such, in addition to the above-mentioned types of flow steps, a data dequantization flow step is required. Since this thesis only employs standard dequantization methods, they are additionally described in Appendix A.3 for completeness.

2.3 Uncertainty in deep learning

Despite the advancements in deep learning, these models often lack robust mechanisms for quantifying and expressing uncertainty in their predictions. This deficiency can lead to overconfident decisions in critical applications where safety and reliability are paramount.

The importance of understanding and quantifying uncertainty in deep learning is threefold. Firstly, it enhances the general safety of AI systems by enabling them to recognize and communicate the limits of their knowledge, which is crucial in high-stakes scenarios such as medical diagnosis and autonomous vehicle navigation. Secondly, uncertainty estimation can improve the model robustness

against noisy, corrupt, or adversarial data inputs. Finally, it provides valuable insights into model behavior, which can guide improvements in model architectures, training procedures, and data utilization.

This section briefly introduces popular methodologies for uncertainty quantification in deep learning. For a detailed introduction, the reader is referred to the overview by Hüllermeier and Waegeman [92]. In this introduction we consider a classification scenario (although the methods can also be extended to segmentation), involving a discrete label space $\mathcal{Y}$ consisting of C classes. The models are denoted as $f : \mathcal{X} \to \Delta^C$ and the output a class probability vector located on the probability simplex Δ^C for any input $x \in \mathcal{X}$. Uncertainty estimators in this context can be categorized into two primary types: distributional methods and deterministic methods.

2.3.1 Deterministic methods

Deterministic methods (non-Bayesian) [93] directly produce a scalar uncertainty estimate $u(\mathbf{x})$, rather than modeling a probability distribution over class probability vectors. This can be achieved with loss prediction techniques [94], [95] which include an additional MLP head that estimates the loss of the network's prediction $f(\mathbf{x})$ on each input $\mathbf{x}$, assuming the loss reflects a notion of (in)correctness. An example for the classification problem, is an uncertainty term predicting correctness (the model estimates $u(\mathbf{x})$) of the likelihood that the predicted class $\hat{y} := \arg\max_{c \in \{1,\dots,C\}} f_c(\mathbf{x})$ is the correct class y, i.e., $p(\hat{y} = y)$.

Another approach for Deterministic Uncertainty Quantification (DUQ) [96] is to learn a latent mixture of Radial Basis Function (RBF) densities on the training/-validation set and outputs $u(x)$, which measures how close an input's embedding is to the mixture means. The Mahalanobis method [97] constructs a similar latent mixture of Gaussians in a post-hoc manner and perturbs inputs in an adversarial manner to train a classifier for separating ID and OOD samples.

Deterministic methods are often more computationally efficient than distributional methods, explaining why they are still widely used, despite potentially lower expressiveness.

2.3.2 Distributional methods

Distributional methods provide an output represented as a probability distribution $q(f(\mathbf{x}) \mid \mathbf{x})$ over all possible class probability vectors. This posterior distribution, commonly abbreviated as $q(f)$, corresponds to a Bayesian hypothesis posterior $p(f \mid \mathcal{D})$ induced by a parameter posterior $p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta)$, where $\mathcal{D}$ denotes the training dataset.

For instance, Spectral-Normalized Gaussian Processes (SNGP) [98] achieve these distributions by approximating a Gaussian process over the classifier output, aided by spectral normalization on all network parameters. The Laplace approximation [99] estimates a Gaussian posterior over network parameters using an efficient Hessian approximation. This post-hoc method is applied to a point estimate network, enabling multiple output samples drawn per input. La-

tent heteroscedastic classifiers (HET-XL) [100] predict a heteroscedastic Gaussian distribution over the pre-logit embeddings and sample multiple embeddings that are then transformed into class probability vectors.

Unlike the above methods, Dropout [101] and Deep Ensembles [102] do not explicitly construct distributions $q(f)$, but instead directly sample from them, either by performing M repeated forward passes or by training M models, respectively. Shallow ensembles [103] provide a lightweight approximation of deep ensembles, utilizing a shared backbone and M output heads (often referred to as "experts"). A single forward pass yields M logit vectors per input. A deterministic network corresponds to a Dirac posterior in the parameter space.

From the above-mentioned methods, none of them is directly adopted, but we draw inspiration from ensemble methods with a light-weight approximation. However, the conceptional direction chosen for implementation in this thesis is to directly train models to fit the uncertainty distributions, from which can be sampled at a later stage. For practical applications utilizing uncertainty estimates, such as threshold-based rejection (OOD detection), a scalar uncertainty output $u(x) \in \mathbb{R}$ is often required instead of a distribution $q(f)$. To obtain this, aggregators compile the distributions into scalar uncertainty estimates $u(\mathbf{x})$. The adopted techniques include calculating the Bayesian Model Average $\tilde{f}(\mathbf{x}) := \mathbb{E}_{q(f)}[f(\mathbf{x})]$ and using its entropy as the uncertainty estimate $\sigma(\mathbf{x})$, or quantifying the variance of $q(f)$. Mucsányi *et al*. [104] extensively discuss and compare many of these aggregation function giving the scalar uncertainty estimates.

2.3.3 Uncertainty disentanglement

Although the previously discussed methods provide a singular, general uncertainty estimate, literature advocates for a more nuanced approach. These methods decompose the posterior distribution $q(f)$, obtained via any of the aforementioned techniques, into multiple estimators. This decomposition aims to quantify different forms of uncertainty, specifically epistemic and aleatoric uncertainties [105].

Epistemic uncertainty, which arises from insufficient data, can be mitigated as more information is gathered. In contrast, *aleatoric uncertainty* originates from the inherent ambiguity in the data-generating process and is fundamentally irreducible [92]. It is crucial for the estimators to accurately disentangle these uncertainties: the aleatoric estimator should exclusively reflect aleatoric uncertainty, while the epistemic estimator should capture only epistemic uncertainty.

This section evaluates two prominent approaches designed to generate such paired estimators by providing formal definitions for obtaining the related uncertainties.

2.3.3.A Information-Theoretical Decomposition

Following the information-theoretical (IT) decomposition [106]–[108], the entropy of the predictive distribution $p(y|x) = \int p(y|x, f)\, dq(f)$ can be decomposed into

an aleatoric and an epistemic component with

$$\underbrace{\mathbb{H}_{p(y|x)}(y)}_{\text{predictive}} = \underbrace{\mathbb{E}_{q(f)}\left[\mathbb{H}_{p(y|x,f)}(y)\right]}_{\text{aleatoric}} + \underbrace{\mathbb{I}_{p(y,f|x)}(y;f)}_{\text{epistemic}}, \tag{2.22}$$

where $\mathbb{H}_p(y \mid x)(y) \equiv \mathbb{H}(Y \mid x)$ is the predictive entropy and $\mathbb{I}_p(y, f \mid x)(y; f) \equiv \mathbb{I}(Y; F \mid x)$ the mutual information. The aleatoric component quantifies the spread of the labels as the average over the plausible model posteriors, while the epistemic component captures only the disagreement among the predictions $p(y \mid x, f)$ across different models f.

This thesis adopts information-theoretical methods within Bayesian frameworks, where interpreting uncertainty across model predictions is fundamental. By leveraging the alignment of these methods with Bayesian and probabilistic interpretations, this approach facilitates a robust quantification of uncertainty.

2.3.3.B Bregman Decomposition

For completeness, Bregman decompositions [95], [109], [110] employ not only the posterior distribution $q(f)$, which each method computes internally, but also consider the ground-truth generative process $p_{\text{gt}}(x, y)$. These methods analyze the expected loss across all possible training datasets. The loss function D_F used in this analysis is a Bregman divergence function, which is more general and can take forms such as the Euclidean distance or the Kullback-Leibler (KL) divergence. This Bregman decomposition is specified by

$$\begin{aligned}\underbrace{\mathbb{E}_{q(f),p_{\text{gt}}(y|x)}\left(D_F\left[y \,\|\, f(x)\right)\right]}_{\text{predictive}} = \underbrace{\mathbb{E}_{p_{\text{gt}}(y|x)}\left[D_F\left(y \,\|\, f^*(x)\right)\right]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{q(f)}\left[D_F\left(f(x) \,\|\, f(x)\right)\right]}_{\text{epistemic}} \\ + \underbrace{D_F\left(f^*(x) \,\|\, f(x)\right)}_{\text{bias}}.\end{aligned} \tag{2.23}$$

The aleatoric uncertainty, as represented by $f^*(x) = \mathbb{E}_{p_{\text{gt}}(y|x)}[y]$ which is the Bayes predictor, corresponds to the Bayes risk (uncertainty in our case) of the generative process. This risk is inherently irreducible and remains unaffected by the posterior $q(f)$. In practical settings, where the generative process is unknown, the aleatoric term is estimated by $\mathbb{E}_{q(f)}\left[H_{p(y|f,x)}(y)\right]$. Epistemic uncertainty, similar in concept to the IT decomposition, is measured as the average distance between the posterior samples $f \sim q(f)$ and their centroid $\bar{f}(x) = \arg\min_z \mathbb{E}_{q(f)}\left[D_F(z \,\|\, f(x))\right]$. Although this average is computed in a dual space, in some instances, it finally is equivalent to the Bayesian Model Average (BMA) [110]. Additionally, to complete the theoretical framework, Bregman decompositions incorporate a third term referred to as the bias. This term accounts for uncertainties related to the function class, extensively discussed by Von Luxburg and Schölkopf [111].

This concludes the background introduction to uncertainty quantification which is extensively utilized in Chapter 4 and Chapter 6.

Chapter 3 CADe in pancreatic ductal adenocarcinoma

3.1 Detecting pancreatic cancer

Pancreatic cancer, particularly pancreatic ductal adenocarcinoma (PDAC), is one of the most aggressive and lethal forms of cancer. Early detection is crucial for improving patient outcomes, yet it remains challenging due to the disease's asymptomatic nature in its initial stages and the pancreas's deep location within the abdominal cavity. The development and implementation of advanced Computer-Aided Detection (CADe) systems can play a pivotal role in enhancing the accuracy and timeliness of PDAC diagnosis.

Early detection and characterization of pancreatic tumors is one of the most promising strategies to improve the prognosis and Overall Survival (OS) of pancreatic cancer patients [112], [113]. This can be attributed to several factors. First, patients diagnosed in early disease stages often have smaller tumors with less vascular involvement. Therefore, they present a much higher 3-year survival rate (82%) compared to patients diagnosed in later disease stages [114].

Second, pancreatic imaging requires specific expertise and radiologists often have to rely on other patterns due to small and iso-attenuating tumors sometimes being barely visible, which might indicate malignant disease. Although CT and MRI generally achieve acceptable sensitivity measures in diagnosing pancreatic cancer, subtle pancreatic changes may be missed on abdominal imaging, especially in asymptomatic patients [115], [116]. Radiologists' sensitivity to detecting small and iso-attenuating PDACs with sizes smaller than 2 cm on CT has been reported to be between 58–77% [117]. Lack of expertise may result in delayed recognition and refrain patients from curative treatment. This may be even more pertinent in hospitals without specific pancreatic expertise [118].

Third, over the years, various studies have reported the presence of visible secondary features, prior to actual diagnosis [119]–[121]. Kang *et al*. demonstrated that secondary signs are present in 88% of the cases. The most common secondary sign was pancreatic duct dilation, and vascular invasion was the most commonly missed [116]. In addition, studies reported that indicative changes of PDAC are visible on imaging 6-18 months prior to actual diagnosis in 50% of patients [120], [122].

Fourth, pancreatic cancer treatment is centralized, which limits the expertise to certain hospitals. Previous studies have shown that patients with non-

metastasized pancreatic cancer had a greater likelihood of receiving surgical treatment when the diagnosis was established in an expert pancreatic cancer center, compared to non-expert hospitals [123]. Centralization of pancreatic surgery may further enhance this discrepancy between expert and non-expert hospitals. Multidisciplinary team meetings may preserve the expertise in various treatment techniques; however, patients still need to be identified before expert assessment can be performed. Due to late recognition, most patients advance to late stages of the disease or even metastases. Pancreatic tumor detection using CT imaging is considered to be the gold standard for the detection of pancreatic cancer [124]. The obtained accuracies of pancreatic ductal adenocarcinoma (PDAC) detection using CT imaging or other radiological imaging techniques largely depends on radiological expertise. Lack of such expertise may result in delayed recognition, which is problematic since only 20% of patients at the time of diagnosis are eligible for resection [125]. Therefore, early detection of pancreatic cancer has significant potential to enable surgical treatment and improve treatment outcomes.

Summarizing the above discussion, the following key aspects are considered as crucial in pancreatic cancer treatment.

- Early detection is vital for safeguarding any survival chance.
- CT remains the gold standard for imaging the disease.
- Pancreatic cancer is not easily visible, experts refer to secondary tumor-indicative signs for early detection.
- Clinical expertise to detect pancreatic cancer is centralized, both in terms of hospitals and medical experts.

Because of the multi-faceted nature of the problem, CADe systems provide an important and interesting solution direction to the medical expert.

- Many developments towards a CADe system for pancreatic cancer detection have been attempted. Understanding their current performances and the research directions will highlight aspects they lack or why the methods have not been adopted in clinical practice. *Providing a broad overview of State-of-the-art (SOTA) techniques for pancreatic CADe systems and approaches is important for further development and creating sufficient background.*
- The input to the CADe system will be CT images of good quality acquired with modern equipment. However, the pancreatic tumor is not directly visible in such images. Following the clinical way-of-working, the CADe system has to incorporate secondary signals to improve detection performance and provide clinicians with reliable and useful input. *Enabling a CADe system to acquire and utilize these secondary signs for improved detection performance is a valuable research question that the research in the chapter aims to solve.*
- Knowing a system that can benefit from such secondary signals and acquiring them still poses a challenge in optimally combining the separate parts. *A complete end-to-end approach is desired, starting with a newly acquired CT scan, capturing the relevant tumor-indicative features, providing them to a detection algorithm to finally make an accurate tumor assessment.*

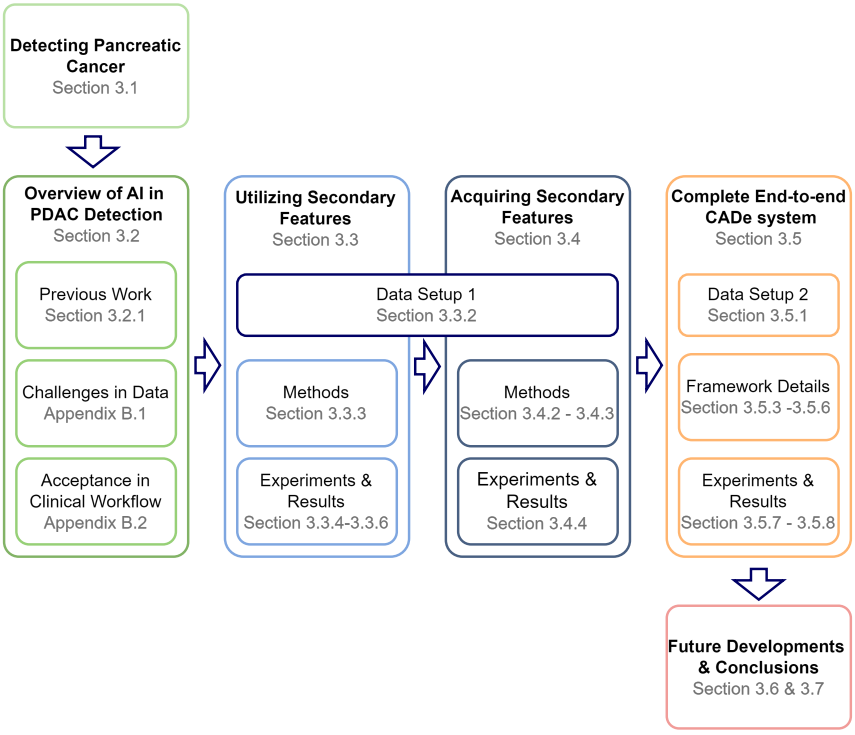


Figure 3.1 *Schematic depicting the development of a CADe PDAC detection framework discussed in this chapter.*

This chapter provides a comprehensive overview of the development of a CADe framework for PDAC detection, as illustrated in Figure 3.1. It begins with an introduction to pancreatic cancer and the clinical context of pancreatic cancer. The significance and challenges of early PDAC detection are discussed, followed by a review of AI-based CADe methods to enhance detection accuracy. The chapter details a novel method utilizing secondary tumor-indicative features for detection, including the data collection, model architecture, and experimental results. With the hypothesis of the value in the secondary tumor-indicative features established, a multi-stage approach for segmentation of the relevant features is explored and robust detection framework using a Residual 3D U-Net architecture is presented. Additional sections cover data collection for algorithm development and the challenges in PDAC detection, concluding with future directions and a summary of the chapter's findings.

3.2 The significance of AI in PDAC detection

Artificial Intelligence (AI) has become an indispensable resource in modern healthcare, particularly in the realm of medical imaging. Initial endeavors to automate pancreatic tumor detection using CT scans have set the groundwork for more sophisticated AI applications. In developing CAD systems for PDAC detection, two

principal methodologies are primarily utilized: deep learning and radiomics. Each approach offers unique benefits and, when integrated, can significantly enhance the performance and reliability of pancreatic tumor detection systems. This synergy between AI techniques holds promise for revolutionizing PDAC diagnostics, thereby improving patient outcomes through earlier and more precise detection.

3.2.1 Previous work in PDAC detection

This section provides a comprehensive overview of the current AI applications in the radiological detection of pancreatic cancer, addresses the existing challenges in clinical implementation, details the state-of-the-art in CAD, and outlines prospective developments in the field.

As a conventional technique, radiomics involves the extraction of a vast array of handcrafted image features, known as radiomics features, from digital images. These features form the foundation for traditional machine learning models to predict and analyze the underlying data.

In recent years, several studies have showcased promising AI methodologies for detecting pancreatic cancer from CT scans. A synthesis of the most impactful studies and their results, along with comparisons of different AI architectures, is presented in Table 3.1. Although these studies have yielded significant findings, it is crucial to recognize the inherent limitations in some research, particularly those focusing solely on binary classification of CT scans. Such studies typically differentiate only between the presence or absence of a tumor, which may not suffice for clinical decision-making where tumor localization is critical.

To provide patients with adequate treatment, clinicians often require additional results, such as the location of the tumor. Moreover, AI-based tools can be leveraged to distinguish PDAC from auto-immune pancreatitis. Various studies have demonstrated impressive results in distinguishing auto-immune pancreatitis from PDAC exploiting both deep learning-based and radiomics-based approaches [126], [127]. Recently, Rigiroli *et al*. [128] took a first step in investigating whether tumor-related and alternative CT radiomic features improve preoperative assessment of arterial involvement in patients with surgically proven PDAC. The model showed a sensitivity and specificity of 0.620 and 0.770, respectively, and a higher performance compared to the radiologist's assessment.

Quantitative MRI, such as T1 and T2 image mapping, allows for accurate tissue characterization and provides early indicators of biological changes [140]. Additionally, it offers a non-ionizing radiation alternative. However, availability of high-quality MRI data is limited and literature on the detection of PDAC using MRI is scarce. Kaissis *et al*. have applied machine learning to MRI images to preoperatively predict survival and molecular subtypes in patients with PDAC [141]–[143]. Their survival prediction model achieves impressive results with a sensitivity and specificity of 0.870 and 0.800, respectively, and an area under the curve (AUC) of 0.90 for the prediction of above-median vs. below-median overall survival (OS) [141]. Liang *et al*. [144] have specifically aimed at develop-

Table 3.1 *List of recently published studies on models for detecting and differentiating PDAC, concluding with the proposed model (N.A.:Not Available).*

Reference	Aim of model	Type of model	Dataset (n)	Sensitivity (%)	Specificity (%)	AUC	Accuracy (%)
Chu *et al*. [129] (2019)	Differentiating PDAC	3D U-Net	156 PDAC, 300 Control	94.1	98.5	N.A.	N.A.
Liu *et al*. [130] (2019)	Detection of pancreatic cancer	Faster R-CNN	338 Patients	N.A.	N.A.	0.963	N.A.
Zhu *et al*. [131] (2019)	Detection of PDAC	3D U-Net	136 PDAC, 303 Control	94.1	98.5	N.A.	57.3
Chu *et al*. [132] (2019)	Differentiating PDAC	Radiomics	190 PDAC, 190 Control	100	98.5	0.999	99.2
Liu *et al*. [133] (2020)	Differentiating tissue	VGG-CNN	370 Cancer, 320 Control	97.3	100	0.997	98.6
Zhang *et al*. [134] (2020)	Detection of pancreatic cancer	Custom CNN	2,890 CT images	83.7	91.7	0.945	90.2
Ma *et al*. [135] (2020)	Differentiating PDAC	Encoder only CNN	222 PDAC, 190 Control	91.6	98.3	0.965	95.5
Si *et al*. [136] (2021)	Detection of pancreatic cancer	ResNet & U-Net	319 Patients	86.8	69.5	0.872	87.6
Qiu *et al*. [137] (2021)	Diagnosis analysis of PDAC	Radiomics	312 Patients	N.A.	N.A.	0.880	81.2
Ebrahimian *et al*. [138] (2022)	Differentiating lesions	Radiomics	103 Patients	84.0	95.0	0.990	91.0
Alves *et al*. [139] (2022)	Detection of PDAC	3D U-Net	119 PDAC, 123 Control	N.A.	N.A.	0.909	N.A.
Proposed Model Section 3.3	Detection of PDAC	3D U-Net	99 PDAC, 97 Control	99.0	99.0	N.A.	N.A.

ing a deep learning algorithm allowing automatic segmentation of gross tumor volume and reported performances, similar to expert radiation oncologists [144]. Over the years, only few other studies have reported MRI-based machine learning models and mainly have focused on identification and characterization of pancreatic abnormalities. An overview of the most notable results is provided by Table 3.2.

From the described tables and overview, it can be observed that CT is the preferred imaging method for developing PDAC detection algorithms. This is likely due to its superior resolution and comprehensive anatomical detail for the pancreas, availability of data for training algorithms, and alignment with the clinical way-of-working. Segmentation methods demonstrate similar performance to classification-based methods in terms of detection accuracy. Moreover, segmentation techniques offer the added benefit of providing precise tumor localization, which is critical for planning follow-up treatments. By generating detailed tumor maps, these methods facilitate more targeted and effective therapeutic interventions, ultimately improving patient outcomes in the management of PDAC. The integration of segmentation methods into clinical practice enhances the ability to monitor tumor progression and response to treatment over time.

In addition to the previously mentioned advantages of segmentation-based approaches for PDAC detection, providing information about secondary tumor-indicative features is highly valuable for clinicians during diagnosis [149]. These features assist radiologists in assessing the presence, size, shape, and potential extent of the tumor's involvement with surrounding tissues. It is hypothesized that these secondary features could also enhance the performance of automated detection algorithms for PDAC. In the next section, we explore this hypothesis by implementing a deep learning-based segmentation algorithm for PDAC that incorporates these secondary features, aiming to improve detection accuracy and provide comprehensive diagnostic information.

3.3 PDAC detection by utilizing clinically-relevant secondary features

Initial diagnosis of pancreatic tumors through CT imaging maintains acceptable sensitivity measures of around 90% for pancreatic cancer diagnosis [150]. In general, pancreatic tumors appear hypodense (darker in the CT image) compared to normal pancreatic parenchyma. However, indeterminate CT findings such as small tumor size, growth pattern, iso-attenuating pancreatic cancer and the difficulty in differentiating from chronic pancreatitis, can make accurate delineation of viable tumor tissue a troublesome task [151]. In addition, pancreatic cancer often causes non-specific symptoms prior to developing into an advanced stage. Therefore, it is important to identify secondary features which may indicate disease to improve early detection of PDAC.

CAD techniques hold great promise in enabling the early detection of PDAC.

Table 3.2 *Studies investigating AI-based pancreatic cancer diagnosis using MRI. WHO: World health organization, PDAC: Pancreatic Ductal Adenocarcinoma, IPMN: Intraductal papillary mucinous neoplasm, MFP: Mass-forming pancreatitis, pNET: pancreatic neuroendocrine tumor.*

Reference (year)	Aim of model	Model type	Dataset (n)	Sensitivity (%)	Specificity (%)	AUC	Accuracy (%)
Gao *et al.* [145] (2019)	Histopath. WHO grade pred. pNET	CNN Encod.	96 patients	N.A.	N.A.	0.885	81.1
Corral *et al.* [146] (2019)	Identify neoplasia in IPMN	CNN feature repres.+ SVM	139 cases	75.0	78.0	0.780	N.A.
Gao *et al.* [147] (2020)	Differentiate pancreatic diseases	Inception v4 (CNN)	398 patients	N.A.	N.A.	0.864	76.8
Deng *et al.* [148] (2021)	Differentiate PDAC from MFP	Radiomics	52 PDAC, 13 MFP	N.A.	N.A.	0.945	79.5

Such a tool allows for expert knowledge to be captured and shared, which can be used when the patient is first screened for the disease. Deep learning-based CAD methods have achieved impressive results in recent years. For these methods to be successfully adopted in the clinical environment, it is necessary to provide more than the standard "black-box" machine learning model [152], [153]. For clinical acceptance of this technology, on top of high detection accuracies, it is essential to provide additional insights into the model's operation.

In this research, we propose a PDAC segmentation model that utilizes the same visual cues in the surrounding anatomy that experts use when looking for the presence of PDAC. This focus and way of working is to maximally leverage easily accessible external information and fully exploit clinical expertise, to ultimately optimize classification and localization performance. Since we start from the radiologists' reasoning, the proposed method becomes more clinically meaningful. For instance, a clinician pays close attention to pancreatic ductal size as a large (potentially dilated) duct could be indicative of tumor. Compared to normal pancreatic tissue in a CT scan, pancreatic cancer appears less visible as an ill-defined mass. It enhances poorly and is hypodense between 75% and 90% of arterial phase CT cases. For this reason, experts utilize secondary features which may be predictive of pancreatic cancer. These include, but are not limited to: ductal dilatation, hypo-attenuation, ductal interruption, distal pancreatic atrophy, pancreatic contour anomalies and common bile duct dilation. For a detailed description of these indicators, we refer to the work by Zhang *et al*. [149].

As these secondary features offer crucial information to experts during analysis, we hypothesize that a deep learning-based CAD method could also explicitly leverage this information. As such, we enrich the input of a 3D U-Net [154] segmentation model with an indication of the external secondary features and observe state-of-the-art results in PDAC detection. In this study, we validate the hypothesis that incorporating secondary features significantly enhances PDAC detection. This model is trained using ground-truth annotations of these secondary features. The performance is then evaluated by testing the model with manually annotated secondary features as inputs. In Section 3.4, we further investigate methods for automatically extracting such secondary features and integrating them into the developed model.

Because of the broad nature of the provided overview, the most-influential and recent papers are recapitulated (Section 3.3.1) as a reference for the proposed method. The data collected to train the proposed methods is discussed in Section 3.3.2. Section 3.3.3 introduces the proposed method for PDAC segmentation and Section 3.3.4 and Section 3.3.5 details the experiments executed to test the proposed approach. Finally, the results of the conducted experiments are presented in Section 3.3.6.

3.3.1 Related work on PDAC detection

Invaluable research towards automated PDAC detection has also been conducted. In addition to the comprehensive overview of AI-based pancreatic cancer detec-

tion methods discussed in Section 3.2.1, we highlight a few recent methods in detail with the same objective as the proposed method.

Recently, both Liu *et al*. [155] and Si *et al*. [156] implemented a patch-based PDAC classification of CT volumes. These classification methods show high accuracy, but clinicians require more interpretable results, such as an indication to the tumor area, as discussed in Section 3.2.1. A Multi-Scale Coarse-to-Fine Segmentation method is proposed by Zhu *et al*. [157] that makes use of three U-Net-like segmentation models at different resolutions in a segmentation-for-classification approach. The output of the three networks are combined using a connected-component graph between the adjacent tumor-positive voxels. Finally, false positive components from the graph are pruned and the tumor voxels are selected based on empirically selected thresholds. We refer to the work by Zhu *et al*. as previous SOTA with a sensitivity of 94.1% and a specificity of 98.5%. Similarly, Alves *et al*. [158] have proposed a segmentation-for-classification approach that makes use of four nnU-Net-based models [13] to ultimately detect the presence of the tumor. Although these methods achieve impressive results, the engineering nature of the solutions lack transparency, sufficient motivation from a clinical perspective, and suffer from long inference times. We propose a more intuitive, clinically-motivated method for PDAC detection. The proposed approach utilizes clinically-relevant cues to realize SOTA detection scores while significantly simplifying the network architecture, making it more suitable for deployment at medical centers.

3.3.2 Data collection for PDAC detection

In this retrospective single-center research study, we collected contrast-enhanced CT images of 97 control CT scans and 99 scans with PDAC located in the pancreatic head from the Catharina Hospital in Eindhoven (CZE), The Netherlands. Patients aged 18 years or above who underwent surgical treatment at the CZE for pancreatic head cancer, were eligible when both a surgical report and a complete pathology report were available. All CT scans were manually annotated in preparation for this research. Relevant anatomical structures (tumor, pancreas, pancreatic duct and common bile duct) were annotated by a surgical resident and supervised by an expert abdominal radiologist, using IntelliSpace Portal [1]. Patients in the control group were derived from a previous randomized control trial in which patients with esophageal cancer were included. These patients all had a CT scan as an initial preoperative action.

The external secondary features play an important role in the expert radiologist's decision-making w.r.t. tumor presence, size and location. As such, significant annotation effort was spent not only on the tumor, but also these indicative features. Two important factors that arose during this process were: (1) how to annotate some of the structures that belong to the same organ (pancreatic duct

[1]Software package available from Philips Healthcare, The Netherlands.

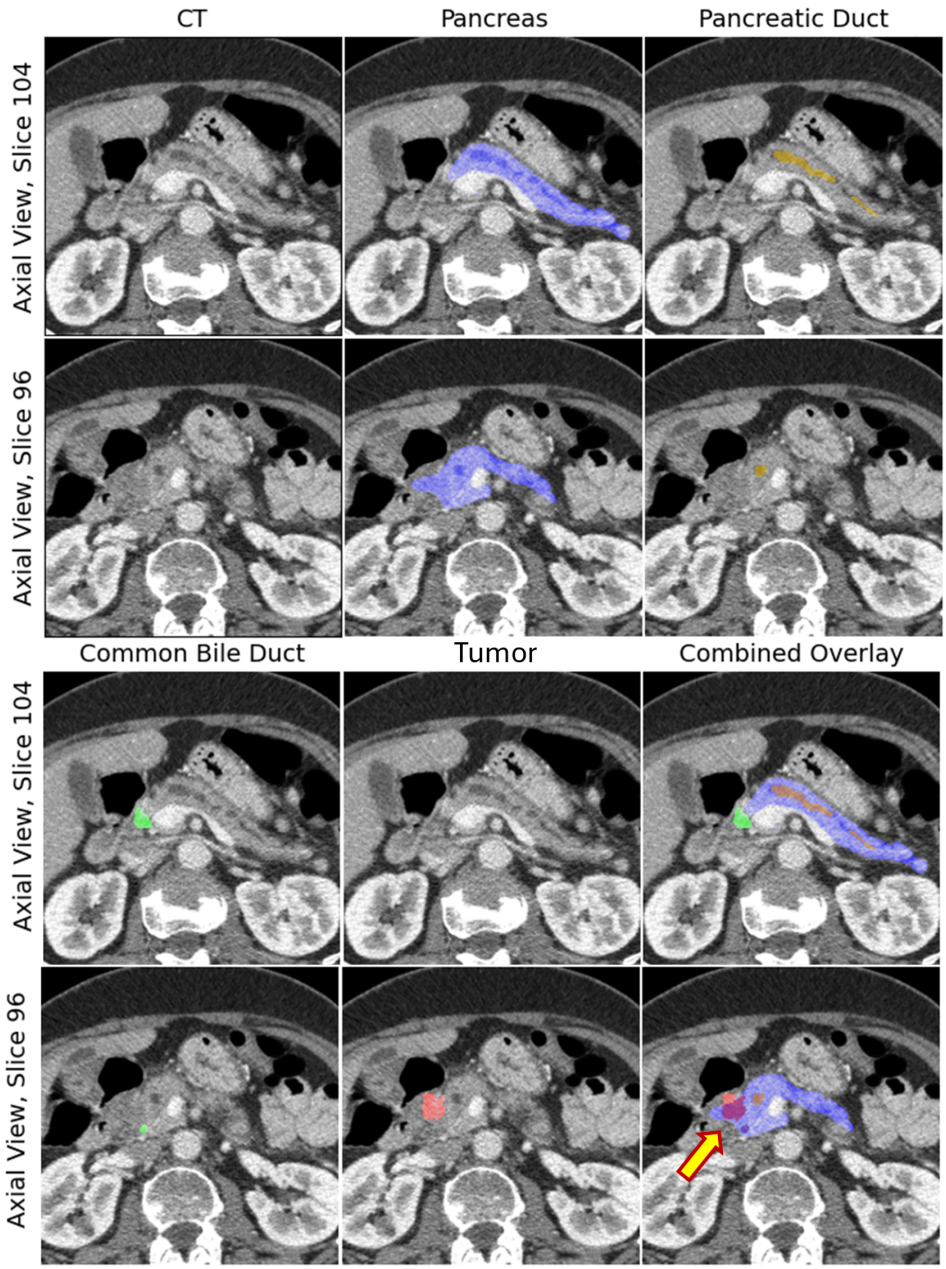


Figure 3.2 *Two slices from a case highlighting the involvement of the different structures and dilated ducts caused by the tumor blockage. These features are indicative of the tumor presence and its location. The pancreas is depicted in blue, dilated pancreatic duct in yellow-ish, dilated common bile duct in green and the tumor, causing the blockage and ductal dilation, is depicted in light red. The bottom-right image shows the involvement between the tumor (indicated by yellow arrow) and pancreas in dark red.*

inside the pancreas), and (2) how to treat cases at locations where a gradual transition from one structure to the other occurs. The latter occurs when the common bile duct enters the pancreas, but importantly, also at the borders of the tumor itself. We decided that each structure should be annotated and stored individually to preserve maximum information. However, this implied that CT voxels could potentially belong to multiple structures simultaneously. Figure 3.2 depicts an example case and corresponding ground-truth annotations. The last image at the bottom-right shows the overlap between the pancreas and the tumor and the pancreatic duct in the pancreas.

In addition to our proprietary dataset, we utilize the publicly available Medical Decathlon [159] (MD) dataset in this research. As part of the MD dataset, Task 07 involves the segmentation of the pancreas and pancreatic masses (intraductal papillary neoplasms, pancreatic neuroendocrine tumors, or pancreatic ductal adenocarcinoma). This dataset consists of patients with often well-developed late-stage disease. As a result, there is a high proportion of large tumors and easily detectable cysts in this dataset. In addition, due to the extensive disease and associated symptoms, many cases contain metal stents, which could incur a bias in a learning algorithm. To the best of our knowledge, this is the only publicly available dataset that aims to detect pancreatic cancer, and although very valuable, it is still a step away from being an ideal dataset for training a deep learning-based CAD system for detection of PDAC. To provide some insight into how the proposed approach competes against other methods on this public benchmark, we have supplemented 10% of this dataset's training set (28 cases) with suspected adenocarcinoma in the pancreatic head with separate annotations for the pancreatic duct, common bile duct, the full pancreas (unobstructed by the tumor) and the tumor. This subset will be used as an additional, extra critical unseen test set in the upcoming experiments[2].

3.3.3 Model architecture

By now, it should be evident to the reader that detecting and accurately delineating PDAC is a complex and challenging task, even for experts. The detailed data and meticulous annotations described in the previous section highlight many of these difficulties, underscoring the need for a robust method to address them. Therefore, the development of advanced techniques to effectively manage these intricacies is crucial for improving diagnostic accuracy and clinical outcomes.

Bearing in mind the subtle nature and development of PDAC, a CAD system should utilize any available information to maximize tumor detection performance, but also provide the necessary assistance in the early diagnosis of pancreatic cancer, in a clinically-interpretable way.

The objective is to develop a segmentation method suitable for both classification and localization. A standard 3D U-Net (depicted in Figure 3.3) is employed that takes the segmentation maps as input, capturing the external indicators, to

[2]Newly annotated data: https://github.com/cviviers/3D_UNetSecondaryFeatures

segment the tumor in the CT volume. These external secondary features are used by expert radiologists to identify and localize the tumor, but can much more easily be obtained and identified in a non-expert setting with minimal effort. In practice, the secondary features can also be obtained by a preceding segmentation model to streamline the process even further. Taking the difficulties related to accurate segmentation of pancreatic tumor into account (even for an expert radiologist), our objective is not to acquire a detailed segmentation map. Instead, we aim for a global indication of where the tumor is located. A detailed, segmentation network or radiologist can then initiate follow-up work. In this chapter, we first implement a coarse detection solution with support of a radiologist, and later in a fully automated manner.

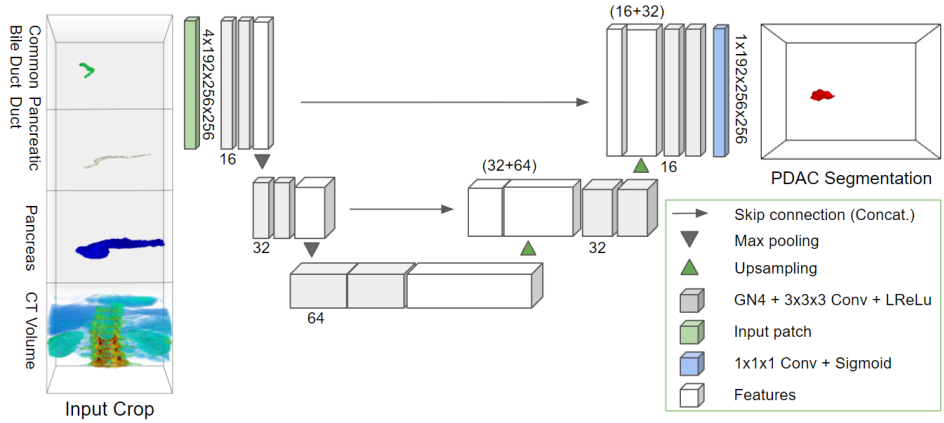


Figure 3.3 *Diagram of the 3D U-Net used for tumor segmentation in abdominal CT scans, provided with detailed external secondary feature segmentation maps. The 3D convolutional filters are reused across the different channels containing the CT and secondary features (Pancreas, Pancreatic Duct and Common Bile Duct) of the input data.*

The 3D U-Net is 3 layers deep with 16, 32 and 64 convolutional filters at each layer. A sigmoid activation function is employed to convert the model predicted logits to tumor confidence values. The model is largely based on the standard U-Net architecture.

3.3.4 Experiments

To evaluate the proposed approach and the extent of the influence of the external secondary features in the tumor detection, the following experiments are conducted. (1) We start by setting the baseline at detecting a tumor using only the CT scan. This baseline is established using the popular nnU-Net [13] (Full-Resolution 3D) and a the already proposed 3D U-Net. (2) In a follow-up experiment, we add the detailed segmentation maps of the pancreas and ducts to the CT scan, in a concatenated channel-wise fashion. The same 3D U-Net is trained to segment the tumor, but now with this additional information derived from the radiologist.

(3) As an ablation experiment, we replace the segmentation maps of the ducts with a Boolean input. The pancreatic and common bile ductal 3D volumes are replaced with unity values if they are dilated. (4) Finally, we apply the models, using the CT scan and detailed segmentation maps, trained and validated on the three datafolds of the proprietary dataset, to ultimately test the model performance with the Medical Decathlon Dataset as test set.

3.3.5 Data preparation and training details

The radiologist starts the investigation for a tumor by localizing the pancreas in the CT scan. Once the pancreas has been located, the radiologist slides through scans looking for the various aforementioned indicative secondary features of the cancer. As such, we preprocess our data in the same flow as the expert's way of working. The radiologist derives a detailed segmentation map of the pancreas, pancreatic duct and common bile duct from the abdominal CT scan and uses them as the secondary features. In practice, this is performed by a prior segmentation model, but since this is outside the scope of this initial step, we use the ground-truth detailed segmentations provided by the expert radiologist as input. We crop the CT scan and corresponding labels, uniformly spaced around the pancreas' center. The crop is shaped within the dimensions [192, 256, 256] in the z, x, y-axes, respectively. Additional resampling and normalization is performed, as described in the work by Isensee et al. [13] prior to cropping. We stack the CT scan, pancreas and two ducts channel-wise along a 4th dimension in preparation for training. Our final dataset is: $\mathbf{D} = \{(\mathbf{X}_1, \mathbf{Y}_1), ...(\mathbf{X}_N, \mathbf{Y}_N)\}$, with N being the dataset size, where $\mathbf{X}_n \in \mathbb{R}^{C \times Z \times W \times H}$ is the 4D volume of input data and $\mathbf{Y}_n \in \mathbb{R}^{Z \times W \times H}$ is the 3D tumor segmentation map.

In our implementation, we perform threefold cross-validation using a random 70/30% training and validation split and report results on the validation sets and the MD dataset as test set. The custom 3D U-Net is implemented in PyTorch and extends on the work by Wolny *et al.* [160]. During training, we only employ a cross-entropy loss, a batch size of 2, an Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. We use extensive data augmentation, consisting of random flipping, random rotation, elastic deformation, contrast adjustment, and additive Gaussian and Poisson Noise. In all experiments, the same crops, hyperparameters and augmentation techniques are used, with a hardware configuration based on a TITAN RTX GPU[3].

3.3.6 Results and discussion

The experimental results are listed in Table 3.3. In all cases, the model outputs are binarized (standard threshold setting of 0.5) and converted to segmentation maps. For the classification metric, if the resulting segmentation prediction overlaps with the ground-truth tumor label, even partially, we consider it a true positive prediction. If there is yes (no) prediction and a tumor without overlap, it is a false

[3]Commercially available from Nvidia Corp., Santa Clara, California, USA

positive (negative). In the case there is no prediction whatsoever and the tumor is absent, we consider it a true negative. The sensitivity, specificity and average DSC values (across all the tumor-positive cases) on the validation sets are reported. We also show the results of the model using the full input (CT scan and detailed segmentation maps) and apply it to the test MD dataset.

Data Input	Model	Sensitivity	Specificity	DSC
CT Only	nnU-Net	0.92 ± 0.02	0.27 ± 0.16	0.42 ± 0.04
CT Only	3D U-Net	0.98 ± 0.03	0.11 ± 0.10	0.40 ± 0.07
Binary Ducts	3D U-Net	0.83 ± 0.24	0.19 ± 0.06	0.16 ± 0.04
Full	3D U-Net	1.00 ± 0.00	0.99 ± 0.02	0.31 ± 0.07
Test MD - Full	3D U-Net	0.99 ± 0.02	N/A	0.31 ± 0.05

Table 3.3 *Results obtained with the nnU-Net and 3D U-Net with different input channel information. Given the limited amount of data, numbers are constrained to two decimals (N/A is not applicable).*

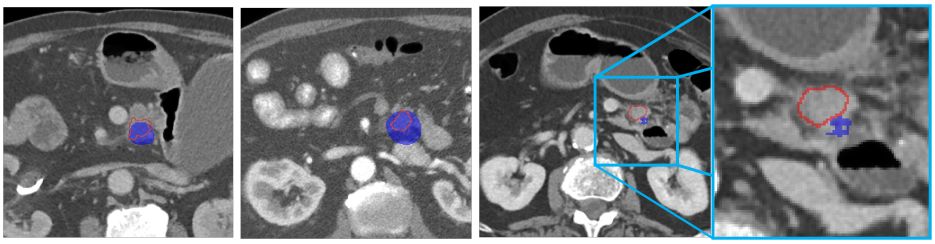
Results on Baseline: The nnU-Net and the proposed 3D U-Net showcase similar performance when trained using only the CT scan as training data. In both models, the network eagerly tries to segment the tumor, even when the tumor is absent, resulting in a low specificity.

Adding Binary Ducts: The segmentation performance does not improve when the model is trained using the additional *binary* labels, indicating the presence of dilated ducts. Hence, duct dilation alone is not a decisive tumor factor and has to be combined with other indicative features from the tumor region.

Detailed Segmentation Maps: The model is provided with the detailed segmentation maps of the ducts and pancreas, along with the CT scan. We observe that the model can learn the connection between these indicative features and the presence of a tumor. The model correctly predicts tumor with an overlap of the ground-truth segmentation in the majority of cases. In a single case, the tumor is predicted to be at a different location than the label (False Positive). We observe a lower DSC value compared to the models with only CT scans as input. These baseline models maximally predict tumors in most cases. This results in a higher DSC value when there is a tumor factually present, at the expense of a large number of false positive predictions. When increasing sensitivity, the model logically locates more tumors, albeit some with low DSC values.

Test Set (MD) Experiments: We observe a very similar, impressive performance when the model is applied to the MD dataset as test set. In two of the three datafolds, the models showcase 100% sensitivity, without missing a tumor. The model from the third fold missed the tumor in one of the cases and made no prediction whatsoever (False Negative). The same tumor was predicted with a relatively high DSC value of 0.40 and 0.26 in the other two models. Averaging the sensitivity across the three models (100%, 100% and 96.43%) explains the 99$\pm$2%

sensitivity at the bottom of the table. Visual example predictions on the test set can be observed in Figure 3.4. Note that the aim of this study is not to achieve maximum segmentation accuracy and rather develop a more effective, clinically-relevant and efficient method for tumor detection. The required inference time using this method is 0.33 s on an RTX 2080 Ti GPU.



(a) *3D DSC: 0.44* (b) *3D DSC: 0.29* (c) *3D DSC: 0.01*

Figure 3.4 *Segmentation performance from three different cases. An example of a low-performing segmentation is visualized (3.4c). The figure is best viewed in color.*

3.3.7 Limitations of the initial PDAC detection model

The secondary features used in this work and provided as external input are acquired from the same CT scan. It is expected that a CNN would be able to extract these embedded spatial features and discover the causality between these features and the presence of the tumor. Unfortunately, this expectation does not hold. Future work should investigate these underlying causal factors and how to enable a CNN to learn this available information.

Additionally, up to this point, the features provided and utilized by the model have been the ground-truth annotations. In the following Section 3.4, we investigate if the same performance can be obtained if the these features are automatically obtained with a segmentation model.

3.3.8 Summary on utilizing secondary features

Despite the eminent success of deep learning networks, even for the detection of PDAC, the method presented in this work demonstrates that external tumor-indicative features can significantly enhance CAD performance. We optimize a segmentation for classification and localization approach, by adding the easily obtainable and clinically valuable external secondary features used by the radiologist, to considerably improve classification performance. The proposed approach consists of a 3D U-Net that takes the CT scan, along with a segmentation map of the pancreas, pancreatic duct and common bile duct as input, in order to finally segment the pancreatic tumor. By integrating these indicative secondary features into the detection process, the proposed method achieves a sensitivity of 99±2% (one case missed), yielding 5% gain over the previous state-of-the-art

method. The proposed method also achieves a specificity of 99% and ultimately requires no sacrifice of specificity in favor of sensitivity. In addition, the method provides further insights into the tumor location and obtains similar segmentation scores on prospectively collected and the Medical Decathlon data. Generally, this research reveals the important value of explicitly including clinical knowledge into the detection model. We suggest that future CAD methods integrate higher orders of feature information, particularly valuable clinical features, into their domain-specific problem to improve performance when such information can be identified. The proposed method paves the way for equipping clinicians with the necessary tools to enable early PDAC detection, with the aim to ultimately improve patient care.

3.4 Automated segmentation of external and clinically-relevant features for improved PDAC detection

In the previous Section 3.3, a neural network is trained for tumor detection through segmentation of early stage PDAC. The model is trained with Computed Tomography (CT) images and manual annotations of the pancreas, the pancreatic duct and the common bile duct. These are indicative clinical features that medical professionals use to assist in the detection of pancreatic tumors. For example, bile duct dilation is a common result of an obstruction in the bile ducts caused by the presence of a tumor. The work indicates that including this pathological response to a nearby tumor as input, improves the tumor detection rates of the AI model.

In the previous work, we have developed a PDAC segmentation model to exploit auxiliary clinical information. To investigate whether such additional clinical features benefit tumor detection, we specifically consider the pancreas, common bile duct and pancreatic duct. Based on this previous work, we develop three-dimensional (3D) segmentation models to segment these clinical features. By segmenting these structures using deep learning models, manual input is not further required to include these features, so that the segmentation becomes fully automated.

In this section, we introduce a multi-stage segmentation workflow designed to automatically gather auxiliary information. The overarching sequential workflow and attributes common to the deep learning models used in this process are detailed in Section 3.4.1. Further specifics on the pancreas segmentation models and ductal segmentation models can be found in Section 3.4.2 and Section 3.4.3, respectively. Finally, the performance of these models and their impact on the final tumor detection models are analyzed and discussed in Section 3.4.4.

3.4.1 Multi-stage segmentation approach

We propose a multi-level coarse-to-fine sequential processing workflow [161] to segment the pancreas and bile ducts, consisting of multiple sequential (initially two) U-Net-based CNNs, to use as input for a tumor segmentation and detection model. For the segmentation models, we use the U-Net with the repository devel-

oped by Wolny *et al*. [160], due to its applicability to various medical image tasks. The models are patch-based networks using patches of size 128×128×128 voxels, with a stride of 32×32×32 voxels, using a unity batch size, trained with the binary cross-entropy loss and the Adam optimizer [162]. The sensitivity, specificity and segmentation accuracy in terms of the Dice Similarity Coefficient (DSC) of each individual model are finally evaluated in the proposed workflow.

In order to limit the search space over the CT scan, the processing sequence commences with coarse pancreas segmentation over the full-image CT scan, which is used to crop the scan to a narrower region that is positioned about the center of the pancreas mask. This crop with a size of 192×256×256 voxels is used as input for fine segmentation of the pancreas through a second model. The crop is refined by the final pancreas segmentation and used as input for segmentation of the bile duct structures. Using the predicted masks of the pancreas and bile ducts, a final model then detects and segments the tumor. Figure 3.5 depicts the proposed complete sequential processing chain and networks.

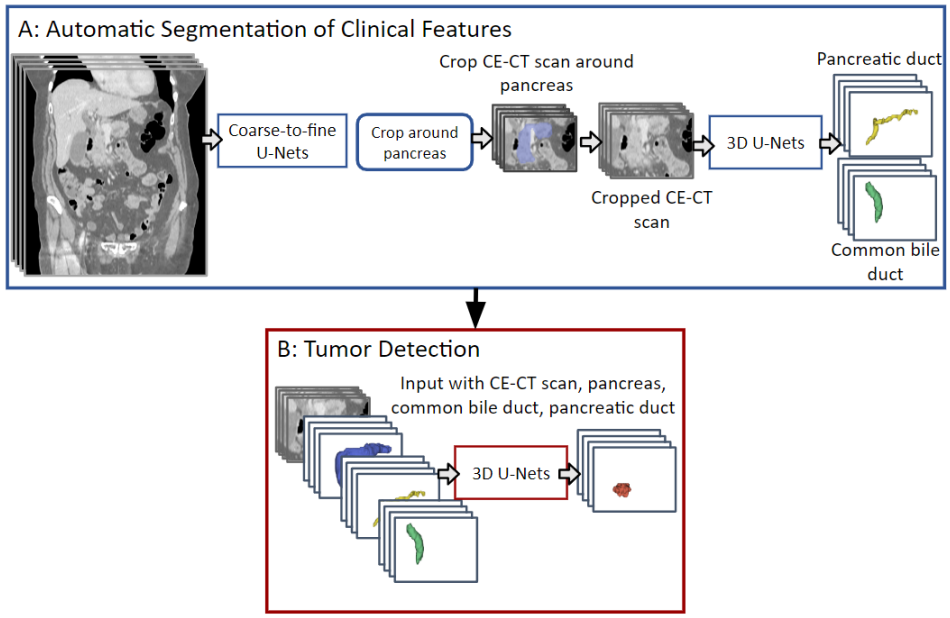


Figure 3.5 *Illustration of the tumor detection processing workflow for processing CT scans with intermediate segmentation of relevant anatomical structures. Block A: The pancreas is segmented using multiple 3D U-Nets in a coarse-to-fine structure. All data is then cropped around the segmented pancreas to a size of 192×256×256 pixels. Using the input of the cropped CT scan, the bile ducts are segmented with another model. Block B: The CT scan and the masks for the pancreas, the common bile duct and the pancreatic duct are then used as inputs for the final tumor segmentation-for-detection model.*

3.4.2 Pancreas segmentation

The pancreas is segmented using two 3D U-Nets in a coarse-to-fine approach. The coarse model processes the full-image CT scan of different sizes and a lower resolution (resampled to twice the target voxel spacing at 2 mm×1.37 mm×1.37 mm), with a patch-based approach of patch size 64×128×128 voxels. The fine pancreas segmentation model is trained with an image crop of the CT scans with the finer spacing. The fine pancreas model also uses a patch-based 3D U-net model trained only over the cropped region of the CT scan. This approach is used to improve the segmentation accuracy of the pancreas. During training, the selection of patches is guided by their relation to the ground-truth mask: if there is no part of the ground truth inside the patch, then the patch is omitted with high probability for further processing.

3.4.3 Bile duct segmentation

For the bile duct segmentation, two different methods are implemented: (1) a multi-label model that segments both bile duct structures and (2) separate single-label models that segment each structure individually. This allows two different ways for segmentation of these structures. Segmentation of both structures with one model potentially provides lower performance than the single-label model. However, the processing time and required memory is lower when using one model.

To reduce the occurrence of bile duct segmentation mask fragments, which are disconnected from the main and largest structure, the *cc3d* package is used to separate segmented components [163]. Components not connected to the largest duct component or without overlap with the pancreas, are categorized as background. In this way, the final bile ducts are localized over the correct region.

3.4.4 Results & discussion

The secondary feature segmentation performance required by the tumor segmentation model to maintain a high detection accuracy, is evaluated. This method, as discussed in Section 3.3, employs manually annotated multi-channel input of the pancreas, pancreatic duct and common bile duct. The performance of the tumor segmentation model is compared using manually annotated and automatically obtained input from the anatomical segmentation models.

Evaluation Criteria: The pancreatic duct and the common bile duct are typically more visible when the ducts are dilated. As a result, for some of the scans in the proprietary dataset, the bile duct is not annotated. Hence, we present the segmentation accuracy in terms of DSC only for the cases with annotated bile ducts. Along with this metric, the specificity and sensitivity with respect to the segmented masks are presented. Since we employ a segmentation-for-detection model for tumor detection, the detection sensitivity and specificity are based on the segmented tumor mask. Compared to Section 3.3.6, here the detection accuracy is computed differently. A positive detection is found if a mask is segmented. If there is a ground-truth mask for the corresponding case, a true positive is ob-

tained, otherwise this would result in a false-positive detection. The predicted segmentation does not need to overlap with the ground-truth label for it to be considered true positive. If no tumor mask is segmented and there is no ground-truth mask, it results in a true negative. However, if there is a ground-truth mask for this image, it is classified as a false negative. Based on the aforementioned cases, the sensitivity and specificity of the tumor detection are determined.

Pancreas & Duct Segmentation: The results of the pancreas segmentation model and the bile duct segmentation models are presented in Table 3.4, in terms of the mean DSC as well as sensitivity and specificity of the bile duct masks, measured over the subset of 28 patients of the MSD dataset. The coarse pancreas segmentation model predicts coarse pancreas masks with a mean DSC of 0.72 ± 0.05, which are used to segment the CT scan for a fine pancreas prediction. The fine pancreas segmentation results have a mean DSC of 0.86 ± 0.03, where the main visible drawback of these predictions is an incorrect shape around the edge of the structure and the mask, as seen in Figure 3.6. The multi-label model bile duct predictions have a higher mean DSC of 0.60 ± 0.08 and 0.71 ± 0.08 for the pancreatic duct and common bile duct, respectively, and detected both structures for all images of the dataset with ground-truth annotations. The multi-label model predicts masks for the pancreatic duct more frequently than the single-label model, which causes it to make more regular predictions where there is no ground-truth mask. The pancreatic duct predictions have a lower mean DSC than the common bile duct predictions. This can be a result of the structure being smaller than the common bile duct, which makes the DSC more sensitive to errors. The specificity and sensitivity values of the bile duct structures indicate whether the models predict bile ducts when there is no bile duct visible (no annotation available). However, the structures that are incorrectly predicted by the model are in a region where there would be a bile duct, so the segmented masks may still be valid. Future research should look into the implications of this effect and aim at more accurate predictions for the use of these masks in other models.

Tumor Detection: The proposed approach, which entails using predicted clinical features as input for the pre-trained PDAC model, achieves an outstanding tumor detection sensitivity of 100% across all the folds on the MSD dataset (see Table 3.5). The tumor detection specificity is not included in this study, as the MSD dataset only contains positive tumor cases. Future studies should focus on investigating the specificity of the model with negative tumor cases. As observed in Table 3.5, tumor segmentation results have a low mean DSC of 0.31. In Figure 3.6, the segmented masks of three slices from a case are highlighted and it is visible that the segmented tumor mask is considerably larger than the ground-truth mask for some cases. During the early PDAC detection stage, obtaining a high detection sensitivity is the primary focus. Although this work employs a segmentation approach, the goal of the model is tumor detection. Accurate delineation of the tumor should be addressed using a refinement segmentation model. For PDAC diagnosis and treatment assessment, the important feature of the tumor is its relation to

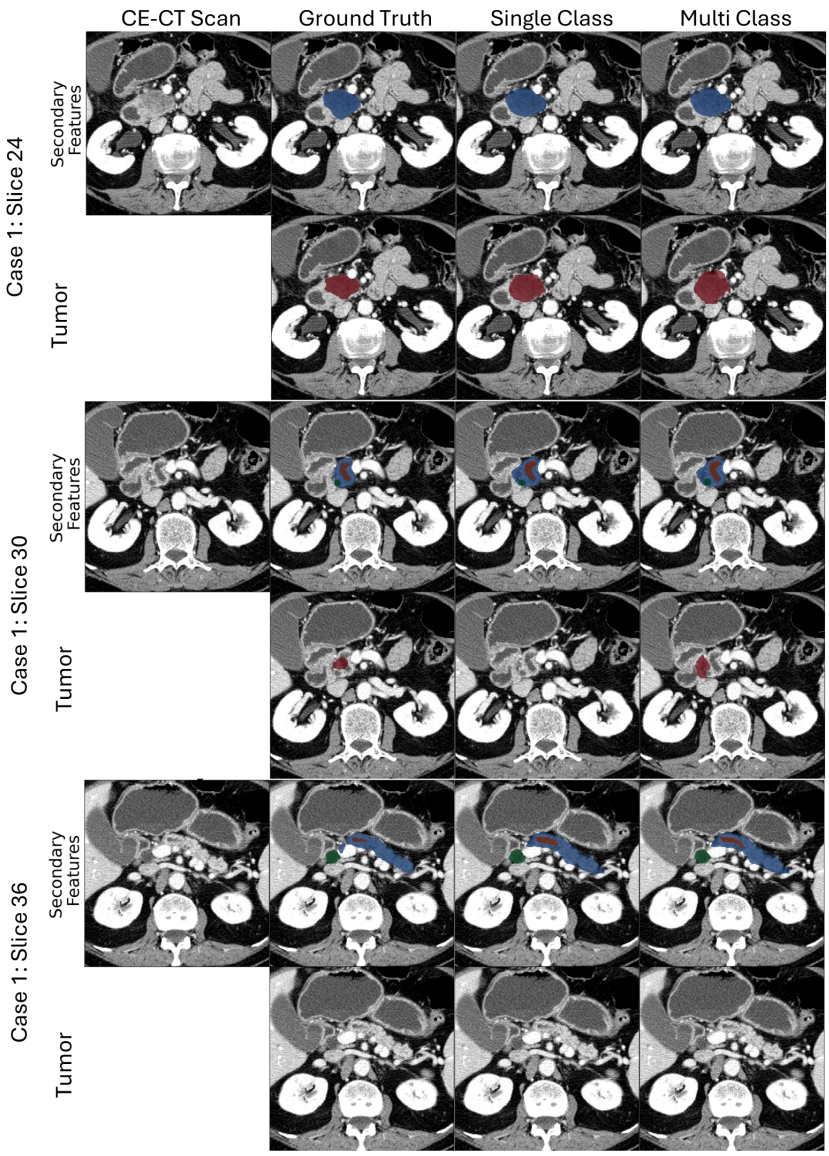


Figure 3.6 *Segmentation of clinical features pancreas (blue), common bile duct (green) and pancreatic duct (orange) as well as tumor (red). Comparison of ground-truth masks, predictions using single-label model and multi-label model for clinical feature segmentation and tumor predictions using the corresponding clinical features. Illustration over three slices. The tumor segmentation obtained with inputs from the single-class model and the multi-class model are compared. It can be seen that the tumor segmentation with input from the multi-class model is more accurate, especially in slice 30.*

nearby blood vessels. This feature is used to determine the resection options for each case. Hence, the follow-up tumor segmentation refinement should have a focus on the relation to nearby structures rather than the absolute correctness in shape.

Table 3.4 *Measured performance of the pancreas and bile duct segmentation models, in terms of mean DSC, detection sensitivity (true positive over all folds / total positive cases over all folds) and specificity (true negative over all folds / total negative cases over all folds), over the MSD dataset.*

Single-class model	DSC	Sensitivity	Specificity
Coarse Pancreas model	0.72 ± 0.05	100%	NA
Fine Pancreas model	0.86 ± 0.03	100%	NA
Pancreatic duct	0.57 ± 0.11	93.3% (70/75)	67% (6/9)
Common bile duct	0.69 ± 0.10	97.5% (79/81)	0% (0/3)
Multi-class model	DSC	Sensitivity	Specificity
Pancreatic duct	0.60 ± 0.08	100% (75/75)	22.2% (2/9)
Common bile duct	0.71 ± 0.08	100% (81/81)	0% (0/3)

Table 3.5 *Performance of PDAC segmentation model in terms of mean DSC over the publicly available MSD dataset. Comparison of methods using manually annotated input of the bile ducts and pancreas and employing bile ducts segmented by single-class and multi-class AI models.*

Method	Tumor DSC	Tumor Sensitivity*
Manual input	0.31 ± 0.05	$99\% \pm 2\%$
Multi-class model input	0.31 ± 0.10	$100\% \pm 0\%$
Single-class model input	0.31 ± 0.09	$100\% \pm 0\%$

* No negative tumor cases in MSD dataset.

The results of this study show the value of using features that are found to be relevant for clinicians in the PDAC detection process. In the review of Elbanna *et al*. [115], several indicative features visible in CT scans are listed, which are used to identify early stage PDAC. Some examples of these features are abruption and dilation of bile ducts, irregular pancreatic contour and vascular encasement or narrowing. Encouraging an learning-based detection algorithms to utilize

these features can potentially improve the detection or segmentation of the disease, since it has proven valuable for clinicians.

3.5 Detection and localization of pancreatic head cancer on CT

This section details the development of an end-to-end system designed to enhance the capabilities of clinicians in diagnosing pancreatic head cancer using computed tomography (CT) scans. The proposed approach leverages a robust sequential processing chain to not only detect, but also localize pancreatic cancer effectively, while ensuring that the results are interpretable in a clinical setting. To achieve this, the system incorporates the analysis of secondary diagnostic signs, such as duct dilatation, which are crucial for improving the interpretation of the imaging results by clinicians. These features help in providing a clearer diagnostic analysis and assist in the differential diagnosis process

Since the design an end-to-end detection system requires a reconsideration of the previous work in this chapter, the upcoming extensive description is divided over multiple subsections. In Section 3.5.1 a critical analysis of the initial PDAC detection approach is discussed and the steps are detailed to enhance the overall robustness and potential clinical efficacy. Section 3.5.2 discusses the improved data collection and labeling strategies for enhanced data quality and diversity, while Section 3.5.3 highlights key clinical features of the new dataset. Section 3.5.4 details the complete PDAC segmentation-for-detection framework. The corresponding deep learning-based model is described in Section 3.5.5 and employs a Residual 3D U-Net architecture. The obtained results for secondary features segmentation are presented in Section 3.5.7. Section 3.5.8 presents the system's PDAC detection performance, followed by a reflection on the results in Section 3.5.9. Finally, Section 3.5.10 addresses current constraints and future improvement areas.

3.5.1 Reconsidering PDAC detection

To ensure clinical efficacy, an end-to-end PDAC detection model must be both mature and rigorously validated. Realizing the limitations in the initial components of the earlier detection system, we reconsider these components and propose a unified PDAC detection framework.

In this framework, we (1) improve the data quality and quantity, (2) employ state-of-the-art architectural model components for increased accuracy and robustness and (3) increase the level of automation of the overall system. This comprehensive system aims to provide a robust tool for the early and accurate diagnosis of pancreatic cancer, facilitating timely and targeted therapeutic interventions.

Enhanced data quantity and quality: Since the objective is the design of a complete end-to-end system, all analysis and guidance should be given by deep learning models. This naturally places an extra strong requirement on the data quantity and quality. The previously employed datasets lacked sufficient diversity and required enhancements in image and annotation quality. The improved and

enlarged datasets include the following aspects.

- *Sufficient dataset size*: The dataset size is increased to ensure that the models are trained on a broad spectrum of cases, thereby improving model robustness and reliability.
- *High annotation quality*: We ensure high-quality annotations by re-annotating previously missed structures to improve the accuracy and consistency of the training data.
- *Diverse patient representation*: A more diverse patient population is incorporated, to better capture variability in the disease presentation, which enhances the generalization of the model.
- *Test sets enable comprehensive metric evaluation*: The inclusion of more patient cases enables a sufficiently large dataset to facilitate an internal test set. This test set is used to measure all relevant metrics, ensuring that the model's performance is thoroughly assessed.

Improved accuracy and robustness: The updated models obtain higher performance and improved robustness in unseen settings, by employing more recent learning algorithms and techniques. These enhancements aspects are as follows.

- *Advanced algorithms and model ensembling*: State-of-the-art algorithms are utilized to improve the model's ability to extract valuable features to the PDAC detection task.
- *Cross-validation*: Rigorous cross-validation techniques are implemented to ensure that the model performs well across different subsets of data.

Increased automation level: The PDAC detection processing workflow is fully automated. This higher level of automation is designed to address the following aspects.

- *Highly reduced human intervention*: Manual input is fully removed to streamline the detection process and reduce the potential for human error.
- *Improved efficiency*: The detection process is accelerated to provide faster diagnostic results, which is valuable for timely therapeutic decisions.
- *Ensured consistency*: The detection results are enhanced in consistency by standardizing the processing steps, which leads to more reliable and reproducible outcomes.

Incorporating the above-mentioned improvements, an end-to-end PDAC detection framework is proposed to enhance early diagnosis and precise localization of pancreatic head cancer using CT scans. This framework leverages deep learning-based segmentation models to effectively detect and localize pancreatic cancer, ensuring suitable clinical interpretation by integrating secondary diagnostic signs like duct dilatation. The above-listed improvements enable the fully automated end-to-end system for PDAC detection and localization.

3.5.2 Additional data collection and labeling

Acknowledging the limitations posed by the initially small dataset in our study, we undertook efforts to expand our data resources and improve the overall quality.

Increased size: We have additionally included a collection of 50 patients with up to two CE-CT images (99 scans) at the CZE, comprising a total of 98 control patients and 99 patients with pathology-proven pancreatic adenocarcinoma. The dataset comprised of medical imaging data and relevant clinical information, including preoperative, intraoperative, and postoperative details obtained from radiology, pathology, and surgery reports. The medical imaging data consisted of CT scans utilizing two separate phases. Therefore, a total of 198 CT scans were included in the PDAC cohort. The included CT scans consist of a portal venous phase and either a parenchymal phase, a late arterial or late liver phase and a portal venous phase, allowing for integration of complementary information when obtaining a cancer prediction for each patient. The CT slice thickness varied between 1.0–3.0 mm, obtained based on accessibility, also aiming to integrate the diversity of scans encountered in standard clinical practice.

Improved annotations: A PhD candidate MD, manually (re-)annotated relevant anatomical structures, consisting of the tumor, pancreas, pancreatic duct, common bile duct, and various arteries (aorta, superior mesenteric artery, celiac axis, common hepatic artery, splenic artery, gastroduodenal artery, aberrant arteries) and veins (vena cava, vena porta, superior mesenteric vein, inferior mesenteric vein, splenic vein). Bile ducts were annotated in both groups when visible. Annotations were performed using IntelliSpace Portal [4] and were supervised by an expert radiologist for pancreatic tumors. Anonymization of subject information occurred during data collection and analysis.

Diverse dataset: The PDAC cohort included individuals aged 18 years and older who underwent surgical interventions for cancer of the pancreatic head at the CZE from 2012 to 2019. Patients were eligible for inclusion if their CT images included at least two phases and were accompanied by both a surgical report and a complete pathology report. Exclusion criteria included patients diagnosed with active pancreatitis at the time of diagnosis, or those with artifacts on CT images, such as metal stents. Meanwhile, the control group was derived from participants in the randomized NUTRIENT-II trial at the same hospital, which included patients diagnosed with esophageal cancer, but who did not have pancreatic tumors. This expanded dataset is expected to enhance the robustness of our findings by improving the representation and diversity of our study sample set [164]. All patients underwent a preoperative CT scan as part of their diagnostic protocol. Section 3.5.3 discusses further clinical characteristics and details of the dataset.

Test datasets: Including the newly collected data, 15% of the complete proprietary dataset is set aside as an internal test set to do evaluation of the developed framework. Additionally, we again use the publicly available Medical Segmenta-

[4]Software package available from Philips Healthcare, The Netherlands.

tion Decathlon (MSD) dataset from the Memorial Sloan Kettering Cancer Center (Manhattan, NY, USA) for model testing [159]. The dataset has already been summarized and explained in Section 3.3.2.

3.5.3 Clinical characteristics

The dataset contains 197 patients with 290 CT volumes. A total of 99 patients are diagnosed with pancreatic head cancer, corresponding with 198 CT volumes. The remaining 98 patients are assigned to the control group with a normal pancreas. Clinical characteristics of the patients, classified by the presence or absence of PDAC, are shown in Table 3.6. For the PDAC cohort, the mean age is 74.9 ± 7.5 years with 52 male and 47 female patients. In total, 21 patients have Stage I PDAC, 55 patients Stage II, 20 patients Stage III, and 3 patients have Stage IV PDAC. A total of 77 patients present with hypoattenuating tumors, 21 with isoattenuating tumors, and 8 with hyperattenuating tumors. Additionally, 52 patients have pancreatic carcinoma, 21 have cholangiocarcinoma, and 24 have ampullary carcinoma. The median tumor size in the dataset is 2.6 cm (range 2.0–3.5 cm).

Table 3.6 *Clinical characteristics of the patients in the PDAC cohort and control cohort. Continuous variables are displayed as mean ± standard deviation or median (interquartile range). The tumor stages are Stage I: T1-2N0 PDAC (medical notation); Stage II: T3 or T1-3N1 PDAC; Stage III: T4 or T1-3N2 PDAC; Stage IV: metastasized. Tumor stages and tumor size are only presented for PDAC patients. N.A. is not applicable, PDAC is Pancreatic Ductal Adenocarcinoma.*

Clinical Characteristics	With PDAC	W/o PDAC
Number of patients (scans)	99 (198)	98 (98)
Age (years)	74.9 ± 7.5	71.2 ± 8.1
Gender (male/female)	52/47	79/19
Tumor size, Median (cm)	2.60 (2.0 – 3.5)	N.A.
Tumor Stage (I/II/III/IV)	21/55/20/3	N.A.
Tumor attenuation on CT Hypo/Iso/Hyper intense	77/14/8	N.A.
Tumor Origin Pancreas/Cholangio/Ampullary	52/21/24	N.A.

3.5.4 PDAC segmentation for detection framework

A multi-stage coarse-to-fine framework is designed to enhance the detection process of PDAC through sequential refinement steps. The proposed methodology delineates the detection into several stages, each aiming to progressively augment the accuracy from a broad to a more detailed analysis. The overall approach is coarsely the same as the steps validated in the earlier sections, but each stage contains a stronger model implementation. Initially, (A-1) the framework employs

a pancreas localization model to determine the coarse position of the pancreas. Subsequently, (A-2) a fine pancreas segmentation model is applied to delineate the organ with higher precision. Following this, (B) the segmentation of the common bile duct and pancreatic duct are conducted, which are crucial for detailed necessary anatomical insights and accurate PDAC localization. The final stage (C) integrates these segmented features into a tumor detection model, which utilizes the refined data to achieve precise tumor segmentation. This hierarchical approach enables systematic progression in detection capabilities, facilitating comprehensive analysis and improved diagnostic accuracy. In addition, it follows the typical clinical workflow when a CT scan is inspected by a clinician for the presence of pancreatic cancer. The following description of these steps is aligned with the indicated steps in Figure 3.7.

A-1: Coarse pancreas segmentation: Starting with a new CT scan, the workflow initiates with the scan resampled to a *coarse* 1.37 mm ×1.37 mm×2 mm voxel spacing. The segmentation process begins with a coarse delineation of the pancreas on this full CT scan. This initial step targets a broad delineation of the pancreas across the full CT scan, functioning to localize the pancreas and yield a subsequent crop to a region specifically centered around this organ. The preliminary segmentation is executed using a patch-based (128×128×64 voxels with a stride of 32×32×32 voxels) 3D U-Net [154] architecture designed for coarse volumetric data analysis.

A-2: Fine pancreas segmentation: Following the coarse localization, the CT scan undergoes a second resampling step to a *fine* resolution of 0.68 mm×0.68 mm×1 mm, preparing it for a more detailed analysis. A cropped volume of 256×256×192 voxels, derived from the location pinpointed in the previous step, is then utilized as the input for a second, more precise pancreas segmentation model. This model is also based on a patch-based 3D U-Net architecture and is specifically trained to perform high-quality, fine-grained segmentation on the finely resampled data. The accurate pancreas-centered crop obtained from this model serves as the basis for subsequent ductal segmentation. Both the coarse and fine models consist of a four-layer deep U-Net, with each layer containing 32, 64, 128, and 256 convolutional filters in a subsequent fashion. However, they are trained on data of different resolutions to optimize their performance at the stage in the processing chain.

B: Secondary feature segmentation: For the segmentation of the bile and pancreatic ducts a multi-class model that segments both duct structures simultaneously are employed. The approach utilize a segmentation strategy that processes patches of size 128×128×128 voxels with a stride of 32×32×32 voxels, thereby promoting significant overlap and ensuring consistency and continuity across the segmented regions. The 3D U-Net consist of a similar four-layer network with each layer containing 32, 64, 128, and 256 convolutional filters. Additionally, the Connected Components 3D package (*cc3d* [163]) is employed to enhance segmentation realism by categorizing non-pancreatic connected components, identified within the bile duct structures as background. This ensures accurate localization of the

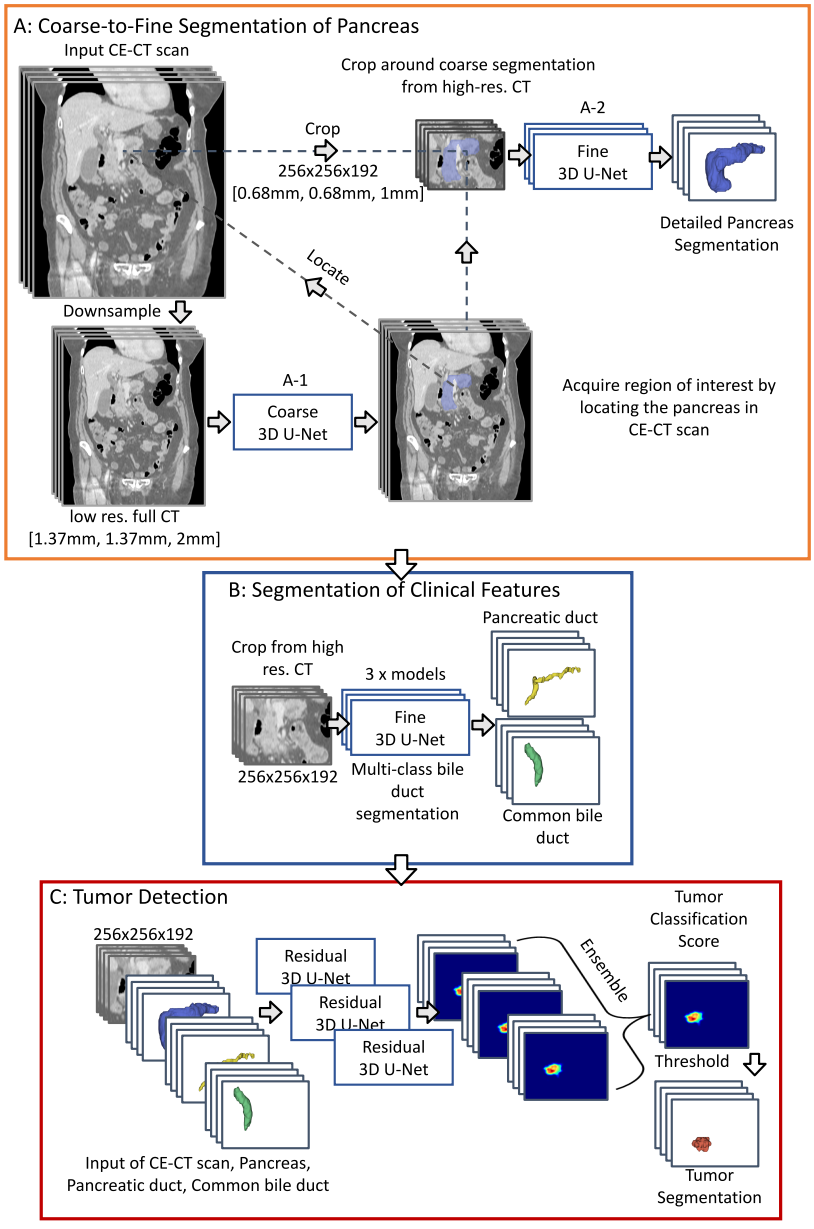


Figure 3.7 *Multi-stage framework of AI models for pancreatic tumor detection (The block numbering matches with the text descriptions).*

ducts within the targeted region. Due to its ease of implementation and improved inference speed and a comparable segmentation performance, a model that segments both classes simultaneously is generally preferred over multiple single-class models (Section 3.4.3).

C: Tumor segmentation: The tumor segmentation model integrates the comprehensive CT crop (256×256×192) with the multi-segmentation of previously delineated anatomical features, conducting a single inference step to segment PDAC if present. This 3D global perspective allows the model to analyze anatomical correlations across all three planes, leading to superior performance compared to alternative, patch-based models. Since we provide the segmentation of the pancreas and ducts as input to the detection model, the model is trained to utilize any valuable feature from these anatomical delineations. Therefore, the model potentially also includes other secondary features such as atrophic pancreas, size variations, ductal interruption and peri-pancreatic infiltration. Compared to the standard U-Nets employed when segmenting the pancreas and ducts, for the tumor a SOTA Residual 3D U-Net architecture is employed to further enhance detection accuracy. The model and implementation details are discussed in Section 3.8. The complete PDAC detection sequential workflow is depicted in Figure 3.7.

General attributes: All models are trained using a threefold bootstrapping method with 70% of the internal dataset (per patient) used to train the model and 15% for validation. This approach enables maximum usage of 85% of the data without involving the test set, preventing any form of data leakage. Additionally, this approach leads to three distinct models for each step of the detection process, all acting as a separate opinion on the specific task. The approach for comprehensive tumor detection integrates a series of models in an ensemble framework, enhancing accuracy and precision. First, the (1) cropped pancreas is supplied into three pancreas segmentation models trained on the different bootstrapped training subsets. This yields (2) three refined segments of pancreas-centered crops, of which each is uniquely tailored to capture the organ's intricate details. Subsequently, these crops are processed through (3) three respective secondary feature models for ductal segmentation, specifically designed for each fold. This step is crucial for generating three accurate sets of delineations, focusing on the pancreatic ducts. The next stage involves stacking these segmentation maps with the original CT crop, forming a composite input for the (4) three tumor segmentation models. From this, three separate tumor prediction crops are derived, each representing a different perspective of the potential tumor. Finally, (5) an ensemble of these predictions is collated, culminating in a robust and reliable final tumor prediction with enhanced confidence. This layered and iterative process not only maximizes the precision of tumor detection, but also leverages the strengths of multiple models to provide a more comprehensive and reliable diagnosis. To test the approach, we have applied the complete detection workflow to a predefined representative test split encompassing 15% of the proprietary data. In addition, the models have been applied to the public MD dataset as a separate and external test set.

3.5.5 Residual 3D U-Net architecture

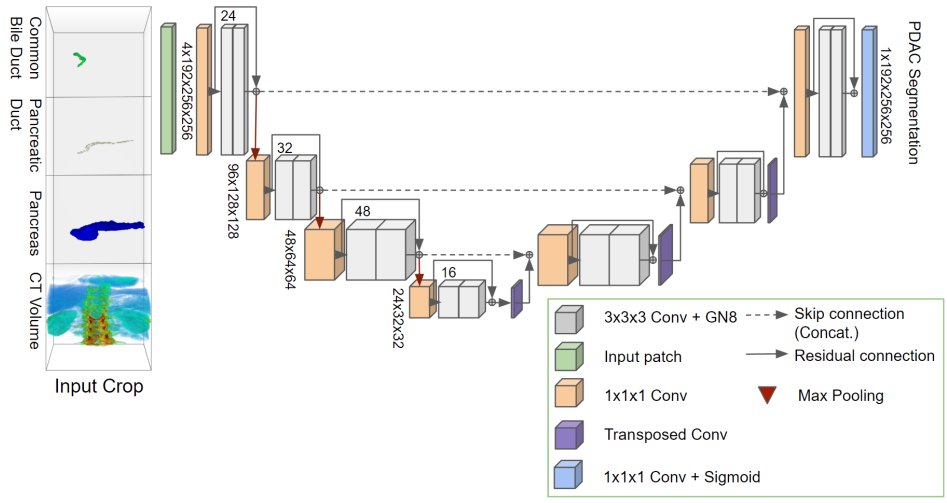


Figure 3.8 *Diagram of the Residual 3D U-Net utilizing secondary features for improved segmentation of pancreatic tumors.*

The Residual 3D U-Net architecture is depicted in Figure 3.8 and enhances the conventional U-Net framework through the integration of residual learning mechanisms. The model adopts the encoder-decoder structure, but is augmented with additional residual connections across convolutional blocks that facilitate the seamless transfer of spatial and contextual information across and deeper in the network.

Residual Learning Blocks: Each residual block within the encoder and decoder consists of two 3D convolutional layers, each employing its own group normalization [165] (an alternative batch normalization when few batch elements are employed during training). Importantly, these blocks include residual connections (acting as a shortcut) that add the input of the block directly to its output, followed by the block's non-linear activation. These residual connections are key to prevent the vanishing of gradients by facilitating unimpeded gradient flow during back-propagation, even in deeper network architectures. These connections are applied at all stages of the network at various layers and in both the encoder and decoder. The application of residual connections in the decoder is adopted for processing consistency [166].

Implementation details for PDAC segmentation: We implement a four-layer deep Residual 3D U-Net (C-Model in Figre 3.7) for segmenting pancreatic tumor. The layers in the network use 24, 32, 48 and 16 convolutional filters in each convolution block at each layer. A LeakyRelu is employed as the the non-linear activation function in the network.

Training details: The model is trained using an AdamW [167] optimizer fol-

lowing a one-cycle learning rate scheduler starting at 1×10^{-5}, that increases to 1×10^{-3} over 5% of the training duration. After this initial increase, the learning rate is slowly decreased to 1×10^{-7} for the remainder of the training process. The model is trained for 300 epochs on each datafold.

3.5.6 Web application for fully automated PDAC detection

To streamline the PDAC detection system, ensure reproducible results and to make it more accessible for clinical use, we have developed a web application that fully automates the complete detection process. This application is designed to be user-friendly, allowing clinicians to easily upload CT scans in the Neuroimaging Informatics Technology Initiative (NIfTI) format and receive comprehensive diagnostic outputs without requiring specialized technical knowledge. This implemented web application has been applied to and in collaboration with the clinicians to determine the final test results and evaluate them appropriately. The web application operates through the following steps.

1. *Uploading CT Scans:* Users upload their CT scans in NIfTI format. The application supports secure and efficient upload mechanisms to handle large medical imaging files, ensuring data integrity and confidentiality.
2. *API Interaction:* Through an API, external applications can interact with the detection pipeline, enabling automated workflows, batch processing of multiple scans, and integration into electronic health record (EHR) systems. Other benefits and features are not further discussed here.
3. *Preprocessing:* The application initiates the preprocessing stage that includes normalization of the image intensities, alignment of the images to a standard anatomical orientation, and resampling to ensure consistent resolution across different scans, as required by the different segmentation models.
4. *Detection and Localization:* The preprocessed images are then supplied into the PDAC detection processing chain and subsequent models (Figure 3.7). The models identify and localize potential pancreatic tumors by analyzing the secondary diagnostic signs and primary tumor characteristics.
5. *Segmentation:* The detected tumor regions are segmented and converted to their original resolution. This high-resolution segmentation preserves the details necessary for accurate diagnosis and treatment planning, providing a clear visualization of the tumor boundaries.
6. *Tumor Scoring:* The application calculates a tumor score based on the detected regions. This score reflects the likelihood of the current CT scan containing cancerous voxels. Guidance is provided on what is considered a tumor-positive case or tumor-negative case, based on the statistical results of the processing chain on the validation dataset.
7. *Result Presentation:* Finally, the application presents the results in the web interface, or as a response from the API. This report includes the tumor score,

detailed segmentation maps of all the segmented organs and the uncertainty maps computed from the model ensembles.

The web application ensures that the entire detection pipeline is automated, from image upload to result generation, thereby minimizing the need for manual intervention and reducing the potential for human error.

3.5.7 Segmentation results of secondary features

The segmentation accuracy of the pancreas segmentation model, as well as the segmentation models for bile duct and pancreatic duct segmentation, are assessed using the DSC metric. The fine pancreas segmentation models achieve a mean DSC of 0.86±0.05 on the proprietary test set and 0.88±0.03 on the MD dataset (Table 3.7). For bile duct segmentation, a multi-class model obtains a mean DSC of 0.61±0.22 for the common bile duct and 0.49±0.25 for the pancreatic duct on our proprietary test set, while it achieves a mean DSC of 0.67±0.19 and 0.51±0.19 on the common bile duct and pancreatic duct for the MD dataset, respectively. The models perform consistently between the proprietary test set and the public MD dataset.

Table 3.7 *Results obtained and evaluated using the DICE Similarity Coefficient (DSC) for the proprietary test set and the Medical Decathlon (MD) Dataset. ("not ann." means that the cases where the ductal structures were small and not annotated were removed from the evaluation).*

Anatomical Structure	Proprietary test set	Medical Decathlon Dataset
Pancreas	0.86 ± 0.05	0.88 ± 0.03
Common bile duct	0.61 ± 0.22	0.67 ± 0.19
Pancreatic duct	0.49 ± 0.25	0.52 ± 0.19
Common bile duct (not ann.)	0.63 ± 0.18	0.69 ± 0.14
Pancreatic duct (not ann.)	0.60 ± 0.11	0.55 ± 0.13

3.5.8 Tumor detection results

All models have been applied to the proprietary and the MD dataset for testing. The detection models are analyzed using the ground-truth annotations (pancreas and ducts) as input and in a fully automated manner, utilizing the predictions from the preceding models as input. Employing the ground-truth annotations, the tumor detection model demonstrates perfect accuracy in identifying all tumors within the proprietary test set, resulting in a sensitivity of 1.00 (Table 3.8). Utilizing the ground-truth segmentation maps, the model achieves a specificity of 0.86 on the proprietary test set. Combining the three different models, further improves the detection performance without false positive predictions, resulting in a perfect sensitivity and specificity of unity. When applying the multi-stage algorithm where the pancreas, common bile duct and pancreatic duct are automatically segmented and provided to the tumor detection model, the specificity drops to 0.64±0.15. However, combining the different models through an ensem-

Table 3.8 *Performance evaluation of the complete PDAC detection and segmentation framework using the DICE Similarity Coefficient (DSC), Sensitivity, and Specificity on the proprietary test dataset. Results are compared of the method using (1) manually annotated inputs of the pancreas and bile ducts and (2) the method employing the automated segmentation of bile ducts and pancreas. GT is ground truth.*

Input type Model type	GT Individual	GT Ensemble	Predicted Individual	Predicted Ensemble	Predicted Ensemble
Data subset	All	All	All	All	Tumors < 2 cm
Sensitivity	1.00 ± 0.00	1.00	1.00 ± 0.00	0.97	1.0
Specificity	0.86 ± 0.10	1.00	0.64 ± 0.15	1.00	1.0
Precision	0.94 ± 0.04	1.00	0.87 ± 0.05	1.00	1.0
F_1 Score	0.97 ± 0.02	1.00	0.93 ± 0.03	0.98	1.0
Accuracy	0.96 ± 0.03	1.00	0.89 ± 0.05	0.98	1.0
ROC	0.96 ± 0.02	0.98	0.97 ± 0.03	0.99	0.98
Mean Dice	0.35 ± 0.04	0.37	0.31 ± 0.02	0.34	0.19 ± 0.24

ble, improves overall performance to a sensitivity of 0.97 (1 case missed) and a specificity of 1.0. Overall, the models offer accurate tumor classification results with an AUROC of 0.99 (depicted in Figure 3.9).

In a subanalysis of tumors smaller than 2 cm within the proprietary test set (12 cases<2 cm vs. control set), the model achieves an impressive AUROC of 0.98. However, it also records a much lower DSC of 0.19±0.24, which highlights the challenge of accurately delineating these small tumors. Using the ground-truth annotations and the multi-stage algorithm as inputs, the tumor detection model demonstrates a perfect sensitivity of 1.00 within the MD test set in both instances (Table 3.9). In addition, the model segmentation of the tumor records a mean DSC of 0.37 in both the proprietary test set and the MD test set. Finally, we visualize the predicted segmentation maps generated by the multi-stage algorithm in the proprietary test set, to understand which aspects contribute to pancreatic tumor detection and to offer insights into the algorithm performance in Figure 3.10.

Table 3.9 *Evaluation using Sensitivity and the Mean DSC on the MD test dataset. Methods are compared using manually annotated input of the pancreas and bile ducts and employing the automated bile ducts and pancreas segmented by AI models.*

Input type Model type	GT Input Individual	GT Input Ensemble	Predicted Input Individual	Predicted Input Ensemble
Sensitivity	1.00 ± 0.00	1.00	1.00 ± 0.00	1.00
Mean DSC	0.37 ± 0.03	0.37	0.37 ± 0.01	0.37

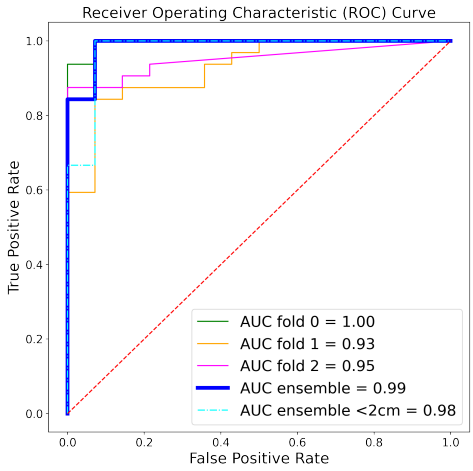


Figure 3.9 *Area Under Receiver Operating Characteristic Curve (AUROC) of the full tumor detection approach (ensemble) on the proprietary test set.*

3.5.9 Discussion

This study presents accurate PDAC detection by employing a multi-stage deep learning framework designed for detection on contrast-enhanced CT scans. Recently, there has been a notable interest in segmentation for classification approaches, as it enables both cancer detection and localization of pancreatic cancer [53], [115], [131], [139], [168]–[171]. This work additionally indicates the value of incorporating secondary features of pancreatic cancer such as a dilated bile duct, to improve pancreatic tumor detection [53].

Pancreas: The proposed pancreas segmentation algorithm performs comparably to SOTA models [172]–[174]. However, it sometimes produces irregular boundaries around the pancreas, but this behaviour has also been recorded in other algorithms from literature. For instance, Huang *et al*. [174] developed a semi-automated DUNet tailored to capture the variable and irregular contours of the pancreas, achieving a DSC of 87.25±3.27. Both methodologies, although highly accurate, occasionally struggle with the pancreas's inherent shape variations, irregular boundaries and the ambiguity of the pancreatic structure in CT.

Bile ducts: The diagnostic usefulness of the framework is further enriched by the integration of adjacent structural information, such as dilated pancreatic and common bile ducts, which elucidates the relationship between these ducts and tumors, where clinician are especially focusing on the bile duct. Considering the ductal segmentation algorithm, despite achieving perfect sensitivity in bile duct identification, the algorithm records a specificity of 0.4±0.067 due to false positives, largely attributed to the smaller size of the pancreatic duct and lowering the DSC values. These metrics indicate substantial variability, primarily because small, sometimes barely visible ducts were not annotated, yet the models establish proper visual segmentation of the structures in these regions.

External validation: Multiple studies have utilized supervised classification net-

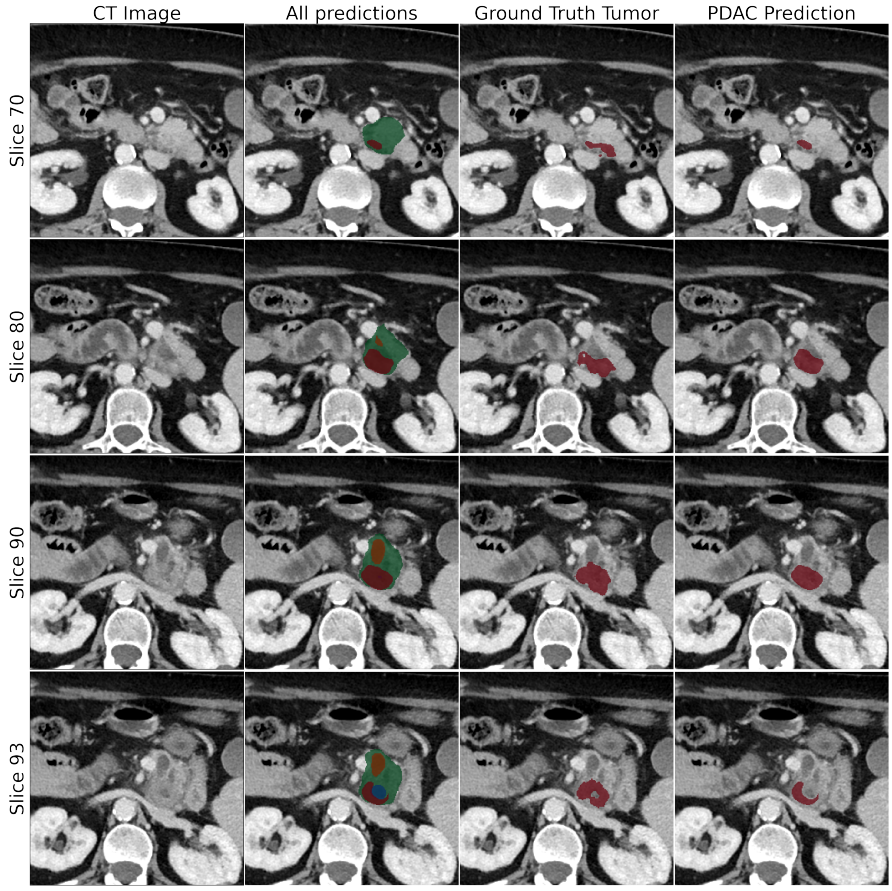


Figure 3.10 *Predicted segmentation results of the proposed PDAC CAD system. The tumor is depicted in red, pancreas in green, pancreatic duct in orange and common bile duct in blue. The ground-truth tumor and PDAC prediction are depicted in consecutive columns. For this patient case, the model achieves a high DICE score of 0.71 on an image from the proprietary test set.*

works, resulting in accurate detection of PDAC and other types of pancreatic cancer on CT scans [130], [133]–[135], [138], [175]. However, only a few studies have externally validated their models, with just two studies utilizing the publicly accessible MD dataset [133], [139] in combination with a tumor negative set, to determine detection performance. Alves *et al*. [139] have introduced an automated framework for PDAC detection, employing three nnU-Net [13] models that in addition to the pancreatic tumor, also predicts secondary features for tumor presence evaluation. The authors have demonstrated the benefit of integrating anatomical information and reported a notable AUROC of 0.91. Liu *et al*. [133] have developed a deep learning model utilizing a modified VGG network to differentiate between pancreatic cancer tissue and non-cancerous tissue. Their model is also evaluated using the MD dataset, achieving a sensitivity of 0.79 and specificity of 0.84, resulting in a maximum AUROC of 0.92 at the patient level.

Internal validation: By employing the ground-truth secondary-feature annotations and segmentation maps from the multi-class algorithm, the tumor detection model showcases exceptional sensitivity. It achieves a perfect sensitivity score of 1.00 on the MSD test set and the internal proprietary test set. With an AUROC of 0.99, the proposed method outperforms the state-of-the-art approaches, particularly in external validation on the MD dataset. This highlights the importance of incorporating secondary tumor-indicative anatomical information and represents a notable improvement in pancreatic tumor detection accuracy. Despite achieving a mean DSC of 0.34, indicating limited tumor localization accuracy, our primary focus is on identifying tumor presence and approximate location, rather than delineation, given the significant benefits of early detection for patient survival [22], [176], [177]. With the scan highlighted by the proposed method as containing tumor and the additional approximate location of the PDAC, a follow-up segmentation model can be employed to refine the segmentation.

Supplementary value: This method stands out for its integration of the secondary features indicating tumor and evaluation of pertinent anatomy within a structured workflow. Literature indicates that 5.4-–14% of pancreatic tumors present as completely iso-attenuating, making them indistinguishable from normal pancreatic tissue [178]. Particularly in such cases, radiologists often rely on alternative patterns suggesting malignant disease. Throughout the years, multiple studies have consistently reported the presence of visible secondary features, prior to actual diagnosis [119], [121]. For instance, Kang *et al*. [116] found that 88% of the cases exhibited such secondary signs. Additionally, imaging has revealed indicative changes associated with PDAC up to 18 months prior to diagnosis in 50% of the patients [120], [122]. The multi-stage detection framework integrates detailed duct and pancreas segmentation maps with CT scans, thereby enabling it to establish connections between these secondary features and tumor presence. Furthermore, it provides valuable insights into the model performance which contributes to solving concerns regarding result explainability. This study offers a quantitative assessment of PDAC detection performance. However, quantifying the diagnostic value of information derived from the segmentation of secondary features remains a challenging aspect.

3.5.10 Limitations

The study exhibits several limitations that are briefly acknowledged below. Firstly, the models are trained on a relatively small dataset that primarily include CT scans from patients with pancreatic head cancer, contrasting against normal pancreatic scans. Consequently, the models have not been exposed to a diverse array of pathological conditions, including other neoplastic disorders or tumors situated in varying locations within the pancreas. This lack of variability may potentially limit the generalization of the model to other forms of pancreatic diseases.

Secondly, the employed dataset is derived from a specifically selected cohort, which does not accurately reflect the broader demographic and disease prevalence found in the general population [179]. This selection bias may lead to a skewed

distribution, affecting the model performance when applied to a more diverse, real-world clinical setting. To address this limitation, there is a clear need for a prospective study that evaluates the model effectiveness across various target populations to ensure its clinical feasibility and robustness.

Lastly, the external validation of the tumor detection models has been performed using the publicly available MD dataset. While this dataset is a valuable resource, it predominantly comprises cases with advanced-stage pancreatic cancer, characterized by larger tumor sizes and the frequent presence of metal stents, which may introduce bias in the performance assessment of the models. Moreover, the MD dataset exclusively contains tumor-positive cases, leading to the absence of specificity metrics in this study since, tumor-negative scenarios are not tested. Future research endeavors should aim to investigate the model specificity by incorporating a balanced representation of tumor-negative cases, to provide a comprehensive evaluation of its diagnostic accuracy.

3.6 Challenges and future directions in PDAC detection

Deep learning-based CAD systems have grown as a high-potential technology that will shape the future of medical healthcare. Deep learning-based image processing techniques such as Convolutional Neural Networks [180], Vision Transformers [8] and Diffusion models [15] have rapidly revolutionized the standard performance expected of automated systems. However, the translation of effective techniques into clinical deployment presents a new opportunity for clinical studies, machine learning research, and human-computer interaction designs.

Considering the current limited successful deployment of deep learning-based CAD systems in the clinical practice, it is considered that the following aspects are becoming increasingly important for further progression of automated pancreatic cancer detection.

First, training deep learning algorithms for difficult tasks requires large and diverse amounts of high-quality labeled data. However, due to the relatively low prevalence of pancreatic cancer, this is one of the biggest hurdles hampering the development of data-driven algorithms. The first indications of the potential of AI applications are clearly visible in retrospective research on small, homogeneous datasets, where these systems demonstrate impressive results in detection of pancreatic tumors. However, datasets are the primary drivers of these algorithms and collaboration and data sharing between expert centers for pancreatic surgery would provide the opportunity to deal with the scarcity of high-quality labeled datasets. Section B.1 provides further discussion on this topic with a focus on data representations, biases and confounders.

Second, for proper evaluation of the model, the results should be presented transparently, following clear standards. To avoid misconceptions, it should be clear which metrics are used to evaluate the performance of the model. Metrics such as accuracy, sensitivity, specificity and area under the receiver operating characteristics should be described [181], [182]. Therefore, clear guidelines can

be used, following transparent reporting of a multi-variable prediction model for individual prognosis or diagnosis (TRIPOD) [183], [184]. Additionally, a targeted calculation of a sample size of the test set, as proposed by Riley *et al*. [185], which is needed to accurately determine the quality of early-stage algorithms, could facilitate unbiased evaluation.

Third, for an algorithm to be of added value in clinical practice, aspects such as reliability, uncertainty and robustness against variations will become increasingly important. This implies novel approaches that require a strong collaboration between clinician researchers and machine learning experts. While detection of cancer is the first essential step, follow-up tasks such as tumor resection planning require extremely detailed segmentation maps of the tumor and surrounding anatomical structures. Current existing approaches do not yet meet this high-accuracy requirement. The current research trend is to design better models to adhere to these shortcomings, while it is possible that there will always be edge cases unaccounted for or different opinions on a correct segmentation. A possible solution to bridge the gap between what is technically possible and clinically necessary is interactive AI (IAI), where a user can exploit AI with additional input. By implementing IAI, the pattern recognition abilities of AI and the domain knowledge of the clinician can be combined, resulting in more accurate and robust results [186]. This could enable clinicians to achieve highly accurate annotations of the structures of interest and can improve effectiveness and implementation of AI algorithms in clinical practice.

Finally, the development and implementation of AI-based CAD systems requires an interdisciplinary approach between technical and medical partners. A strong collaboration between medical doctors and engineers is necessary to create a common understanding of the possibilities and limitations of systems. Supported by findings in a study of diabetic retinopathy screening [187], it has already been shown that humans assisted by AI performed better than either standalone AI or the independent clinician. Understanding human-algorithm interfacing and the importance of human-centered AI will be critical in future adoption of AI applications. The AI model operation, as well as the user-interface (UI) through which the physician and AI interact, should be optimized for this team performance. Choices in the AI development and UI design influence how physicians work with AI to leverage the collective intelligence of the physician and the AI [188].

3.7 Conclusions

This chapter has explored Deep learning-based CADe for PDAC detection. The key to the proposed framework is the integration of external, tumor-indicative features to enhance the performance of PDAC CAD systems. The method incorporates clinically-relevant secondary features into a 3D U-Net, which processes CT scans alongside segmentation maps of the pancreas, pancreatic duct, and common bile duct to segment PDAC more reliably. This approach not only achieves a high detection sensitivity (99$\pm$2%) and specificity (99%), but also provides in-

sights into tumor locations, maintaining robust performance across both a small prospectively collected dataset and the Medical Decathlon dataset.

A key finding of the research in this chapter is that the PDAC detection framework can be fully automated, including the localization of the pancreas, detailed pancreas segmentation, detailed bile duct segmentation maps and ultimately accurate tumor detection. The experiments indicate that the same high sensitivity and specificity on an independent test set can be obtained, even with the fully automated framework. The complete setup has been rigorously tested on a larger, prospectively collected proprietary dataset and the publicly accessible Medical Decathlon dataset.

The value of developing an additional dataset for more detailed training and testing has proven a high importance, despite its proprietary nature. The joint development of this dataset incurred close collaboration between clinical and technical experts and has improved the discussion on the relevance of primary and secondary tumor-indicative features.

To increase further clinical relevance, the proposed framework with its promising results needs to be corroborated through extensive, multi-center studies to ensure its reliability and effectiveness across various clinical environments. This validation is essential to confirm the robustness of the models in diverse settings and to substantiate its role in revolutionizing the diagnosis and treatment of pancreatic cancer. On the longer term, the use of deep learning-based algorithms for the early detection of pancreatic cancer holds the potential to substantially increase the number of patients who are eligible for curative treatments, thereby improving their chances of survival.

In the next chapter, the research on a segmentation framework generalizes from the specific PDAC case and focuses more on quantifying uncertainty in the segmentation of ambiguous structures in medical imaging. This discussions in this chapter already indicate substantial uncertainties appear typically at the boundaries of organs and withing complex structures like tumor growths. Quantifying this uncertainty can provide vital information to clinicians in curative treatment planning and strategies.

Chapter 4

2D/3D Uncertainty quantification in medical image segmentation

4.1 Introduction

Medical image segmentation is a crucial step in diagnostic radiology that involves delineating anatomical structures from the acquired image. Clinicians employ detailed segmentation of anatomical structures in medical images to enhance their diagnostic accuracy and the treatment planning by precisely localizing abnormalities, measuring volumes, and assessing functional parameters. Segmentation guides surgical navigation and radiotherapy, thereby ensuring precise targeting while minimizing harm to healthy tissue. The segmentation facilitates medical research by enabling anatomical studies and the development of accurate models for simulation and training. Additionally, segmentation improves communication through clear visualizations and supports the development of automated analysis tools, thereby increasing efficiency and consistency in clinical practice.

Despite its numerous benefits, the segmentation process in medical imaging faces several limitations. Manual segmentation is time-consuming and highly dependent on the clinician's expertise, leading to variability in results. Automated deep learning-based segmentation methods, although being faster, often struggle with the complexity and variability of human anatomy, resulting in inaccuracies, especially in cases of abnormal or pathological anatomy. The quality of the segmentation can also be affected by the resolution and noise levels of the imaging modality, as well as the presence of artifacts.

As a result of the considerable advances in machine learning research over the past decade, computer-aided diagnostics (CADx) using deep learning is rapidly gaining attention. Convolutional neural network (CNN)-based approaches have been adopted in a large number of CADx applications and especially in semantic segmentation. In this approach, the objects of interest are localized by assigning class probabilities to all pixels of the image. In the medical domain, and especially in the context of lesion segmentation, the exact edges or borders of these lesions are not always readily clearly defined (even for physicians) and ground-truth annotations are associated with a high interobserver variability. Hence, in the case of multiple assessors, clinicians may disagree on the boundaries of the localized lesions, based on their understanding of the surrounding anatomy. Furthermore, it is often impossible to define exact tumor boundaries in medical imagery, as registration with histopathology is not applicable. However, the exact edges or

borders of these areas of interests often play a critical role in the diagnostic process. For example, when determining decisions on surgery for a patient or the surgical planning, knowledge about the invasion of a tumor into local anatomical structures is crucial. Generally, the above discussion indicates that CADx semantic segmentation models suffer from the variable imaging quality, the limited capabilities of the model architectures or lack of sufficient training data (to name a few aspects), all of which can introduce various sources of ambiguity and uncertainty. Thus, multiple forms of uncertainties come into play when employing semantic segmentation-based approaches for CADx. As such, accurately quantifying these uncertainties have become an important factor in CADx. Specialized doctors provide ground-truth segmentation maps for models to be trained, based on their knowledge and experience. When this is done by multiple individuals per image, often discrepancies in the annotations arise, resulting in ambiguities in the ground-truth labels.

In the transition towards more automated and assistive systems using advanced computer vision techniques, the role of learning-based methods in medical image segmentation has become significantly pronounced. However, together with this transition, a critical challenge arises of managing uncertainties inherently occurring in the processing steps. These uncertainties can substantially affect the reliability of CADx-based diagnostic outcomes. These uncertainties in machine learning can be broadly classified into two categories: *epistemic uncertainty* and *aleatoric uncertainty* (introduced in Chapter 2.3). Similarly, these uncertainties are present in deep learning-based segmentation models. The epistemic uncertainty is the result of lack of knowledge about the model parameters and the underlying data distribution. The aleatoric uncertainty is due to limitations in the image acquisition process or, as mentioned above, as introduced with ambiguous ground-truth annotations.

The purpose of this chapter is to address the ambiguity present in medical image segmentation tasks. An optimal segmentation architecture is considered where aleatoric uncertainty can be expressed explicitly. Expressing this uncertainty accurately may contribute to minimizing interventions and maximizing automation in the processing. These considerations lead to the following research questions.

- A significant challenge exists of dealing with variability in clinical annotations and intrinsic noise in imaging data when training deep learning-based models. *How can we accurately segment structures under ambiguous ground-truths, while aiming to capture and express the aleatoric uncertainty?*
- The Probabilistic U-Net (PU-Net) is an effective method for modeling the aleatoric uncertainty. However, restricting the modeled distribution to be normally distributed limits its accuracy. *Does augmenting the strictly Gaussian posterior in the PU-Net with Normalizing Flows improve aleatoric uncertainty quantification?*

- In transitioning from 2D to 3D image segmentation, the (PU-Net) model is expected to capture spatial context more effectively, potentially leading to higher consistency and an increased accuracy of segmentation outcomes. *Does extending the PU-Net to 3D processing and modeling improve the consistency, efficiency and accuracy in aleatoric uncertainty quantification?*

This chapter discusses uncertainty and its potential impact in medical image segmentation. The Probabilistic U-Net [189] has previously been developed for segmenting ambiguous structures, however, the approach and the architecture have several limitations that could encumber its accuracy. The chapter details an improved model that specifically allows for a more flexible posterior distribution and, as such, captures the aleatoric uncertainty accurately. These improvements are then extended to 3D processing and modeling, utilizing information from the full 3D data volume to more accurately capture and express the segmentation aleatoric uncertainty. This inquiry not only advances the field of diagnostic radiology, but also contributes to the overarching objectives of uncertainty and interobserver variability modeling in the ambiguous medical imaging domain to ultimately improve patient-specific interventions.

4.2 Related work

4.2.1 Types of uncertainty in medical image segmentation

Uncertainty in machine learning can be broadly classified into two categories: aleatoric uncertainty and epistemic uncertainty (introduced in Chapter 2.3). Similarly, these uncertainties are present in deep learning-based segmentation models.

4.2.1.A Epistemic uncertainty

Epistemic uncertainty, or model uncertainty, arises from the lack of knowledge about the model parameters and the underlying data distribution. This type of uncertainty is reducible and can be mitigated by acquiring more data and improving model architecture and training processes [190]. In the context of medical imaging, epistemic uncertainty reflects the limitations of the segmentation model itself, such as inadequate training data or sub-optimal model complexity. Epistemic uncertainty affects the generalization capability of segmentation models. Techniques such as Bayesian neural networks, ensemble methods, and active learning are employed to quantify and showcase the epistemic uncertainty, thereby improving model robustness [102].

In the case of multi/single-assessor annotated data, the epistemic uncertainty stems from the preferences, experiences, knowledge (or lack thereof) and other biases of the multiple/single assessors specifically. This epistemic uncertainty from the assessor(s) manifests into a different type of uncertainty when providing annotations and training a learning-based model on that data.

4.2.1.B Aleatoric uncertainty

Aleatoric uncertainty, also known as statistical or data uncertainty, arises from the inherent noise and variability in the data. In medical imaging, this type of uncertainty is often due to factors such as low-resolution images, poor contrast between structures, motion artifacts and, importantly, inter-observer variability [191]. Aleatoric uncertainty is irreducible and cannot be mitigated by acquiring more data. Instead, it can be addressed through robust statistical modeling and noise-resistant algorithms.

Aleatoric uncertainty significantly impacts the performance of segmentation models. For instance, in the presence of high noise levels or low-contrast regions, the model predictions become less reliable. A segmentation model should show increased uncertainty levels under such conditions to indicate which part of the segmentation is less reliable or needs to be handed over to a clinical expert for evaluation. Addressing aleatoric uncertainty involves techniques such as probabilistic modeling and leveraging advanced imaging modalities, to enhance image quality [192].

Assessor's epistemic uncertainty can manifest into aleatoric uncertainty. Due to the inherent ambiguity that exists in the imaging data, clinicians annotate the structures of interest following their own experience (clinician's epistemic uncertainty) and available time. Different assessors could have different opinions and invest different amounts of time in annotating a target structure, resulting in multiple plausible annotations. Training supervised deep learning-based segmentation models then encounter multiple ground truths.

The intersection of these uncertainties in deep learning-based medical image segmentation demands a robust framework that not only acknowledges, but also quantifies or mitigates these uncertainties. Quantifying uncertainty in segmentation models can lead to more transparent, reliable, and interpretable tools. For instance, understanding the degree of uncertainty associated with a segmented tumor boundary can influence clinical decisions.

4.2.2 Methods for quantifying uncertainties in image segmentation

Several methods have been proposed to quantify uncertainties in medical image segmentation. Many of the methods introduced Chapter 2.3 in an image classification context, also extends to segmentation applications. Notably, Kendall and Gal [51] presented some of the first research to capture uncertainty and introduced a unified framework based on Bayesian deep learning. The framework distinguishes between the types of uncertainty, where aleatoric uncertainty is captured through model output variance, while epistemic uncertainty is estimated via the model weight distributions. Since its inception, many methods have been proposed to capture the segmentation uncertainties either more accurately or efficiently. In recent work by Zou *et al*. [193], an overview of these methods along with their applications are highlighted.

4.2.3 Probabilistic U-Net for segmenting ambiguous images

The probabilistic U-Net (PU-Net) [189] is a method for segmenting ambiguous images which significantly advances the field of medical image analysis, by addressing the challenge of inherent ambiguities in image segmentation tasks. The core of the contribution lies in the development of a generative model that integrates the strengths of the U-Net [194] with the probabilistic modeling capabilities of a conditional Variational Autoencoder (cVAE) [195]. The U-Net serves as a powerful encoder-decoder network that processes image data to produce object features, while the cVAE captures the distribution of possible segmentation outputs that account for the inherent ambiguities. Figure 4.1 depicts the training and test stages of the complete model. The arrows represent the flow of operations, blue blocks are feature maps and the heatmap (in grey) represents the probability distribution in the low-dimensional latent space $\mathbb{R}^N$ (e.g., N=6 in the experiments). For each forward-pass of the network, one sample $\mathbf{z}$ from $\mathbb{R}^N$ is drawn and combined with the image features to predict one segmentation mask. The green block shows an N-channel feature map resulting from broadcasting sample $\mathbf{z}$. The number of feature map blocks depicted in the figure is reduced for clarity of presentation.

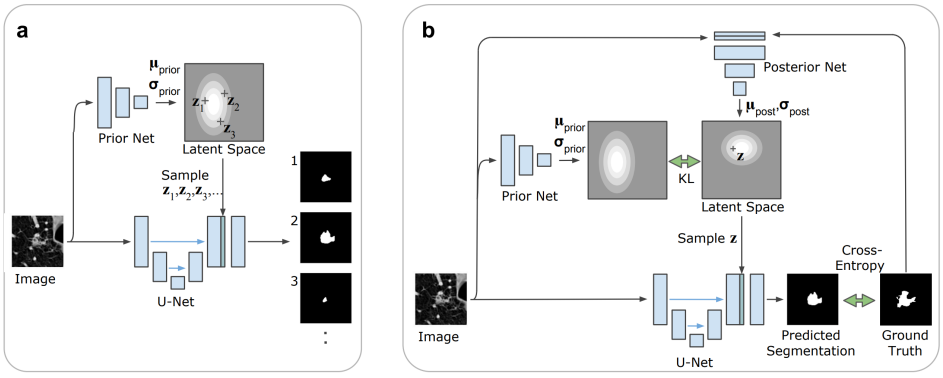


Figure 4.1 *Diagram of the Probabilistic U-Net. (a) Test-time sampling process. (b) Training process illustrated for 1 training example. The image is based on the original paper [189].*

There are several attributes that set the PU-Net apart from prior work.

- *Explicit density modeling:* The uncertainty is captured by a down-sampled axis-aligned Gaussian prior that is updated through the KL divergence of the posterior during training.
- *Latent space sampling:* The model samples from this low-dimensional latent representations that captures the distribution of segmentation hypotheses. This sampling is guided by a "prior net" that predicts the distribution of latent variables conditioned on the input image, enabling the generation of diverse segmentation hypotheses.

- *Efficient hypothesis generation:* By integrating sampling mechanisms directly into the network, the model can efficiently produce multiple plausible segmentation masks for a single input, which forms a significant improvement over traditional deterministic methods.

The model is employed to segment ambiguous structures such as lung abnormalities from CT images (LIDC-IDRI dataset [196]) and in urban scenes (Cityscapes dataset [197]). These datasets will be discussed in more detail in Section 4.3.2, 4.4.3 and Appendix C.1. The performance is evaluated using intersection over union (IoU) and a novel application of the squared generalized energy distance (GED). These metrics help quantify how well the model's output distribution matches the distribution of ground-truth segmentation masks, with a particular focus on capturing both common and rare segmentation variants. The model is shown to outperform existing methods like Bayesian SegNet [198], various ensemble approaches and methods employing multiple decoders (M-Heads).

In medical diagnostics, ambiguities in image data can lead to vastly different treatment decisions. By providing multiple plausible segmentation masks, the model allows clinicians to consider a range of possible diagnoses or further testing strategies. The succeeding chapter in this thesis delves into one such use case. The model can be used to suggest additional diagnostic tests that may resolve ambiguities, thereby supporting more informed clinical decision-making.

4.2.4 Limitations of the literature

While the PU-Net model demonstrates impressive capabilities, the authors also acknowledge certain limitations that pave the way for improvements in model-based segmentation.

- *Complexity of ambiguities:* The model is currently tailored to scenarios where ambiguities are known and captured through the annotated labels. Extending this approach to scenarios with unknown or unmodeled ambiguities may enhance the model applicability.
- *Accuracy:* The model handles multiple hypotheses well, however, the extent thereof can be improved. Various architectural aspects suggest that improved performance can be obtained with adjustments to the overall model.
- *Efficiency:* The current implementation of the Probabilistic U-Net utilizes a 2D approach, which may not fully leverage the information available in medical applications where data typically come from 3D volumes. Extending the model to handle 3D data can significantly improve the consistency and efficiency of uncertainty quantification, making use of the rich spatial information inherent in 3D medical scans.
- *Scope:* While the Probabilistic U-Net is adept at segmenting ambiguous structures, its approach to uncertainty is primarily confined to generating multiple plausible segmentation masks rather than quantitatively capturing and representing the underlying uncertainty in those masks. Follow-up work

should attempt to explain what the ambiguity represents.

Overall, the probabilistic U-Net effectively demonstrates how the integration of probabilistic generative models with deep learning segmentation architectures can address significant challenges in segmenting ambiguous tasks. The method not only advances the state-of-the-art in medical image segmentation, but also opens up new avenues for research in handling ambiguities in other domains of computer vision and medical imaging. The above-listed limitations are addressed in the succeeding sections on probabilistic segmentation.

4.3 Improving aleatoric uncertainty quantification in multi-annotated medical image segmentation with NFs in 2D images

In medical image analysis, automated segmentation results are crucial for vital decision-making, making it essential to quantify the uncertainty in the segmentation output. In the supervised multi-annotation setting (introduced under Point B in Section 4.2.1), compelling attempts have been made in quantifying the uncertainty in image segmentation architectures. By using a probabilistic segmentation model, this work attempts to learn a distribution of possible annotations. However, existing work in this field restricts these learnt densities to be strictly Gaussian. It is important to enable expressiveness of the probability distributions to sufficiently capture the variability in the data. In this multi-assessor settings, the adoption of rich and multi-modal distributions is desired.

In this work it is shown that by using invertible bijections, also known as Normalizing Flows (NFs), we can obtain more expressive distributions to adequately deal with the disagreement in the ground-truth information. The Probabilistic U-Net (PU-Net) [189] is employed as the base model and subsequently improved, by adding a planar and radial flow to render a more expressively learnt posterior distribution. This enables the learnt densities to be more complex and facilitate more accurate modeling for uncertainty quantification. The qualitative as well as quantitative evaluations show a clear improvement on two large public datasets: the multi-annotated and LIDC-IDRI and single-annotated Kvasir-SEG segmentation datasets. The improvements are mostly apparent in the quantification of aleatoric uncertainty and the increased predictive performance of up to 14%. This result strongly indicates that a more flexible density model should be seriously considered in architectures that attempt to capture segmentation ambiguity through density modeling.

Additional background and motivation: Selvan *et al*. [199] used an NF on the posterior of a cVAE-like segmentation model and showed that this augmentation increases sample diversity. The increased sample diversity resulted in a better score on the GED metric and a slight decrease in DSC score. The authors reported significant gains in performance. However, it is argued that this claim requires more evidence to confirm this positive effect, such as training with K-fold

cross-validation and evaluating confirmation using additional metrics. Additional insight into the reasons for the obtained improvements are not provided and critical details of the experiments are missing, such as the number of samples used for the GED evaluation. Shi Hu *et al*. [200] showed how this ambiguity can be interpreted as uncertainty. This work aims to provide a more comprehensive argument and show clear steps towards improving the quantification of aleatoric uncertainty.

This section delves into enhancing aleatoric uncertainty quantification in medical image segmentation using normalizing flows. It is shown that the segmentation ambiguity can be interpreted as aleatoric uncertainty and that this uncertainty can also be captured in single-assessor settings. The proposed model architecture and integration of NFs into the segmentation framework are presented in Section 4.3.1. The datasets used, baseline experiments conducted and training details are described to establish a comparative foundation for the proposed approach (Section 4.3.2 - 4.3.4). Finally, the results of the experiments are presented and improvements in aleatoric uncertainty quantification (Section 4.3.5) are discussed. This model section concludes with insights gained from this work in Section 4.3.6.

4.3.1 2D Model architecture

The proposed method builds on the PU-Net extended with an NF, as is shown in Figure 4.2. A key element of the architecture is the posterior network Q, which attempts to encapsulate the distribution of possible segmentations, conditioned on the input image $\mathbb{X}$ and ground truth $\mathbb{S}$ in the base distribution. The flexibility of the posterior is enhanced through the use of an NF, which warps it into a more complex distribution.

During training, a sample from the posterior distribution (Q) is combined with the features from the reconstruction network (U-Net), to generate a new segmentation. The prior P is updated with the evidence lower bound (ELBO [47]), which is based on two components: (1) the KL divergence between the distributions Q and P, and (2) the reconstruction loss between the predicted and ground-truth segmentation. The use of NFs is motivated by the fact that a Gaussian distribution is too limited to fully model the input-conditional latent distribution of annotations. An NF can introduce complexity to Q, e.g. multi-modality, in order to more accurately describe the characteristics of this relationship.

The training objective of the PU-Net extends on the standard ELBO (Section 2.2.1) and is defined as

$$\mathcal{L} = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x})}\big[\log p(\mathbf{s}|\mathbf{z},\mathbf{x})\big] + \mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x}) \,||\, p_\psi(\mathbf{z}|\mathbf{x})\big), \tag{4.1}$$

where the latent sample $\mathbf{z}$ from the posterior distribution is conditioned on the input image $\mathbf{x}$, and ground-truth segmentation $\mathbf{s}$. We proceed by extending the PU-Net training objective and explain the associated parameters in detail. To

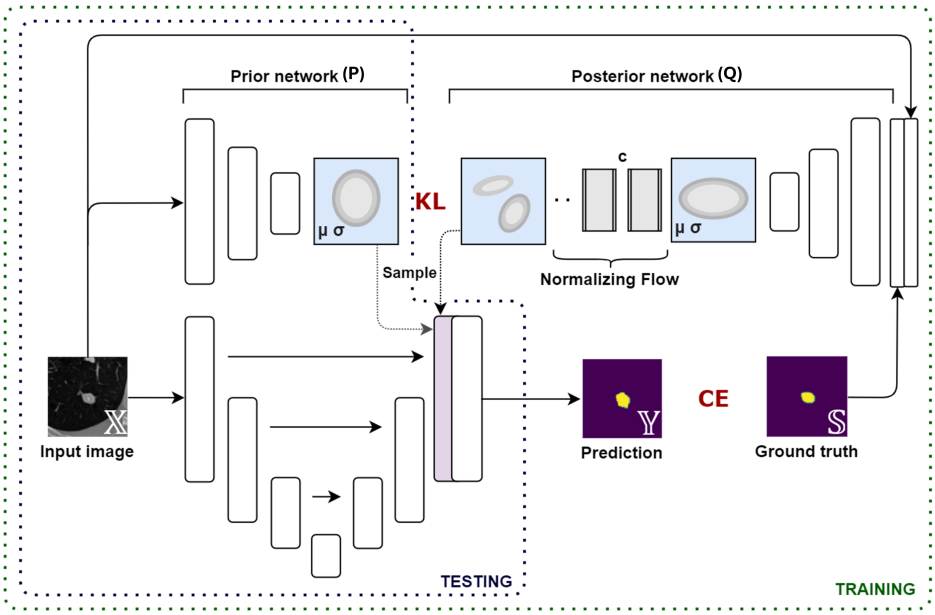


Figure 4.2 *Diagram of the PU-Net with an NF posterior. The names of the distributions align with the textual descriptions.*

this end, we make use of the NF-likelihood objective (see Section 2.2.2.B) with transformation $f : \mathbb{R} \mapsto \mathbb{R}$ to define the posterior as

$$\log q(\mathbf{z}|\mathbf{s},\mathbf{x}) = \log q_0(\mathbf{z}_0|\mathbf{s},\mathbf{x}) - \sum_{i=1}^{K} \log\left(\left|\det \frac{df_i}{d\mathbf{z}_{i-1}}\right|\right), \tag{4.2}$$

to obtain the final objective

$$\begin{aligned}\mathcal{L} = &- \mathbb{E}_{q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x})}\left[\log p(\mathbf{s}|\mathbf{z},\mathbf{x})\right] \\ &+ \mathrm{KL}\left(q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x}) \,||\, p_\psi(\mathbf{z}|\mathbf{x})\right) - \mathbb{E}_{q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x})}\left[\sum_{i=1}^{K} \log\left(\left|\det \frac{df_i}{d\mathbf{z}_{i-1}}\right|\right)\right].\end{aligned} \tag{4.3}$$

The input-dependent context vector $\mathbf{c}$, is used to obtain the posterior flow parameters. During training, the posterior flow is used to capture the data distribution with the posterior network $Q(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{c}|\mathbb{X}, \mathbb{S})$, followed by sampling thereof to reconstruct the segmentation predictions $\mathbb{Y}$. At the same time, a prior network $P(\boldsymbol{\mu}, \boldsymbol{\sigma}|\mathbb{X})$ only conditioned on the input image, is also trained through constraining its KL divergence with the posterior distribution. The first term in Eq. (4.3) entails the reconstruction loss, in this case the cross-entropy function as mentioned earlier. At test time, the prior network produces the latent samples to construct the segmentation predictions.

4.3.2 Data and 2D baseline experiments

In this work, extensive experimental validation is performed using the Vanilla Probabilistic U-Net with 2-step or 4-step planar-flow and radial-flow variants on processed versions of the LIDC-IDRI (LIDC) [196] and the Kvasir-SEG [201] datasets. The preprocessed LIDC dataset [199] transforms the 1,018 thoracic CT scans with four assessors into 15,096 patches of 128×128-pixels, according to prior work [189], [202]. Each image has 4 annotations. The Kvasir-SEG dataset contains 1,000 polyp images of the gastrointestinal tract from the original Kvasir dataset [203]. The images are resized to be 128×128 pixels and converted to 8-bit depth grayscale images. Example images of the datasets can be found in Appendix C.1. The employed NFs include the planar flow, which conforms to related work [199], [204] and additionally, we experiment with the radial flow. These flows are often chosen because they are computationally inexpensive transformations that possess the ability to expand and contract the distributions along a direction (planar) or around a specific point (radial) and is fully discussed in Chapter 2.

4.3.3 Performance evaluation on 2D experiments

For evaluation, the *Squared Generalized Energy Distance* (GED) (also known as the *Maximum Mean Discrepancy*) is considered as a metric. This metric is defined as

$$D^2_{GED}(P_{pr}, P_{out}) = 2\,\mathbb{E}\left[d(\mathbb{S}, \mathbb{Y})\right] - \mathbb{E}\left[d(\mathbb{S}, \mathbb{S}')\right] - \mathbb{E}\left[d(\mathbb{Y}, \mathbb{Y}')\right], \tag{4.4}$$

where $\mathbb{Y}$, $\mathbb{Y}'$ and $\mathbb{S}$, $\mathbb{S}'$ are independent samples from the predicted distribution and ground-truth distributions P_{pr} and P_{gt}, respectively. Here, parameter d is a distance metric, in this case, unity minus the 2D Intersection over Union (1-IoU). When the predictions poorly match the ground truth, the GED is prone to simply reward diversity in samples, instead of accurate predictions because the influence of the $\mathbb{E}\left[d(\mathbb{Y}, \mathbb{Y}')\right]$ term becomes dominant. Therefore, we also evaluate the Hungarian-matched IoU, using the average IoU of all matched pairs for the LIDC dataset. We duplicate the ground-truth set, hence matching it with the sample size. Since the Kvasir-SEG dataset only has a single annotation per sample, we simply take the average IoU from all samples. Furthermore, when the model correctly predicts the absence of a lesion (i.e. no segmentation), the denominator of the metric is zero and thus the IoU becomes undefined. In previous work, the mean excluding undefined elements was taken over all the samples. However, since this is a correct prediction, we award this with full score (IoU= 1) and compare this approach with the method of excluding undefined elements for the GED.

To qualitatively depict the model performance, we calculate the mean and standard deviation with Monte-Carlo simulations (i.e. sampling reconstructions from the prior). All evaluations in this work are based on 16 samples to strike a right balance between sufficient samples and a justifiable approximation, while maintaining minimal computational time.

For quantitative evaluation, the squared Generalized Energy Distance (GED) is adopted, as this metric is also used in previous work on this topic. We hypothesize

that this commonly used metric is prone to some biases, such that it rewards sample diversity rather than predictive accuracy. Therefore, we also evaluate on the average and Hungarian-matched IoU[1] for the single/multi-annotated data, respectively, as is also done by Kohl *et al*. [205].

To qualitatively evaluate the ability to model the intervariability of the annotations, we present the means and standard deviations of the segmentation samples reconstructed from the model. In this work, we exploit of the multi/single-annotated LIDC-IDRI (LIDC) and Kvasir-SEG datasets, thereby handling limited dataset size and giving insights on the effects of the complex posterior on hard-to-fit datasets.

4.3.4 Training details

The training procedure entails tenfold cross-validation using a learning rate of 10^{-4} with early stopping on the validation loss, based on a patience of 20 epochs. The batch size is chosen to be 96 and 32 for the LIDC and the Kvasir-SEG dataset, respectively. The number of dimensions in latent space is set to L=6. The dataset is split according to the ratio of 90/10 (train+validation/test) and is evaluated on the test set using the proposed metrics. All experiments are executed on an 11-GB RTX 2080TI GPU.

4.3.5 Results and discussion

Quantitative evaluation: The models are referred to by their posterior structure, either unaugmented (Vanilla) or with their n-step Normalizing Flow (NF). The results of the conducted experiments are presented in Table 4.1. In line with literature, it is shown that the GED improves with the addition of an NF. This hypothesis is tested using both a planar and radial NF and it is observed that both have a similar effect. Furthermore, both the average and Hungarian-matched IoU improve with the NF. It is seen that the 2-step radial (2-radial) NF is slightly better than other models for the LIDC dataset, while for the Kvasir-SEG dataset the planar models tend to perform better. The original PU-Net introduces the capturing of the variability of annotations into a Gaussian model. However, this distribution is not expressive enough to efficiently capture this variability. The increase in GED and average IoU performance from the experiments confirm the hypothesis that applying NF to the posterior distribution of the PU-Net improves the accuracy of the probabilistic segmentation. This improvement occurs because the posterior becomes more flexible and can thus provide more meaningful updates to the prior distribution.

Including/excluding correct empty predictions does not result in a significant difference in the metric value when comparing the Vanilla models with the posterior NF models. The results show that the choice in the NF has minimal impact on the performance and suggest practitioners to experiment with both NFs. Another publication in literature [199] has experimented with more complex posteriors

[1]This way of matching checks all possible combinations between two sample sets.

such as GLOW [206], leading to no additional performance improvement. In the conducted research, we have found that even a 4-step planar or radial NF (which are much simpler in nature) can already be too complex for our datasets, yielding no additional performance. A possible explanation is that the variance in annotations captured in the posterior distribution only requires a complexity that manifests from two NF steps. This degree of complexity is then most efficient for the updates of the prior distribution. More NF steps would then possibly introduce unnecessary model complexity as well parameters for training, thereby reducing the efficiency of the updates. Another explanation could be that an increase in complexity of the posterior distribution does in fact model the annotation variability in a better way. Nevertheless, not all information can be captured by the prior, as it is still a Gaussian distribution. In this case, a two-step posterior is close enough to a Gaussian for meaningful updates, yet complex enough to be preferred over a Gaussian distribution. We consider that for similar problems, it is better to adopt simple NFs with only a few steps. However, in cases where the varying nature in the ground truth follows different characteristics, e.g. encompassing non-linearities, the need for a more complex NF should also be considered.

Dataset	Posterior	GED ↓		IoU ↑	
		Excl.	Incl.	Avg.	Hungarian
LIDC	Vanilla	0.33 ± 0.02	0.39 ± 0.02	—	0.57 ± 0.02
	2-planar	0.29 ± 0.02	0.35 ± 0.03	—	0.57 ± 0.01
	2-radial	$\mathbf{0.29 \pm 0.01}$	$\mathbf{0.34 \pm 0.01}$	—	$\mathbf{0.58 \pm 0.01}$
	4-planar	0.30 ± 0.02	0.35 ± 0.04	—	0.57 ± 0.02
	4-radial	0.29 ± 0.02	0.34 ± 0.03	—	0.57 ± 0.01
Kvasir-SEG	Vanilla	0.68 ± 0.18	0.69 ± 0.17	0.62 ± 0.07	—
	2-planar	0.62 ± 0.05	0.63 ± 0.05	$\mathbf{0.71 \pm 0.01}$	—
	2-radial	$\mathbf{0.63 \pm 0.03}$	$\mathbf{0.64 \pm 0.03}$	0.66 ± 0.06	—
	4-planar	0.63 ± 0.06	0.67 ± 0.05	0.71 ± 0.04	—
	4-radial	0.65 ± 0.04	0.67 ± 0.05	0.65 ± 0.07	—

Table 4.1 *Test set evaluations with the GED and IoU metrics based on 16 samples. Further distinction in the GED is made on whether the correct empty predictions are included. The IoU is evaluated with the Hungarian-matching algorithm and averaged with the LIDC and Kvasir-SEG dataset, respectively.*

The Vanilla, 2-planar and 2-radial models are compared by depicting their GEDs based on sample size (Appendix C.2). As expected, the GED scores decrease as the number of samples increase. It is also evident that the variability in metric evaluations is less for models with an NF posterior. The NF posteriors consistently outperform the Vanilla PU-Net for the LIDC and Kvasir-SEG datasets. *Qualitative evaluation:* The pixel-based mean and standard deviation based on 16 segmentation reconstructions from the validation set is shown in Figures 4.3 and 4.4. Ideally, it is expected to obtain minimal uncertainty at the center of the segmentation, because annotations mostly agree on the center area in the em-

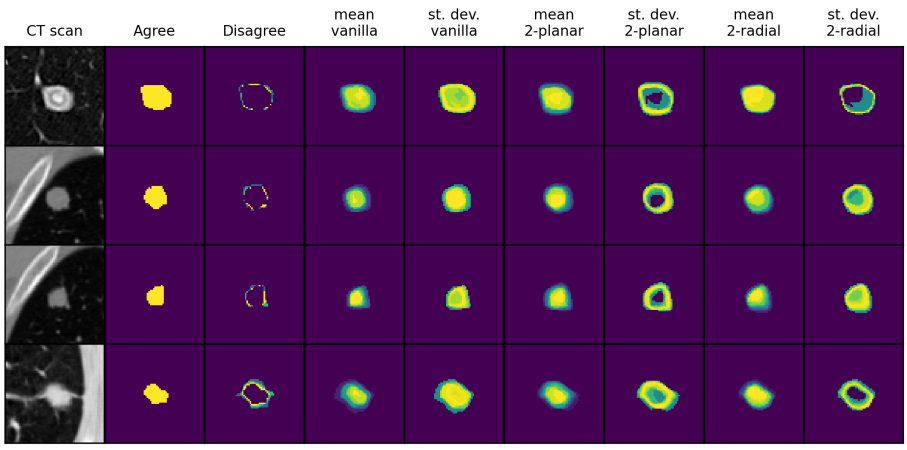


Figure 4.3 *Reconstructions of the LIDC test set.*

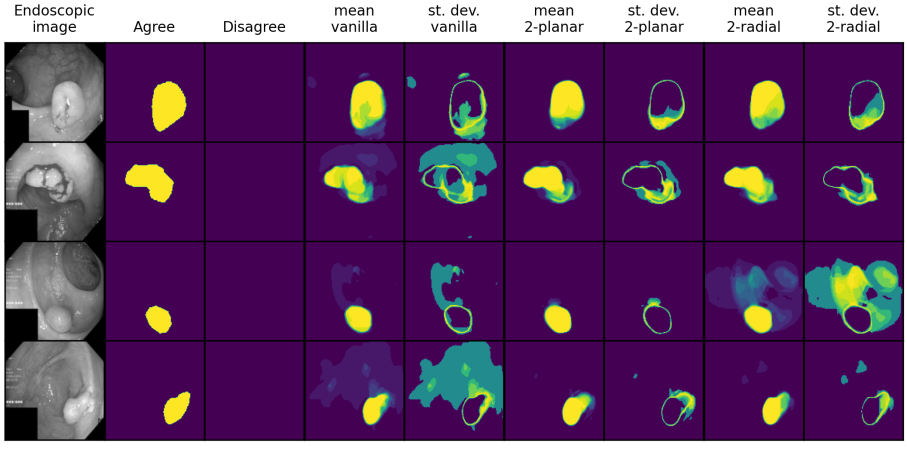


Figure 4.4 *Reconstructions of the Kvasir-SEG test set.*

ployed datasets. This also implies that the mean of the center should be high because of the agreement of the assessors. For both datasets, the means of the sampled segmentation maps match well with the ground truths and have high values in the center areas corresponding to good predictions. Furthermore, the PU-Net without an NF shows uncertainty at both edges and segmentation centers. In contrast, for all NF posterior PU-Net models, the uncertainty is mostly on the edges alone. A high uncertainty around the edges is also expected, since at those areas the assessors almost always disagree. From this, we can conclude that NF posterior models are better at quantifying the aleatoric uncertainty of the data. Even though there is no significant quantitative performance difference between the NF models, there is a clearly distinguishable difference in the visual analysis. In almost all cases, it can be observed that the planar flow is better than the radial NF posterior in learning the agreement between the segmentation centers. The

prior distribution is also investigated to determine if it captures the ambiguity that exist in the input image. In Appendix C.3, the prior distribution variances for different test set input images are shown. Qualitative observations reveal that with increasing variance, the subjective assessment of the annotation difficulty increases. This suggests the possibility of obtaining an indication of the uncertainty in a test input image without sampling and evaluating the segmentation reconstructions.

Follow-up work: The prior distribution is an area that needs to be further explored, since this is still assumed Gaussian. It is hypothesized that augmenting the prior with an NF could result in further improvements. Future work should also include an investigation into the correlation between the prior and segmentation variance. A limiting factor of the proposed model is the use of only a single distribution. We consider that when using flexible distributions at multiple scales, the overall model will further improve. Finally, the complete approach is based on 2D images, that in the case of the lung nodules, comes from 3D volumes. Utilizing the full 3D information can yield additional performance improvements.

4.3.6 Conclusions on aleatoric uncertainty quantification in 2D images

Quantifying uncertainty in image segmentation is important for decision-making in the medical domain. This work has proposed to use the broader concept of NFs for modeling both single/multi-annotation data. This concept allows more complex modeling of the aleatoric uncertainty in the image segmentation task. Modeling of the posterior distribution by Gaussians is too restrictive to model variability contained in the data. By augmenting the model posterior with a planar NF or radial NF, up to 14% improvement in GED and 13% in IoU is obtained, resulting in an improved quantification of the aleatoric uncertainty.

Density modeling with normalizing flows (NFs) should be explored in various ambiguous contexts within the medical domain, as this approach is expected to yield valuable insights for future research. This work reveals that significant improvement can be obtained by only augmenting the posterior distribution with NFs, whereas little-to-none investigations have been made into the effect of additionally augmenting the prior distribution. These improvements are realized with an approach based on 2D images derived from 3D volumes containing lung nodules. Utilizing the full 3D data may further enhance performance which is explored in the next section.

4.4 Probabilistic 3D segmentation for aleatoric uncertainty quantification

Deep learning-based semantic segmentation methods using convolutional neural networks have successfully been adopted as CAD methods for a wide range of medical imaging modalities. While research has been conducted towards quantifying the types of uncertainty occurring when using a segmentation model, most of this work is limited to the quantification of the uncertainty in two-dimensional

(2D) slices or images, where the latter often originate from a 3D volume such as in CT and MRI. However, these approaches fail to exploit the rich 3D features that may help in resolving ambiguities in the volume.

This study focuses on lung nodule segmentation as its primary application domain. Given the noisy nature of CT imaging, the low contrast of the nodules and the significant variability in their locations, a substantial ambiguity is contained within the data, which has to be considered by clinicians when segmenting lung nodules. As in discussed 2D case, we employ the LIDC-IDRI lung CT dataset [196], which makes use of multiple ground-truth annotations per lung nodule. As described in Section 4.2.1, the epistemic uncertainty – i.e. preferences, experiences and knowledge – of the assessors manifests into aleatoric uncertainty when providing annotations as ground-truth data. The different annotations per nodule adds ambiguity during training of a segmentation network. During the annotation process, the radiologist typically annotates on a single 2D plane of the 3D volume. However, whilst annotating, full access to the other two views of the CT scan are typically available on the same screen. This allows the assessor to correct the annotation if it does not align with the other two views and as a result, the assessor creates a true 3D annotation. In the LIDC-IDRI dataset (see Section 4.4.3) assessors were allowed multiple rounds of annotation, thereby potentially increasing the quality of annotations by exploiting the available full 3D information.

In recent research, various methods have been proposed to quantify the uncertainty arising in segmentation models or resulting from images. One increasingly popular approach is the Probabilistic U-Net [189], as proposed by Kohl *et al*. and extensively discussed in Section 4.2.3. They propose combining a 2D U-Net with a conditional variational autoencoder (VAE) capable of learning a distribution over the possible annotations and ultimately construct a generative segmentation model. The Probabilistic U-Net provides compelling results in resolving the ambiguity in an image. In Section 4.3 improvements to this model have been proposed by adding a Normalizing Flow to the posterior network of the Probabilistic U-Net. This allows the model to move away from modeling the ambiguity as strictly axis-aligned Gaussian and, instead, allows for a learnt posterior distribution of varying complexity.

Extending on these advancements, the extension to 3D provides the following contributions. First, in Section 4.4.1 a 3D probabilistic framework which builds upon the research from Kohl *et al*. is proposed. The model exploits the full-3D spatial information to resolve the uncertainty in the original CT volumes (Section 4.4.3). Second, in Section 4.4.5 it is shown that similar to the 2D case, more diverse 3D segmentation maps are obtained when the posterior distribution is enhanced by a Normalizing Flow. Third, the proposed method's ability to capture uncertainty on the LIDC-IDRI lung nodule datatset is tested and the results in the 3D version of the GED metric are presented. It is shown that a high segmentation accuracy is obtained using a Hungarian-matched 3D IoU. In Section 4.4.6, these findings are discussed, which suggest that such modeling enables capturing the uncertainty more accurately.

4.4.1 3D Model Architecture

This research extends the Probabilistic U-Net to the 3D domain and addresses a key limitation by augmenting the posterior network with a Normalizing Flow (NF). The network consists of a 3D U-Net, a 3D Prior network, 3D Posterior network enhanced with an NF, and a Feature Combination network. All the 2D operations of the PU-Net are replaced with their 3D equivalents. By combining the 3D spatial features extracted by the U-Net with samples taken from a latent distribution encapsulating the solution space, a set of diverse yet plausible segmentation maps can be generated. The standard deviations across these predictions can be interpreted as the aleatoric uncertainty. The U-Net [12] and 3D U-Net [154] have regularly shown their ability to segment structures of interest at state-of-the-art performance. Although we employ the 3D U-Net to obtain the relevant 3D spatial information, this approach is generic and allows for any other segmentation network to be used. A deep 3D CNN conditioned on the input CT scan is used to model a low-dimensional axis-aligned Gaussian latent space, representing the segmentation variants (prior distribution). Another CNN-based axis-aligned Gaussian encoder (posterior network) that is conditioned on both the query CT scan and a ground-truth segmentation, is utilized during training to model a posterior low-dimensional latent space. Earlier in Section 4.3, we have pointed out the shortcomings in modeling the posterior distribution to be strictly Gaussian. As such, the posterior network is augmented with either a two-step planar or radial flow [88], to potentially increase the complexity of the captured posterior distribution, thereby providing more meaningful updates to the prior network during training.

Figure 4.5 portrays a detailed diagram of the proposed network architecture. During training, the probabilistic 3D U-Net makes use of the posterior network, prior network, U-Net and the feature combination layers. Samples are taken from the image-label conditional distribution, captured by the posterior network and combined with the features extracted from the U-Net through the feature combination network. The loss is then computed using Equation (4.5). The prior network follows the posterior network during training, as enforced by the KL-divergence, and thus learns to capture this image-label conditional distribution from the image alone. At test time, the posterior network is discarded and instead, samples are taken from the prior network. It should be noted that there is only one forward pass through the U-Net (for image feature extraction) and the prior network (to capture the image-conditional distribution). However, multiple passes through the feature combination network are made, in order to combine a new sample from the prior distribution with the image features. The code for the proposed probabilistic 3D U-Net is publicly available [2].

[2]Code available at https://github.com/cviviers/prob_3D_segmentation

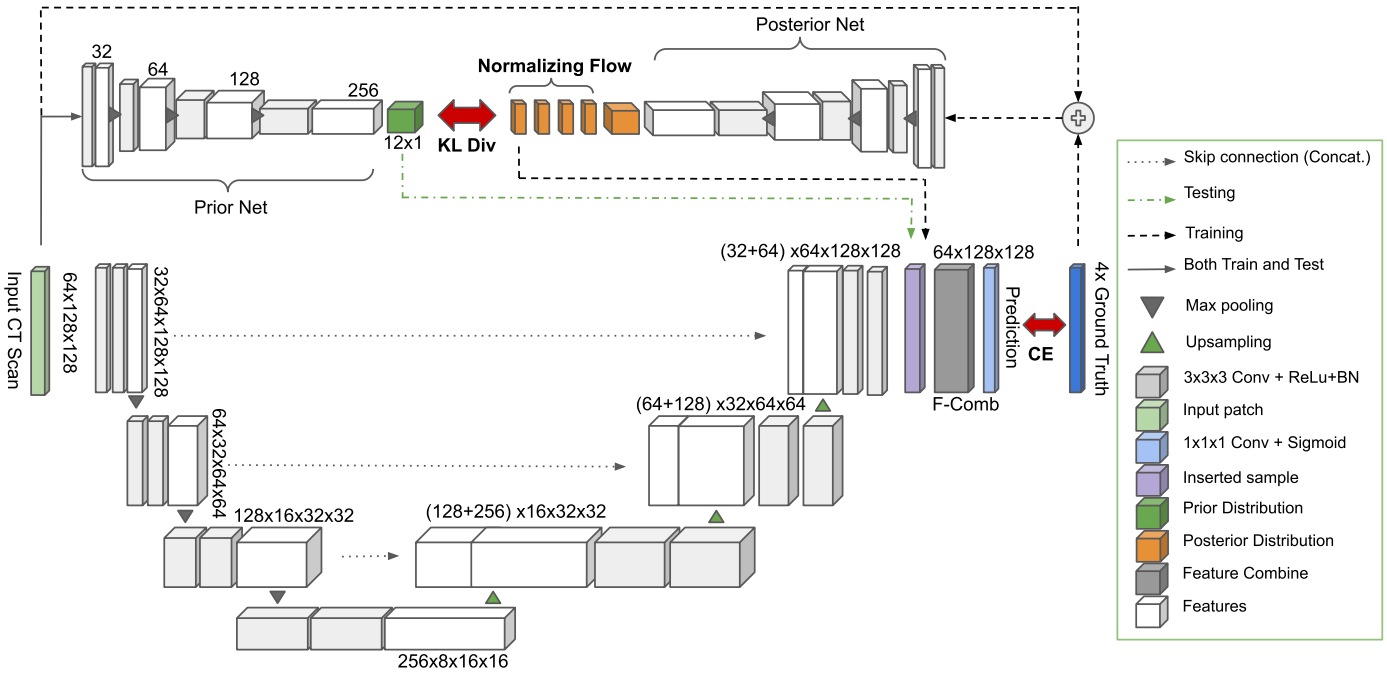


Figure 4.5 *Diagram of the 3D Probabilistic U-Net with an augmented-flow posterior. The bottom network depicts the 3D U-Net, the prior and posterior networks are shown at the top and a feature combination network at the right combines samples taken from the captured distributions. The diagram additionally depicts both the training and testing configuration. The legend depicts the individual operations and the dimensions of the respective features (after the applied operations) shown in the diagram.*

4.4.2 3D loss function and evaluation criteria

In line with previous work on conditional variational autoencoders, the training objective consists of minimizing the variational lower bound [47]. This entails minimizing (1) a cross-entropy difference (in our case) between the ground-truth segmentation (**y**) and a prediction (**s**), (2) the Kullback-Leibler (KL) divergence between the posterior distribution (p_ϕ) and the prior distribution (p_θ), and finally, (3) a correction term for the density transformation through the Normalizing Flow [204]. Given a query image (**x**) and a posterior sample (**z**), the feature combination network combines the sample with the features extracted by the U-Net to generate a plausible segmentation (**s**). This loss term can formally be specified by

$$\begin{aligned}\mathcal{L}(\mathbf{y},\mathbf{x},\theta,\phi,\psi) = &- \mathbb{E}_{p_\phi(\mathbf{z}|\mathbf{y},\mathbf{x})}\left[\log p_\psi(\mathbf{y}|\mathbf{z},\mathbf{x})\right] \\ &+ \beta \cdot \left(\mathrm{KL}\left(p_\phi(\mathbf{z}_0|\mathbf{y},\mathbf{x}) \,||\, p_\theta(\mathbf{z}|\mathbf{x})\right) - \mathbb{E}_{p_\phi(\mathbf{z}_0|\mathbf{y},\mathbf{x})}\left[\sum_{i=1}^{K} \log\left(\left|\det \frac{df_i}{d\mathbf{z}_{i-1}}\right|\right)\right]\right).\end{aligned} \tag{4.5}$$

These loss terms are combined and the second large term is weighted with hyperparameter β [207], [208]. A detailed derivation of the ELBO loss [47] is presented in Section 2.3. Furthermore, how it is used in the context of PU-Net [189] and the NF-likelihood objective [56], [88], [204] is performed in the same way as in the 2D case and shown in Section 4.3.1. The metric Squared Generalized Energy Distance (GED) has become the *de-facto* metric in the context of uncertainty quantification and the quantification of the distance between distributions of segmentation maps. This GED metric is defined as

$$D^2_{\mathrm{GED}}(P_{\mathrm{GT}}, P_{\mathrm{Out}}) = 2\mathbb{E}\left[d(\mathbb{S},\mathbb{Y})\right] - \mathbb{E}\left[d(\mathbb{S},\mathbb{S}')\right] - \mathbb{E}\left[d(\mathbb{Y},\mathbb{Y}')\right], \tag{4.6}$$

where d is a distance measure which equals $1-\mathrm{IoU}_{3\mathrm{D}}(x,y)$, hence the 3D version of the IoU in this implementation. The parameters $\mathbb{S}$ and $\mathbb{S}'$ are independent samples from the predicted distribution P_{Out}. The parameters $\mathbb{Y}$ and $\mathbb{Y}'$ are the 4 samples from the ground-truth distribution P_{GT}. In addition to the GED, we also report the Hungarian-matched IoU. This compensates for a shortcoming in the GED that when the predictions are relatively poor, the metric rewards sample diversity by definition. We duplicate the ground-truth set (4 annotations) to match the desired sample number when computing the Hungarian-matched 3D IoU. This measure calculates the distance between two discrete distributions by determining an optimal coupling between the ground-truth and prediction set subject to the IoU metric (involving all sample combinations).

4.4.3 Dataset and 3D data preparation

To evaluate the proposed method's ability to capture the ambiguity in the data, we use the popular LIDC-IDRI dataset [196]. This dataset contains the lung CT scans from 1,010 patients with manual lesion annotations from up to 4 experts. In total, there are 1,018 CT scans potentially containing multiple lung nodules of different levels of malignancy. In this work, we have used the annotations from a second reading, in which the radiologists were presented an anonymized version of the annotations from other experts and were allowed to make adjustments to their

own annotations. Contrary to previous work, we use every nodule in the dataset if it has been annotated by at least one radiologist (potentially missed by three), regardless of the shape or severity of the nodule. We preprocess the CT scans by clustering all nodule annotations for a scan through a computation of a distance measure between the annotations. If an annotation is within one-voxel spacing of that particular CT scan from another annotation, it is grouped to belong to the same nodule. The scan is resampled to 0.5 mm along the x, y-dimensions and 1 mm along the z-dimension to obtain uniform voxel spacing between all samples. This is followed by cropping the CT scan and resulting annotations based on the center of the first assessor's mask with a dimension of $96 \times 180 \times 180$ voxels in the z, x, y-dimensions. Finally, if the nodule does not have at least four annotations, the ground-truth (GT) masks are filled with empty annotations. This addition is made to be consistent with previous work on this dataset [56], [189] and to capture the difficulty in detecting a nodule. This results in a total of 2,651 3D patches, each containing a nodule and four annotations. Visual examples of the nodule in the CT scan and the four ground-truth annotations are depicted in Figure 4.6.

4.4.4 Experiments

To compare the proposed approach against prior work, we conduct six experiments. We train the (1) original Probabilistic 2D U-Net and the (2) Radial NF-augmented Probabilistic 2D U-Net on 2D axial slices of the 3D volume. In practice, we filter the slices based on the presence of at least one positive annotation from any of the annotating experts and use them for training, to avoid a heavily imbalanced training set. Further experiments include (3) the 3D U-Net, (4) Probabilistic 3D U-Net (3D PU-Net) and an (5-6) NF-augmented (Radial and Planar) 3D PU-Net, which are trained on the 3D patches. In contrast to prior work where the 3D lesion was sliced and split into 2D images, where some 2D slices potentially land in the training set and some in the validation/test set, we conduct the experiments on a per-lesion basis. This avoids any potential model bias caused by the splitting and makes the proposed approach more clinically relevant, since we can present the uncertainty for each lesion.

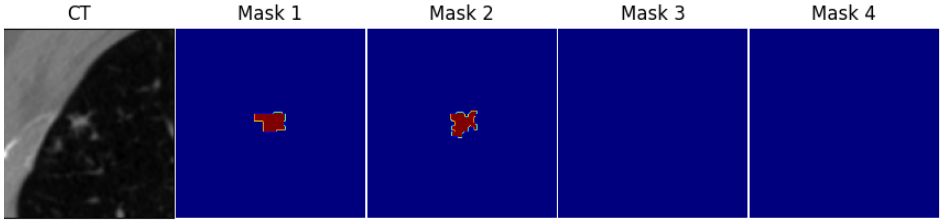


Figure 4.6 *Example nodule in a slice from the CT scan and the four ground-truth annotations.*

The nodule data are split in a 70/15/15 training and validation/test split. During training, a random sample of one of the four annotations is drawn to be used as ground-truth segmentation and the CT volume and label is cropped to

64×128×128 voxels. In line with previous work, for the 3D PU-Net, the dimensionality of the latent space is set to L=6. The proposed framework is implemented in PyTorch and extends on the work conducted by Wolny *et al*. [160]. We have trained using a batch size of 32 in the 2D case and 4 in the 3D case. An Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} is used. The learning rate is reduced by a factor of 0.2 if the validation loss does not decrease after 20 epochs. The parameter β is controlled using a cosine cyclical annealing strategy, as described by Fu *et al*. [207]. In all the 3D PU-Net experiments, we use the same hyperparameters and a hardware configuration with an RTX 3090Ti GPU[3]. Training to completion takes about 2 days on the average. For performance evaluations, we report results using the broadly available RTX 2080Ti GPU (at the time of publication).

4.4.5 Results on 3D experiments

The results of the conducted experiments are shown in Figure 4.7, Figure 4.8, Figure 4.9 and Table 4.2. In Figure 4.7, example predictions from all the models used in the experiments are showcased for qualitative evaluation. Here, μ GT refers to the mean segmentation of the four raters and μ Pred is the mean of the predictions. This mean prediction is the segmentation recommended by the Probabilistic 3D U-Net. Additionally, the figure depicts the variation in the predictions. More specifically, the standard deviation of the ground-truth labels (σ GT) and the logits (after sigmoid activation) resulting from the model predictions (σ Pred) are depicted. In the figure it can be observed that this deviation across the predictions can be interpreted as the uncertainty. We scale the uncertainty heat map visualization to the the maximum standard deviation of the predictions of a particular model. Additionally, the figure depicts a rather conservative segmentation of a part of the lesion from the deterministic 3D U-Net segments, while the other models are capable of producing a more accurate segmentation results. Figure 4.8 depicts the predictions for the 2D and 3D Prob.U-Net for 2 slices from a nodule in the test set. It can be observed that the 2D model misses the nodule in Slice 34, while its 3D counterpart correctly detects it. Although the 2D model has some uncertainty about the presence of the nodule, it is rather low.

In Figure 4.9, multiple consecutive slices are depicted of a CT scan from the test set and Prob. 3D U-Net predictions for a nodule. Slice 25 displays some uncertainty from the model about the presence of a lesion, although no rater indicated its existence yet. In the next slice (26), the lesion is clearly delineated by the raters and the model captures and displays the uncertainty in a similar fashion as the disagreement between the raters. Slices 27-31 and 33-35 are not shown, since the model correctly segments and captures the uncertainty in comparison to the raters. Slice 38 reveals the large lesion as shown by the annotations from

[3]available from Nvidia Inc. Santa Clara, CA, USA

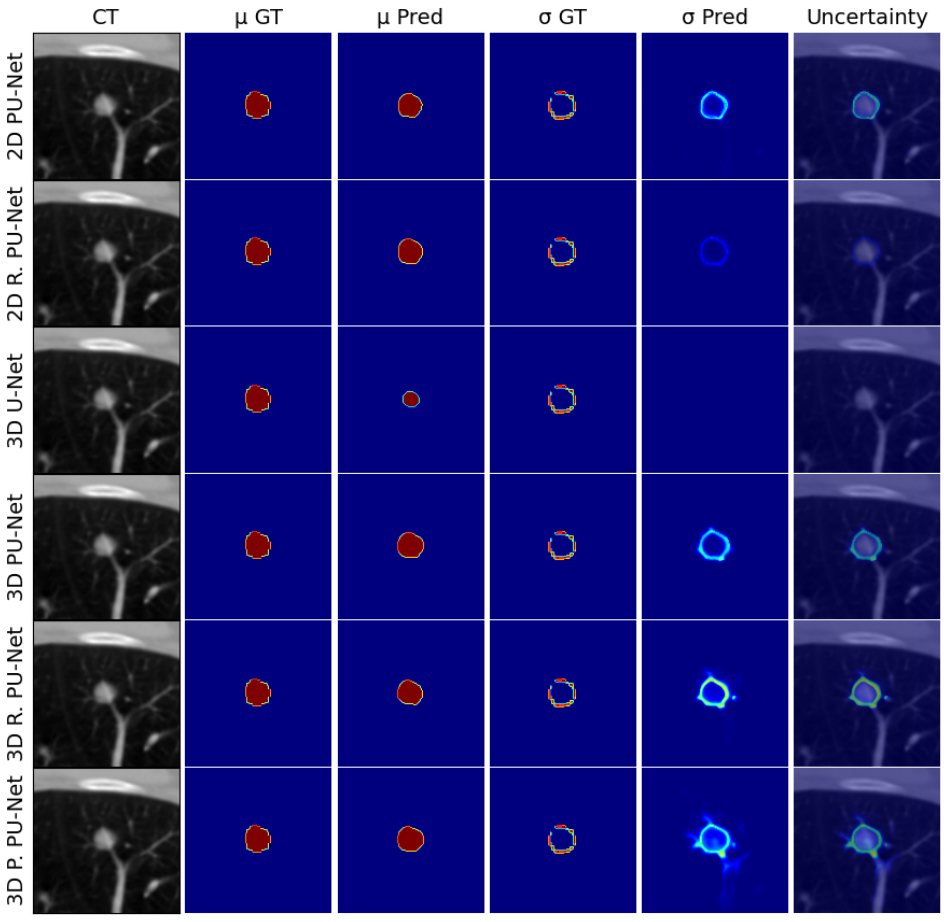


Figure 4.7 *Example predictions for the same data slice with a nodule from the 2D Prob.U-Net, 2D Prob.U-Net with Radial flow, deterministic 3D U-Net, 3D Prob.U-Net, 3D Prob.U-Net with Radial flow and 3D Prob.U-Net with Planar flow in the test set. This same slice is used for fair model-performance comparison.*

the raters, but it rapidly disappears towards Slice 39. However, the model still segments the lesion in Slice 39 and expresses high uncertainty.

Table 4.2 presents a quantitative comparison of the proposed approach with the 2D counterparts, aiming to resolve the ambiguity in the LIDC-IDRI dataset. We compute the 2D GED and 2D Hungarian IoU on a per-slice basis and take the average across all the slices of the lesion, ignoring slices with empty ground-truths and predictions. The 3D GED and 3D Hungarian IoU are immediately computed at a per-case level and then averaged across the test set. In the case of the 2D models, during the forward pass of a single 2D slice, an image-based 2D prior distribution is computed (see the green block at the top left in Figure 4.5). We then randomly draw 16 samples from this distribution. For the next slice in the series of

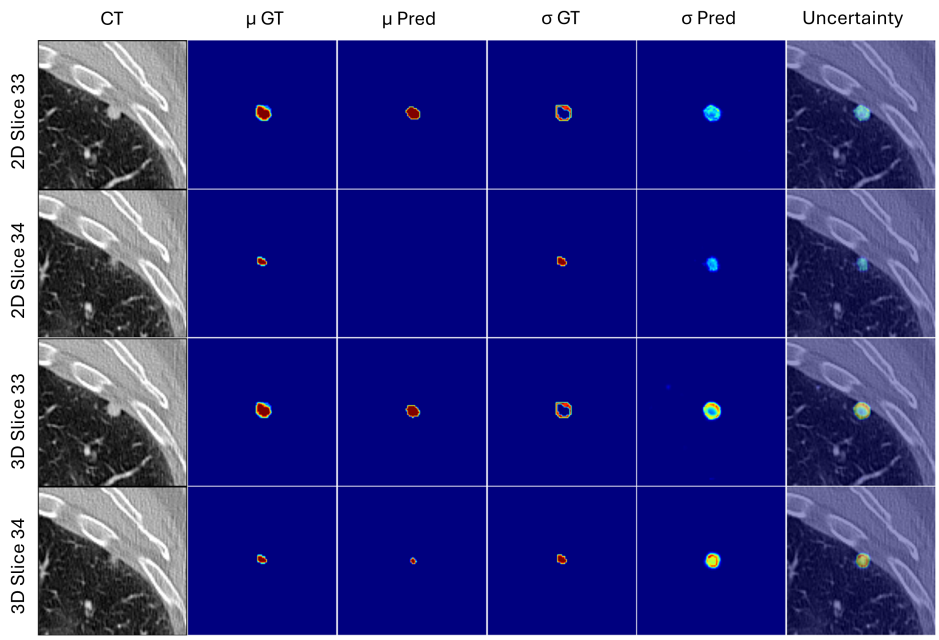


Figure 4.8 *Zoomed example predictions from the 2D & 3D Prob.U-Net for 2 subsequent slices from a nodule in the test set. Slice 34 forms a more difficult case for segmentation and uncertainty quantification.*

the lesion, a completely new 2D prior distribution is presented at inference time. As such, the uncertainty captured by this prior distribution is inconsistent over the individual slices of the lesion and it is not possible to reconstruct a consistent and true 3D segmentation using this approach (certainly not with 3D metrics such as 3D GED and 3D IoU calculation). For the 3D Probabilistic U-Net, we report the 2D Hungarian-matched IoU and 2D GED (2D IoU distance averaged along the z-axis) and the 3D Hungarian-matched IoU and GED.

Model	2D↓ GED	2D↑ IoU	3D↓ GED	3D↑ IoU
Kohl *et al.*	0.445	0.473	N/A	N/A
Valiuddin *et al.*	0.441	0.481	N/A	N/A
3D U-Net	1.283	0.332	1.263	0.383
3D Prob.U-Net	0.427	0.510	0.422	0.457
+ Planar Flow	**0.417**	0.511	**0.393**	0.465
+ Radial Flow	0.429	**0.520**	0.401	**0.468**

Table 4.2 *Evaluations of the baseline, 3D U-Net and Prob. 3D U-Net models on the LIDC-IDRI test set (15%) using the GED and Hungarian IoU metric based on 16 samples.*

For the complexity analysis, we compare the inference time of the 3D U-Net and the Prob. 2D and 3D U-Net per nodule volume (64×128×128 voxels). We

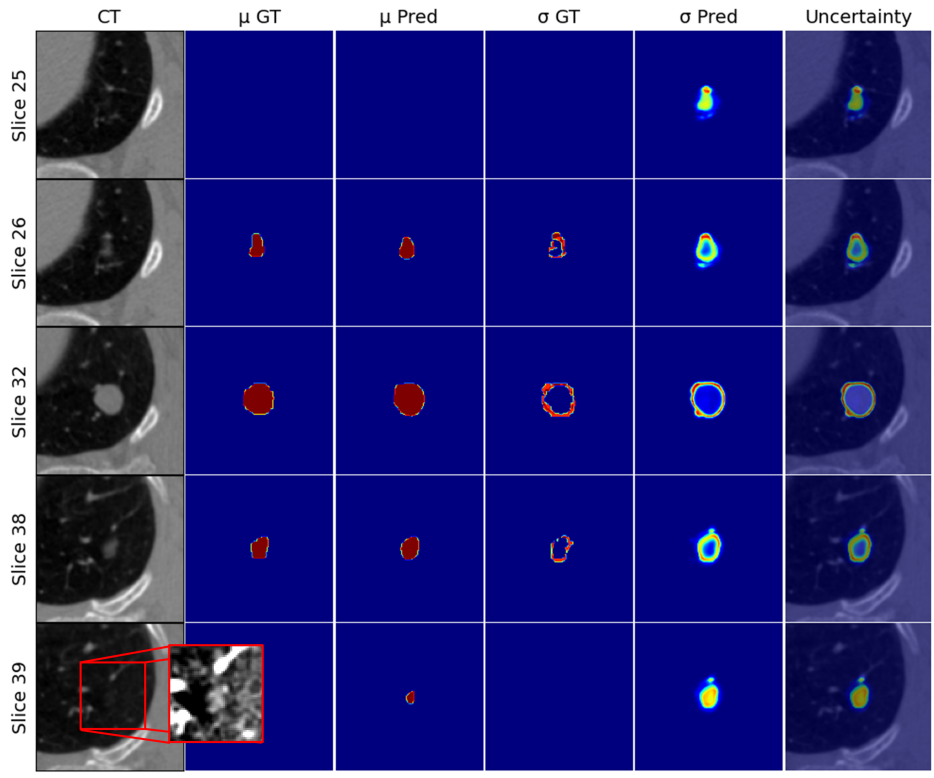


Figure 4.9 *Example 3D Prob.U-Net predictions for multiple slices from a nodule in the test set. For Slice 39 a postprocessed and enlarged view of the nodule is depicted at the adjacent right for improved visibility to show that the lesion is still present.*

do not include additional results on the models with NF-augmented posteriors networks, since the two-step low-dimensional bijective transformation has a negligible computational time in comparison to the rest of the network. Table 4.3 showcases the computation time per operation for the different models and with different batch sizes (BS). It can be noted that the Prob. 3D U-Net has a shorter inference time for the above-given volume, compared to its 2D counterpart.

Operation	Deterministic	Probabilistic		
Model	3D U-Net BS=1	2D PU-Net BS=1	2D PU-Net BS=64	3D PU-Net BS=1
Forward pass	2.34 ms	5.75 ms×64	397.27 ms	124.31 ms
Sample + F-comb	N/A	0.51 ms×64×16	157.09 ms×16	8.44 ms×16
Total	2.34 ms	892.29 ms	2910.51 ms	259.35 ms

Table 4.3 *Inference time (per operation) of the utilized models, 16 samples from the Probabilistic 2D and 3D U-Net per nodule (64×128×128 voxels). BS is the Batch Size. One CPU core and the GPU was available during execution time computation.*

4.4.6 Discussion

This research extends the Probabilistic U-Net to the 3D domain to utilize the rich 3D spatial information when resolving the uncertainty. We have introduced the Probabilistic 3D U-Net and have employed recent improvements in the 2D Probabilistic U-Net, by adding either a Planar or Radial flow to the posterior network. This augmentation with NFs enables capturing distributions of various complexity, thereby relaxing the strictly axis-aligned Gaussian constraint previously employed. To test the model's ability to capture the aleatoric uncertainty, we have used the LIDC-IDRI dataset for benchmark tests.

General segmentation performance: Section 4.4.5 displays the results of the conducted experiments. For qualitative evaluations, Figure 4.7 showcases example predictions from all the applied models. It can be noted that all models perform well on this clearly defined lesion, except for the 3D U-Net. The 3D U-Net delineates the lesion in a conservative manner, possibly due to being exposed to many empty ground-truth labels during training and not being able to capture this ambiguity in a meaningful way. The 3D PU-Net with Planar flow expresses more uncertainty about various parts of the lesion.

2D vs. 3D segmentation performance: Figure 4.8 highlights example predictions of the 3D PU-Net where information from the complete 3D volume is used to detect, segment and resolve the uncertainty about the lesion in the axial Slice 34 of the CT scan. The same nodule is missed by the 2D PU-Net, due to lack of information from prior slices. These results are also reflected in the 2D GED and Hungarian IoU, as shown in Table 4.2, with the 3D models outperforming the 2D models.

Limitations of 2D annotations: Interestingly, in Figure 4.9, the 3D spatial awareness of the model is showcased through the uncertainty expressed in Slice 25. A rather large lesion is coming up (iterating through the CT slices in an ascending order) and the model expresses uncertainty about the exact starting position, since the raters have not annotated a lesion followed by a large lesion in consecutive slices. The same phenomenon can be seen moving from Slice 38 to 39, although here the model incorrectly (according to the raters) presents a lesion segmentation of the lesion while it is, in fact, still partially visible.

Execution complexity: By the performance evaluations presented in Table 4.3, it can be observed that it is most efficient to present the uncertainty with a 3D PU-Net. For a volume of $64\times128\times128$ voxels, the 64 2D slices can be passed through the 2D PU-Net in a large batch, but it scales poorly in comparison to a single-slice forward pass. The forward pass of the 2D PU-Net with a batch size of 64 takes 397.27 ms to compute. Drawing 16 samples from the Prior distribution and combining it with the 2D U-Net features through the Feature Combination network (F-Comb in Figure 4.2) takes 2513.44 ms ($157.09\text{ ms}\times16$). In total this approach will take 2910.51 ms for execution, compared to the approximate $10\times$ speed improvement that is required for calculating the uncertainty with the 3D PU-Net (259.35 ms). It should be noted that significant computational time drawbacks

occur when using the 3D PU-Net in comparison to the the standard 3D U-Net (2.34 ms for inference), although this is not a realistic alternative because uncertainty cannot be expressed.

4.4.7 Conclusions on probabilistic 3D segmentation

In CAD methods, it is important to provide clinicians with an accurate measure of uncertainty when they evaluate and plan their procedures. Accurately capturing and presenting segmentation uncertainty will increase clinical confidence in model predictions and facilitate more informed decision-making. Existing CT-based segmentation methods aim to realize this by quantifying the uncertainty from 2D image slices, whereas the true uncertainty resides in the full 3D CT or MRI volume. A novel 3D probabilistic segmentation model is proposed that is capable of resolving and presenting the aleatoric uncertainty in 3D volumes through diverse and plausible nodule segmentation maps. The model consists of a Deep 3D U-Net and a 3D conditional VAE that is augmented with a Normalizing Flow (NF) in the posterior network. Since NFs allow for more flexible distribution modeling, we have alleviate the strictly Gaussian posterior distribution that was previously enforced. We have tested the approach on the LIDC-IDRI lung nodule CT dataset.

This is among the first implementations that presents the 3D Squared Generalized Energy Distance (GED) and 3D Hungarian-matched IoU for lung nodule segmentation and uncertainty prediction. We have quantified the uncertainty prediction performance and achieved a GED of 0.401 and a Hungarian-matched 3D IoU of 0.468 with the Radial 3D PU-Net. This approach also outperforms the 2D counterpart on the 2D GED and 2D Hungarian IoU. In addition, since the model uses the full original 3D volumes, it is a step closer to the practical application of accurately delineating and presenting uncertainty in 3D CT data. Finally, we present the aleatoric uncertainty, computed as the standard deviation across the model predictions, in a visual manner. This enables an interpretable expression of the uncertainty and is potentially providing clinicians additional insight into data ambiguity and allowing for more informed decision-making.

4.5 Overall Conclusion

In this chapter, we have explored various methods for uncertainty quantification in medical image segmentation, a critical aspect for enhancing decision-making in the medical field. Initially, the types of uncertainty encountered in medical image segmentation are discussed, specifically epistemic and aleatoric uncertainties. The focus is set on aleatoric uncertainty, which is an inherent ambiguity in the data. Various approaches for quantifying this type of uncertainty are highlighted, amongst which the preferred model is the Probabilistic U-Net (PU-Net). The PU-Net is examined as a method for handling ambiguities in image segmentation. Despite its innovative approach, the model limitations have been identified, paving the way for improvements and further research.

The presented research is continued by focusing on improving aleatoric uncertainty quantification in multi-annotated medical image segmentation using Normalizing Flows (NFs). By augmenting model posterior distributions with Planar or Radial NFs, notable improvements in quantifying the uncertainty are achieved. This is evident from the enhancements in Generalized Energy Distance (GED) and Hungarian-matched Intersection over Union (IoU) metrics, as well as various qualitative improvements in nodule segmentation. This method demonstrates the ability to model more complex segmentation uncertainties and suggests further exploration of NFs in various medical imaging contexts.

The chapter has introduced a novel approach also for probabilistic three-dimensional (3D) segmentation to address the inherent limitations of 2D slice-based methods. By leveraging a Deep 3D U-Net and a 3D conditional VAE augmented with NFs, successful modeling and aleatoric uncertainty in 3D CT volumes are presented. This approach results in improvements in various metrics, including a 3D GED of 0.401 and a Hungarian-matched 3D IoU of 0.468, which proves the potential for 3D processing and modeling for practical applications in clinical settings.

In conclusion, the advancements discussed in this chapter, particularly the use of NFs and the transition to 3D modeling, represent significant steps forward in the accurate quantification and presentation of uncertainty in medical image segmentation. These developments not only improve model performance, but also enhance the interpretability and confidence of clinicians in model predictions, ultimately contributing to more informed and reliable medical decision-making.

The 3D modeling and discussion on uncertainties is further explored in the succeeding chapter, yet using different research directions for extensions. The chapter focuses on prediction of PDAC resectability, where the uncertainty about the tumor and surrounding vessels are extensively modeled. The variations in the structures in this pancreatic setting is much more complex than the contour of a lung nodule. Therefore, the variation and interaction in the structure contours are utilized to compute a PDAC resection metric. This metric aligns with the clinical way-of-working in pancreatic cancer treatment.

Chapter 5 Prediction of PDAC resectability from 3D segmentation

5.1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most aggressive malignancies with a dismal prognosis and an overall five-year survival rate of less than 10% [209]. Despite recent advancements in the field of oncology, pancreatic cancer often goes undetected until it has progressed into an advanced stage. As a result, the majority of patients have advanced or metastatic disease, leading to limited treatment options and poor outcomes [210]. With its high mortality rate and limited treatment options, identifying the optimal approach for pancreatic cancer patients remains a crucial area of clinical concern. In recent years, the concept of *resectability* [211]–[213] has emerged as a pivotal factor in determining the appropriate treatment strategy, emphasizing the importance of accurately assessing the feasibility of curative surgical resection.

Pancreatoduodenectomy (PD) is the cornerstone for surgical treatment of pancreatic cancer. However, this procedure poses significant technical challenges and is associated with a considerable morbidity rate, ranging from 20% to 30% [214]. However, only 20% of the patients are considered eligible for resection upon initial diagnosis [125]. Therefore, it is essential to carefully evaluate vascular involvement of the tumor and identify potential arterial anatomical variations during preoperative assessment. These factors play a critical role in determining the feasibility of surgical resection [215]. Currently, multi-phase contrast-enhanced

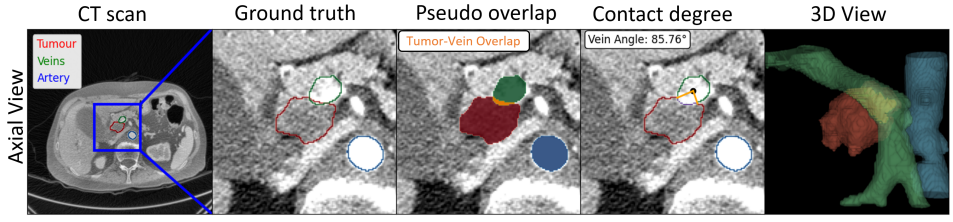


Figure 5.1 *Slice from CT scan depicting the involvement between the tumor and vein, a pseudo overlap label and the computed angle of involvement based on contact pixels (purple). Other labeled structures are omitted.*

multi-detector computed tomography (MDCT) is the gold standard for evaluating pancreatic cancer and determining the resectability. Standardized resectability criteria are used to tailor the need for neoadjuvant therapy and select patients for (minimally-invasive) surgical resection [211]. Resectability is graded as either *resectable*, *borderline resectable*, or *irresectable*, based on the degrees of contact between the tumor and surrounding vasculature [212], [213], [216]. However, determining surgical resectability based on CT scans only can be difficult, especially after neoadjuvant treatment. Tumor regression after neoadjuvant treatment is rarely visible on CT and the amount of vascular involvement tends to be overestimated [212], [217]–[219]. Moreover, existing literature demonstrates significant interobserver variability, even among highly experienced clinicians [220], [221]. This motivates why clinicians have severe difficulties in accurately assessing tumor resectability [222]. Further discussion on methods for resectability assessment is provided in Appendix B.2.

By leveraging a deep learning-based clinical decision-support system (CDSS), there is a potential for significant improvement in resectability assessment, assisting clinicians, enhancing the overall accuracy, while incorporating interobserver variability of the process. This research proposes a workflow to acquire a focused region of interest containing PDAC and the surrounding anatomical structures. Three deep learning-based segmentation architectures are implemented to delineate the structures of interest to ultimately present multiple levels of relevant clinical information. The initial segmentation maps are used to assess tumor size and its location with respect to the surrounding anatomy. Sufficiently accurate segmentation of the tumor and surrounding vessels enables the estimation of two aspects: (1) determining *if* there is any involvement of the tumor with vessels and (2) the extent of the involvement. Addressing these points will already provide information on the resectability of the tumor. Moreover, each of these steps carry additional clinical value and further insights into patient treatment options. Finally, we present the ambiguity captured by each of the models and show how this ambiguity can facilitate in the decision-making process of tumor resectability.

This chapter aims to advance the understanding and application of automated resectability prediction in PDAC. Building on the groundwork laid in Chapters 3 on various segmentation techniques and the uncertainty quantification methods discussed in Chapter 4, this chapter develops and evaluate a workflow and deep learning-based segmentation models to automatically assess tumor-vessel involvement, which is essential for determining tumor resectability. In this context, we present the following research questions.

- *Accurate segmentation* of pancreatic ductal adenocarcinoma (PDAC) and the surrounding vasculature is crucial as it forms the foundation for subsequent analyses, including the assessment of tumor size, location, and involvement with nearby blood vessels. The effectiveness of segmentation directly impacts the overall success of the clinical decision support system (CDSS). *Can we accurately segment the PDAC and relevant vasculature?*

- *Determining resectability* involves analyzing the spatial relationships between the tumor and the surrounding vasculature to understand the extent of vessel involvement. This step is critical in clinical practice as it informs surgical decisions and treatment planning. *How do we automatically determine if the PDAC is resectable?*
- *Uncertainty in model predictions* can arise from various sources, including data variability, model limitations, and inherent ambiguities in medical images. By capturing and presenting this uncertainty, the CDSS can provide more nuanced information to clinicians, helping them make better-informed decisions. *How can we utilize model uncertainty to provide clinically-relevant resectability predictions?*

The proposed workflow in this chapter involves processing CT scans to segment the tumor and surrounding vascular structures, followed by analyzing the spatial relationships and extent of vascular involvement. This method mirrors the expert radiologists' approach to PDAC assessment. Three segmentation architectures, *i.e.* nnU-Net [13], 3D U-Net [154], and Probabilistic 3D U-Net [223], are utilized to achieve high accuracy in the segmentation of veins, arteries, and tumors. The segmentation maps enable automated detection of tumor involvement with high sensitivity and specificity, as well as the computation of the degree of tumor-vessel contact. Additionally, we address the significant inter-observer variability in assessing these structures by presenting the uncertainty captured by each model, thereby providing clinicians with a clearer indication of tumor-vessel involvement. This approach facilitates more informed decision-making for surgical interventions, offering a valuable tool for improving patient outcomes, personalized treatment strategies, and survival rates in pancreatic cancer.

The chapter is divided into the following sections. Section 5.2 provides an overview of related work on PDAC detection, segmentation, and resectability prediction, contextualizing the proposed approach within the existing research landscape. Section 5.3 details the methods, including data collection, model architectures and training of the segmentation models. It further discusses the computation of vessel involvement, distinguishing between aleatoric and epistemic uncertainty, and examine the effect of ambiguity on tumor-vessel involvement. Section 5.4 presents the results of the experiments, showcasing the performance of the utilized models in predicting resectability and their implications for clinical decision-making. Finally, Section 5.5 concludes the chapter, summarizing the findings and their significance.

5.2 Related work on PDAC detection, segmentation and resectability prediction

Segmentation and detection. Deep learning-based methods have demonstrated significant potential in the detection of pancreatic cancer on CT scans. Several studies have employed classification networks and achieved high accuracy in detecting PDAC and other types of pancreatic cancer [224]–[232]. Recently, segmenta-

tion for the classification of pancreatic cancer has garnered significant attention, since it both detects and localizes cancer [233]–[237]. Notably, Alves *et al*. [236] and Viviers *et al*. [53] have proposed a similar segmentation-for-classification framework, leveraging the surrounding anatomy and secondary tumor-indicative features, such as the common bile duct and pancreatic duct, to enhance tumor segmentation and improve detection accuracy.

Resectability. Obtaining an automated detailed segmentation map of the tumor provides high clinical value. As such, Mahmoudi *et al*. [238] have proposed a hybrid 2D-3D segmentation-based approach for detailed segmentation of the tumor mass and surrounding vessels in tumor-only cases. While they showcase good segmentation accuracy (DSC: 0.61 PDAC, 0.81 Artery and 0.73 Vein), they note that a full 3D method will further improve results and will be essential for determining tumor-vessel involvement. Recently, Yao *et al*. [239] presented a multicenter, retrospective study in which they construct an imaging-derived prognostic biomarker, called DeepCT-PDAC, for overall survival (OS) rate prediction. They train segmentation (nnU-Net) and prognostic models (CE-ConvLSTM and Tumor-vascular Involvement 3D CNN) to model the spatial relations between the tumor and anatomy. The 3D predictions of PDAC, the portal vein and splenic vein (PVSV), superior mesenteric vein (SMV), superior mesenteric artery (SMA) and truncus coeliacus (TC) are used in a CNN branch modeling the tumor-vascular involvement in 3D. Contact area features are predicted to be used in a final risk score or OS prediction. While the research presents impressive results for the accuracy of OS predictions, intermediate steps leading to the final prediction remain unclear at a clinical level. This lack of information could inhibit the adoption of this technique as a co-pilot or assistive tool to oncologists. Instead, CAD models should present clinically-relevant information based on the current way of working and allow the oncologist to assess each point and then finally decide on patient treatment options. Despite the remarkable progress in utilizing deep learning models for PDAC segmentation, the achieved accuracy is still relatively low and may not be sufficient for determining PDAC resectability.

5.3 Methods

This section details the data collection (Section 5.3.1), the segmentation model architectures (Section 5.3.2), and the procedures for data preparation and training details (Section 5.3.3). We further discuss how we compute and assess vessel involvement (Section 5.3.4), distinguish between aleatoric and epistemic uncertainty (Section 5.3.5), and examine the effect of ambiguity on tumor-vessel involvement (Section 5.3.6).

5.3.1 Data collection

This retrospective single-center research study investigates PDAC resectability in 99 patients specifically located in the pancreatic head. Determined by radiological assessment, a group of 50 patients have PDAC without vascular involvement,

while 49 patients have PDAC with potential involvement of critical adjacent vasculature. We employ contrast-enhanced CT images obtained from the Catharina hospital Eindhoven, The Netherlands. Each patient underwent a multi-phase pancreatic protocol CT scan, including (at least) the portal-venous phase, parenchymal phase, arterial phase, or late liver phase. Consequently, a total of 195 CT scans are included in the analysis. Prior to conducting the research, all CT scans were meticulously annotated. Under supervision of an expert abdominal radiologist, a surgical resident manually annotated all the relevant anatomical structures at voxel-level, including the tumor, pancreas, pancreatic duct (PD), common bile duct (CBD), aorta, superior mesenteric artery (SMA), celiac trunk, hepatic artery, splenic artery, splenic vein, superior mesenteric vein (SMV), portal vein, gastroduodenal artery (GA) and inferior vena cava. For model training purposes, we aggregated the different arteries into a single arterial structure and, similarly, all the veins into one venous structure.

The determination of tumor resectability requires careful consideration of tumor presence, size and its relationship with surrounding anatomical structures. Particularly, the extent of contact between the tumor and neighboring veins and arteries plays a crucial role. This degree of contact is typically computed after the clinican made the segmentation delineations where each of the structures are (or could be based on their best knowledge). Consequently, CT voxels have the potential to belong to multiple structures simultaneously. Figure 5.1 illustrates an example along with corresponding ground-truth annotations of the involvement. Due to the inherent ambiguity in the data and low contrast in some phases, segmentation maps and the derived tumor-vessel involvement varies between subsequent scans of the same patient. The reported results are on a per-scan basis.

5.3.2 Segmentation models

This research employs three segmentation models to delineate the tumor and surrounding anatomy. (1) We train the 3D nnU-Net to automatically create the segmentation maps of the structures of interest in 3D. The six different structures are layered from the least to the most important: pancreas, common bile duct, pancreatic duct, arteries, veins, tumor. To determine overlap, a 7^{th} and 8^{th} pseudo-structure is created for the tumor-artery and tumor-vein overlap. (2) A custom 3D U-Net is developed to segment the structures in multi-channel 3D, alleviating the need for pre-computed pseudo-labels and enabling direct overlap prediction. The model is set up to be identical to that of the default nnU-Net, except for a final sigmoid activation, instead of softmax probabilities in the nnU-Net, that enable mutually independent class predictions. This approach was also chosen to have a fair indication of the effect of the proposed novel overlap loss (5.2). (3) The third approach, the Probabilistic 3D U-Net follows the same segmentation scheme as the 3D U-Net and is utilized to express the aleatoric uncertainty in the structures of interest by presenting multiple plausible segmentation hypotheses. In Figure 5.2, these three approaches are visualized along with the initial tumor

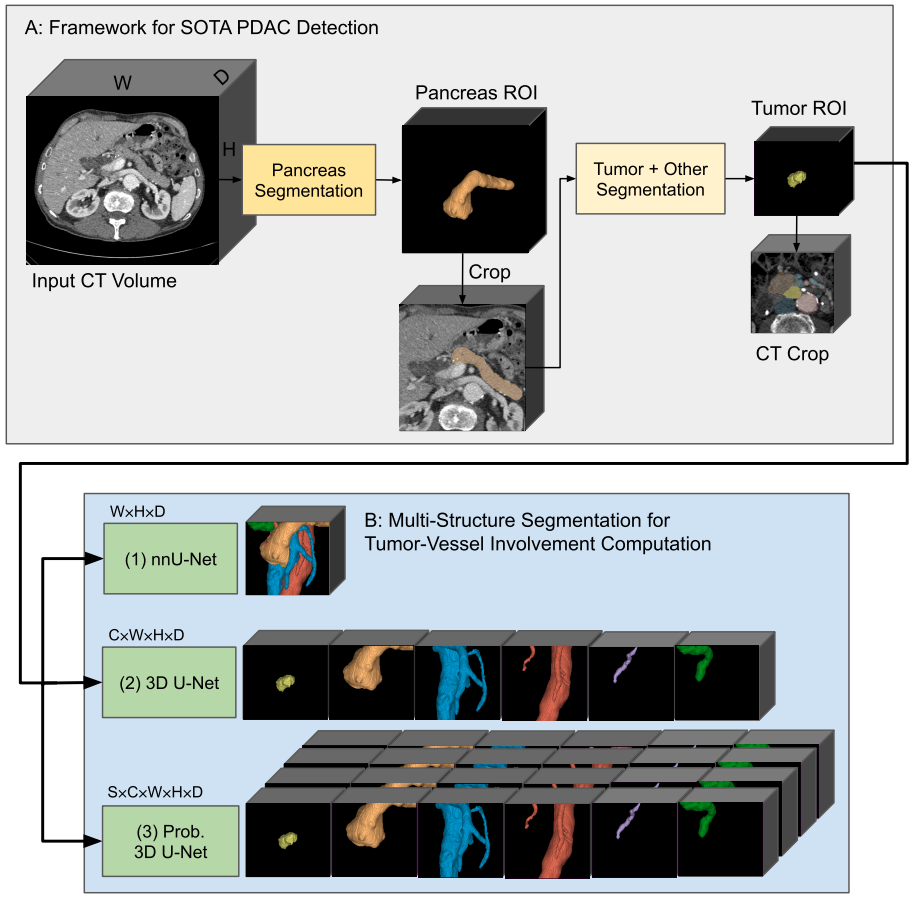


Figure 5.2 *Workflow illustration for tumor segmentation from a CT scan. Using the CT scan, the pancreas segmentation maps is created and cropped for automated tumor segmentation [53], [236]. Having the tumor detection and segmentation, another crop is taken around the tumor and provided to the proposed three models (Block B) to determine tumor-vessel involvement.*

detection processing chain.

5.3.3 Data preparation and training details

The data for resectability prediction are prepared according to the workflow depicted in Figure 5.1. In Section 5.2, various methods are presented that achieve high PDAC detection accuracy with reasonable segmentation accuracy. As such, we continue by cropping the CT around the tumor center. This is implemented based on the ground-truth tumor labels. However, in practice, this is performed by a prior segmentation model (as depicted in Block A in Figure 5.2). We crop the CT scan and corresponding labels of the tumor, pancreas, pancreatic duct, common bile duct and an aggregate of all the arteries and veins (as mentioned

in Section 5.3.1). For the nnU-Net implementation, two additional overlapping pseudo-labels are created. The dataset is resampled to the mean dataset size ([1 mm, 0.67 mm, 0.67 mm] in the z, y, x-axes) and cropped to [64, 128, 128]-voxels in the z, y, x-axes, respectively.

The patient dataset is split into a 70%/15%/15% for train/validation/testing. The test split is chosen by a surgical resident to be representative of the distribution of the tumor size, location and involvement present in the dataset. From the 85% train/validation data we perform threefold bootstrapping using random patient splits and report results on the validation and test splits. The full-resolution 3D nnU-Net is employed as reported publicly [1] without any modifications. The custom 3D U-Net is implemented in PyTorch and extends on the work by Wolny *et al*. [160] [2]. The Probabilistic 3D U-Net is adapted from the implementation by Viviers *et al*. [223] and available online [3]. During training, the loss function as introduced in Eq. (5.3) is employed and all other U-Net model parameters are chosen to be consistent with that of the nnU-Net where possible. The U-Nets are 5 layers deep with 32, 64, 128, 256, and 320 filters at each layer, respectively. A cosine annealing strategy is employed to modulate the Probabilistic 3D U-Net ELBO parameter β between 1 and 10. All the models are trained for 1000 epochs with the Adam optimizer and a linearly decaying learning rate scheduler. During training, the model weights with the best validation performance are chosen.

Recent advancements in semantic segmentation for medical applications have demonstrated the effectiveness of combining binary cross-entropy (BCE) and DSC loss functions to enhance performance [13]. The cross-entropy loss is proficient in capturing global context and penalizing misclassifications, while the DSC loss emphasizes spatial overlap and similarity. Although optimizing these objectives contributes to determining the overlap between tumors and vessels, we propose a specific loss function, called the Overlap Loss (OLL).

More formally, assume $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ as random variables taking values in $\mathbb{R}^{Z \times H \times W}$ and $\mathbb{R}^{C \times Z \times H \times W}$, representing the input images (Z image depth, H-height and W-width) and ground-truth masks, respectively. Accordingly, We define random variables $\mathbf{T} \in \mathcal{T}$, $\mathbf{A} \in \mathcal{A}$ and $\mathbf{V} \in \mathcal{V}$, representing tumor, artery and vein segmentation masks, respectively, which are elements in subsets of $\mathcal{Y}$. We denote the pseudo-overlap labels α and ν for the tumor-artery and tumor-vein pairs, which are defined as

$$\alpha = \mathbf{T} \odot \mathbf{A}, \ \nu = \mathbf{T} \odot \mathbf{V}, \tag{5.1}$$

where $\odot$ implies the element-wise product of the tensors. Then, the overlap loss

[1] https://github.com/MIC-DKFZ/nnUNet

[2] https://github.com/wolny/pytorch-3dunet

[3] https://github.com/cviviers/prob_3D_segmentation

term OLL is specified by

$$H_o = H(\hat{\alpha}, \alpha) + H(\hat{\nu}, \nu), \tag{5.2}$$

where H is the element-wise BCE and H_o the summation of the two BCE loss terms. The hat notation ($\hat{\alpha}$) is utilized to differentiate sigmoid-activated logit predictions from ground-truth (non-hat) masks. It is important to note that sigmoid activation precedes the creation of the pseudo labels.

The OLL directly aims to optimize the predictions of the overlapping structures. Although the objective to accurately predict the degrees of involvement is computed based on contact/adjacent pixels (see Section 5.3.4), we conjecture that accurate overlap prediction will result in precise delineation of the contact area. By introducing this direct objective, we anticipate further improvements in the accuracy of segmentation results in the areas of interest and related analysis. The complete training objective (CLL) for the 3D U-Net and the reconstruction loss for the Probabilistic 3D U-Net can then be formulated as

$$\begin{aligned} \text{CLL} = \alpha \cdot \Big[\beta \cdot H(p,q) + (1-\beta) \cdot \text{DSC}(p,q)\Big] \\ + (1-\alpha) \cdot H_o(p,q), \end{aligned} \tag{5.3}$$

where α and β are weighting factors between the different loss components. The addition of the OLL loss components further emphasizes the importance of accurate segmentation of the overlapping and contact areas of the tumor and veins/arteries. In the conducted experiments, we have empirically found that $\beta = 0.5$ and $\alpha = 0.8$ work well.

5.3.4 Computing and assessing vessel involvement

The degree of tumor-vessel involvement is computed on a per (2D) axial-slice basis. The contact area between the tumor and the vessel is calculated based on adjacent pixels, which is followed by computing the vessel centroid and the distance of each of the contact pixels to the vessel center. The angle between each pixel and the center is calculated using the 4-quadrant arctangent function. The difference between the maximum and minimum angle is then used to determine the degree of involvement. An example result can be observed in Figure 5.1, where vertical upside of the image is the reference and the axis origin at the vessel center. It is important to note that while clinicians do not have automated tools to perform this, the degree of involvement is clinically assessed and measured in a similar way.

We introduce a classification metric of predicting involvement with the arteries and veins: if the resulting tumor and artery/vein segmentation predictions have involvement (degree > 0), even at the wrong location compared to the GT, while the GT also has involvement somewhere, we consider it a true positive (TP) prediction. If there is involvement prediction and the GT has no involvement, it is a false positive (FP). In the case of no predicted involvement whatsoever and the

GT also has no involvement, we consider it a true negative (TN) and vice-versa for false negatives (FN). For the scan-level sensitivity and specificity, we follow the same procedure (if either the artery or vein involvement is a TP, then at scan-level it is a TP and so forth). Additionally, the sensitivity and specificity results are measured for the clinically relevant superior mesenteric vein (SMV), portal vein (PV), superior mesenteric artery (SMA) and Truncus. In our evaluation, we remove the arterial and venous segmentation predictions that have an overlap with the pancreas (simply remove overlapping voxels) to strictly separate vessels/arteries from the pancreas. The remaining segmentation predictions are then compared to the GT aggregate of the SMV and PV (venous) or SMA and Truncus (arterial). The Dutch Pancreatic Cancer Group (DPCG) [213] classifies tumor resectablity based on the degrees of tumor-vessel contact. The tumor is considered resectable if SMV and PV have $\leq 90^\circ$ contact, borderline resectable if the contact is between 90° and 270° and irresectable if the contact is $> 270^\circ$. For the arterial vasculature, the DPCG considers tumor as borderline resectable for $\leq 90^\circ$ contact and any amount of involvement more than 90° deems the tumor irresectable.

5.3.5 Distinguishing between aleatoric and epistemic uncertainty

Distinguishing between aleatoric and epistemic uncertainty is valuable for designing effective segmentation models. By quantifying and analyzing these uncertainties individually, researchers can identify whether the model limitations stem from data noise or model deficiencies.

While the theoretical distinction between aleatoric and epistemic uncertainty is crucial for model development and research, its practical value in clinical applications of medical image segmentation models remains a topic of debate. In real-world clinical settings, the primary concern is often the total uncertainty associated with model predictions rather than its decomposition into aleatoric and epistemic components. As long as the total uncertainty is accurately quantified and effectively communicated to clinicians, this provides sufficient information for informed decision-making. There is currently limited empirical evidence demonstrating that distinguishing between these types of uncertainties yields significant clinical benefits. Instead, focusing on the overall uncertainty can streamline the implementation of segmentation models in clinical workflows, thereby ensuring that the confidence levels of model predictions are clearly understood and appropriately acted upon by healthcare professionals [92].

Given the above context and the proven ability to capture different types of uncertainty, we do not differentiate between aleatoric and epistemic uncertainty when evaluating their impact on tumor-vessel involvement. The primary goal is to provide an accurate resectability assessment with well-defined deviation margins, regardless of the source of uncertainty. By focusing on the total uncertainty, we ensure that the system delivers comprehensive and actionable information to clinicians, thereby facilitating informed decision-making and enhancing the reliability of the resectability predictions.

5.3.6 The effect of ambiguity on tumor-vessel involvement

Accurate estimation of uncertainty is vital in image segmentation tasks to assess the reliability of the predicted segmentation maps. In this study, we compute the tumor-vessel involvement directly from the segmentation results and, as such, any variation in the resulting segmentation can have a large impact on the involvement prediction and the extent (degree) thereof. We propose a comprehensive approach that (a) ensembles different model folds providing the epistemic uncertainly and (b) a probabilistic modeling approach to capture both epistemic and aleatoric uncertainty in the resulting segmentation maps.

To capture epistemic uncertainty, we construct an ensemble of the segmentation models. Each model in the ensemble is trained with different weight initialization procedures and training dataset folds. By considering the disagreement among the ensemble members as samples from the model weight distribution, we can effectively capture the model's epistemic uncertainty regarding the segmentation [240], [241]. This is further detailed in the paragraph below. If so, the ensemble for 3D nnU-Nets and 3D U-Nets are thus capable of expressing the epistemic uncertainty. To address aleatoric uncertainty associated with ambiguity in the image data, we employ a probabilistic U-Net [56], [189]. The probabilistic U-Net explicitly models the uncertainty within data by learning a lower-dimensional latent distribution of plausible variations in the output. This enables to capture the inherent variability and ambiguity in voxel-level predictions [56]. By integrating the ensemble of probabilistic U-Nets, we obtain a holistic uncertainty estimation framework that captures both epistemic and aleatoric uncertainty. This combined uncertainty estimation approach enhances the interpretability and reliability of the segmentation results.

To address the practical implementation aspect of expressing the disagreement between model ensemble members for epistemic uncertainty for the nnU-Net and 3D U-Net, we exploit the mean and standard deviation of the predicted segmentation probabilities across the three folds as the mean prediction and the epistemic uncertainty. The predicted probabilistic U-Net sigmoid probabilities can be written as $\mathcal{Y} \in \mathbb{R}^{S \times C \times Z \times H \times W}$, where S denote the samples. Computing $\sigma(Y)$ represents the aleatoric uncertainty and $\mu(\sigma(Y_0), \sigma(Y_1), \sigma(Y_2))$ specifies the mean aleatoric uncertainty across the three model folds. The epistemic uncertainty can be computed as the variance of the individual mean predictions, hence $\sigma(\mu(Y_0), \mu(Y_1), \mu(Y_2))$. Figure 5.4 depicts the sum of the aleatoric and epistemic uncertainty for the probabilistic U-Net.

5.4 Results and discussion

The experimental results are listed in Table 5.1, Table 5.2 and Figure 5.4. The mean and standard deviations are reported of the sensitivity, specificity and DSC (across all cases) on the validation sets, the three different models (from the folds) on the test set and an ensemble of the models' folds predictions on the test set. We do not include any results on segmentation performance of the pancreas, pancreatic

duct or common bile duct, since it does not directly contribute to the tumor-vessel involvement focus of this study. As presented in Table 5.2 and Figure 5.3, an R^2 score is provided on how well the predicted maximum involvement correlates to the GT maximum involvement.

We have produced this plot by taking the maximum involvement angle at any slice (computed using the GT) and compare it with the maximum predicted angle of involvement. The maximum angle of involvement is one of the most important criteria used by clinicians in determining the treatment plan. Figure 5.4 showcases the performance of the three segmentation models (in the three dual bottom rows) on a scan from the test set (top row). The particular model ensemble prediction (second vertical column), the overlapping structure, either predicted or derived (third vertical column), and computed degrees of involvement are presented (fourth vertical column). In the uncertainty (second) row for each model, the associated uncertainty is showcased (computed as described in Section 5.3.6). The uncertainty heat map is on a standard 0-0.5 scale and clipped below 0.01 to enable visualization of the background. The segmentation maps derived by subtracting 1, adding 1 and adding 2 voxel-level standard deviations are presented in the next columns to illustrate the effect that the uncertainty can have on the final tumor-vessel involvement prediction.

Validation results: In terms of segmentation accuracy (Table 5.1, the nnU-Net outperforms the other models and achieves tumor, artery and vein DSC scores of 0.67±0.03, 0.88±0.02 and 0.87±0.02. However, the overlap DSC scores for artery and vein are low, indicating difficulty in delineating these pseudo-structures. The OLL used in the 3D U-Net hardly affected the DSC scores of the overlapping structures compared to the nnU-Net. The sensitivity and specificity values for artery and vein segmentation vary, with the vein sensitivity showing higher performance compared to artery sensitivity. However, the OLL significantly enhances both artery (0.48±0.17) and vein (0.86±0.07) sensitivity of the 3D U-Net at the cost of a few FP predictions.

Test results: The obtained segmentation accuracies (Table 5.1 align closely with the validation results. Specifically, the proposed OLL approach exhibits enhanced generalization ability, with the 3D U-Net model showing slight improvements in the DSC scores for overlapping structures (due to the few pixels belonging to this overlapping structure). The artery overlap segmentation achieves a DSC score of 0.02±0.01 compared to the baseline score of 0, while the vein segmentation yields a DSC score of 0.14±0.03, surpassing the baseline score of 0.08±0.01. These low scores can be attributed to very small structures, large GT variability and general difficulty in accurate delineation due to a lack of contrast. The sensitivity and specificity values for artery and vein segmentation on the test set are generally consistent with the validation results. However, it is worth noting that the 3D U-Net model displays larger standard deviations, indicating that some model folds correspond well with the test set, while others demonstrate certain discrepancies.

Metric	3D nnU-Net	3D U-Net OLL	Prob. 3D U-Net OLL
Validation			
Tumor DSC	0.67 ± 0.03	0.65 ± 0.02	0.50 ± 0.03
Artery DSC	0.88 ± 0.02	0.83 ± 0.03	0.84 ± 0.03
Vein DSC	0.87 ± 0.02	0.85 ± 0.03	0.86 ± 0.02
Artery Overlap DSC	0.05 ± 0.04	0.04 ± 0.03	0.05 ± 0.03
Vein Overlap DSC	0.16 ± 0.08	0.17 ± 0.05	0.14 ± 0.04
Artery Sensitivity	0.35 ± 0.18	0.48 ± 0.17	0.49 ± 0.12
Artery Specificity	1.00 ± 0.00	0.90 ± 0.08	0.85 ± 0.05
Vein Sensitivity	0.79 ± 0.15	0.86 ± 0.13	0.85 ± 0.05
Vein Specificity	0.87 ± 0.11	0.87 ± 0.07	0.73 ± 0.25
Scan Sensitivity	0.81 ± 0.14	0.87 ± 0.11	0.87 ± 0.03
Scan Specificity	0.92 ± 0.11	0.85 ± 0.03	0.74 ± 0.20
Test			
Tumor DSC	0.65 ± 0.01	0.63 ± 0.01	0.49 ± 0.08
Artery DSC	0.86 ± 0.00	0.86 ± 0.01	0.86 ± 0.01
Vein DSC	0.90 ± 0.00	0.87 ± 0.00	0.88 ± 0.01
Artery Overlap DSC	0.00 ± 0.00	0.02 ± 0.01	0.01 ± 0.00
Vein Overlap DSC	0.08 ± 0.01	0.14 ± 0.03	0.12 ± 0.02
Artery Sensitivity	0.23 ± 0.05	0.53 ± 0.17	0.05 ± 0.08
Artery Specificity	0.94 ± 0.02	0.91 ± 0.04	0.83 ± 0.08
Vein Sensitivity	0.85 ± 0.06	0.81 ± 0.09	0.73 ± 0.08
Vein Specificity	0.77 ± 0.06	0.75 ± 0.18	0.60 ± 0.11
Scan Sensitivity	0.81 ± 0.00	0.83 ± 0.08	0.77 ± 0.07
Scan Specificity	0.74 ± 0.07	0.74 ± 0.13	0.67 ± 0.09
Test Ensemble			
Tumor DSC	**0.66**	**0.66**	0.56
Artery DSC	**0.86**	**0.86**	0.87
Vein DSC	**0.91**	0.88	0.89
Artery Overlap DSC	0.00	**0.01**	**0.01**
Vein Overlap DSC	0.07	**0.15**	0.13
Artery Sensitivity	0.20	0.30	**0.40**
Artery Specificity	0.91	**1.00**	0.95
Vein Sensitivity	0.81	**0.88**	0.75
Vein Specificity	0.81	**0.81**	0.62
Scan Sensitivity	0.81	**0.88**	0.79
Scan Specificity	0.79	**0.86**	0.72
Test Ensemble Predictions with only the SMV, PV, SMA, Truncus involvement			
SMA or Truncus Sensitivity	**0.50**	**0.50**	**0.50**
SMA or Truncus Specificity	0.90	**0.93**	0.90
SMV or PV Specificity	**0.92**	**0.92**	0.77
SMV or PV Specificity	0.79	**0.89**	0.68
Scan Sensitivity.	**0.92**	**0.92**	0.79
Scan Specificity	0.79	**0.89**	0.68

Table 5.1 *Segmentation and overlapping scores obtained with the 3D nnU-Net, 3D U-Net with OLL and Prob. 3D U-Net with OLL across 3 validation folds. These three models and an ensemble of these models are applied to the test set.*

Metric	3D nnU-Net	3D U-Net	Prob. 3D U-Net
Validation			
Artery R^2	-0.07 ± 0.26	-0.17 ± 0.09	-0.55 ± 0.56
Vein R^2	0.34 ± 0.39	0.16 ± 0.40	-1.95 ± 2.76
Test			
Artery R^2	-0.24 ± 0.01	0.12 ± 0.22	-0.24 ± 0.17
Vein R^2	0.37 ± 0.21	0.42 ± 0.13	-0.04 ± 0.44
Test Ensemble			
Artery R^2	-0.27	-0.24	-0.06
Vein R^2	0.52	0.42	0.31
Test Ensemble with only the SMV, PV, SMA, Truncus involvement			
Artery R^2	-0.14	-0.01	-7.53
Vein R^2	0.54	0.44	0.60
Test Ensemble vascular criteria SMV, PV, SMA, Truncus involvement			
$0° =$ Involvement	(19/19), (27/30)	(19/19), (28/30)	(19/19), (27/30)
$0° <$ Involvement $\leq 90°$	(5/5), (1/1)	(5/5), (1/1)	(5/5), (1/1)
$90° <$ Involvement $\leq 270°$	(0/7), (0/1)	(2/7), (0/1)	(4/7), (0/1)
$270° <$ Involvement	(1/1), (0/0)	(0/1), (0/0)	(1/1), (0/0)
Accuracy Vein, Acc. Artery	0.781, 0.875	0.813, 0.906	0.906, 0.875
Overall Accuracy	0.828	0.860	0.891

Table 5.2 *Correlation between ground-truth and predicted involvement. Attending to the bottom rows of the table and following the DPCG criteria [213], the tumor-vein (first set of brackets) and tumor-artery (second set of brackets) involvement is categorized and assessed. The first number in the bracket indicates the model prediction and the second number indicates the GT number of in the corresponding involvement condition. In the last two rows on vascular criteria, the table depicts the accuracy obtained for each topic of interest (vein, artery) and the overall accuracy with each model. The listed numbers are computed from the involvement conditions indicated directly above them.*

Test ensemble: Predictions from the three folds are combined, resulting in a minor segmentation performance increase across all the models (Table 5.1). The overlap DSC scores for artery and vein remain low, however, the 3D U-Net shows segmentation improvements over the nnU-Net that result in larger detection performance improvements for both involvement with the artery and vein.

Degree of involvement: Table 5.2 and Figure 5.3 show that there is moderate agreement between the predicted angle of involvement and the ground-truth angle for the venous structures specifically in the validation, test and test ensemble for the critical structures. The final test ensemble showcases slightly better alignment for the Prob. U-Net (R^2 0.60) for angles in the superior mesenteric vein (SMV) or portal vein (PV). Very low agreement for the degree of involvement with the arteries is shown. The test set only contains one case with involvement with the superior mesenteric artery (SMA) and one with the Truncus. The remaining cases

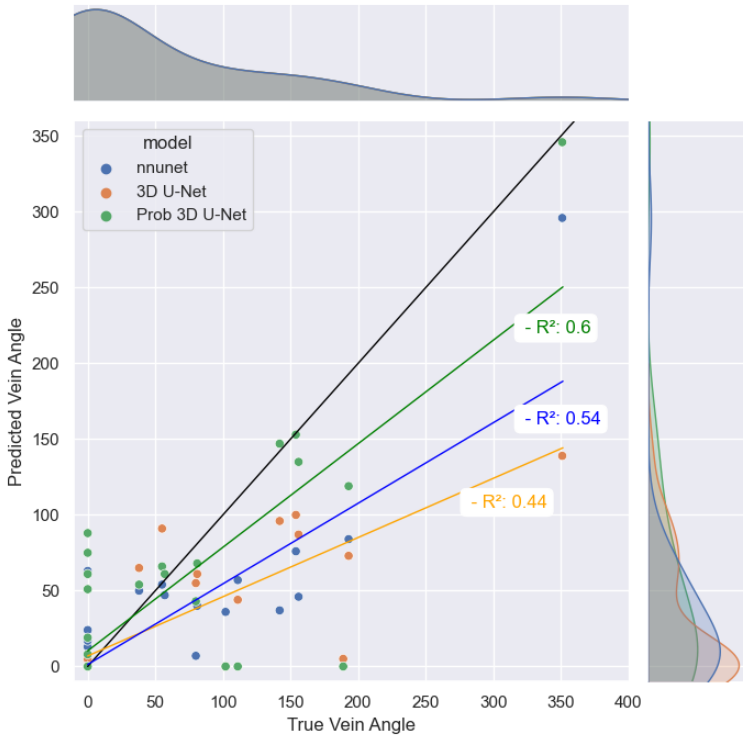


Figure 5.3 *Maximum SMV or PV degrees of involvement.*

either have no involvement or involvement with the gastroduodenal artery (see Figure 5.4), which can easily be ignored by removing arterial predictions within the pancreas. Table 5.2 depicts the degree of involvement according to the DPCG criteria. Despite not achieving perfect accuracy in predicting the exact angles (a challenging task to begin with), the models provide clinically-relevant evaluations that are sufficiently accurate for practical use as per the four involvement classes. All three models capture the extent of involvement almost perfectly for smaller degrees of involvement and the Prob. 3D U-Net performs slightly better than the other models at larger degrees of involvement. It is worth mentioning that the models tend to underestimate the involvement for larger degrees of involvement, indicating a potential limitation in capturing extensive involvements accurately. This can be attributed to the absence of cases with extensive involvement in the dataset. The Prob. 3D U-Net (third column in Table 5.2) consistently outperforms the other models, particularly at larger degrees of involvement, as evidenced by its superior overall accuracy (0.891) and balanced performance across venous and arterial structures, thereby making it the most reliable model for clinically-relevant resectability evaluations.

Test ensemble uncertainty: The models' uncertainties are presented in Figure 5.4 for a scan from the test set that, in this slice, appears to be borderline resectable

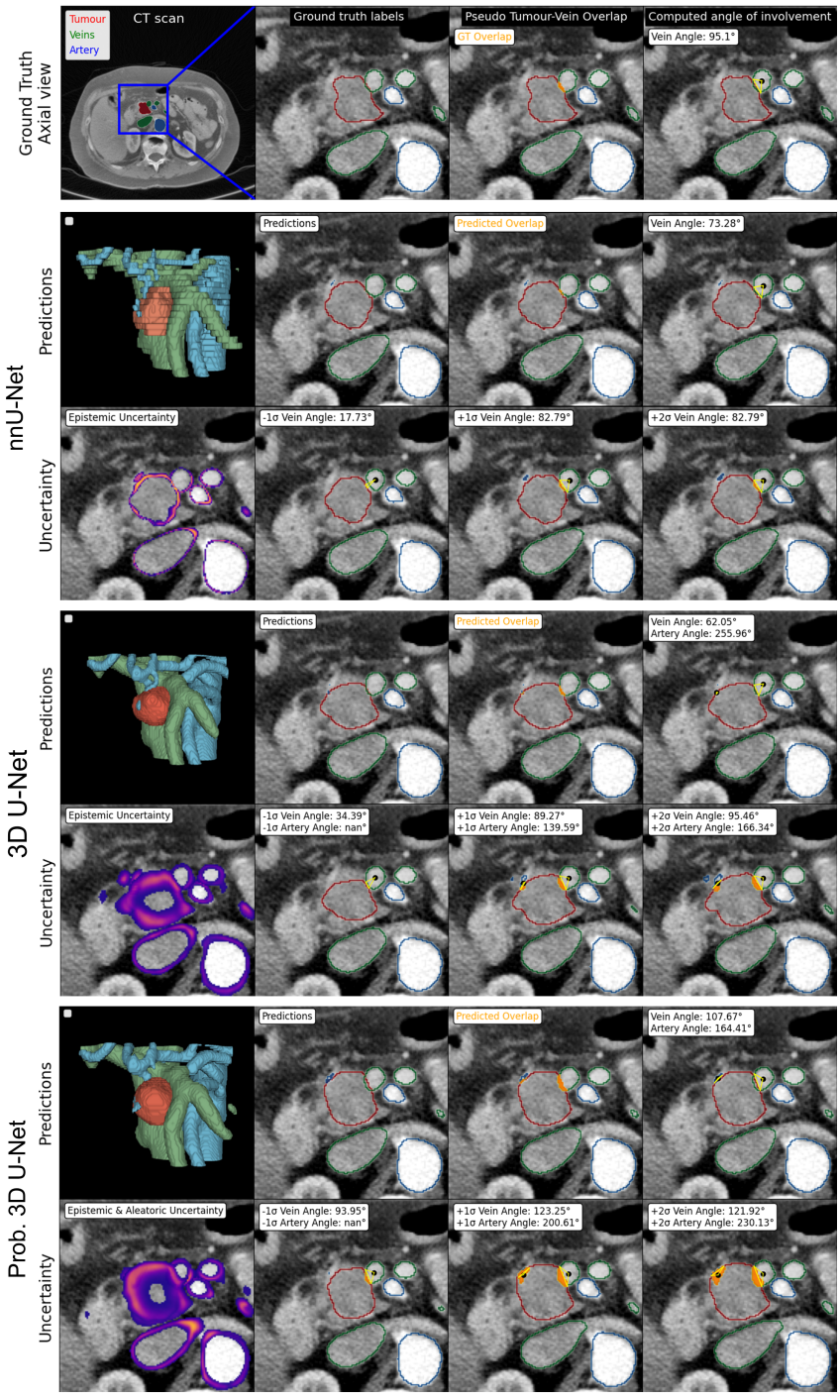


Figure 5.4 *Ground truth (top) and predictions of the three models from a test set case. The meaning of the various sub-pictures is explained in the text.*

(90°<Involvement≤ 270°), due to the involvement with the SMV. Incorporating the uncertainty in the segmentation predictions allows for a likelihood-based evaluation of the tumor-vessel degree of involvement. In the example, the nnU-Net underestimates the tumor size and involvement and predicts a resectable tumor. With very defined uncertainty regions and taking uncertainty steps (-1 σ, +1 σ and +2 σ) does not change the degree of involvement prediction by much and therefore, the treatment strategy will not be affected. The 3D U-Net initially underestimates the involvement, but with +2 σ steps, the borderline-resectable margin is crossed, indicating a potentially larger tumor with more involvement and the correct treatment approach. The Probabilistic 3D U-Net already predicts the correct response (borderline resectable) with a larger degree of involvement (107.67°). While the model is capable of expressing all the uncertainty, in this case it does not affect the predicted resectability.

All segmentation models obtain good segmentation accuracies for the desired structures of veins, arteries and the pancreatic tumor. Although the deep learning models demonstrate promising segmentation results, there is still room for improvement, particularly in the tumor segmentation and capturing the area of overlap between the tumor and vessels, which, we conjecture correlates with the contact area and ultimately the degrees of involvement. The lower tumor DSC can be connected to the lack of visual information, both in texture and contrast, in CT volumes concerning the tumor. This can be readily understood, since this overlap measurement is a secondary step after the primary step of obtaining sufficient segmentation accuracy of the individual components. Presenting the R^2 metric, the effect of OLL and valuable uncertainty estimates is a first attempt at accurately quantifying the amount of overlap and extent of tumor-vessel involvement. Obtaining a more accurate measurement and a metric that incorporates uncertainty in the involvement assessment is future work. As for the primary objective, we obtain high resectability classification accuracy from the segmentation maps, clearly predicting tumor-vessel (sensitivity 88% and specificity 86%) and tumor-critical vessel (sensitivity 92% and specificity 89%) involvement. These results are of high clinical value and very encouraging because it is achieved by following the clinical way of working from deriving tumor-vessel contact based on the previously mentioned segmentation results.

Limitations: The above-mentioned findings are based on a small dataset, particularly concerning tumor-artery involvement. Accurate assessment of tumor-vessel involvement heavily relies on precise segmentation, which needs to exactly match the annotations provided by experts. However, achieving such accuracy is challenging, considering the inherent ambiguity associated with tumor visibility on CT, which is openly discussed among experienced clinicians. This work is one of the first to facilitate and contribute to this challenge that clinicians have to face on a daily basis with potentially severe patient consequences in decision-making.

5.5 Conclusion

This chapter has presented a robust workflow for predicting tumor-vessel involvement and tumor resectability in pancreatic ductal adenocarcinoma (PDAC) using CT scans. Building on the methodologies and findings developed in Chapters 3 and 4, this study has concentrated on the application and evaluation of three state-of-the-art deep learning-based segmentation architectures: the 3D nnU-Net, a 3D U-Net with overlap loss (OLL), and the Probabilistic 3D U-Net with OLL. The overlap loss (OLL) introduced in this chapter, although giving marginal improvements in the DSC score of the overlapping structures, has a large impact on the obtained sensitivity (the most important metric). These models have been implemented to automate the segmentation of PDAC while involving the surrounding critical anatomical structures. The extensive validation with the three models has established solid baselines and makes the obtained results more reliable for further use and clinical consideration.

The groundwork laid in Chapter 3 on PDAC detection and segmentation provided the necessary foundation for the accurate identification of tumor and secondary features, while the exploration of uncertainty quantification in Chapter 4 informed the proposed approach to modeling and interpreting the segmentation uncertainties. Leveraging these insights, we have successfully delineated key anatomical structures with three automated segmentation models, achieving high accuracy for veins (DSC 0.88), arteries (DSC 0.86), and pancreatic tumors (DSC 0.66) with the 3D U-Net. This enables precise prediction of tumor-vessel involvement by deriving degrees of involvement from the segmentation maps. Particularly, the implemented approach to compute the degrees of contact towards the center points of the segmented vessels enables to derive an angular difference rather than absolute values, which makes the prediction more robust.

The proposed approach automates the detection and quantification of tumor-vascular contact from the predicted segmentation maps. Although accurately measuring the degree of involvement remains challenging, the proposed approach yields compelling results in classifying resectability. The models demonstrate high accuracy in determining any tumor involvement, with the 3D U-Net achieving the highest sensitivity (0.88) and specificity (0.86). Although the accuracy in computing the exact degrees of involvement is low across all the models, it is accurate enough to sufficiently classify the extent of involvement according to DPCG criteria. For this task, the Prob. 3D U-Net performs best with a high 89.1% accuracy, offering clear and reliable indications to clinicians. Additionally, the system provides clinicians with valuable uncertainty intervals of involvement, thereby enhancing the reliability of the resectability assessments. This advancement facilitates more informed decision-making for surgical interventions and supports the development of personalized treatment strategies, ultimately contributing to improved patient outcomes in PDAC management.

Chapter 5

Chapter 6 Out-of-distribution detection in medical imaging

6.1 Introduction

Out-of-Distribution (OOD) detection is a pivotal aspect of machine learning that enables detecting samples that deviate from the normal and In-Distribution (ID) set. Detecting these anomalous samples can either be the main task or OOD detection can be employed as a safeguard to a subsequent system to ensure the reliability and robustness of models when deployed in real-world settings.

The primary challenge addressed by OOD detection is the identification of inputs that differ significantly from the data encountered during the training phase. Failure to accurately identify such inputs can lead to erroneous predictions and potential system failures, particularly in critical applications such as medical diagnosis, autonomous driving, and security systems. To this end, this chapter focuses on two critical aspects of OOD detection: semantic OOD detection and covariate shift detection. Semantic OOD detection aims at identifying inputs that differ from the training data in terms of class or semantics. In contrast, covariate shift detection identifies shifts in the high-level input distribution while maintaining consistent semantics. These two experiments explore OOD detection from both perspectives, ensuring that both high-level and low-level changes in data can be captured.

As machine learning models become more sophisticated, their deployment environments grow increasingly complex, often presenting data with characteristics that deviate from the norm. Conventional models, trained under the assumption that the test data distribution will match the training data distribution, are poorly equipped to handle such deviations. This mismatch, known as a distribution shift, can be categorized into two primary types: semantic shifts and covariate shifts. Semantic shifts occur when the outlier data belong to entirely different classes than those processed during training, whereas covariate shifts arise from changes in the input distribution, while the conditional distribution of the output given the input remains unchanged.

The first part of this chapter delves into semantic OOD detection in the context of melanoma diagnosis, a domain where early and accurate detection is crucial. Melanoma, a severe form of skin cancer, presents unique challenges due to the inherent imbalance in datasets – malignant cases are relatively rare compared to benign cases. This imbalance complicates the training of supervised classification

models, leading to potential misdiagnosis. This imbalance indicates that an unsupervised Out-of-Distribution (OOD) detection approach is more appropriate. Specifically, the distribution of the abundant benign melanoma images can be modeled, while deviations from this distribution can be flagged as potentially malignant.

Generative models have been widely applied to the OOD detection problem. However, each model type possesses its own inductive biases, which may affect its suitability for specific scenarios. Normalizing Flows (NFs) are particularly attractive for OOD detection due to their capability for explicit density modeling and evaluation. However, it has been demonstrated that NFs tend to model local pixel correlations rather than semantic content. Additionally, they are computationally inefficient due to their fully bijective transformations. To address these issues, recent research has proposed modeling the coefficients of a wavelet decomposition with NFs to alleviate the significant computational cost constraint. The following research questions are considered in the semantic OOD detection of melanoma. *Research questions for semantic shift:*

- Conventional Normalizing Flows exhibit limitations in capturing and evaluating the semantic context. *Can domain-specific knowledge of skin melanoma be utilized to improve the detection performance of likelihood-based OOD detection of malignant images?*
- When analyzing the malignant nature of melenoma from an image, the semantic context and lower frequency components are more important. *How can wavelet-based NFs be exploited to improve the detection of OOD melanoma images, particularly in the context of imbalanced datasets?*

To address these challenges, a novel generative model based on wavelet-based NFs is designed to learn the benign data distribution and detect OOD malignant images through density estimation. NFs are particularly suited for this task because of their capability to compute exact likelihoods. However, their conventional implementations tend to focus on apparent graphical features rather than the semantic context, limiting their effectiveness. To address these limitations, we enhance NFs with wavelet transforms, integrating domain-specific knowledge of melanoma. This approach not only improves the likelihood-based detection of malignant images, but also reduces the number of parameters required for inference, making it feasible for deployment on edge devices. The proposed method demonstrates a significant improvement in detection performance, achieving an increase of 9% in the Area Under Curve of the Receiver Operating Characteristic (AUROC) curve. This advancement is promising for assisting medical professionals in the accurate and early diagnosis of skin cancer and potentially improves patient outcomes.

The second part of this chapter addresses the detection of covariate shifts, an area that has garnered less attention compared to semantic OOD detection. Covariate shifts entail the change in distribution of high-level image statistics (covariates) subject to consistent low-level semantics. For example, lowering the

applied dose in an X-ray system will exhibit an increase in noise in the resulting image, leading to a shift in the covariance and high-level statistical image components. A more intriguing example of covariate shift is the modeling of the ID covariate factors under known imaging conditions with the intention to detect a change in these factors due to a failure in the imaging system or processing chain. Detecting these shifts is critical for maintaining the performance and reliability imaging systems and of machine learning models consuming these images in dynamic environments. This research into covariate shift detection calls for the following questions. *Research questions for covariate shift:*

- Generative models, with their ability to capture the underlying distribution of In-Distribution (ID) data in an unsupervised manner, are capable of detecting deviations from this learned distribution. *Can generative models effectively detect and quantify covariate shifts in natural images?*
- Covariate shift affects the distribution of high-frequency signal-dependent and independent details. *How to explicitly model the high-frequency heteroscedastic image components and does this lead to improved OOD covariate shift detection performance?*

For this second problem statement, we introduce CovariateFlow, a novel method for OOD detection tailored to covariate heteroscedastic[1] high-frequency image components using conditional Normalizing Flows (cNFs). This approach focuses on modeling the high-frequency signal-dependent and independent details, which are crucial for identifying sensory anomalies and deviations in global signal statistics. Extensive analyses on datasets such as CIFAR10 vs. CIFAR10-C, ImageNet200 vs. ImageNet200-C and a new X-ray dataset demonstrate the efficacy of CovariateFlow in accurately detecting covariate shifts. This work enhances the fidelity of imaging systems and supports the robustness of machine learning models in the presence of distribution shifts.

In summary, this chapter provides a comprehensive exploration of generative OOD detection methodologies, addressing both semantic and covariate shifts. In Section 6.2 and Section 6.3, we propose and validate wavelet-based NFs for semantic OOD detection of malignant melanoma. Furthermore, Section 6.4 and Section 6.5 introduces CovariateFlow and validate its effectiveness in OOD covariate shift detection. The proposed methodologies offer robust frameworks for improving OOD detection in medical and general imaging contexts. These advancements highlight the potential for generative models to enhance the reliability and robustness of machine learning systems in diverse real-world applications.

[1]The variance of the residuals is dependent on a signal over a range of measured values

6.2 Semantic case: Efficient OOD detection with wavelet-based NFs

Melanoma is a serious form of skin cancer with high mortality rate at later stages. Fortunately, when detected early, the prognosis of melanoma is promising and malignant melanoma incidence rates are relatively low. As a result, datasets are heavily imbalanced which complicate training current state-of-the-art supervised classification AI models. We propose to use generative models to learn the benign data distribution and detect Out-of-Distribution (OOD) malignant images through density estimation. Normalizing Flows (NFs) are ideal candidates for OOD detection due to their ability to compute exact likelihoods. Nevertheless, their inductive biases towards apparent graphical features rather than semantic context hamper accurate OOD detection.

In this section, we aim at using these biases with domain-level knowledge of melanoma, to improve likelihood-based OOD detection of malignant images. The obtained results are encouraging and demonstrate potential for OOD detection of melanoma using NFs. The proposed method achieves an increase of 9% in Area Under the Curve of the Receiver Operating Characteristics by using wavelet-based NFs[2]. This model requires significantly less parameters for inference, making it more applicable and suited for edge devices. The proposed methodology can aid medical experts with diagnosis of skin-cancer patients and continuously increase survival rates. Furthermore, this research paves the way for other areas in oncology with similar data imbalance issues.

Section 6.2.1 discusses the difficulties in melanoma detection and introduces NF-based OOD detection as a potential solution direction. In Section 6.2.3 OOD detection using Wavelet Flow is proposed as a means to overcome the limitations of NFs by focusing on the low-frequency wavelet coefficients as a proxy for distinguishing between the semantics of benign and malignant melanoma.

6.2.1 Introduction to melanoma detection

Melanoma, a form of skin cancer, develops in the melanocytes of the skin [242]. Symptoms can develop in the form of changing moles or growth of new pigmentation. Non-cancerous growth of the melanocytes is referred to as benign melanoma and is not harmful, while malignant melanoma is harmful. It is essential to recognize the symptoms of malignant melanoma as early as possible to classify its malignancy in order to avoid late diagnosis and ultimately an increased mortality rate [243]. To classify melanoma malignancy, experts consider indications of the skin pigmentation such as asymmetrical shapes, irregular borders, uneven distribution of colors and large diameters (relative to benign melanoma) [244]. These clinical properties involve characteristics related to the texture and graphical details on the skin.

Since most cases of melanoma are benign, the number of malignant melanoma

[2]Code available at: https://github.com/A-Vzer/WaveletFlowPytorch

images are still relatively low. This data imbalance can negatively influence the predictions of machine learning (ML) models aiming to classify melanoma malignancy. Furthermore, most state-of-the-art supervised ML models are not calibrated, which poses the question on their validity for reliable skin-cancer detection [245]. Ideally, query images are assigned a calibrated confidence score, which can be interpreted as a probability of malignancy. Given these circumstances, a sensible option is to perform likelihood-based Out-of-Distribution (OOD) detection with the abundant benign data available.

Yielding tractable distributions, Normalizing Flows (NFs) serve as an excellent method for the aforementioned application. NFs are a family of completely tractable generative models that learn exact likelihood distributions. However, OOD detection with NFs is notoriously difficult. This is caused by its inherent learning mechanisms that result in inductive biases towards graphical details, such as texture or color-pixel correlations rather than semantic context in images [246]. As such, OOD data is often assigned similar or higher likelihoods than the training data. In the first half of this chapter, it is shown that with domain-level understanding of melanoma, OOD detection of melanoma is improved using NFs. Since the dominant features for indicating the malignancy of melanoma are described by their size and texture, wavelet-based NFs are employed. To this end, we implement Wavelet Flow [247] for OOD detection of malignant melanoma and realize a performance gain of 9% in Area Under Curve (AUC) of the Receiver Operating Characteristics (ROC). The number of parameters can be significantly reduced when applying Wavelet Flow for OOD detection, enabling implementation on smaller devices.

The proposed methodology presents the potential of NFs for aiding in reliable diagnosis of melanoma. Normalizing Flows for OOD detection and its inductive biases are discussed in Section 6.2.2. Thereafter, the approach and method are discussed in Section 6.3.1. The results are presented in Section 6.3.2.

6.2.2 Background to NF-based OOD detection

This section introduces the background of NFs (Subsection 6.2.2.A) and NF-based OOD detection (Subsection 6.2.2.B). The inductive biases in NFs are discussed in Subsection 6.2.2.C and finally the Wavelet Flow architecture is introduced in Subsection 6.2.3.

6.2.2.A Normalizing Flows

Normalizing Flows (NFs) are a sequence of bijective transformations, typically starting from a complex distribution, transforming into a Normal distribution. Using Section 2.2.2 as detailed introduction to NFs, here we discuss specific properties of NFs that are addressed in the research of this semantic shift section.

Formally, the log-likelihood $\log p(\mathbf{x})$ of a sample from the Normal distribution

subject to an NF transformation $f_i : \mathbb{R} \mapsto \mathbb{R}$ is computed with

$$\log p(\mathbf{x}) = \log p_{\mathcal{N}}(\mathbf{z}_0) - \sum_{i=1}^{K} \log \left(\left| \det \frac{df_i}{d\mathbf{z}_{i-1}} \right| \right), \tag{6.1}$$

where the latent sample $\mathbf{z}_i$ is from the i-th transformation in the K-step NF and parameter $p_{\mathcal{N}}$ the base Normal probability distribution. Due to the bijective relation of the transformations, Eq. (6.1) can be used to sample from $p_{\mathcal{N}}$ and construct a visual image with known probability. This transformation is referred to as the generative direction. Alternatively, an image can also be transformed in the normalizing direction (towards $p_{\mathcal{N}}$) to obtain a likelihood on the Normal density. Training in the normalizing direction is performed through Maximum Likelihood Estimation (MLE). Recently, many types of NFs have been proposed [89], [248]–[251]. Better flows are generally more expressive, while having an computationally inexpensive Jacobian determinant. For a more comprehensive explanation on NFs, the reader is referred to Section 2.2.2. A widely used choice of an NF step is coupling flows, such as RealNVP and Glow [91], [206]. The latter has been chosen as a baseline because it was SOTA at the time of the conducted experiments.

6.2.2.B Out-of-Distribution detection

The properties of NFs make them ideal candidates for OOD detection. Maximizing the likelihood of the data distribution $p(\mathbf{x})$ through a bijective transformation on $p_{\mathcal{N}}$ shifts likelihoods of OOD data, when the density is normalized. Nevertheless, NFs assign similar likelihoods to train and (in-distribution) test data, indicating that flows do not overfit. This also suggests that not all OOD data receive low likelihoods. Ultimately, the assigned likelihoods are heavily influenced by the inductive biases of the model. Many NFs have inductive biases which limit their use for OOD detection applications [246], [252].

6.2.2.C Inductive biases in coupling flows

Inductive biases of a generative model determine the training solution output and thus OOD detection performance. The input complexity plays an important role in OOD detection. Likelihood-based generative models assign lower likelihoods to more textured images, rather than simpler images [253]. The widely accepted *affine coupling* NF is used in this research study. Kirichenko *et al.* [246] show that structural parts such as edges can be recognized in the latent space. This suggests that this type of flow focuses on visual appearance such as texture and color of the images, as opposed to the semantic content. Furthermore, the authors present coupling-flow mechanics that cause NFs to fail at OOD detection. This concept is briefly discussed below to keep the section self-contained, but the reader is encouraged to address the original work.

Given image $\mathbf{x}$, coupling flows are masking the image partly (x_m) and update

it with parameters depending on the non-masked part x_{res}, which is described by

$$x_m = (x_m + T(x_{\text{res}})) \cdot e^{S(x_{\text{res}})}, \tag{6.2}$$

where S and T are parameterized functions that output the scale and translation parameters, respectively. The log-Jacobian determinant in Eq. (6.1) for coupling flows is calculated by

$$\log\left(\left|\det \frac{df_i}{d\mathbf{z}_{i-1}}\right|\right) = -\sum_{i=1}^{D} S_i(x_{\text{res}}), \tag{6.3}$$

where i iterates over the image dimensionality D. Naturally, function S is defined to predict high values in Eq. (6.2) to maximize the log-likelihood in Eq. (6.1). To compensate for this, function T should predict values that are an accurate approximation of $-x_m$. Therefore, the NF assigns high likelihoods to images when the flow can accurately predict the masked part of the image. This can enable solutions that assign high likelihoods to any structured image, regardless of its semantic contents. Two mechanisms are identified to drive the accurate prediction of masked pixels and therefore assign higher likelihoods to OOD data. These mechanisms are: (1) learning local color-pixel correlations and (2) information on masked pixels encoded in previous coupling layers, known as *coupling layer co-adaptation*. For the latter, different masking strategies such as cycle masking can be used to deprive the model from information in previous coupling-layer iterations [246]. Hence, experiments are performed with masking strategies to counteract coupling-layer co-adaptation. An example with the opposite effect to cycle masking, is checkerboard masking [91]. Masking in this manner means that the predicted pixels are conditioned on its direct neighbouring pixels. Continuously, this encourages the NF to leverage local pixel correlations and further hinders semantically relevant OOD detection.

6.2.3 Wavelet Flow

Yu *et al.* [247] introduced the Wavelet Flow architecture (Figure 6.1) for efficient high-resolution image generation. Instead of learning the image pixel likelihoods, the network models the conditional distribution with a coupling NF, specified by

$$p(\mathbf{x}) = p(L_0) \prod_{i=0}^{N-1} p(D_i|L_i), \tag{6.4}$$

where D_i and L_i are the detail and low-frequency components of the Haar decomposition, respectively, while N represents the number of decomposition stages. During inference, an independent sample from $p(L_0)$ is upscaled with the inverse Haar transform, using the predicted wavelet coefficients. At the time of research, this architecture was not yet tested for OOD detection. Modeling the wavelet coefficients further guides the model to consider the characteristic details of the image.

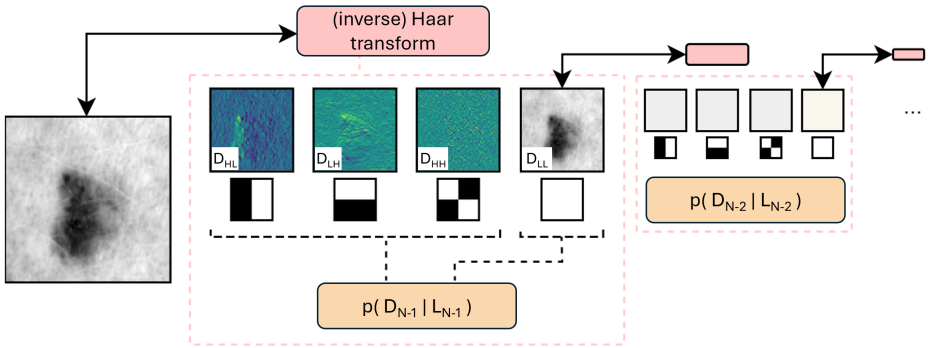


Figure 6.1 *Conceptual diagram of the Wavelet Flow architecture. At each decomposition level, the likelihood of the high-frequency wavelet coefficients are learned conditioned on the low-frequency decomposition. Probability density $p(L_0)$ is modeled unconditionally.*

As discussed in Section 6.2.1, melanoma can be distinguished by the texture of the skin. As a result, this inductive bias can improve OOD detection of melanoma. Furthermore, the high-frequency (detail) coefficients of the image makes it easier to distinguish between highly textured malignant and less structured benign melanoma. This can facilitate better OOD detection, since NFs tend to assign higher likelihoods to smoother images.

6.3 Semantic case: Method to OOD melanoma detection

Subsection 6.3.1 details the proposed approach to melanoma OOD detection and in Subsection 6.3.2 the results of the conducted experiments are discussed.

6.3.1 Melanoma detection with Wavelet Flow

As discussed in Section 6.2.2.C, the inductive biases of coupling NFs restrict their OOD detection capabilities. Given this information, we improve this by changing the data and model architecture. The proposed approach is tested on the ISIC dataset [254]. In this application, it can be beneficial that generative models assign higher likelihoods to less complex images, because benign melanoma are less textured and smaller in radius [244]. Initially, the RGB images are downscaled to 128×128 pixels and trained on the GLOW architecture naively, in a multi-scale setting, with default parameters K=32 and L=3. The AUCROC is used to evaluate the model performances. The color channels are heavily correlated and influence the likelihoods adversely, as discussed in Section 6.2.2.C. Therefore, we use grayscale images to hinder exploitation of local color-pixel correlations, as well as to reduce training complexity. Thereafter, the Wavelet Flow concept is employed. This shifts the optimization from the image pixels to their wavelet coefficients. This optimization will further bias the model towards the detailed semantic appearance of the images, since the tumor malignancy will become even more distinguishable by its textural description. Additionally, we experiment with different masking strate-

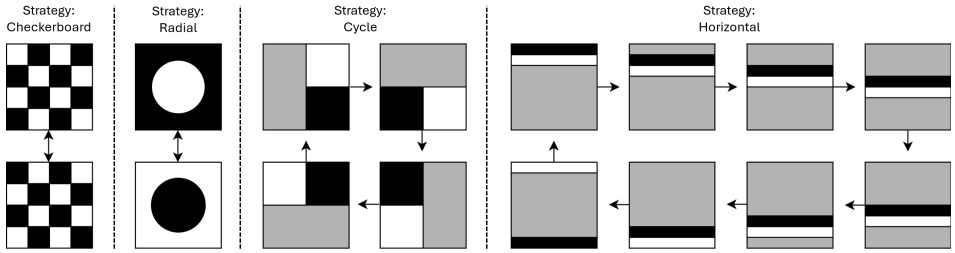


Figure 6.2 *Illustration of various masking strategies employed in the coupling-flow transformation. The masks vary at each coupling-flow step. The white area indicates the input of the S, T-network, which predicts parameters for the masked areas (indicated in black color). Grayscale areas are disregarded in the coupling-flow process.*

gies (see Figure 6.2). With the Wavelet Flow framework, we obtain a likelihood and thus an AUROC score per decomposition scale. The individual likelihoods are averaged over all scales that contain sufficient information about the original contents of the image. In this case, these are wavelet coefficients from 4×4 pixel dimensions up and until the highest decomposition level. As a suggestion for future work, it may be beneficial to select only particular scales with good AUROC values. However, this concept would require supervision, *i.e.* access to the malignant class, which is beyond the scope of this research.

6.3.2 Results and discussion on melanoma detection

Table 6.1 presents the AUROC curves for the various tested models. Likelihood distributions of the GLOW architecture trained on color images are shown in Figure 6.3. Firstly, it can be observed that the train and test sets coincide well, indicating the absence of over-fitting. When comparing the benign test to the malignant likelihoods, we obtain an AUROC of 0.73. This solution is suboptimal since many benign images are assigned a high negative likelihood score (long

Architecture	K	L	Channels	Masking	AUROC ↑	No. parameters ↓
GLOW	32	3	RGB	Affine	0.73	159M
GLOW	32	3	Gray	Affine	0.74	9.51M
GLOW	32	1	RGB	Affine	0.72	3.47M
GLOW	32	1	Gray	Affine	0.75	2.57M
Wavelet Flow	32	1	Gray	All	0.78	2.50M
Wavelet Flow	32	1	Gray	Checker	0.78	2.87M
Wavelet Flow	32	1	Gray	Horizontal	0.78	2.87M
Wavelet Flow	32	1	Gray	Cycle	0.78	2.87M
Wavelet Flow	32	1	Gray	Radial	0.78	2.87M
Wavelet Flow	16	1	Gray	All	**0.78**	**1.25M**

Table 6.1 *Test set results of the GLOW and Wavelet Flow models trained on the ISIC dataset for various hyperparameters. For Wavelet Flow, the number of parameters are that of the highest decomposition level as each level can be trained independently. Gray indicates grayscale image input. The best scores are printed in bold.*

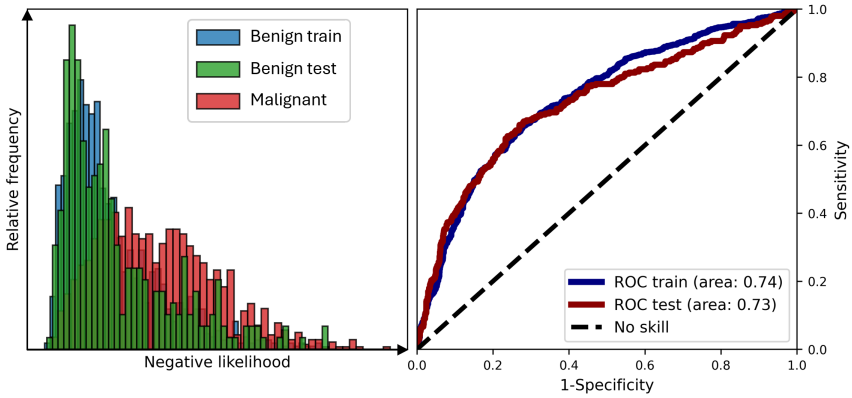


Figure 6.3 *Likelihood distribution and ROC curve of the trained GLOW architecture (baseline model).*

tail). In the same likelihood range, most of the malignant images are present as well. This is because the model learns color-pixel correlations, which can be used to leverage accurate predictions of the masked latent variables in the coupling layers. As a result, this leads to lower negative likelihoods assigned to OOD data.

When training on the wavelet coefficients with Wavelet Flow, there is a substantial improvement on several decomposition scales (see Figure 6.4). At all decomposition scales, besides the level seven (corresponding to the highest image resolution), an improvement in test evaluation is observed. The best AUROC values are found from the 3rd up until the 6th decomposition scales. At these levels, the wavelet coefficients represent the most relevant frequency components of benign and malignant melanoma. As expected, the lowest decomposition scales contain almost no relevant information on the malignancy of melanoma and have very low AUCROC values. In Figure 6.5, we average the likelihoods over the relevant decomposition scales and obtain a higher 0.78 AUROC score on the test set. In a separate evaluation, we have performed OOD detection using only the magnitude of the wavelet coefficients in which we observed acceptable AUROC values on individual scales. However, in contrast with Wavelet Flow, averaging over the decomposition scales is working adversely, making this approach infeasible. Furthermore, the different masking strategies do not improve performance.

Figure 6.6 illustrates the model's ability to distinguish between benign and malignant melanoma samples based on averaged negative likelihood values. As can be observed in the figure, benign and malignant samples exhibit distinct patterns along the likelihood axis. Notably, at smaller negative likelihood values, the visual features of malignant samples resemble those of benign melanomas, especially in terms of pigmentation and texture. This suggests that early-stage malignancies can appear similarly to benign samples, making it challenging for

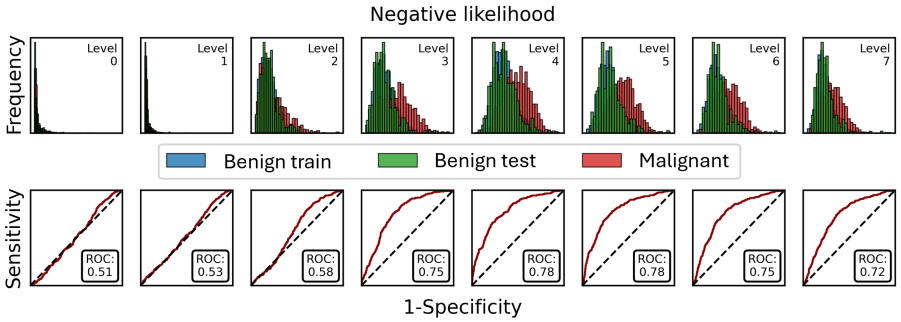


Figure 6.4 *Likelihood distributions per Haar wavelet decomposition level*

the model to identify them as OOD.

As the negative likelihood increases, more prominent texture patterns emerge, often characterized by darker, splattered pigmentation and increased hair coverage. These traits align with known visual indicators of malignancy, particularly irregular pigmentation and surface texture. While the presence of larger area of pigmentation alone is not necessarily indicative of malignancy, darker and more complex textures are key features that the model successfully identifies. The model assigns higher negative likelihood values to these malignant cases, recognizing them as OOD.

Interestingly, in cases where hair significantly obscures the lesion, the model can misclassify benign samples, assigning them higher negative likelihoods. This can occur due to the increased high-frequency details in the wavelet domain, which the model interprets as complexity, thus inflating the likelihood score. Such misclassifications align with previous findings in OOD detection, such as those presented by Serrà *et al.* [253], where likelihood estimates are impacted by the

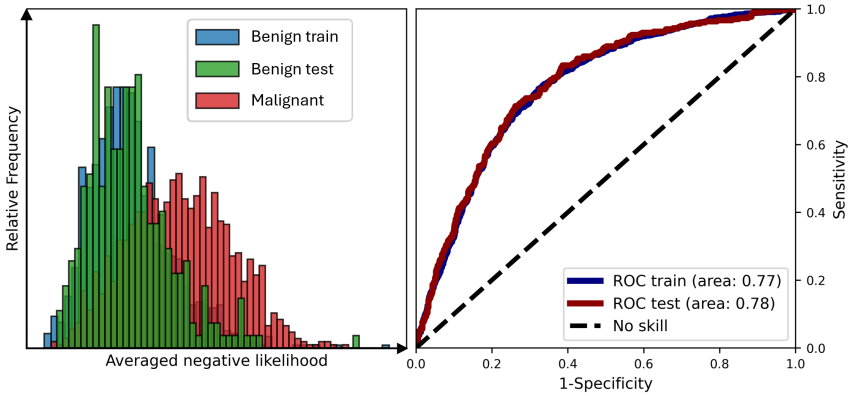


Figure 6.5 *The likelihood distribution and AUROC curve of the trained Wavelet Flow architecture, averaged over the decomposition scales*

complexity of the input. To correct for such instances, a complexity adjustment that accounts for the presence of hair could help reduce false positives. Hairy benign images, which may currently receive inflated negative likelihood scores due to their texture complexity, would be shifted to more appropriate smaller negative likelihood ranges. For more clearly visible malignant melanoma samples, the model relies on features beyond pigmentation size and hair presence, indicating that its inductive biases allow it to effectively capture key diagnostic information. This property equips it with the ability to differentiate between ID benign samples and OOD malignant melanoma with improved accuracy, despite occasional challenges posed by occlusion or image complexity.

Overall, the model's performance demonstrates that it can distinguish between benign and malignant patterns in the wavelet domain, although it may benefit from further refinement in handling cases where hair or other factors obscure the underlying melanoma. The inclusion of a complexity term could potentially improve the model's robustness in such cases, while still leveraging its ability to capture critical melanoma features that define malignancy.

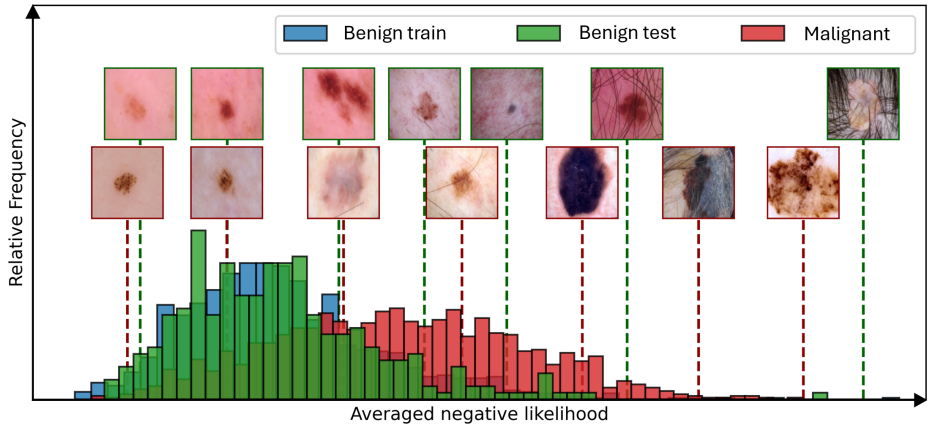


Figure 6.6 *Images of benign and malignant melanoma at various likelihoods. Note that lower likelihoods are either malignant or highly textured benign melanoma.*

In conclusion, late diagnosis of melanoma poses high risks for patients with skin cancer. Early detection of malignant melanoma with machine learning is highly valuable, but is difficult due to data imbalance caused by its relatively low occurrence. We have learned the benign image data distribution with Normalizing Flows to perform Out-of-Distribution (OOD) detection. It is shown that with knowledge on melanoma and the inductive biases of Normalizing Flows, we can improve likelihood-based OOD detection with wavelet-based Normalizing Flows. Furthermore, we have demonstrated that memory requirements for OOD detection can be reduced significantly with Wavelet Flow, enabling the deployment on edge devices. It is recommended to include a term in the likelihood calculations that corrects for the presence of hair in future work. The proposed

methodology focuses solely on melanoma, however, it is suggested that further research should aim at facilitating exact likelihood-based OOD detection for other areas of oncology featuring large data imbalances to improve detection accuracy.

6.4 Covariate case: Generative models for OOD covariate shift detection

6.4.1 Introduction to covariate shift

Identifying abnormal image statistics is critical for deploying precise sensing technology and reliable machine learning. Out-of-Distribution (OOD) detection methods are modeling the available data or a set of In-Distribution (ID) features, to identify test examples drawn from a different distribution. Notably, generative models offer an unsupervised paradigm to model distributions without making explicit assumptions on the form of the OOD data. With a plethora of possible covariates (abnormal variations in high-level image statistics) and potential downstream machine learning image applications, unsupervised generative modeling is a promising approach for general OOD detection. The prevailing approaches for OOD detection predominantly focus on the semantic contents of the image data, with little to no consensus over covariate shift. Therefore, this study elucidates covariate shifts, *i.e.* the change in distribution of high-level image statistics (covariates) subject to consistent low-level semantics.

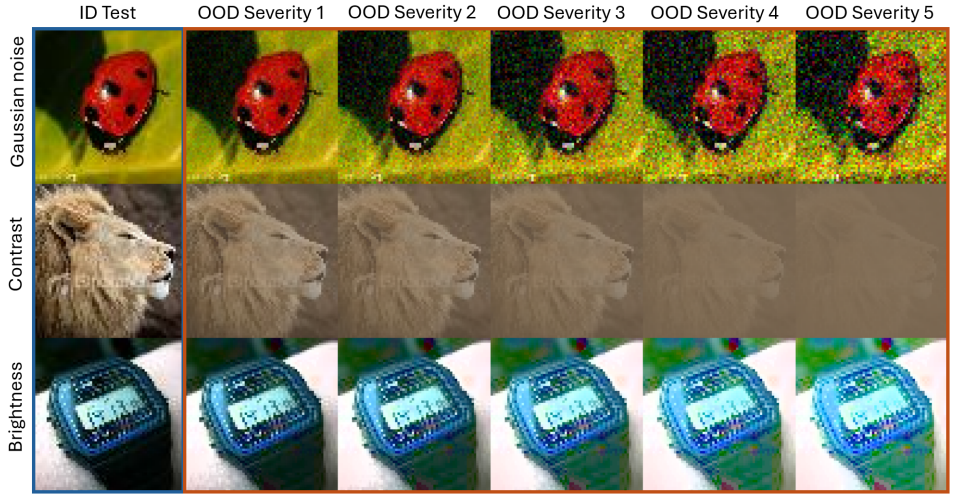


Figure 6.7 *In-Distribution ImageNet200 samples and different degradations and with growing severity levels of ImageNet200-C to the right as OOD.*

Likelihood-based methods, such as Normalizing Flows (NFs), offer an intuitive way of OOD detection by evaluating the likelihood of test samples. However, as evidenced in previous research [252], NFs have exhibited limitations in effective OOD detection, often assigning higher likelihoods to OOD samples. Various

works have explored this phenomenon and proposed alternative methods to direct likelihood estimation [255], [256]. Recent theoretical investigations [257] indicate that these methodologies are inherently susceptible to certain types of OOD data. Moreover, the metrics employed for evaluation exhibit a predisposition towards specific categories of OOD data, suggesting an intrinsic limitation of the current approach to OOD detection. In this study, we explore this phenomenon while improving the covariate OOD detection capabilities of NFs. Additionally, this shortcoming is addressed by proposing to unify the log-likelihood (LL)-based metric with the *typicality* score [258] in a simple Normalized Score Distance (NSD). Other generative models have been applied in various contexts to the task of semantic OOD detection, ranging from density-based methods [91], [259], [260] to different reconstruction-based models [261]. However, OOD covariate shift within the context of generative models remains largely unexplored.

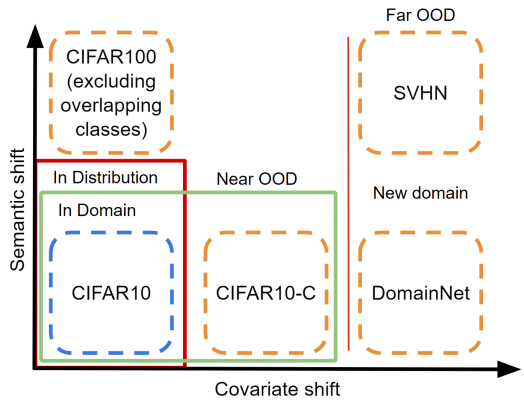


Figure 6.8 *Illustration of in-domain covariate shift (Near OOD) vs. covariate shift across domains (Far OOD). The shift is depicted in terms of popular computer vision datasets.*

We indicate two branches of covariate shift: (1) domain covariate shift, such as images in different styles (*e.g.* natural vs. sketch) and (2) domain-specific covariate shift (also known as sensory anomalies [260]), images under different sensing conditions (*e.g* different lighting, cameras or sensor-level degradation (Figure 6.7). Covariate shifts are recognized for their potential to significantly degrade the predictive performance of the model, where in some specialized imaging applications it can indicate system failure. Detecting these covariate factors and the distribution shifts under consistent semantic content [262] will enhance the safety and reliability of imaging systems in diverse fields and the machine learning systems being trained with these images [263]–[265]. This necessitates covariate shift detection, and if possible, the quantification of its severity. To this end and to the best of our knowledge, we are the first to implement unsupervised, domain-specific OOD covariate shift detection.

Images across different applications can demonstrate complex noise patterns and variability due to factors such as equipment variations, environmental condi-

tions, and the specific nature of the imaged objects or scenes [266]. A novel and effective strategy for improving OOD detection should utilize the data-dependent (heteroscedastic) noise that is present in the signal. This inherent noise serves as a rich source of information that can be exploited to differentiate between ID and OOD samples. In fact, the noise patterns in images can encode subtle differences that may not be apparent from the semantic content in the image alone.

To address these challenges and leverage the nuanced information encoded in noise patterns across various imaging applications, this work proposes a streamlined approach that models the conditional distribution between low-frequency and high-frequency signal components. This method contrasts with conventional techniques that attempt to model the entire signal distribution, which may inadvertently obscure critical covariate details. We employ a simple filtering approach that decomposes the image into distinct low-frequency and high-frequency components. By focusing on the interaction between these frequency components, the proposed approach effectively detect covariate shifts.

To define the research direction of this section, the original direction at the beginning of the chapter are recalled here. The research aims to determine if generative models can effectively detect and quantify covariate shifts in natural images. Additionally, it is known that covariate shift affects the distribution of high-frequency signal-dependent and independent details and, as such, how can these high-frequency heteroscedastic image components explicitly be modeled to improve OOD covariate shift detection performance.

This section delves into the concept of OOD covariate shift detection with a focus on generative models. A background on covariate shift is presented in Section 6.4.2, which additionally discusses semantic OOD detection techniques and introducess Normalizing Flows (NFs) and the concept of *Typicality*. The proposed approach to OOD Covariate Shift Detection is presented in Section 6.5, where the definition of covariate shift is revisited, and *CovariateFlow* (Section 6.5.2) is introduced. This section also discusses how to unify Log-likelihood and Typicality (Section 6.5.2.A) and provides an overview of the datasets used. Section 6.5.3 details the evaluation metrics and models employed, followed by an analysis of covariate shift in CIFAR10 and ImageNet200. The chapter then provides a discussion on covariate shift in natural images (Section 6.5.4), exploring its implications and challenges. The final sections cover the future Work and limitations (Section 6.5.5), and a specialized study on covariate shift in X-ray images (Section 6.5.6), including details on the X-ray dataset (Section 6.5.6.A) and results from the experiments (Section 6.5.7). The chapter concludes with a comprehensive summary on the key findings and contributions from both the semantic OOD detection and covariate shift detection research.

6.4.2 Background on covariate shift

Since detecting covariate shift is under explored, in Subsection 6.4.2.A state-of-the-art methods for semantic OOD detection are discussed. Subsection 6.4.3 introduces covariate shifts and hightlights the different application areas for which detecting

covariate shift is an attractive option. Subsection 6.4.4 briefly introduces Normalizing Flows and in Subsection 6.4.5 the concept of *Typicality* is introduced as an alternative approach to detecting semantic OOD concepts.

6.4.2.A Semantic Out-of-Distribution detection

Approaches to OOD detection are generally divided into two categories: *supervised*, which necessitates labels or OOD data, and *unsupervised*, which relies solely on ID data. Although semantic OOD detection does not constitute the core focus of this study, we nevertheless provide a concise overview of the recent developments, since these methodologies hold the potential to translate to covariate OOD detection. For an in-depth exploration of OOD detection methodologies, the reader is referred to the comprehensive review by Yang *et al*. [260].

Explicit density methods

A straightforward method for OOD detection involves the use of a generative model, $p(\mathbf{x};\theta)$ parameterized by θ and trained to fit a given distribution over data $\mathbf{x}$. The process evaluates the likelihood of new, unseen samples under this model with the underlying assumption that OOD samples will exhibit lower likelihoods compared to those that are ID. The Evidence Lower Bound (ELBO) employed in Variational Auto Encoders (VAEs) [259] can be used for OOD detection by evaluating a lower bound on the likelihood of a test sample. Plumerault *et al*. [267] introduced the Adversarial VAE – a novel approach that marries the properties of VAEs with the image generation quality of Generative Adversarial Networks (GANs), thereby offering a robust auto-encoding model that synthesizes images of comparable quality to GANs, while retaining the advantageous characteristics of VAEs.

Unlike VAEs, Normalizing Flows (NFs) [49] offer exact and fully tractable likelihood computations. With the introduction of coupling layers [91], NFs can be arbitrarily conditioned and seem to be excellent contenders for conditional OOD detection. However, as evidenced in previous research [268], NFs have exhibited limitations in effective OOD detection, often assigning higher likelihoods to OOD samples. This limitation has been associated with an inherent bias in flow model architectures, which tends to prioritize modeling local pixel correlations over the semantic content of the data [252].

Exploration by Gratwohl *et al.* [256] and Nalisnick *et al.* [255] posits that this phenomenon can be attributed to the fact that ID images are not high-likelihood samples but, rather, constituents of the *typical set* of the data distribution. Consequently, the investigation into methods that assess the *typicality* [258] of data instances, as an alternative to direct likelihood estimation, has gained traction. Despite empirical evidence demonstrating the efficacy of typicality in OOD benchmarks [258], recent theoretical investigations [257] indicate that these methodologies have inherent susceptibilities to specific OOD types and an evaluative bias towards particular OOD categories, thereby underscoring the complexity of OOD detection.

Image reconstruction-based methods
These OOD detection methods are based on the principle that models are less effective at accurately reconstructing data that significantly deviates from the training distribution. Graham *et al*. [269] improve on an innovative approach to OOD detection that leverages the potent generative prowess of recent denoising diffusion probabilistic models (DDPMs) [15], [270]. Unlike prior reconstruction-based OOD detection techniques that necessitated meticulous calibration of the model's information bottleneck [271]–[273], their method utilizes DDPMs to reconstruct inputs subjected to varying degrees of noise. This work implements this DDPM method as baseline for OOD covariate shift detection.

6.4.3 Previous work in detecting covariate shift

In essence, covariate shift refers to the phenomenon where images share consistent semantic content (*i.e.* similar subjects), and yet, are captured under varying imaging conditions. The degree of variation in these conditions signifies the magnitude of the shift. For example, a minor shift may involve images of a subject under varying lighting conditions, while a more substantial shift, such as transitioning from natural images to graphical sketches of the same subject, exemplifies a transition towards domain shift (Figure 6.8). This section concentrates on in-domain covariate shifts, as these scenarios represent instances where machine learning silently fails [263]–[265], [274].

In related work on covariate shift detection, Averly *et al*. [275] adopt a model-centric approach to address both covariate and semantic shifts, suggesting a methodology for identifying instances that a deployed machine learning model, such as an image classifier, fails to accurately predict. This strategy implies that the decision to detect, and potentially exclude a test example, is dependent on the specific model in question. While being effective for well-established machine learning models, this method inherently links the detection of shifts to the peculiarities of the individual model, resulting in each model having its unique set of criteria for rejecting data, which may vary broadly, even when applied to the same dataset. A significant drawback of this approach is its reliance on a robust pre-trained model, which poses a challenge for scenarios where identifying covariate shift is the primary objective, leaving such cases without a viable solution.

Generalized ODIN [276] is another direction of work that adopts the model-centric approach. This method replaces the standard classification head, and instead, decomposes the output into scores to behave like the conditional probabilities for the semantic shift distribution and the covariate shift distribution. This approach is then only evaluated on out-of-domain covariate shift such as the DomainNet [277] benchmark. Follow-up work by Tian *et al*. [262] further explores calibrating the confidence functions proposed in [276], which realize improvements on both semantic and covariate OOD detection. They additionally apply their refinement on in-domain covariate shift, such as CIFAR10 vs. CIFAR10-C.

Besides the above-mentioned work, covariate shift has been studied predomi-

nantly from a robustness perspective [263] or in a domain adaptation setting [278], [279]. The defense against adversarial attacks [265] is another research direction that falls in the domain of covariate shift. This perspective stems from the recognition that adversarial examples, by nature, often represent data points that deviate significantly from the distribution observed during model training, thereby inducing a form of covariate shift. Researchers have leveraged insights from adversarial robustness [265] to devise methods that can identify and mitigate the effects of such shifts, focusing on enhancing model reliability and security against deliberately crafted inputs designed to deceive. Fortunately, the shift introduced is completely artificial and typically a shift targeted at a specific model.

6.4.4 Normalizing Flows (NFs)

Consider an image sampled from its intractable distribution as $\mathbf{x} \sim P_{\mathbf{X}}$. Additionally, we introduce a simple, tractable distribution $p_{\mathbf{Z}}$ of latent variable $\mathbf{z}$ (which is usually Gaussian). Normalizing Flows utilize K consecutive bijective transformations $f_k : \mathbb{R}^D \to \mathbb{R}^D$ as $\mathbf{f} = f_K \circ \ldots \circ f_k \circ \ldots \circ f_1$, to express exact log-likelihoods by

$$\log p(\mathbf{x}) = \log p_{\mathbf{Z}}(\mathbf{z}_0) - \sum_{k=1}^{K} \log \left| \det \frac{\mathrm{d} f_k(\mathbf{z}_{k-1})}{\mathrm{d}\mathbf{z}_{k-1}} \right|, \tag{6.5}$$

where $\mathbf{z}_k$ and $\mathbf{z}_{k-1}$ are intermediate variables, and $\mathbf{z}_0 = \mathbf{f}^{-1}(\mathbf{x})$.

Numerous bijections have been introduced which balance expressivity and have a simple evaluation of the Jacobian determinant in Equation (6.5). Specifically, coupling flows have seen much success [91], [280]. Since they can be parameterized through arbitrary complex functions, we explore conditioning the flow on the frequency components of an image.

6.4.5 Typicality

Examining sequences of N independent and identically distributed (i.i.d.) datapoints $\mathbf{x}_n$, the *typical set* comprises all $\mathbf{x}_n$ that satisfy

$$H(\mathbf{X}) - \epsilon \leq -\frac{1}{N} \sum_{n=1}^{N} \log_2 p(\mathbf{x}_n) \leq H(\mathbf{X}) + \epsilon, \tag{6.6}$$

where ϵ represents an arbitrary small value and $H(\mathbf{X})$ denotes the Shannon entropy of the dataset. In other words, the empirical entropy of the set approaches the entropy of the source distribution. Leveraging the Asymptotic Equipartition Property (AEP), it is deduced that

$$\frac{1}{N} \sum_{n=1}^{N} \log_2 p(\mathbf{x}_n) \to H(\mathbf{X}) \quad for N \to \infty, \tag{6.7}$$

leading to the conclusion that the probability of any *sequence* of i.i.d. samples of sufficient length approaches unity. Thus, despite the typical set representing

merely a small subset of all potential sequences, a sequence drawn from i.i.d. samples of adequate length will almost certainly be considered typical [281].

In various studies, indications have emerged that NFs perform poorly when the likelihood is utilized as a metric for detecting OOD samples [252], [257], [268], [282]. It can be argued that datasets are a typical sequence of samples, rather than high in likelihood, also known as the Typical Set Hypothesis (TSH). Therefore, in the recent work by Nalisnick *et al*. [255], an innovative approach is proposed for OOD detection that leverages typicality as an evaluation metric in lieu of likelihood. This methodology has been further refined in subsequent studies [256], introducing *approximate mass*. Motivated by the fact that typical samples are localized in high-mass areas on the PDF, the metric evaluates the gradient of the LL w.r.t. the input data, also known as the *score*. The value of this score can be expressed mathematically as

$$\text{Typicality}_{\text{score}} = \|\partial L(\mathbf{x};\theta)/\partial \mathbf{x}\|\,, \tag{6.8}$$

where $\mathbf{x}$ denotes the input, L is the evaluated LL by the model parameterized by θ, and $\|.\|$ represents the Euclidean norm. Despite some criticism on TSH [257], this metric demonstrates superior performance in OOD detection across various benchmarks [256], [258].

6.5 Covariate case: Method to OOD covariate shift detection

In this section a formal definition for covariate shift in terms of low-frequency and high-frequency components are provided (Subsection 6.5.1), followed by an overview to the proposed CovariateFlow in Subsection 6.5.2. Subsection 6.5.3 details the experiments conducted to validate the proposed approach and in Subsection 6.5.4 the results from these experiments are discussed. Subsection 6.5.5 suggest future avenues for research and improvements to covariate shift detection in natural images. The same CovariateFlow approach is extended to the X-ray domain to detect perturbations in imaging statistics in Subsection 6.5.6 and the results from these experiments are discussed in Subsection 6.5.7.

6.5.1 Definition of Covariate Shift

Formally, semantic- and in-domain covariate shifts can be delineated as follows. Consider samples from the training distribution, $\mathbf{x} \sim P_{\mathbf{X}}$, and anomalous data from an OOD source $\hat{\mathbf{x}} \sim P_{\hat{\mathbf{X}}}$. These are subject to a low-pass filter (g_{L}, $g_{\text{L}} : \mathbb{R} \to \mathbb{R}$) to obtain the low-frequency components, $\mathbf{x}_{\text{L}} = g_{\text{L}}(\mathbf{x})$ and the high-frequency components $\mathbf{x}_{\text{H}} = \mathbf{x} - g_{\text{L}}(\mathbf{x})$. *Semantic shift* is characterized by a discrepancy in the marginal probability distributions, $P_{\mathbf{X}_{\text{L}}} \neq P_{\hat{\mathbf{X}}_{\text{L}}}$, when the conditional probability distributions of high-frequency components remain consistent, $P_{\mathbf{X}_{\text{H}}|\mathbf{X}_{\text{L}}} \approx P_{\hat{\mathbf{X}}_{\text{H}}|\hat{\mathbf{X}}_{\text{L}}}$. Conversely, *covariate shift* is identified when the conditional probability distributions diverge, $P_{\mathbf{X}_{\text{H}}|\mathbf{X}_{\text{L}}} \neq P_{\hat{\mathbf{X}}_{\text{H}}|\hat{\mathbf{X}}_{\text{L}}}$, but the marginal probability distributions of the low-frequency components remain the same $P_{\mathbf{X}_{\text{L}}} \approx P_{\hat{\mathbf{X}}_{\text{L}}}$. Furthermore, these definitions hold with in the supervised setting with predefined targets ($\mathbf{Y}$).

6.5.2 CovariateFlow

In the development of methodologies for detecting covariate shift within datasets, several critical factors should be meticulously considered to ensure efficacy and accuracy. Firstly, the process of resizing images can significantly alter the distribution of high-frequency statistics, potentially obscuring key data characteristics. Secondly, the inherent nature of encoding architectures, which essentially function as low-pass filters [283], may constrain their capacity to fully capture the complex distribution of noise present within the data. This limitation is particularly relevant as covariate shifts often manifest through alterations in the general image statistics, thereby necessitating a method capable of discerning such nuances. Thirdly, the utilization of *only* log-likelihood-based evaluation in NFs, has proven a predisposition towards low-level semantics and is more sensitive to high-frequency statistics [252]. An effective method should be sensitive to covariate shifts affecting all frequency bands, from noise degradations to contrast adjustments.

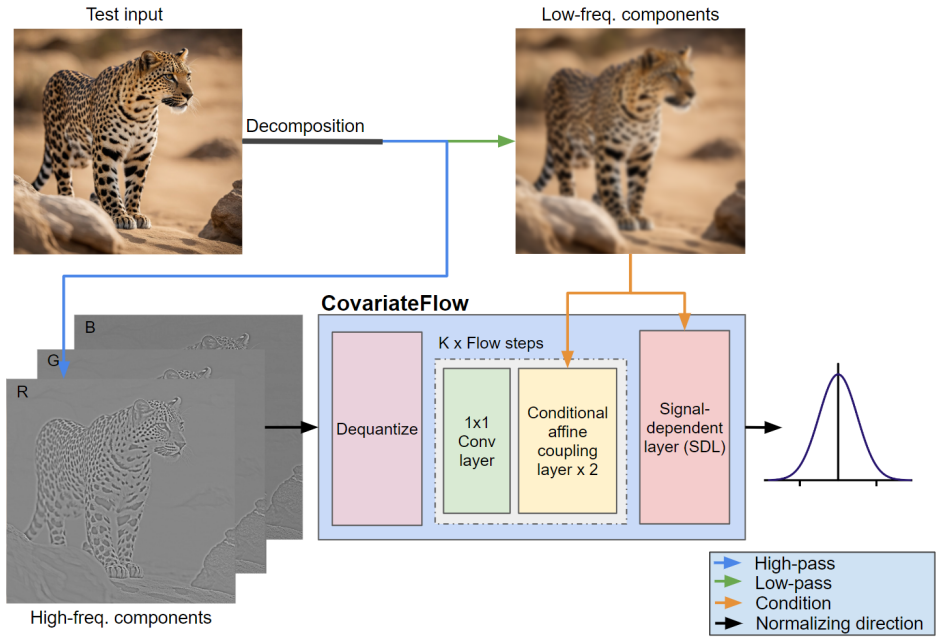


Figure 6.9 *High-level diagram of the CovariateFlow model architecture introduced in this section. The test image is decomposed into low-frequency and high-frequency components using a Gaussian filter. The high-frequency components are transformed into a Normal distribution through a series of flow steps that are conditioned on the low-frequency image components.*

In light of the above considerations, Normalizing Flows (NFs) emerge as a particularly suitable candidate for modeling the imaging features essential for detecting covariate shift. NFs are distinct as they abstain from any form of downsampling or encoding processes to preserve their bijective property. It is also recog-

nized that NFs prioritize pixel correlations over semantic content [252]. However, given the expectation that covariate shift involves changes in high-frequency image statistics, accurately modeling the complete image distribution (including both low-frequency semantics and high-frequency components) presents significant challenges. This is especially due to the relatively limited capacity of NFs compared to more recent generative models [15], [270], [284].

To re-position our focus on leveraging generative models for covariate shift detection rather than image generation, two key approaches can be considered. (1) Utilize state-of-the-art methods with enhanced modeling capacity, which could maximize the model's overall performance. (2) Prioritize simplifying the objective by concentrating solely on the components essential for effective covariate shift detection. The latter approach reduces complexity and tailors the model specifically to the task at hand, potentially offering a more efficient solution. To detail the proposed approach, we introduce a novel method that simplifies the modeling of components informative for covariate shift detection. The proposed approach involves a filtering strategy that divides the image into separate low-frequency and high-frequency components, thereby allowing the detection system to concentrate specifically on the high-frequency elements to improve detection capabilities.

More formally, consider an input signal $\mathbf{x}$ and the low-pass filter g_{L}, the high-frequency components of $\mathbf{x}$ are $\mathbf{x}_{\mathrm{H}} = \mathbf{x} - g_{\mathrm{L}}(\mathbf{x})$, given the low-frequency components are computed by $\mathbf{x}_{\mathrm{L}} = g_{\mathrm{L}}(\mathbf{x})$. By recognizing that certain high-frequency components are correlated with low-frequency signals, we can model this relationship conditionally. Based on this premise, we develop the CovariateFlow model (Figure 6.9), which constitutes a novel approach of modeling the conditional distribution between high-frequency and low-frequency components using conditional NFs specified by

$$\log p(\mathbf{x}_{\mathrm{H}}|\mathbf{x}_{\mathrm{L}}) = \log p_{\mathbf{Z}}(\mathbf{z}_0) - \sum_{k=1}^{K} \log \left| \det \frac{\mathrm{d} f_k(\mathbf{z}_{k-1}, \mathbf{x}_{\mathrm{L}})}{\mathrm{d}\mathbf{z}_{k-1}} \right|. \tag{6.9}$$

This formulation sets the foundation for a detection system that is finely tuned to the nuances of covariate shift, enhancing its ability to identify and respond to shifts in high-frequency image statistics. The proposed model is predominantly defined by (1) a signal-dependent layer (SDL) [285], (2) conditional coupling flow [280], (3) an unconditional 1×1 convolutional (conv.) layer [206] and (4) uniform dequantization. The SDL layer and conditional coupling layer are specifically conditioned on $\mathbf{x}_{\mathrm{L}}$. The 1×1-convolution and conditional coupling flow is repeated K times, depending on the dataset at hand. We employ a Gated ResNet [286] as f_Θ and a checkerboard masking strategy [91] in our coupling layers. Figure 6.9 depicts a high-level overview of the CovariateFlow model architecture. We employ a simple Gaussian filter for g_{L}, to decompose the signal into low-frequency and high-frequency components. To minimize any assumptions about the high-frequency components, we use a conventional Gaussian kernel. A kernel with a

standard deviation (σ) of unity has empirically proven to yield the best performance. The coupling layers are depicted in Figure 6.10 and the model involves a mere 945,882 trainable parameters with K=8 resulting in 16 coupling layers (Figure 6.10 is employed 8 times). For a detailed description on training details, ablation experiments and inference time comparisons, the reader is referred to Appendix D.7 in the supplementary part of this thesis. The code for the model is publicly available[3].

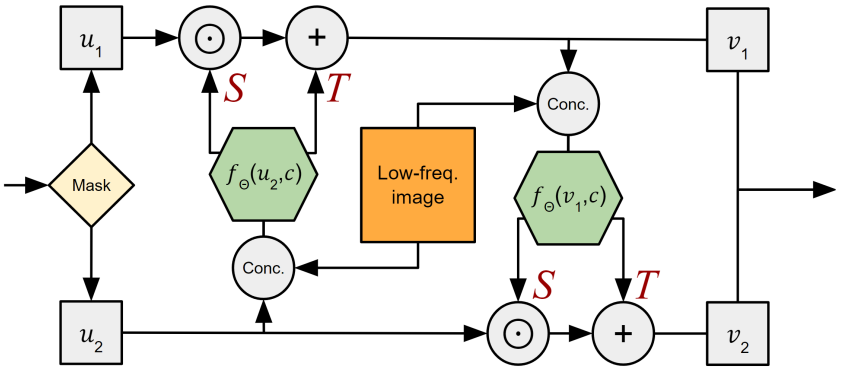


Figure 6.10 *Processing diagram of the conditional affine coupling layer (2 layers depicted, in total 8 of such stages) in the third block of Figure 6.9. Function $f_\Theta(\mathbf{x}, \mathbf{c})$ details the neural network transforming the image. In the figure, u_1 and u_2 refer to the different masked parts of the incoming image and v_1 and v_2 their transformed variants. (Conc. = concatenation)*

Chapter 6

6.5.2.A Unifying Log-likelihood and Typicality

The inductive bias of NFs towards structural complexity when evaluating with a log-likelihood (LL) function has been discussed in Section 6.4.5. As an alternative, evaluation on typicality using the gradient of the LL w.r.t. the input data, has shown improvements in semantic OOD detection over using LL solely [258], [287]. However, it has been found in literature that the metric and model are similarly biased towards certain categories of data [257]. As such, we propose to combine the LL evaluation with the Typicality score (Eq. 6.8) to overcome the limitations of each individual approach. The proposed approach normalizing both the LL and the typicality scores in terms of their respective training statistics to zero mean and unity variance. After normalization, we can transform each metric into an absolute distance from the expected mean. The LL distance and Typicality score distance can then simply be added to obtain a unified distance. In this manner, the evaluation is sensitive to all statistical deviations, rather than only being lower in score, thereby reducing the effect of the biases of the respective metrics.

The above-mentioned discussion on computing and normalization of the the LL and Typicality distances is presented more formally here. Again, we consider

[3]https://github.com/covariateflow/CovariateFlow

a sample $\mathbf{x} \sim P_{\mathbf{X}}$ with log-likelihood $\log p(\mathbf{x})$. Furthermore, the magnitude of the gradients is denoted as $||\nabla_{\mathbf{x}} \log p(\mathbf{x})||$, i.e. the typicality score. The means for the empirical likelihoods are determined through $\mu_L = \mathbb{E}_{P_{\mathbf{X}}}[\log p(\mathbf{x})]$, and of the Typicality scores with $\mu_T = \mathbb{E}_{P_{\mathbf{X}}}[\,||\nabla_{\mathbf{x}} \log p(\mathbf{x})||\,]$. Similarly, the variances of the LL and Typicality are denoted by $\sigma_L^2 = \mathbb{E}_{P_{\mathbf{X}}}[(\mathbf{x} - \mu_L)^2]$ and $\sigma_T^2 = \mathbb{E}_{P_{\mathbf{X}}}[(\mathbf{x} - \mu_T)^2]$, respectively. Finally, the Normalized Score Distance (NSD) is then obtained for a new sample $\mathbf{x}^*$ as the summation of the standardized L1-norms by computing

$$\text{NSD}(\mathbf{x}^*) = \left| \frac{\log p(\mathbf{x}^*) - \mu_L}{\sigma_L} \right| + \left| \frac{||\nabla_{\mathbf{x}} \log p(\mathbf{x}^*)|| - \mu_T}{\sigma_T} \right|. \tag{6.10}$$

Figure 6.11 depicts the individual scores (LL and Typicality) and then the resulting NSD for two separate degradation types on CIFAR10 and using the GLOW architecture (Baseline architecture from the beginning of this chapter). This experiment is conducted to visualize the limitations of each individual metric and illustrate how the proposed NSD overcomes these limitations, utilizing the strengths of each metric.

6.5.2.B Datasets

CIFAR10(-C) & ImageNet200(-C): CIFAR10 [288] and ImageNet200 with their respective corrupted (C) counterparts, CIFAR10-C [289] and ImageNet200-C [289], serve as exemplary datasets for developing and evaluating unsupervised covariate shift detection algorithms. CIFAR10 and ImageNet200 provide a collections of images that encompass a broad range of in-distribution covariate shifts, ensuring a suitable level of diversity. However, the corrupted versions introduce real-world-like (undesired) degradations, such as noise, blur, weather, and digital effects. Figure 6.7 depicts 3 of the 15 effects employed in the ImageNet200-C dataset. Images are utilized in their original resolution at 64×64 pixels. CIFAR10-C consists of 19 corruptions in total with images at 32×32 pixels. This setup enables testing the detection performance of covariate shift across multiple distortion types and severity levels.

In all the conducted experiments, we train the models only on the original dataset training set and then test it against *all* corruptions at every severity level. For CIFAR10, this is the original test set of the dataset (ID test) and CIFAR10-C's 19 corruptions at 5 severity levels (95 OOD test sets). Similarly, we treat the ImageNet200 test set as ID test and the 15 corruptions at 5 severity levels from ImageNet200-C as 75 OOD test datasets. The datasets follow the OpenOOD [290] benchmarks[4].

6.5.3 Experiments

This section describes the conducted experiments and presents the key results obtained in the investigation. Further detailed experimental results can be found

[4]https://github.com/Jingkang50/OpenOOD

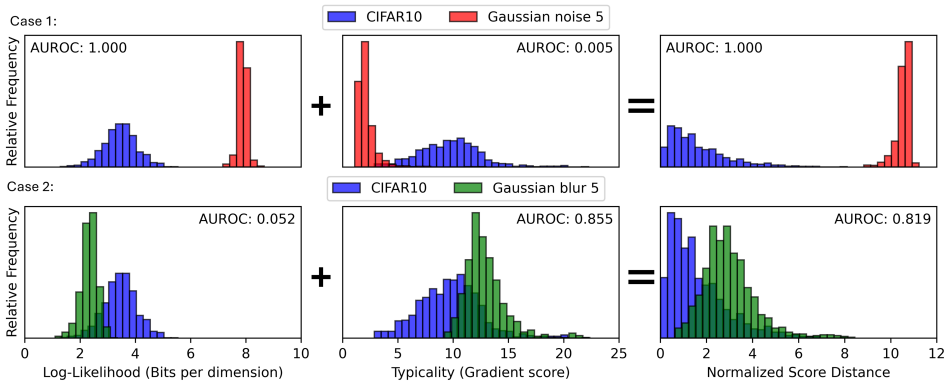


Figure 6.11 *Visual diagram of graphically adding Log-likelihood (Bits per dimension) and Typicality (Gradient) score, resulting in the NSD value at the right. The top row depicts CIFAR10 + Gaussian Noise 5 and the bottom row CIFAR10 + Gaussian Blur 5. In the first column (LL), the first metric separates best noise-based degradation, but fails at blur, while in the second column (typicality) fails at noise but succeeds at detecting blur. NSD successfully distinguishes between CIFAR10 test and both degradation types.*

in the supplementary materials, specifically, results on CIFAR10 (Appendix D.3), ImageNet200 (Appendix D.4) are presented. Extensive ablation experiments with the proposed CovariateFlow model are described in Appendix D.7.

6.5.3.A Evaluation Metrics & Models

To evaluate the model ability to detect OOD covariate shifts, commonly found metrics from related work are utilized: the Area Under the Receiver Operating Characteristic (AUROC) curve and the False Positive Rate (FPR) at a 95% True Positive Rate (TPR). In all the experiments with CIFAR10(-C) and ImageNet200(-C), we use the designated test set (10k samples) to compute each metric.

The contributions of this complete section include contextualizing the VAE, AVAE, GLOW evaluated with log-likelihood and the DDPM with the reconstruction loss, within OOD covariate shift as baseline models. Furthermore, we evaluate GLOW using typicality and the proposed NSD metric and the CovariateFlow model with all the aforementioned metrics. Most models are trained from scratch on the ID data. For the VAE-FRL [291], a method leading in semantic OOD detection, the available pretrained CIFAR10 weights[5] are utilized. A detailed description of the implemented models can be found in Appendix D.1 of the supplementary materials in this thesis.

6.5.3.B Covariate Shift in CIFAR10 and ImageNet200

Table 6.2 showcases various models and their averaged AUROC values across all the degradations per CIFAR10-C/ImageNet-C severity level. While some models

[5]https://github.com/mu-cai/FRL

excel in handling specific types of degradation, only the overall performance is truly relevant, since it is difficult to predict the type of perturbation that will occur in real-world settings. A detailed breakdown of the results per perturbation is shown in Appendix D.3 of the supplementary materials.

In Table 6.2 it can be observed that models preserving the data dimension and maintaining the high-frequency signal components, such as the DDPM and NF-based approaches, perform best. ImageNet200-C contains fewer noise-based degradations than CIFAR10-C. The NF models evaluated with LL generally perform well on noise perturbations (Table D.7 and Table D.15) and because of this disparity in the types of degradations present in the datasets, the LL evaluation exhibits a drop in average performance going from CIFAR10 to ImageNet200. The VAE-FRL is designed to focus on semantic content and thus fails to accurately detect a change in general image statistics. It can be observed that CovariateFlow with NSD consistently outperforms the other methods at every severity level, realizing an average improvement of 5.6% over GLOW on CIFAR10 and 7.8% over GLOW on ImageNet200 when evaluated with the NSD metric. This shows the strength of the proposed NSD metric, consistently improving over just LL or Typicality on both the GLOW and CovariateFlow models. Figure 6.11 highlights an example of how NSD consistently performs well under different degradations.

Appendix D.3 and Appendix D.4 present a comprehensive evaluation of various methods for every type of OOD covariate shift between the CIFAR10(-C) and ImageNet200(-C) datasets. Table D.1 focuses on the model performances across three specific degradations (Gaussian Noise, Gaussian Blur, and Contrast) at five severity levels, which summarize the general results seen across all degradations. ImageNet200-C does not contain Gaussian Blur, but in general, the same trend can be observed between the two datasets for all the employed models. A complete comparison between all the models and their average performance per degradation type (averaged over severity levels) can be found in Appendix Table D.13.

To evaluate the impact of filter kernel size on performance, we have conducted an experiment using CovariateFlow. Figure 6.12 illustrates the average AUROC score achieved with varying Gaussian filter sizes employed when train the CovariateFlow model. The sigma values in the table indicate the size of the Gaussian kernel used in this process. The results indicate that a smaller filter (Sigma=1) yields the highest average performance. Example evaluations from the CovariateFlow model with NSD are presented in Figure 6.13. Notably, the evaluated scores increase with each severity level, although the rate of increase is not linear or consistently increasing between the different degradation types. The CovariateFlow model is fully invertible and, as such, can generate heteroscedastic high-frequency components. Figure D.11 in the Appendix depicts an example with sampled high-frequency components, the reconstructed image and a comparison between the reconstructed image and the original image.

CIFAR10 ID Data	CIFAR10-C Data OOD Severity Levels					Average AUROC↑
Models	1	2	3	4	5	/FPR95↓
Reconstruction						
DDPM [269] (T150: LPIPS)	55.1	59.9	63.6	66.5	70.5	63.1 / 83.9
DDPM [269] (T20: LPIPS+MSE)	58.2	63.8	69.0	71.0	75.6	67.5 / 75.2
Explicit Density						
Vanilla VAE [259]	48.3	47.8	48.8	50.3	49.5	48.9 / 83.3
AVAE [267]	53.6	58.0	60.2	63.9	65.2	60.2 / 73.1
VAE-FRL [291]	51.0	56.4	55.8	59.3	63.6	57.2 / 76.3
GLOW [206] (LL)	60.7	57.5	58.4	58.7	57.7	57.7 / 69.5
GLOW [258] (Typ.)	41.9	42.9	41.2	40.7	41.2	41.6 / 85.8
GLOW (NSD)	63.1	67.7	68.9	70.9	75.6	69.3 / 65.7
CovariateFlow (LL)	59.8	56.6	57.3	58.5	59.1	58.3 / 63.5
CovariateFlow (Typ.)	44.5	46.1	46.1	45.1	45.7	45.5 / 83.8
CovariateFlow (NSD)	**65.9**	**72.9**	**75.5**	**78.6**	**81.7**	**74.9** / **61.7**
ImageNet200 ID Data	ImageNet200-C Data OOD Severity Levels					Average AUROC↑
Models	1	2	3	4	5	/FPR95↓
Reconstruction						
DDPM [269] (T20: LPIPS+MSE)	48.6	56.9	65.1	69.7	74.0	62.9/75.8
Explicit Density						
Vanilla VAE [259]	31.5	36.1	40.2	42.6	45.7	39.3 / 92.9
AVAE [267]	34.7	37.9	40.8	42.3	44.9	40.1 / 92.7
GLOW [206] (LL)	35.2	38.4	37.0	35.8	34.7	36.2 / 81.7
GLOW [258] (Typ.)	50.7	48.8	49.9	51.7	53.8	51.0 / 79.8
GLOW (NSD)	52.3	61.6	66.4	69.8	72.4	64.5 / 65.6
CovariateFlow (LL)	18.7	23.7	27.7	28.6	29.0	25.5 / 86.9
CovariateFlow (Typ.)	**65.6**	64.1	60.9	61.4	62.0	61.8 / 73.1
CovariateFlow (NSD)	64.2	**64.7**	**74.6**	**78.0**	**80.0**	**72.3** / **60.1**

Table 6.2 *Average AUROC scores of various methods on detecting the different severity levels of OOD covariate shift with the CIFAR10(-C) and ImageNet200(-C) dataset.*

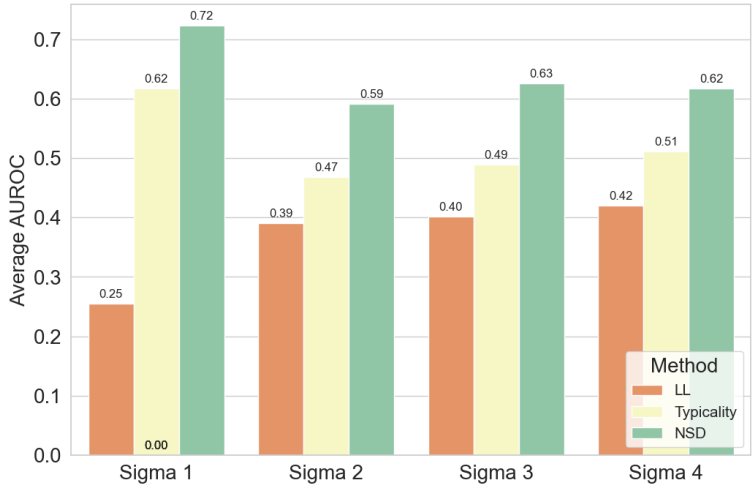


Figure 6.12 *The average AUROC obtained with CovariateFlow model on ImageNet200 vs. ImageNet200-C (all corruptions) at different filter sizes. The figure depicts the score obtained with evaluating using log-likelihood, typicality and the proposed NSD.*

6.5.4 Discussion on covariate shift in natural images

The findings from the analyses validate the hypothesis that OOD covariate shifts can be effectively identified by explicitly modeling the conditional distribution between low-frequency and high-frequency components. The proposed CovariateFlow model is designed to specifically capture this distribution, thereby surpasses other methodologies in detecting covariate shifts in CIFAR10 and Ima-

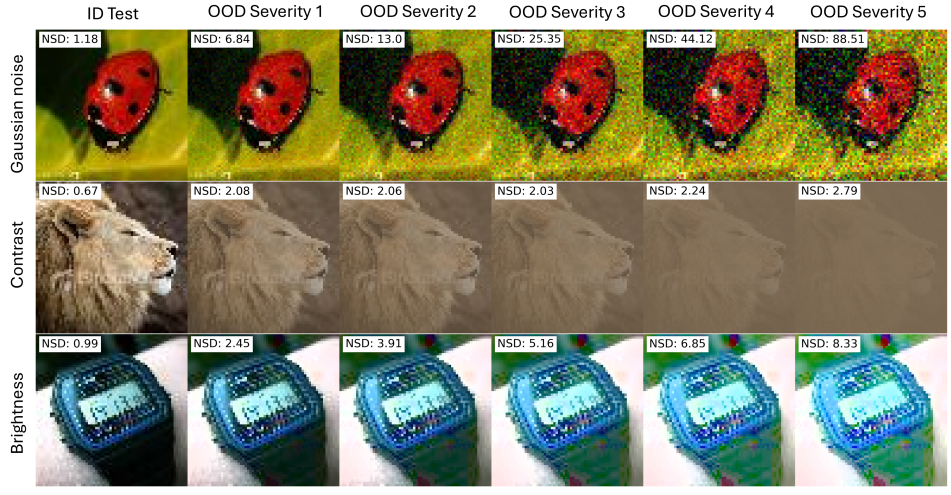


Figure 6.13 *Example CovariateFlow model (NSD) predictions for images from the ImageNet200 ID test set and corresponding covariate shifted images from ImageNet200-C. The severity level of the covariate shift increases from left to right.*

geNet200. Given the diverse array of subjects and covariate conditions within the corrupted datasets, focusing on this conditional distribution streamlines the task of the model, allowing it to concentrate on the most relevant distribution for the detection process.

Using the analysis in Table D.1 in Appendix D, the VAE-based models show adequate performance in detecting noisy degradations, due to their inductive bias towards modeling low-frequency image components. However, the model falls short for this exact reason when exposed to any blurring or color degradations in the images. The DDPM with the LPIPS+MSE metric, present strong performance on noise and blurring-based covariate shift, but struggles when exposed to color shift. This is likely due to color reconstructions happening earlier in the reconstruction schedule. Consistent with existing literature [252], the NF-based methods evaluated using LL are extremely sensitive to noisy degradations. However, any blurring or color shift is evaluated as being highly probable under the modelled distribution, highlighting the bias of LL-based evaluation towards lower textural content. Employing the newly proposed *typicality* metric shows the exact opposite behaviour. Both GLOW and the proposed CovariateFlow, fail at detecting noise-based covariate shift, but show remarkable improvements on both blurring and color-based covariate shifts when evaluated with *typicality*. Combining typicality and LL in the newly proposed NSD metric accentuates the strengths of each, enabling strong detecting performance across most of the covariates with CovariateFlow. NSD enhances the OOD detection capabilities of both the standard GLOW model and the proposed CovariateFlow, establishing it as a general and robust metric for OOD detection in NF-based models. On the higher resolution images from ImageNet200, the model also shows some effectiveness in distinguishing JPEG compression as OOD, a difficult perturbation to detect.

When to use CovariateFlow: Despite GLOW evaluated with LL slightly superior performance in general noise detection, the CovariateFlow model leveraging NSD as metric, proves to be better overall. This provides a clear and general recommendation for its applicability: LL is preferred in case strictly increasing noise-based shifts are expected. Without *a-priori* knowledge on the OOD shift type (which is usually the case), the CovariateFlow model with NSD is optimal. This work demonstrates that it is possible to detect (even slight) perturbations in a target domain without introducing biases or prior knowledge of these perturbations into the model, unlike some contrastive learning approaches [292]. It only assumes access to a sufficiently large dataset that captures the *in-distribution covariate shifts* and aims to detect any covariate shift outside of this distribution.

6.5.5 Future work and limitations

Some concerns can be raised about the complexity of the *typicality* computation, since test time inference requires a forward pass to compute the LL followed by a backpropagation computation per sample. This increases the memory requirements when deploying the model and decreases the overall inference speed. However, in scenarios where accurate OOD covariate shift is essential, the Covari-

ateFlow model provides the best accuracy vs. speed trade-offs (see Appendix D.3).

This work primarily focuses on detecting covariate shift, with explicit covariate shifts introduced to assess performance. Many publicly available datasets exhibit both semantic and potential covariate shifts. Although the proposed approach demonstrates effectiveness in CIFAR10 vs. SVHN (Table D.23), future work should explore domain-specific datasets with limited ID covariate conditions to test the sensitivity of the proposed approach. As depicted in Figure 6.13, the scores acquired through evaluation with the CovariateFlow model and NSD metric correctly increase with each severity level, however, not at the same rate for each degradation type. Future work should explore the connection of OOD scores with image-quality metrics [293] for a comparable ranking of image degradations vs. quality. If this connection could be described, it would pave the way for unsupervised image-quality assessment.

6.5.6 Additional case: Covariate shift in X-ray

The accurate and reliable detection of out-of-distribution (OOD) data is paramount to diagnostic accuracy and the overall reliability of medical imaging systems, ultimately ensuring patient safety while offering analysis of disease. Faulty systems displaying incorrect imaging statistics can adversely affect clinical diagnostic accuracy. Additionally, it is well established that modern deep learning-based Computer-Aided Diagnosis and Computer-Aided Detection (CAD) systems are vulnerable to distribution shifts, which can lead to erroneous predictions.

In a concluding experiment to the work in this chapter on covariate shift detection, we apply the developed methods to the X-ray imaging case. By focusing on detecting faulty systems through deviations in imaging statistics, we aim to enhance the overall reliability of X-ray systems and automate various system tests. This approach not only ensures the integrity of diagnostic results, but also helps to maintain the consistency and quality of medical imaging over time.

6.5.6.A Additional experiment: X-ray dataset

Medical X-ray images, pivotal for diagnostic purposes, contain various noise sources that can influence both image quality and diagnostic accuracy [294]. We aim to develop a method to detect changes in the imaging system (faulty behaviour) that present themselves as OOD covariate shift in the images. Since we do not have access to such artifacts, we capture *real covariate shift* in the data through altering the imaging settings. If we can detect the subtle variations due to the changes in imaging settings, we hypothesize that the proposed models will be able to measure faulty OOD covariate shifts as an effect of system errors. To this end, we have acquire a new dataset of X-ray images containing a standard test object (clock) using an Azurion Image-Guided Therapy (IGT) system[6]. Images are captured in Dicom format at 12 bits/pixel for different imaging settings, en-

[6]available from Philips Healthcare, Best, The Netherlands

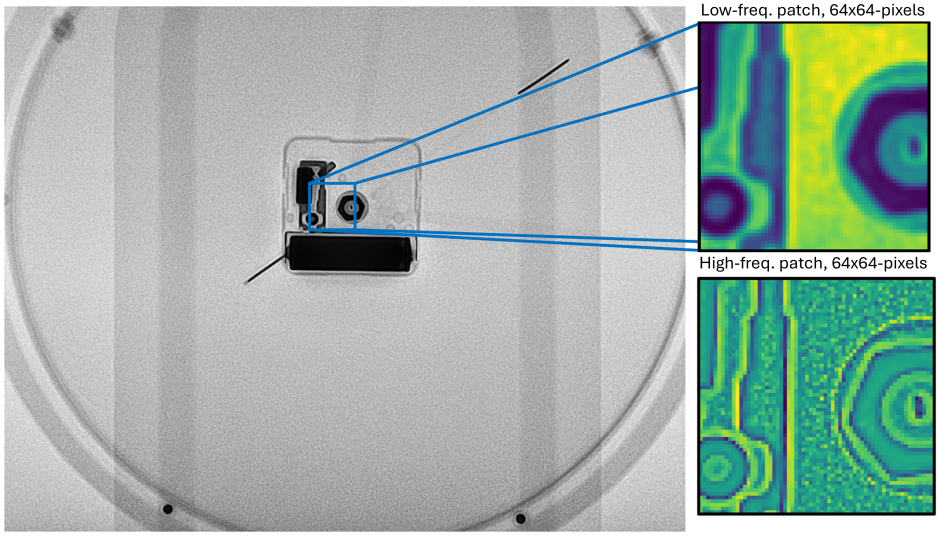


Figure 6.14 *Image from the X-ray dataset depicting the standard clock test object.*

compassing distinct dose levels using both pulsed fluoroscopy and full radiation modes. In addition, these dose variations are operated at a varying source-image distance (SID).

Figure D.9 in the supplementary material depicts images of the 6 modes that are employed for evaluation: Mode 0 (exposure with a normal dose at 110-cm SID), Mode 1 (exposure with a low dose at 110-cm SID), Mode 2 (exposure with a normal dose at 90-cm SID), Mode 3 (exposure with a low dose at 90-cm SID), Mode 4 (fluoroscopy with a normal dose at 110-cm SID), and Mode 5 (fluoroscopy with a low dose at 90-cm SID). Assuming Mode 0 for obtaining ID data, it is expected that there is progressive covariate shift from Mode 0 to Mode 5 in orders of magnitude, where Mode 5 represents data being most OOD. The full dataset consists of 18 Modes and 2 environments. For clarity, we focus on the selected few modes presented only.

Since image resizing changes the noise distribution, we adopt a large-patch-based approach (128×128-pixel patches with a 10% overlap) for dealing with images of different sizes. Furthermore, the proposed models are prepared on data from Mode 0 with approximately 100k image patches for training, 20k for validation during training and set aside, another 20k patches for testing. For all the other modes, we use the same number of patches for testing.

6.5.7 Results on covariate shift in X-ray images

The results in Table 6.3 showcase the performance of the generative OOD detection methods in the X-ray setting with varying acquisition parameters, which can influence heteroscedastic noise in the signal. In contrast to the CIFAR10 benchmark datasets, covariate shifts in the X-ray setting are intractable and usually not

X-Ray Mode 0 ID Data	X-Ray Other OOD Modes					Average
Method	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	AUROC↑ / FPR95↓
Reconstruction						
DDPM [269] (T250: LPIPS)	86.4	88.5	84.2	89.3	96.5	89.0 / 69.4
DDPM [269] (T250: LPIPS+MSE)	79.9	83.0	81.1	89.0	95.8	85.8 / 67.3
Explicit Density						
Vanilla VAE [259]	69.7	96.1	88.9	91.3	98.1	88.8 / 25.0
AVAE [267]	65.5	94.8	86.5	89.5	97.2	86.7 / 28.9
GLOW [206] (LL)	89.3	98.8	99.8	99.9	100.0	97.6 / 8.70
GLOW [258] (Typicality)	30.4	16.5	15.1	10.4	11.1	16.7 / 99.8
GLOW (NSD)	79.0	93.5	94.5	95.3	96.3	92.0 / 17.3
CovariateFlow (LL)	68.5	92.8	98.1	99.2	100.0	91.7 / 17.3
CovariateFlow (Typicality)	72.5	97.2	99.3	99.7	100.0	93.7 / 14.7
CovariateFlow (NSD)	70.0	96.5	99.4	99.7	100.0	93.1 / 15.6

Table 6.3 *AUROC scores of various methods on detecting OOD covariate shift on the X-Ray dataset.*

visible to the non-specialist. Regardless, it can be observed that almost all methods are still able to detect a shift in the high-level image statistics.

CovariateFlow exhibits robust performance in OOD covariate shift detection, demonstrating the effectiveness in strictly modeling the high-frequency distribution conditioned on low-frequency components. Considering that the X-ray dataset is semantically less involved in nature than CIFAR10 and ImageNet-200, accurate modeling should be relatively easier with high-parameter architectures such as GLOW (44 million parameters versus 1 million for CovariateFlow, as detailed in Table D.22 in the Appendix). This enables GLOW to capture the full-image distribution, including accurate high-frequency statistics, and explains its strong performance. Moreover, the covariate patterns observed are predominantly high-frequency-based, which elucidates the preference for LL evaluation. However, while the CovariateFlow model performs adequately on detecting the shift in noise through LL, it is also observed that the model accurately detects the subtle variations in image contrast through *typicality* where similar evaluations with GLOW failed. In light of these findings, LL and Typicality exhibit a bias towards different covariates. As such, an argument in favor of a more general approach can be made and thus consider evaluating the models with the proposed NSD. When assessed with the comprehensive NSD metric, CovariateFlow excels (93.1% vs. 92% with GLOW), offering an overall 1.1% performance enhancement over GLOW, especially in scenarios of pronounced OOD shifts. Remarkably, this high level of performance is achieved with a significantly smaller model size (approximately 1 million parameters), enhancing its efficiency and speed. These findings underscore the robustness of the proposed method in identifying OOD samples, particularly within the X-ray imaging contexts, and its elevated sensitivity to variations in noise conditions.

6.6 Conclusion

This chapter has explored two distinct approaches to OOD detection: semantic OOD detection using wavelet-based NFs, and covariate shift detection using the CovariateFlow model. Each experiment aims at different aspects of OOD detection, with semantic methods focusing on class differentiation and covariate methods addressing distribution shifts in imaging statistics. These methodologies offer robust frameworks to improve OOD detection across various domains, including X-ray and natural imaging.

Firstly, for semantic OOD analysis, we have introduced a wavelet-based normalizing flow (NF) approach for the semantic OOD detection of melanoma images. By integrating wavelet transforms with NFs, domain-specific knowledge is leveraged to overcome the inductive biases of traditional NFs, which tend to prioritize high-frequency features over semantic content. The proposed method demonstrates large improvements in detecting malignant melanoma images within imbalanced datasets, achieving a notable increase in the AUROC curve. This advancement not only aids in the early and accurate diagnosis of skin cancer, but has also broader implications for other medical fields facing similar data-imbalance challenges.

Secondly, we have addressed the less-explored area of covariate shift detection in images. To this end we have proposed the CovariateFlow model, a novel methodology utilizing conditional Normalizing Flows (cNFs) to model high-frequency image components, thereby effectively identifying sensory anomalies and deviations in global signal statistics. The extensive analyses on datasets such as CIFAR10 vs. CIFAR10-C (74.9% AUROC), ImageNet200 vs. ImageNet200-C (72.2% AUROC) and a newly introduced X-ray dataset validated the effectiveness of the CovariateFlow model in detecting covariate shifts. Our analysis reveals that by meticulously modeling the conditional distribution between low-frequency and high-frequency components, the CovariateFlow model outperforms existing models, particularly when employing the Normalized Score Distance (NSD) metric, which is a synthesis of log-likelihood and typicality evaluations. This work not only highlights the critical importance of addressing covariate shifts for enhancing the fidelity of imaging systems, but also underscores the potential of unsupervised generative models in improving the adaptability and robustness of machine learning models in dynamic environments where distribution changes are frequent.

In conclusion, these contributions illustrate the versatility and potential of generative models in OOD detection. By addressing both semantic and covariate shifts, the research provides a comprehensive framework for improving the robustness of machine learning systems in diverse real-world applications and thereby improving the reliability of imaging systems through detection of distribution shifts. The methodologies presented in this chapter not only advance the field of OOD detection, but also pave the way for future research aimed at further refining and expanding these techniques (e.g. extending OOD scores to image-quality metrics).

Chapter 7

Object pose estimation in X-ray imaging

7.1 Introduction

Fluoroscopy-guided minimally invasive interventions have greatly improved patient outcome from trauma, orthopedic or cancer surgeries. These image-guided surgeries largely rely on repeated acquisition of standard projections for instrument guidance and monitoring. Instrument maneuvering is typically performed manually by the clinician's hand (through trial and error) and without additional assistance, requiring multiple and extended sessions of fluoroscopy at the expense of additional radiation to the patient. Procedures are complex and due to an often very limited spatial configuration, surgical results are error-prone and highly surgeon-dependent.

Recently, various methods have been proposed to improve instrument positioning during interventional surgeries. Joint expertise in interventional radiology and image guidance has expanded the treatment options for bone surgeries such as pedicle screws placement in the thoracic and lumbosacral spine [295]. State-of-the-art (SOTA) practice for pedicle screw placement employs an intraoperative cone-beam computed tomography (CBCT) scan and combines it with an external navigation system. The intraoperative 3D augmented reality surgical navigation (ARSN) system uses external optical video cameras to augment the surgical field and assist the clinician in the navigation path for screw placement. Screw placement is then confirmed with an additional postoperative CT scan [296], [297] and manual validation. In line with previous approaches in this field [298], this comes at the expense of extensive external equipment and alters the clinical way of working, which inhibits adoption. Although these methods demonstrate progress in screw placement by indicating a path, they do not provide any guidance or validation through actual screw tracking.

Providing surgical guidance by extracting semantic information from the X-ray images alone is extremely appealing with benefits for several applications. Cardiac interventions have utilized this and improved the visualization of both catheter-based devices and soft-tissue anatomy by co-registering X-ray fluoroscopy (XRF) images with echocardiography through Transesophageal echocardiography (TEE) probe pose estimation from the X-ray image alone [299]. Screw placement surgery is another example of a complex procedure that can greatly benefit from extracting visual information available in the X-ray image for surgical assistance. Through

pose estimation via accurate 6 Degrees of Freedom (DoF) of the surgical instruments from a single X-ray image, additional guidance to clinicians is provided during image-guided procedures and instrument placement is determined without the need for additional external navigation systems or postoperative CT scans.

Motivated by the need for automated robot operation, autonomous driving and VR & AR applications, methods for accurate 6-DoF pose estimation of rigid objects have extensively been studied [300]. While most existing methods assume a fixed image acquisition geometry, which is sufficient for many applications, some domains, such as X-ray imaging or space satellite pose estimation, require the imaging geometry to constantly change during its operation. Adjustment of the focal length (zooming) or the detector field of view (X-ray image size and dose control) are common changes in such a framework. Naturally, it is also evident that pose estimation methods should include the intrinsic camera parameters if the methods will be used across different cameras, otherwise the designer faces manual adjustment for each camera (re-training and data collection in case of learning-based methods).

Therefore, in these domains it is crucial for pose estimation methods to incorporate the changing imaging geometry to accurately recover the pose of the target object(s). Perspective-n-Point (PnP) deep learning-based methods, which use the intrinsic camera parameters to estimate the object's 6-DoF pose, could be readily applied to these domains. However, the accuracy of these approaches, as originally proposed by Tekin *et al*. [301], is limited by the YOLOv2 architecture's inability to accurately regress 2D image locations of the projected vertices of the object's 3D bounding box. Other two-staged methods such as EPro-PnP [302], employ an initial object detection method followed the final object pose estimation, making them computationally less efficient. Recent advancements in the YOLO object detection series suggest that 6-DoF pose estimation can benefit from these improvements to efficiently achieve high pose accuracy under variable acquisition geometry. The above discussion and highlighted limitations culminate in the following research questions.

- Acquiring accurate instrument pose data in X-ray settings is challenging, with prior methods often relying on simulations. *How can a general-purpose method be established to accurately and efficiently acquire data for 6-DoF pose estimation in X-ray imaging?*
- Existing 6-DoF pose estimation methods are often too slow or inaccurate for precise guidance in the medical domain. *What advancements can be made to develop a general-purpose method that is both accurate and fast for 6-DoF pose estimation?*
- Medical procedures in the image-guided therapy setting require real-time adjustment of image acquisition parameters, presenting a challenge for 6-DoF pose estimation methods. *How can X-ray imaging geometry be effectively incorporated into the 6-DoF pose estimation process, and what impact does this have on the accuracy and performance of the model?*

As a solution direction, the object pose is acquired through predicting 2D keypoints for the instrument's virtual 3D bounding box and resolving the pose through a Perspective-n-Point (PnP) algorithm [80] under consideration of the acquisition geometry. This attribute enables the transition to the X-ray domain, where the acquisition geometry is constantly changing during a procedure and across systems. Additionally, we address generalization from our training domain to a clinically relevant setting through a series of extensive data augmentations. The proposed method shows robustness and high accuracy for 6-DoF pose estimation of a surgical screw in a variable intraoperative setting.

In this work, we propose a (medical) instrument pose-estimation method that is general-purpose and is addressing technical challenges that have limited such technology from being incorporated in practice. We further extend our prior work [61] and introduce a novel YOLOv5-6D pose modeling architecture for more accurate and fast object 6-DoF pose estimation. In addition, to address the difficulty in acquiring data, a data collection method is introduced for automatic data labeling that generalizes across all cone-beam X-ray geometries and object types.

This chapter is organized as follows. Section 7.2 discusses the related work on object 6-DoF pose estimation for both the color and X-ray domain. Section 7.3 introduces the proposed approach to X-ray-based object pose estimation. The results of these experiments are presented in Section 7.4. Finally, a discussion on the obtained results and possible future directions are included in Section 7.5 and 7.6.

7.2 Related work

Recent advancements in deep learning have improved the accuracy at which systems can estimate the position (3-DoF) and orientation (3-DoF) of rigid objects. This progress is largely driven by applications for the metaverse, VR & AR, robot operation and intelligent driving. Zhu *et al.* [300] provide an extensive review of methods for 6-DoF pose estimation. Extending this review, we briefly consider related work in object pose estimation in the RGB and X-ray domain. We omit a detailed discussion of methods dependent on depth information (RGB-D), such as RCVPose [303] and PVN3D [304], as well as RGB-D-based, model-free methods like FS6D [305] and the more recent FoundationPose [306], since this depth modality is unavailable in our X-ray setting.

7.2.1 6-DoF pose estimation in RGB

The majority of research efforts in object 6-DoF pose estimation determine the object pose from RGB images with knowledge of the object of interest. These methods commonly utilize the object 3D models followed by task-specific model training. More recently there has also been growing interest in generalizable model-free methods (hereafter refereed to as model-free methods), that do not require additional training to predict the pose of novel 3D objects [307]–[309].

While this does alleviate large constraints on employing the method, they do still fall behind in terms of both speed and accuracy (Table 7.8). In many applications, with the object 3D model often obtainable, accuracy and speed is required over ease of implementation.

The state-of-the-art methods employing object-specific knowledge during training can roughly be categorized as methods that (1) directly regress object pose from the color image (referred to as *direct* methods), (2) employ a PnP algorithm to compute the object pose from 2D predicted keypoints of a corresponding 3D model (referred to as *PnP* methods) and (3) either (1) or (2) followed by an iterative refinement procedure.

Direct pose estimation involves directly regressing the object pose with, typically, a deep convolutional neural network (CNN) from the RGB image in an end-to-end fashion. Bukschat *et al.* proposed EfficientPose [310] that employs the EfficientNet [311] backbone and a BiFPN-net [312], to regress the object pose from RGB images at different scales. EfficientPose regresses the pose of single objects from RGB images in the LINEMOD benchmark at 36.43 ms/image (27.45 Frames Per Second (FPS)) and an average 97.35% ADD(-S) accuracy at $0.1 \cdot d$[1]. Methods in this category do not explicitly consider the camera acquisition geometry and these parameters are thus considered static and fixed per camera/model pair. Recently, Xu *et al.* have developed RNNPose [313] that starts with an initial pose from any method (tested with a direct [314] and PnP method [315]) and iteratively refines the object pose, based on the estimated correspondence field between the reference (2D render of 3D model) and target images. This iterative re-projection strategy considers the intrinsic camera parameters, but comes at the cost of increased computation time. This method currently achieves the highest accuracy on the public LINEMOD benchmark at 97.37% ADD(-S), but with an inference time (4 rendering cycles and 4 recurrent iterations each as per the paper) of 308.35 ms/image (3.24 FPS), excluding the initial pose prediction step.

Tekin *et al.* have proposed a 2D-3D correspondence-based method (SingleShotPose [301] also known as YOLO-6D in Figure A.1 in Appendix A) for 6-DoF pose estimation. The model simultaneously performs a single-shot object detection and the 6D pose prediction from an RGB image. This is realized by predicting the 2D image locations of the projected vertices of the object's 3D bounding box. Using a Perspective-n-Point (PnP) algorithm and known acquisition parameters, the 6D pose of an object can be estimated (we mention the relationship here, but will discuss it in detail in Section 7.3). As a feature extraction network, the model is using the Darknet19-448 backbone, first proposed in YOLOv2 [316] for object detection. Since its release, there have been considerable improvements in YOLO object detection series [317]–[320]. We leverage these advances in the development of the YOLOv5-6D pose estimation model.

[1]This accuracy is related to 10% of the object diameter d. Later in this chapter, we provide a detailed explanation and more versions of the metric. If the distance is not specified, it is assumed to be 10%.

Prior to our work, other PnP-based 6-DoF pose estimation methods have been developed [301], [315], with the best-performing method being EPro-PnP [302], proposed by Chen *et al.*. This two-stage approach achieves a high 96.36% ADD(-S) accuracy, enabled by the dense correspondences extracted form the object image crop and the proposed differentiable PnP layer. The pose estimation step is also computationally efficient (see Section 7.4.2). While the pose estimation step and the PnP layer can be integrated with any architecture, their work employs CDPN [321], a dense correspondence network for 6-DoF pose estimation from object-specific image crops. An initial method is thus required for object detection and cropping on the target images which typically consumes majority of the compute budget. Section 7.4.2 provides an in-depth run-time analysis of these methods.

7.2.2 Object pose estimation in X-ray

Methods for 6-DoF pose estimation in the X-ray domain have been proposed for applications ranging from industrial product inspection, C-arm repositioning for surgical assistance, to surgical tool pose estimation. Presenti *et al.* propose a series of methods [322]–[324] to recover manufactured object pose from X-ray images for defect inspection. Their approach assumes fixed acquisition geometry and displays sub-optimal results when only one image is used [324], compared to methods employing PnP. Similarly, X-Ray-PoseNet [325] has been proposed by Bui *et al.* to directly regress the the translation (3 degrees) and rotation (4 quaternions) of industrial objects with respect to the X-ray system. Their approach is based on a custom CNN architecture and assumes fixed X-ray acquisition geometry, while being trained on simulated X-ray images. Kausch *et al.* [326] developed a C-arm re-positioning pipeline to suggest C-arm imaging angles for assistance during spinal implant placement. It uses the patient spine as reference and suggests a new C-arm position through a series of features extracted from the X-ray image, using multiple U-Net-like models.

Despite the interest in surgical tool guidance, few attempts have been made to directly recover the pose of the instrument used during the treatment. Registration between X-ray fluoroscopy (XRF) and transesophageal echocardiography (TEE) for structural heart interventions relies on accurate pose-estimation of the TEE probe. TEE-probe pose estimation through 2D/3D registration methods based on iterative refinement such as Direct Splat Correlation (DSC) and Patch Gradient Correlation (PCG) have been implemented[299]. In contrast, instrument pose estimation from 3D ultrasound data volumes has received substantially more attention [327], [328].

In one particular case, Kügler *et al.* developed i3PosNet [329], a method for surgical instruments pose estimation using a VGG [330]-based CNN architecture. The network predicts object-specific keypoints from localized patches. While considering the geometric landmarks of fiducials (virtual keypoints) during pose estimation, the method does not account for the image acquisition geometry, limiting its application across different systems and geometries. i3PosNet is designed for pose estimation of symmetrical objects and lacks effectiveness for asymmetrical instru-

ments, as it only estimates a 5-DoF pose. This method is developed and trained on simulated data and finally tested on manually annotated real X-ray images, which introduces a time-consuming setup and potential human errors. Finally, the multi-stage approach employed, including image variety reduction, image information extraction followed by pose reconstruction from pseudo-landmarks, hinders its real-time applicability.

7.2.3 Approaches for 6-DoF pose estimation data acquisition

The LINEMOD dataset [331] is the most commonly used dataset for 6-DoF object pose estimation in the RGB(-D) domain. Images of the 15 objects are collected in sequence under different illumination and large viewpoint changes in a heavily cluttered environment with mild occlusions. The ground-truth poses (labels) are obtained using calibrated cameras and a calibration pattern.

Unfortunately, labeled intraoperative X-ray training data for object 6-DoF pose estimation has neither been described nor published. All of the above-mentioned methods rely on simulated training data that require highly accurate simulations and extensive CAD modeling. The methods then train on these simulations, aiming to generalize to the test domain. This domain gap introduces a challenge when transferring to real-world applications. Kügler *et al.* [329] acquire real data of object poses through tedious manual annotation effort, which involves projecting their object as an outline on the X-ray image and then interactively translating and rotating the object to match the X-ray image. The applied data in their approach are also captured using fixed X-ray acquisition geometry.

Summarizing the outcome of this detailed review, we position the proposed work as a general-purpose 2D/3D correspondence method for instrument pose estimation from a single X-ray image. Building on our prior work [61], the method takes the X-ray acquisition geometry into account, enabling it to generalize to new systems. This generalization will be discussed in Section 7.3.1. To address the difficulty in data acquisition, we present a general method for capturing real X-ray data of any object.

7.3 Approach

7.3.1 X-ray pose estimation

X-ray imaging systems are available in a range of different sizes with varying detector shapes, depending on the needs of the application. In addition, modern X-ray systems allow for the acquisition geometry to change at run-time to improve image quality of the area of interest. In brief, this results in varying acquisition parameters such as the detector size, detector field of view (FOV) and most commonly, the source-image distance (SID). All of these variables have a direct effect on the resulting X-ray image. Computer-aided image-guided methods influenced by these changes need to request fixed acquisition parameters, or incorporate their variation in order to present accurate results. While several methods requesting

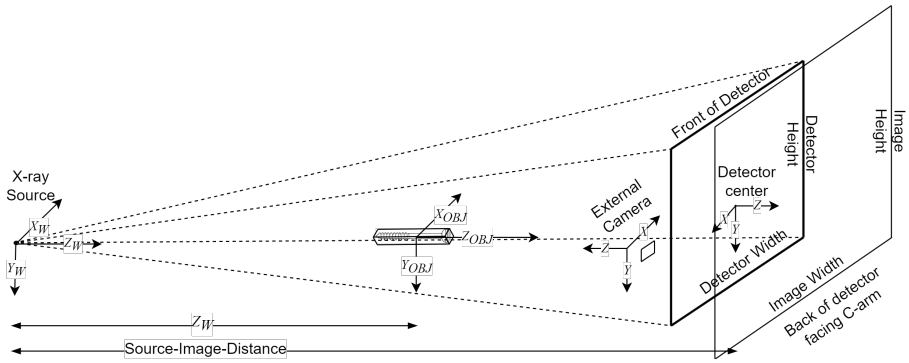


Figure 7.1 *X-ray projection model depicting the X-ray source, a surgical screw, detector with an attached grayscale optical camera, the detector panel and the captured X-ray image. The frame of reference for each point of interest is also depicted.*

fixed acquisition geometry have been adopted, they have limited applicability or require extensive additional preparation effort for each new system. Object pose estimation is fundamentally connected to the image acquisition parameters and, as such, we incorporate them in the proposed method to allow a single trained model to generalize to a wide range of acquisition geometries.

7.3.2 Data acquisition setup

Acquiring labeled data for 6-DoF pose estimation tasks is difficult, due to the inherent limitation of human observers to accurately determine an object's 6-DoF pose. When possible, manual labeling, even in the case of projected keypoints, is prone to errors and extremely laborious. Therefore, we draw inspiration from data collection methods in the optical domain [331], [332] and devise a setup for accurate and automatic data acquisition and labeling for 6-DoF pose estimation in X-ray without corrupting (or introducing a learnable bias to) the X-ray image with external markers.

In our data acquisition setup, we attach an external optical camera to the X-ray detector. The optical cues from the camera that are transparent to the X-ray, can be utilized to assist in the pose estimation task. The complete method consists of (1) a ChArUco board [333] with (2) the object of interest at a known location on the board, (3) the optical camera capturing images of the board, whilst (4) the X-ray system captures X-ray images of the object. The 2D projection of the object's 3D bounding box onto the X-ray detector can then be acquired through the optical pose estimation of the ChArUco board and a series of frame transformations to the X-ray source coordinate system.

The proposed concept allows for fully automated data acquisition through automated movement of the X-ray C-arm and patient table to a diverse set of positions. In contrast to previous work, the method does not rely on accurate rotation or translation sensors from the X-ray system, so that it can be used across a wider range of X-ray systems and still recover accurate data labels. In addition,

the labeled images are void of any external cues that can be utilized to determine the object pose.

In this work, we employ OpenCV [334] for acquiring the pose of the board in the optical camera frame. This is achieved through the detection of the ChArUco markers in the optical image and combined with the knowledge of their physical flat-panel location on the printed ChArUco board on the patient table. Provided with the set of 2D-3D correspondences, the camera pose in the table coordinate system can be obtained by solving the PnP problem.

7.3.3 X-ray acquisition model & calibration

This research adopts the pinhole X-ray acquisition model to traverse between the 3D object frame and the 2D X-ray projection. The model is used during data acquisition and for the PnP pose calculation. Figure 7.1 depicts the X-ray acquisition model and Equation (7.1) formalizes it. The pinhole camera model is formally specified by

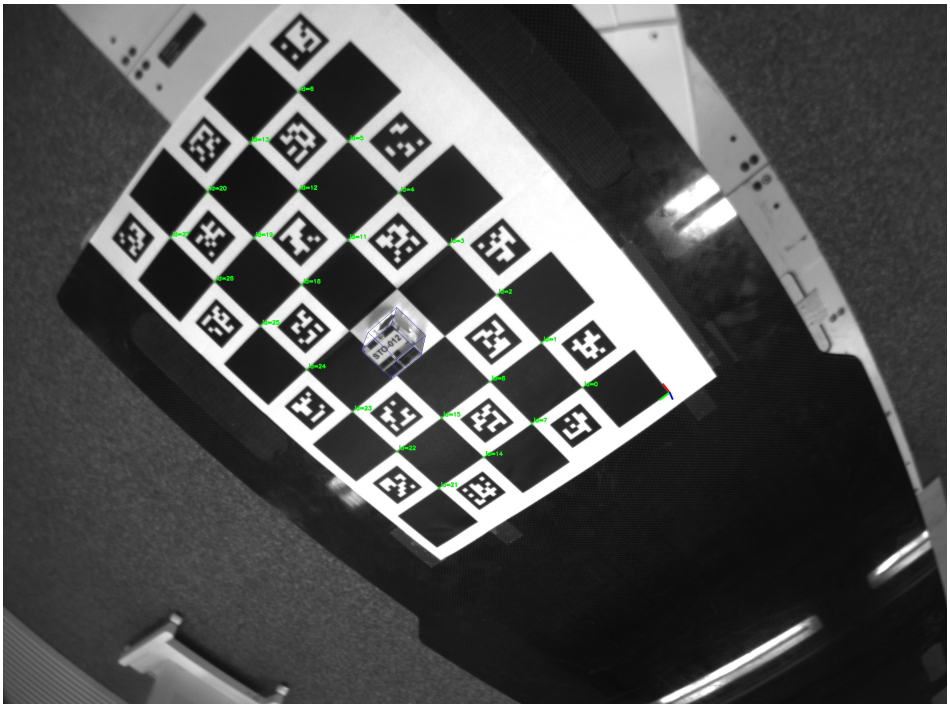
$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{K} & \mathbf{O}_3 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{C} \\ \mathbf{O}_3^T & 1 \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix}, \tag{7.1}$$

where the intrinsic parameter matrix K is defined as

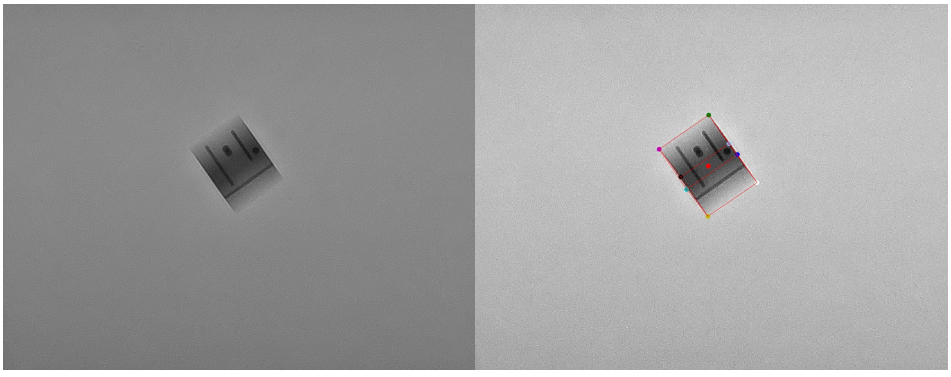
$$\mathbf{K} = \begin{bmatrix} k_u f & 0 & k_u x_0 \\ 0 & -k_v f & k_v y_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{7.2}$$

In the above expressions, matrix $\mathbf{K}$ represents the intrinsic system parameters, which can change during the clinical operation and across different systems. These intrinsic parameters consist of the horizontal (k_u) and vertical (k_v) density of pixels. The pixel density can change depending on the detector and image size combination, e.g. changing the field of view or zooming on an image. The offset of the principal point to the detector center is described by coordinates (x_0, y_0). The source-image distance (SID), or focal length, is represented by parameter f. The extrinsic parameters $\mathbf{R}$ and $\mathbf{C}$ are the rotation and translation matrices to be solved.

The intrinsic parameters of the optical cameras are calibrated using a flat plate with a pattern of circles. Optical images of the plate are captured from different angles and the camera parameters are adjusted by minimizing the reprojection error of the circles. The resulting optical coordinate system is subsequently linked by capturing both optical and X-ray images of a dome-shaped calibration object, consisting of white plastic cylinders embedded in black foam. On the optical images, only the circular sides of the cylinders are visible, of which the center is

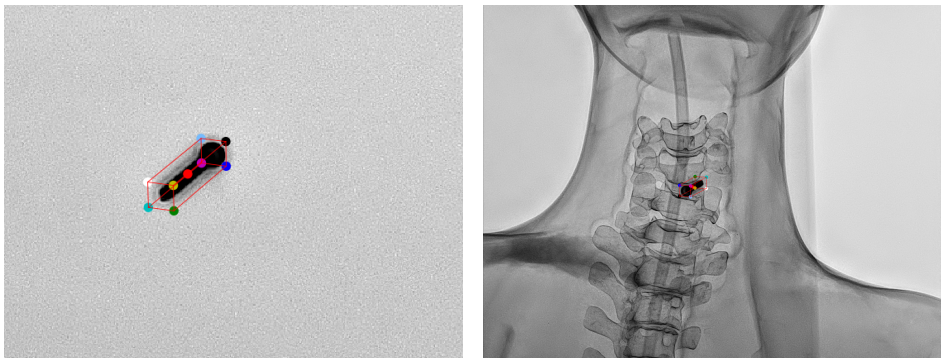


(a) *Grayscale image of the ChArUco board and test cube on the patient table taken from the optical camera on the detector. The 2D projection of 3D cube outline can be seen in blue.*



(b) *Dicom cube image* (c) *Projected 3D bounding box*

Figure 7.2 *(a) Grayscale image showcasing the setup used to automatically acquire the 6-DoF pose of various objects. (b) Corresponding X-ray Dicom image of the cube. (c) Projected 3D bounding box and virtual corner coordinates.*



(a) *Zoomed X-ray image of the surgical screw used for training and validation.* (b) *X-ray image of the surgical screw with spine phantom used as test set.*

Figure 7.3 *Examples from the applied screw train, validation and test datasets. Each image also showcases the projected 3D bounding box of the screw.*

computed. The X-ray images are used to create a 3D reconstruction, on which the cylinders are segmented and the same points are computed, as observed in the optical images. The two point clouds are matched, thereby linking the optical and X-ray coordinate systems.

7.3.4 Datasets

7.3.4.A LINEMOD

In line with previous work on 6-DoF pose estimation, we also evaluate the proposed YOLOv5-6D architecture on the popular LINEMOD dataset. The dataset consists of 13 different objects, each with approximately 1,200 images that are placed in various scenes. In this benchmark there is a predefined division scheme for training and test images which we also adopt. The training set varies between 15-30% of the object's dataset, making it a very small fraction of the dataset. This aspect makes it particularly interesting and relevant for the medical domain, in the sense that approaches for this benchmark need to learn and generalize from scarce data.

7.3.4.B Cube

Using our data collection method described in Section 7.3.2, we have composed a dataset, henceforth referred to as the cube dataset, to test the adaptation of the proposed approach and the YOLOv5-6D network to the X-ray domain. Figure 7.2a depicts the grayscale image of the 30×30×30-mm perspex cube, embedded with metal markers placed on the ChArUco board. Since the 3D bounding box exactly matches that of the cube's physical dimensions, the cube is the ideal test object because one can visually determine the accuracy of the bounding box fit, whereas other objects might have a virtual 3D bounding box. Figures 7.2b and 7.2c depict the X-ray image of the cube and the corresponding 2D projection of the 3D bounding box. Along with the DICOM X-ray image and the 2D projected

coordinates, we also capture the original 6-DoF cube pose and a binary mask of the cube in the X-ray image for training purposes. Table 7.1 lists the X-ray system's acquisition parameters and geometry used to capture the Cube dataset, which is repeated for every side of the cube. In line with prior work [335], X-ray/optical image pairs are taken in 10-degree intervals across the geometrical rotation range of the X-ray system, to ensure a uniform viewing distribution of the object. The SID, translation and FOV parameters are uniformly sampled from the allowed range and automatic gain control manages the applied X-ray dose at a constant K rate of 1.88 mGy/min. In total, we have acquired 1000 images ($r_x \in [-45°, -35°, -25°, ..., +45°]$ with r_y and r_z being in the same range) per cube side at a 960×742 image resolution.

Table 7.1 *C-arm acquisition and table parameters (w.r.t. its starting position) used during the data collection.*

Rotation (degrees °)	Translation (mm)
$r_z \in [-45°, 45°]$	$t_z = 700 \pm 40$
$r_y \in [-45°, 45°]$	$t_y = 0 \pm 40$
$r_x \in [-45°, 45°]$	$t_x = 0 \pm 40$
SID (mm)	FOV (mm) diagonal
$[950.0, 1230.0]$	$[156, 484]$

7.3.4.C Screws

To demonstrate its clinical potential, we also evaluate the proposed approach for 6-DoF pose estimation of surgical screws for potential spine surgeries. The screw is a standard 3.5-mm cannulated cancellous screw, often used during orthopedic surgeries. The screw is 34.3 mm long and has a head with a diameter of 6.88 mm. We have created a 3D model of the screw to be used during the projection onto the grayscale and X-ray image. The screw is inserted into a polystyrene block to enable precise placement on the ChArUco board. The same data collection method as described in Section 7.3.2 has been followed to construct the so-called Screw dataset (Figure 7.3a) for training and validation of the work. In addition to this dataset, we have also constructed a Screw test dataset. The Screw test dataset is set up to test the generalization of the proposed approach to a more realistic setting. We have attached the surgical screws to the spine of a human phantom, similar to their usage during a spine surgery. An example image of the screw and spine phantom can be seen in Figure 7.3b. While this setting is still different from an actual clinical intervention, it does enable to determine whether the pose estimation method can generalize to a more complex domain. The Screw dataset and screw with human phantom dataset each contain 1000 images, acquired following the parameters listed in Table 7.1 and as further specified in Section 7.3.9.

7.3.5 YOLOv5-6D Pose

This research largely draws inspiration from the YOLO6D model [301] for object pose estimation and enhances it by incorporating recent advancements in the YOLO object detection series [320]. As such, this single-shot approach enables simultaneous detection and 6-DoF pose estimation of objects in RGB and X-ray images. The model predicts the 2D image locations of the projected vertices of the 3D bounding box of the object. Using these 2D/3D correspondences and the current acquisition parameters, the 6D pose of the object is then solved using a PnP algorithm [80], in our case specifically ePnP [336]. Figure 7.4 depicts the YOLOv5-6D model architecture. The model follows a standard backbone, neck and head architecture. The backbone is based on the CSP-Net [337], first proposed in the work by Wang *et al.* for improved object detection. For the model neck, the BiFPN [312] introduces a top-down pathway to fuse multi-scale features with an additional bottom-up pathway. The complete architecture, as depicted in Figure 7.4, can be subdivided into different building blocks at a level of processing stages, indicated by different colors. These stages consist of (1) ConvBNSilU - convolution, batch normalization, Silu activation, (2) BottleNeck 1 - Two ConvBNSilU operations followed by a residual connection to the input, (3) BottleNeck 2 - Two ConvBNSilU operations, (4) C3 - ConvBNSilUs and a BottleNeck block (BT1 or BT2) (5) SPFF - represents a pyramid structure through max-pooling operations, and (6) Conv - convolution.

We adjust the model head for keypoint prediction at different scales (three in our experiments). More precisely, three scales produce $W \times H$ grid cells and n_a anchor boxes (also three in our experiments) responsible for detecting the objects. Given the LINEMOD input images of size 640×480 pixels, the network produces 18,900 predictions (80×60×3 + 40×30×3 + 20×15×3). Every cell and anchor-box combination predicts output tensor $\mathbf{T_o}$, which is the 2D location of the object center and 8 corners of the projected 3D bounding boxes in the image. More formally, $\mathbf{T_o} = (b_{x0}, b_{y0}), 8 \times (b_x, b_y), conf, n_{class}$, where (b_{x0}, b_{y0}) are the object center coordinates, (b_x, b_y), the projected 3D bounding-box coordinates, $conf$ the cell confidence of it containing the object and n_{class}, the class-specific confidence. Hence, the model output comprises 19 predicted values, as we only capture one class. Additionally, we apply a scaled sigmoid function specified by

$$f(\cdot) = (2(\sigma(\cdot)) - 0.5) + c_{\text{offset}}), \tag{7.3}$$

to the object-center coordinates prediction for easier predictions when the object center is close to the edge of a grid cell compared to the original single sigmoid function. Finally, the prediction with the highest cell-specific object confidence is chosen for evaluation. In Equation (7.3), σ is the sigmoid activation function and c_{offset} is the offset to the top-left corner of the particular grid cell.

In contrast to YOLO6D, our model incorporates a more advanced feature extraction backbone, utilizing CSP-Net over Darknet 19-448, and integrates an additional 'neck' network, BiFPN. This enhancement enables the feature extrac-

tion across multiple scales, as opposed to YOLO6D's single-scale approach. The features from these different scales are rasterized into the 18,900 cell predictions, compared to 845 cells in YOLO6D. This enables the network to make accurate predictions for much smaller and larger objects. These architectural improvements along with further refinements of the training objective leads to a significant accuracy increase at the cost of a minor speed decrease, as shown in Table 7.8.

7.3.6 Training objective at different scales

We introduce various technical improvements to enable efficient model training with the new architecture. The confidence function proposed by Tekin *et al*. [301] is adjusted to support the multi-scale model and variable input image dimensions. Most notably, we change the distance threshold used in the confidence function in Equation (7.4), based on the output layer grid size instead of a fixed 2D Euclidean distance. The confidence function, $c(\mathbf{x}, W, H)$, dynamically determines a grid cell's object confidence value for the current predicted 2D points ($\mathbf{x}$) based on its distance $D_T(\mathbf{x})$ from the target 2D points. Since the grid-space size changes at different output layers determined by the image resolution and aspect ratio, the confidence function is adjusted accordingly. This function can be formalized by

$$c(\mathbf{x}, W, H) = \begin{cases} e^{\alpha(1-\frac{D(\mathbf{x})}{d_T(W,H)})}, & \text{if } D(\mathbf{x}) < d_T(W,H) \\ 0, & \text{otherwise,} \end{cases} \tag{7.4}$$

where $d_T(W, H) = \beta\sqrt{W^2 + H^2}$, the diagonal of the grid and β a hyperparameter set to an empirically determined value of 0.2. The sharpness of the exponential function is determined by the hyperparameter α and $D(\mathbf{x})$ is the L_1 distance between the predicted point and the ground-truth point in grid-space coordinates. The complete loss function consists of $\mathcal{L} = \lambda_{\text{points}} L_{\text{points}} + \lambda_{\text{conf}} L_{\text{conf}}$, where λ_{points} and λ_{conf} are scaling hyperparameters to control the influence of the loss between points and the confidence loss, respectively.

The process of target (object) prediction in the proposed model involves a critical step of matching each target with the most suitable anchor, ensuring a close match between the widths of the target (determined by object-specific keypoints that are farthest apart –after augmentation– in the vertical and horizontal direction) and the anchor to determine the optimal scale for prediction. Following this, the model identifies the specific grid cell responsible for the prediction, based on the target location. The primary cell for prediction is the one containing the target center point, but adjacent cells may also participate, depending on the target position within the cell. During training, grid cells are trained to predict targets in various positions. This process equips each grid cell with the ability to make accurate predictions for a range of target positions, thereby ensuring robustness and versatility in detecting different types of targets across various locations.

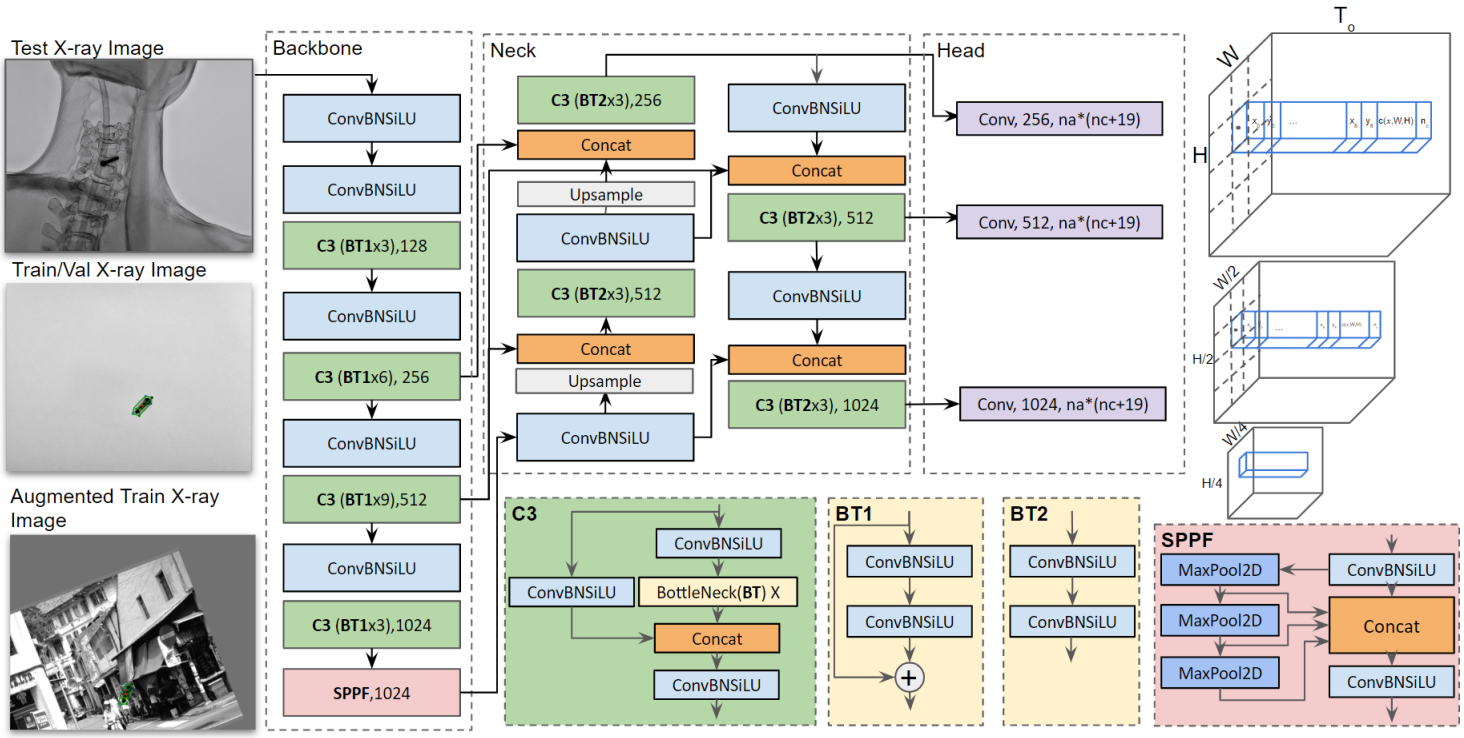
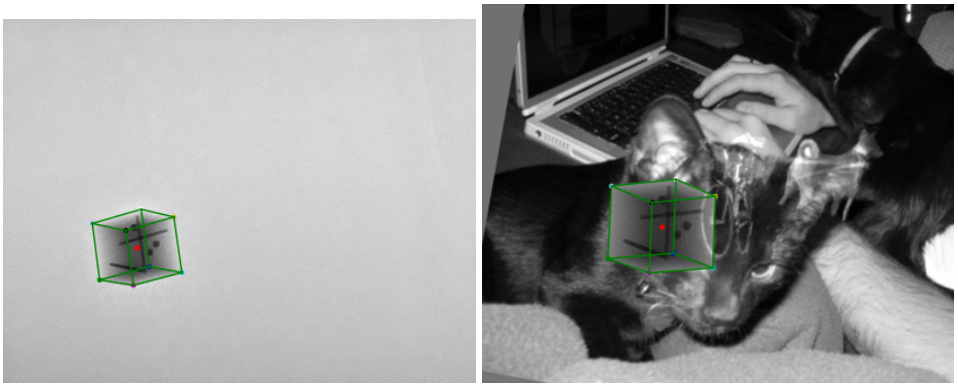


Figure 7.4 *Overview of the YOLOv5-6D architecture. The model backbone is based on the CSP-Net, the neck consists of the BiFPN architecture and the newly proposed model head predicts object keypoints at different scales. The model input is an image containing the object of interest and predicts object-specific keypoints that are used to estimate the object pose. The subblocks C3, BT1, BT2 and SPPF are depicted at the bottom in an enlarged view for further detail.*

7.3.7 Data augmentation and training details

A small dataset of the object of interest is collected and through extensive data augmentation, we establish the YOLOv5-6D model for 6-DoF pose estimation to generalize the model to a new, unseen and more complex domain. A series of data augmentation techniques are employed and refined for accurate keypoint detection in both the RGB and X-ray domain. For the X-ray data loading, all data is processed as a one-channel image, compared to the normal three-channel RGB domain. The proposed augmentation process consists of: (1) replacing the image background with a random image from the PASCAL VOC dataset [338] using the object mask, (2) color-space HSV augmentation and contrast, brightness and noise adjustment for the grayscale images, (3) scaling (30%), zooming ($\pm$ 30%), translation (30%), rotation ($\pm 180°$), sheering augmentation (2°) and finally, (4) an image overlay and occlusion strategy to randomly occlude (RGB images), or reduce the intensity (X-ray domain) of an area about or on top of the object of interest (X-ray "occlusion"). Many of the occlusion augmentations are adapted from the work by Sárándi *et al.* [339]. Figure 7.5 depicts this augmentation applied to an image from the cube dataset. The Cube and Screw datasets are randomly split into 70%/30% train/validation splits. YOLOv5-6D is branched from the



(a) *X-ray cube image before any augmentation is applied.*

(b) *The same image after augmentation.*

Figure 7.5 *Example image from the applied cube datasets from before and after augmentation for training. Notably, a bicycle partially "occludes" the cube in the center of the augmented image, the background is changed and the image is scaled.*

YOLOv5 repository [320] and adjusted for 6-DoF pose estimation, instead of object detection. We exploit many of the training techniques in line with those used in object detection. The model is trained with an ADAM optimizer with a warm-up and cosine learning-rate scheduler. An L_1 loss is employed for keypoints and a Cross-Entropy loss for the objectiveness confidence. Model weights are initialized with the COCO-pretrained weights [320], [340] where possible. The

above implementation is in PyTorch 1.7.0 and is shared for reproducibility [2]. The models are trained on two RTX 3090Ti GPUs and all of the performance tests are carried out on a system with a more readily available RTX 2080Ti and an i9-9900KF CPU, operating at 3.60 GHz for comparison.

7.3.8 Evaluation criteria

We adopt the evaluation metrics from prior work on 6-DoF pose estimation. The commonly used 3D distance of model vertices are employed, often referred to as the average distance difference (ADD) and ADD-S (symmetric objects) metric [335], [341], [342], as the main method of evaluation, while also providing further insight into model performance through the 2D reprojection error, average angle error and the translation error. The ADD metric can be equated as

$$\mathbf{ADD} = \frac{1}{|\mathcal{M}|} \sum_{x \in \mathcal{M}} \|(\mathbf{R}x + \mathbf{t}) - (\bar{\mathbf{R}}x + \bar{\mathbf{t}})\|_2, \tag{7.5}$$

and computes the average 3D distances between a set $\mathcal{M}$ of 3D points (the 3D model vertices) brought about the ground-truth rotation ($\mathbf{R}$) and translation ($\mathbf{t}$) and the predicted rotation ($\bar{\mathbf{R}}$) and predicted translation ($\bar{\mathbf{t}}$). Averaging is done over the cardinality of $\mathcal{M}$. For symmetrical objects, we use the ADD-S metric defined as

$$\mathbf{ADD\text{-}S} = \frac{1}{|\mathcal{M}|} \sum_{x_1 \in \mathcal{M}} \min_{x_2 \in \mathcal{M}} \|(\mathbf{R}x_1 + \mathbf{t}) - (\bar{\mathbf{R}}x_2 + \bar{\mathbf{t}})\|_2, \tag{7.6}$$

capturing the smallest distance of the possible 3D distances. The 3D distance is converted into a binary metric, based on a maximum object diameter threshold of 10%, 5% and 2%. In Section 7.4.2, we also extensively measure and report model inference time.

7.3.9 Clinical context

During X-ray acquisition, attenuations along the beam direction are accumulated and depth information is lost, potentially yielding ambiguous overlays of structures depending on the viewing direction. This is especially important when attempting to recover the pose of an instrument of interest. Consequently, we consider the deployment conditions and working positions of the X-ray system to determine if viewing angles can be constrained to avoid ambiguities, or if the ambiguous images even have an impact on the object pose whatsoever (as is the case with symmetrical objects). With our application of spinal screw-placement surgeries in mind, the patient is typically in a prone position with the clinician performing the spinal surgery with a superior approach (from above the patient). Any screws being placed will be attached with the screw head upwards. This

[2] Code publicly available at: `https://github.com/cviviers/YOLOv5-6D-Pose`

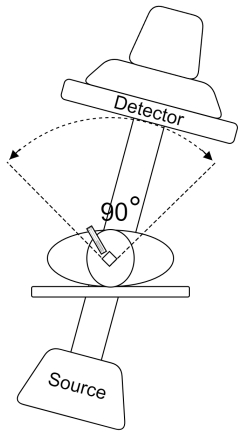


Figure 7.6 *Typical C-arm working positions during spinal screw-placement surgeries [296].*

natural working condition can be exploited and ambiguities are directly avoided by limiting the viewing angles of the screw to be from above the patient as in Figure 7.6. Although the rotational symmetry around the screw z-axis still remains, the physician will always check initial mounting to the correct vertebrae and will be concerned only about 5 degrees of freedom, explicitly the translation in x, y, z directions (connecting point to the bone) and the orientation about the x-axis and the y-axis (tilting angles), because no z-rotation is used. Computer-aided pose estimation methods can in turn also be conditioned to these viewing angles, by strictly acquiring training data from the expected working positions. We employ this conditioning in the Screw datasets by only using images captured with a rotation in range of $r_x \in [-45°, +45°]$, $r_y \in [-45°, +45°]$, $r_z \in [-180°, +180°]$ from the starting position of the X-ray system. Finally, in practice, a projection of the object 3D model will be rendered instead of the bounding box along with strictly relevant transformation axes to further reduce ambiguities and present clinically-relevant information.

7.4 Results

This section presents the results of the proposed approach on the various datasets used during the development of a method for accurate 6-DoF pose estimation in X-ray.

7.4.1 Object pose estimation in RGB images

The quantitative results of the accuracy of the experiments on the LINEMOD dataset are presented in Table 7.3 and Table 7.2, while the qualitative results are shown in Figure 7.7. We compare the proposed approach against seven competitive object pose estimation methods on the LINEMOD dataset: YOLO6D [301],

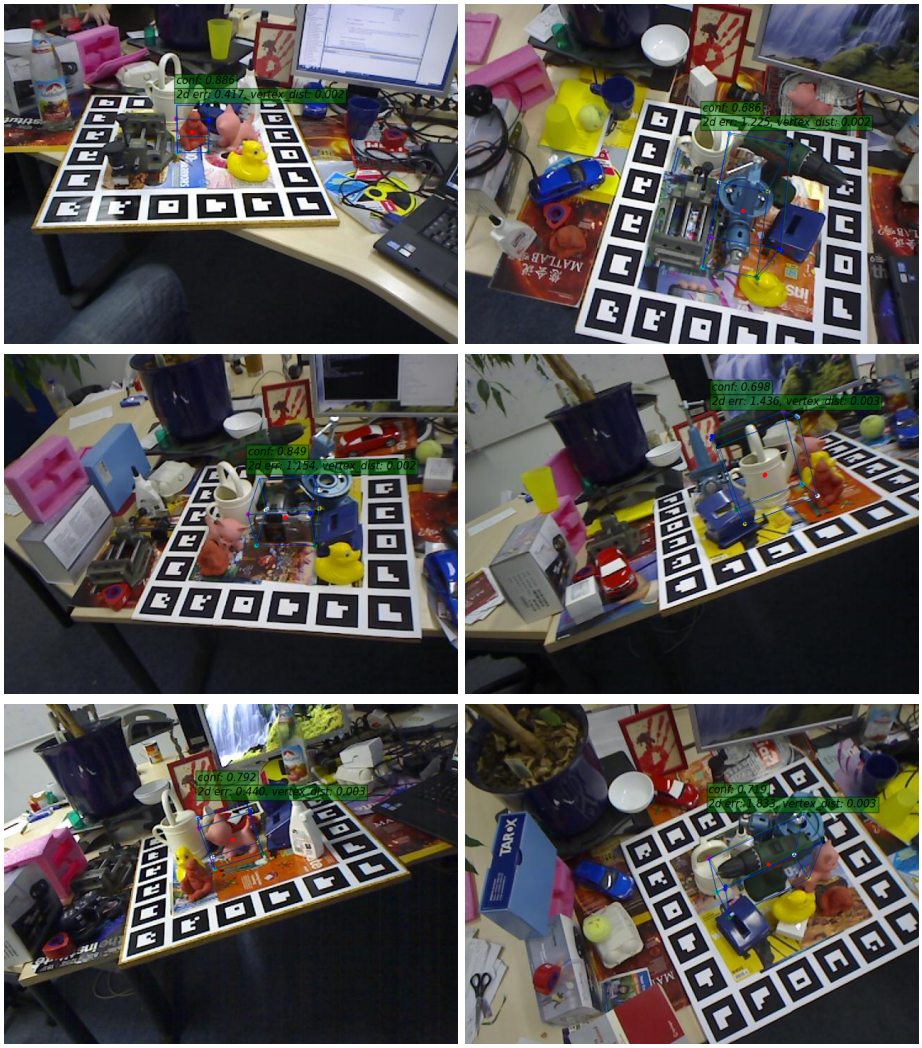


Figure 7.7 *(a) Example predictions of different objects for qualitative evaluation of the proposed YOLOv5-6D model on the LINEMOD test dataset. Green 3D bounding boxes visualize ground-truth poses while the estimated poses are highlighted by blue boxes. The objectiveness, 2D reprojection error and 3D vertices distance are depicted in the green floating annotated text.*

Figure 7.7 *(b) Example predictions of different objects for qualitative evaluation of the proposed YOLOv5-6D model on the LINEMOD test dataset. Green 3D bounding boxes visualize ground-truth poses while the estimated poses are highlighted by blue boxes. The objectiveness, 2D reprojection error and 3D vertices distance are depicted in the green floating annotated text.*

Method	YOLO6D	PoseCNN	PVNet	Gen6D (Model Free)	EfficientPose	RNNPose		EPro-PnPv2	YOLOv5-6D (Proposed)
Type **1st. Stage**	PnP	Direct	PnP	Direct Volume Refine Det. & View point select	Direct	PnP Refinement PoseCNN	PVNet	PnP Faster-RCNN Det.	PnP
Object									
Ape	21.62	25.62	43.62	-	87.71	**88.19**	85.62	-	87.81
Benchvise	81.80	77.11	99.90	77.03	99.71	**100.0**	**100.0**	-	**100.0**
Camera	36.57	47.25	86.86	66.67	97.94	98.04	**98.43**	-	97.45
Can	68.80	69.98	95.47	-	98.52	99.31	**99.51**	-	99.31
Cat	41.82	56.09	79.34	60.68	**98.00**	96.41	96.41	-	96.21
Driller	63.51	64.92	96.43	67.39	**99.90**	99.70	99.50	-	99.11
Duck	27.23	41.74	52.58	40.47	**90.00**	89.30	89.67	-	86.57
Eggbox*	69.58	98.50	99.15	-	**100.0**	99.53	**100.0**	-	**100.0**
Glue*	80.02	94.98	95.66	-	**100.0**	99.71	97.30	-	**100.0**
Holepuncher	52.24	42.63	81.92	-	95.15	**97.43**	97.15	-	95.34
Iron	74.97	70.17	98.88	-	99.69	**100.0**	**100.0**	-	99.19
Lamp	71.11	70.73	99.33	89.83	**100.0**	99.81	**100.0**	-	**100.0**
Phone	47.74	53.07	92.41	-	97.98	98.39	**98.68**	-	97.89
Average	55.95	63.26	86.27	67.01	97.35	**97.37**	97.10	96.36	96.84

Table 7.2 *Comparison of the proposed approach with alternative methods on LINEMOD using the 10% ADD (Average Distance Difference at a 10% object diameter threshold) and ADD-S(*) metric. The best performing model(s) per object is depicted in bold.*

PoseCNN [314], PVNet [315], Gen6D [307], EfficientPose [343], RNNPose [313] and EPro-PnPv2 [302]. The accuracies reported in the respective papers are used for comparison. In addition, Table 7.2 and Table 7.6 categorize the methods as described in Section 7.2 and based on our findings. Table 7.2 indicates the type of network employed during the *1st Stage* of the multi-stage methods and the *Type* of approach (direct or PnP) utilized for obtaining the object pose.

As can be observed from Table 7.3, YOLOv5-6D realizes an average increase of 9.07% on the 2D reprojection performance metric over the YOLO6D model. On the ADD(-S) metric (Table 7.2), YOLOv5-6D shows a strong performance increase (40.98%) over its predecessor and realizes competitive results against SOTA alternative methods, while being much faster (see Table 7.8). Comparisons to the seven SOTA alternative methods are added to set strong baselines and for completeness, as we evaluate YOLOv5-6D as a new architecture in general.

Table 7.3 *Comparison of YOLOv5-6D on LINEMOD data in terms of the 2D reprojection metric.*

Object	YOLO6D	YOLOv5-6D
Ape	92.10	99.24
Benchvise	95.06	99.61
Cam	93.14	99.71
Can	97.44	99.80
Cat	97.41	99.80
Driller	79.41	98.61
Duck	94.65	99.16
Eggbox	90.33	99.34
Glue	96.53	99.61
Holepuncher	92.86	99.91
Iron	82.94	99.59
Lamp	76.87	98.85
Phone	86.07	99.52
Average	90.37	99.44

7.4.2 Inference time

To assess the real-time performance of the proposed YOLOv5-6D model, aimed at achieving 30 FPS, we have conducted a comparative analysis of its inference time against other leading 6-DoF pose estimation methods. This comparison is carried out under uniform hardware conditions (see Section 7.3.7) to ensure fairness, unless otherwise stated. We have employed each method as described in the respective research papers and as made publicly available. In all cases, we use the LINEMOD cat dataset (640×480×3 pixel images) with corresponding pre-trained weights for the cat object, with the exception of GEN6D (no object specific model is required). The text below summarizes our findings.

- **YOLO6D** achieves a total inference time of 17.9 ms per frame (55.8 FPS), which is 5 FPS faster than originally reported. This includes image loading

to GPU (0.6 ms), model forward pass (4.5 ms), and filtering the predictions (12.8 ms).

- **PVNet** yields an inference time of 32.4 ms (30.9 FPS), encompassing data loading time (3.5 ms), PVNet model forward pass (17.6 ms), and the RANSAC-based voting scheme (11.3 ms) used to obtain the reported accuracy.
- **EfficientPose (ϕ=0)** demonstrates a pose prediction time of 36.8 ms (27.15 FPS), comparable to the reported 27.45 FPS. This includes data preprocessing (14.1 ms) and model inference (22.7 ms).
- **RNNPose** utilizes initial poses from PVNet (no end-to-end solution is developed), with refinement inference time depending on the employed number of recurrent and rendering cycles. Each recurrent iteration involves correspondence field (CF) estimation (11.0 ms), pose optimization (0.4 ms), and CF rectification (4.4 ms) for a total execution time of 15.8 ms. The rendering cycle includes reference image rendering (9.4 ms), 3D feature rendering (3.6 ms), image feature encoding (2.6 ms), followed by the earlier mentioned recurrent iterations. As depicted in the RNNPose paper (Figure 5 & Table 2), approximately four rendering cycles (4 total) with each running four recurrent iterations (16 total) are required to obtain the SOTA LINEMOD performance reported in the paper. In addition to the rendering cycles modules (15.6×4=62.5 ms) and recurrent iterations (15.8×16=253.0 ms), a once-off data loading time (2.9 ms) and running the 2D-3D Hybrid Net (2.1 ms) brings the refinement execution time to 320.5 ms and the total execution time (with the addition of PVNet initial poses without RANSAC voting) to 341.6 ms (2.93 FPS).
- **EPro-PnP** exhibits a rapid inference time of 10.1 ms (98.5 FPS), requiring 0.3 ms for image-crop data loading, 4.8 ms for the model forward pass, 1.1 ms for postprocessing, and 3.9 ms for the PnP calculation. However, the 2-stage approach requires an earlier model to detect the objects of interest and provide exact crops to the EPro-PnP part. The research largely improves on the earlier work of Li *et al.* called CDPN [321], which utilizes the same 2-stage approach. While EPro-PnPv2 employs Faster-RCNN [75], a relatively old and slower object detection algorithm, no investigation has been conducted into how the model performs based on the provided input crop. Alternatively, in the CDPN approach (Table 3 & 4) the authors show that by using YOLOv3 they obtain slightly lower performance (ADD(-S) 89.80 with YOLOv3 vs. ADD(-S) 89.86 with Faster-RCNN), but with a significant speed improvement (30 ms vs. 76 ms). Since neither a detection implementation is discussed, nor provided along with EPro-PnPv2, we have employed the YOLOv3-based performance reported in CDPN [321] in this comparison. This enables a total EPro-PnPv2-based pose prediction in 40.2 ms (24.9 FPS).
- **Gen6D**, a 3D object model-free and generic estimation model, achieves a novel object pose prediction at 427.26 ms per frame (2.34 FPS), including

initial object detection (125.7 ms), viewpoint selection (37.1 ms), and pose refinement (3×88.0 ms=256.1 ms).

- **YOLOv5-6D** realizes single-shot object 6-DoF pose estimation at 41.88 FPS (inference time of 23.88 ms per frame). Table 7.4 depicts the exact execution time per module of the YOLOv5-6D model for both the LINEMOD and the newly introduced X-ray datasets.

For a comprehensive comparison, we include Figure 7.8 to illustrate the speed versus average accuracy trade-off on the LINEMOD dataset. The proposed YOLOv5-6D enables single-shot, real-time object 6-DoF pose estimation, demonstrating its efficacy on both the LINEMOD and the new larger X-ray datasets. This comparison highlights the balance between speed and accuracy of 6-DoF pose estimation methods and underscores the efficiency of the proposed model. Finally, the results of all findings are summarized in Table 7.6.

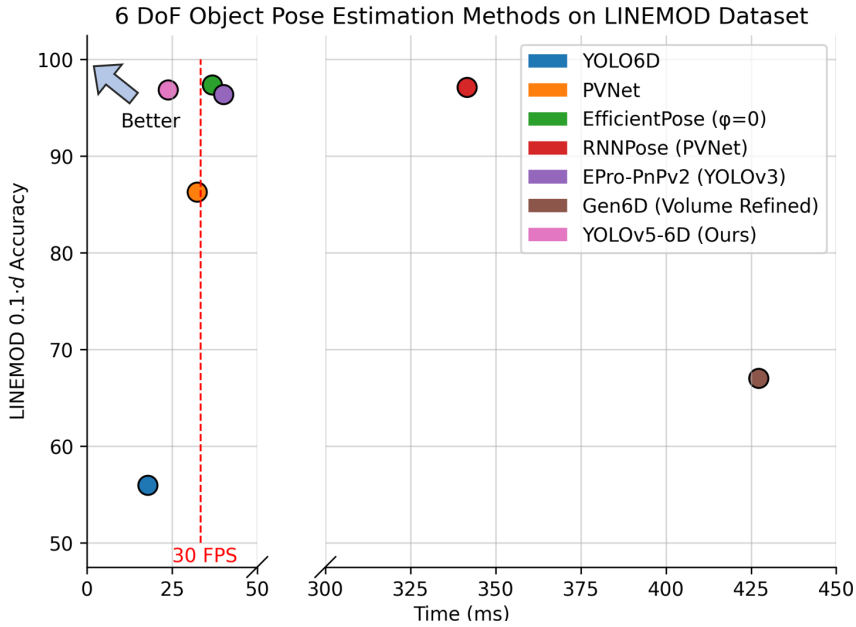


Figure 7.8 *Accuracy and inference-time comparison of YOLOv5-6D and competitive alternative methods. Measurements are obtained with a unity batch size.*

7.4.3 X-ray pose estimation

Section 7.1 and Section 7.3.1 presents the hard requirements for an object 6-DoF pose estimation method to be successful in the medical X-ray domain. In summary, the method needs to be (1) very accurate, (2) incorporate image acquisition geometry and (3) fast, to enable real-time analysis. Given these strict require-

Table 7.4 *YOLOv5-6D inference time on the LINEMOD and X-ray datasets. Measurements are obtained with a unity batch size.*

Operation Image size	LINEMOD 640×480×3	X-ray 960×742×1
Tensor to cuda	0.22 ms	0.20 ms
Predict	23.03 ms	29.82 ms
Filter predictions	0.52 ms	0.42 ms
ePnP	0.11 ms	0.07 ms
Total time	23.88 ms	30.51 ms
Frame rate	41.88 FPS	32.78 FPS

ments and the analysis of the results of the various methods on the LINEMOD dataset, YOLOv5-6D presents itself as the only viable candidate in the X-ray domain. To test this assertion, we conduct experimental analysis of YOLOv5-6D and EfficientPose(ϕ=0) in the X-ray domain. The quantitative results of the experiments on the X-ray datasets can be observed in Table 7.5 and corresponding qualitative results are shown in Figure 7.9.

Table 7.5 *Performance of YOLOv5-6D and EfficientPose on the X-ray Cube and X-ray Screw datasets at various distance thresholds. It should be noted that the employed parameter d is 30 mm for the Cube dataset and 34.3mm for the Screw datasets.*

Model Metric [mm]	Eff.Pose Cube Val.	YOLOv5-6D Cube Val.	 Screw Val.	 Screw Test
ADD(-S) 0.1·d	0.0	99.27	96.87	92.41
ADD(-S) 0.05·d	0.0	97.08	87.50	81.01
ADD(-S) 1.0 mm	0.0	93.43	75.0	55.70
ADD(-S) 0.02·d	0.0	82.48	65.62	43.04
3D Transl. error [mm]	13.8 ± 4.5	0.35±0.21	0.82±0.43	1.27±0.47
3D Angle error [deg.]	33.7 ± 8.3	1.45±1.29	3.18±1.72	3.79±2.72

The conducted experiments show that the YOLOv5-6D model can predict relevant 2D keypoints for accurate 6-DoF pose estimation, notably also in challenging scenarios like the X-ray datasets, where the focal length varies by up to 28 cm and the object undergoes translation and rotation. In contrast, EfficientPose tends to converge to a mean pose present in the Cube dataset, reflecting its low performance in such settings. This is expected due to the ambiguity present in the pose, if the method does not have access to camera intrinsic parameters during training.

Specifically, for the X-ray datasets featuring two small instruments, the YOLOv5-6D model achieves a high accuracy of 99.27% for the asymmetrical cube and 96.41% for the symmetrical bone screw. At a 1-mm distance threshold, the pose of the asymmetrical cube is estimated with a 93.43% accuracy. Similarly, at a 1-mm threshold, the pose of the symmetrical bone screw is accurately acquired in 75%

of the validation cases. The same model trained on images only containing the screw (and heavy augmentation) is then applied to the screw and the spine phantom dataset (Screw Test). The model shows comparable and high accuracy at the $0.1{\cdot}d$ (3.43 mm) and $0.05{\cdot}d$ (1.72 mm) threshold, but experiences a large drop in performance at the smaller distances. We do not report the 2D reprojection error in the X-ray datasets, because the symmetry around the z-axis of the screw allows for multiple plausible 2D keypoint predictions that will resolve the correct object pose. This is visually proven and illustrated in Figure 7.9. Lastly, Table 7.4 depicts the inference time of the YOLOv5-6D model for the two domains.

7.5 Discussion

A novel YOLOv5-6D method for accurate 2D keypoint prediction and associated pose estimation is developed, while considering the image acquisition geometry. Prior keypoint-based methods are not fast or accurate enough for real-world applications, especially in the medical domain where accuracy is essential. The YOLOv5-6D model builds on advancements in the YOLO object detection series, to improve prediction accuracy of 2D keypoints and enable correct pose estimation through solving pose with PnP using the 2D/3D object bounding-box correspondences. In addition, we have presented a new data capturing method for 6-DoF tasks that utilizes an optical camera attached to the X-ray detector. The approach allows for data acquisition across all X-ray geometries and objects, without adding image artifacts (such as AruCo markers or calibration domes) to the final X-ray image, or relying on accurate X-ray system sensors to acquire object-pose labels. The YOLOv5-6D method generalizes across domains and imaging systems. This

Table 7.6 *Pose estimation method properties based on the LINEMOD dataset.*

Method	Task-agnostic	Cam. Intrinsic	Real-Time	≥ 90 Acc.
YOLO6D	-	✓	✓	-
PoseCNN	-	-	-	-
PVNet	-	✓	✓	-
Gen6D	✓	-	-	-
EfficientPose	-	-	-	✓
RNNPose	-	✓	-	✓
EPro-PnPv2	-	✓	-	✓
YOLOv5-6D	-	✓	✓	✓

generalization is evident from (1) its application to the RGB images, (2) X-ray images obtained with different acquisition geometries and (3) different levels of semantic complexity in the X-ray image contents.

With respect to the first (1) aspect of generalization, the YOLOv5-6D model shows competitive results on the public LINEMOD RGB dataset with an average ADD(-S) score of 96.84% compared to the current SOTA (RNNPose [313]) with

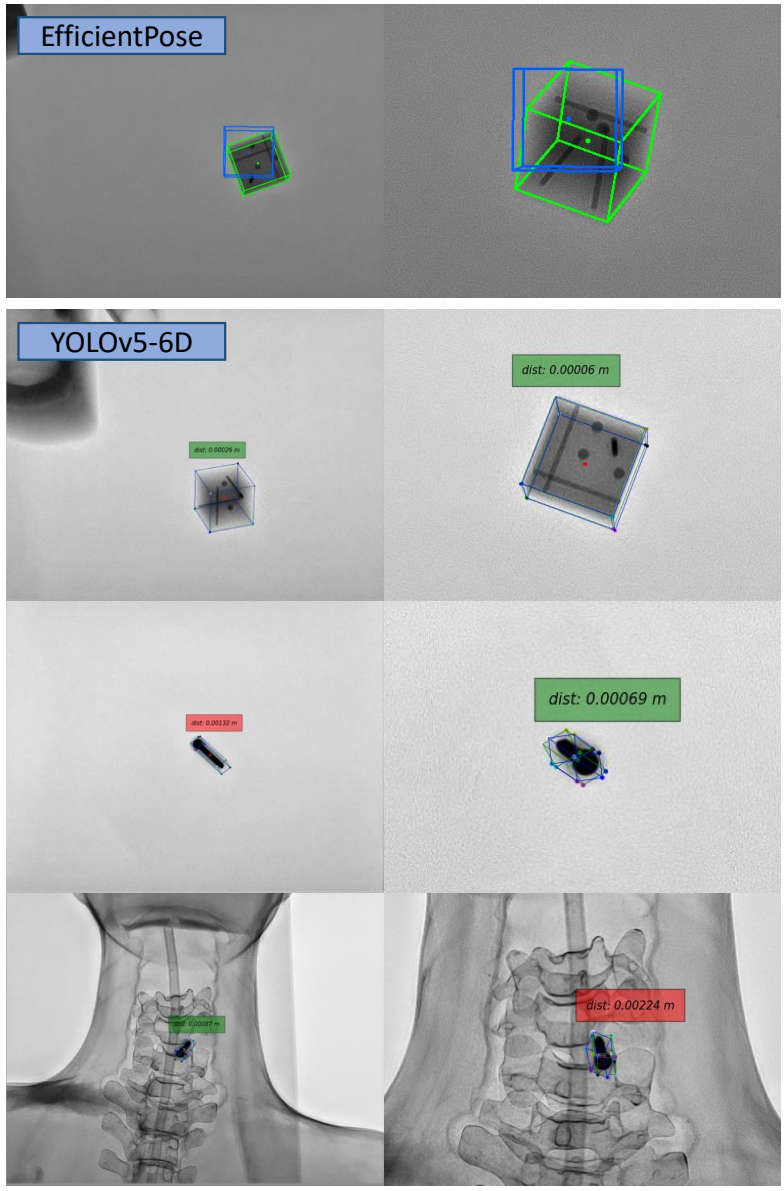


Figure 7.9 *Example predictions for qualitative evaluation on the Cube Validation, Screw Validation and Screw Phantom Test datasets. The first column presents images in the original resolution and the second highlight a zoomed area. Green 3D bounding boxes visualize ground-truth poses, while the estimated poses are highlighted by blue boxes. The average 3D vertices distance for YOLOv5-6d is shown in the floating text box with a color representing pass (green) or fail (red) by the 1-mm ADD(-S) metric. Images are rendered with the respective code bases.*

97.37%, as depicted. in Table 7.2. However, the proposed method is considerably faster (41.88 FPS vs. 27.15 FPS of EfficientPose) in execution at this level of accuracy and leverages the imaging geometry, as summarized in Figure 7.8 and Table 7.6. These characteristics make the method appealing for real-time instrument pose estimation in the X-ray domain.

The second (2) aspect is addressing generalization towards different X-ray geometry. Here, images are obtained with various hardware configurations using higher input image resolution. The images are without depth information and typically contain low contrast of the objects of interest. The proposed YOLOv5-6D successfully predicts relevant 3D bounding-box keypoints for both X-ray objects included in this research, enabling highly accurate pose estimation (Table 7.5) with a translation error of only 0.35 mm$\pm$0.21.

The third aspect (3) is about generalizing across different levels of semantic complexity. The model trained for screw pose estimation in a simple training environment generalizes to the new and more clinically-relevant domain containing the human phantom. This generalization is evident from Table 7.5. Most notably, the proposed approach is able to accurately estimate the pose of a small cannulated cancellous bone screw up to an impressive 75.0% at 1 mm by the ADD(-S) metric on the validation set. In addition, the same model generalizes well to the new and more complex test set containing a spine phantom. Here, we observe a similarly high 92.41% ADD(-S) score at 0.1·d. As a very hard final test, we evaluate the pose accuracy at 1 mm ADD(-S), which shows a drop to 55.70% in comparison to the validation set. This performance drop is expected due to the stringency of the test, but we consider that it can be rather traced to inaccuracy in the labels. For example, prior to the labeling process, the offset from the instrument to the ChArUco frame is manually and precisely measured. However, with the screw being placed in a spine phantom, this measurement becomes considerably more difficult and error prone. Our measurements for acquiring the ground-truth labels of the test set are likely to be off by $\pm$1 mm, which results in a lower performance of the proposed model at these distances. The performance of the proposed method on these 4 datasets expresses the generalization ability of the method and future research can further elaborate on testing with other instruments.

This research is one of the first to propose tracking the actual screw, instead of the surgical path or screw placement tools for assisted clinical guidance. The YOLOv5-6D method enables accurate and fast 6-DoF pose estimation of the screw with respect to the X-ray detector or source. By combining this method with a spine tracking system, such as the one proposed by Manni *et al*. [344], the screw pose can be determined with respect to the target location on the spine, enabling precise screw placement and its validation without the need for postoperative CT.

7.6 Future work and limitations

In this study, we have focused on single-object and single-class pose estimation and have not collected data to investigate multi-object and multi-class pose esti-

mation. However, simultaneous multi-object pose estimation is an intriguing area of future research in the context of spinal screw placement, as multiple screws are typically used in this procedure. We conjecture that the proposed YOLOv5-6D model can be leveraged for estimating the pose of multiple similar-sized screws without the need to retrain the model. In scenarios where multiple identical screws of different sizes (multi-class) are used, the ill-posed nature of X-ray imaging may hinder the ability to distinguish these objects individually. Nonetheless, it is worth noting that the screws used during the clinical procedure are known beforehand, and the corresponding 3D screw model can be manually linked to accurately determine the pose of each screw, regardless of its size. This presents an exciting opportunity for future research to explore the feasibility and effectiveness of this approach in the context of spinal screw placement.

Furthermore, we show that the model successfully estimates the screw pose outside of its training distribution. However, it is still evaluated in a rather simple context and future work needs to fully explore the limitations of the approach under more clinically relevant conditions.

While we investigate object pose estimation under variable X-ray imaging geometry, similar challenges arise outside of the medical domain, such as optical cameras with adjustable focal lengths used for zooming. In these zooming cases, this change in imaging geometry should be carefully considered and should facilitate accurate object pose estimations from the captured image, in order to enable re-use of the proposed framework. Moreover, satellite pose estimation [345], [346] is a domain that already faces this challenge and can straightforwardly benefit from the proposed approach.

7.7 Conclusion

In this chapter, we have discussed the development and evaluation of a novel YOLOv5-6D model for accurate 6-DoF instrument pose estimation in X-ray imaging, with a focus on providing clinical guidance under varying image acquisition geometries. The proposed model leverages recent advancements in the YOLO object detection series to significantly enhance the accuracy of pose estimation, addressing the stringent requirements of real-time medical applications.

The motivation for this work stems from the complexity and error-prone nature of fluoroscopy-guided minimally invasive interventions, which rely heavily on repeated acquisition of standard projections for instrument guidance. Standard methods, while effective, often require extensive external equipment and fail to provide real-time validation through actual screw tracking. The proposed method aims to overcome these limitations by offering a deep learning-based approach for instrument tracking that incorporates the changing imaging geometry, which is critical for applications like X-ray imaging and space satellite pose estimation.

The chapter has introduced a novel data collection method using an external optical camera for automated and precise data labeling across diverse X-ray geometries. The need for a fast, accurate and generalizable 6-DoF pose estimation

method is addressed, by excluding the X-ray imaging geometry from the keypoint prediction task and then utilizing the intrinsic camera parameters in the pose estimation process. Through this process, the model has been designed to acquire accurate poses under changing acquisition geometries. The model demonstrates generalization capabilities through successful performance on datasets of varying complexity, including transitioning from a controlled lab environment to a clinically relevant setting with a spine phantom, thereby proving its robustness and applicability in real-world medical scenarios.

In the conducted experiments, the YOLOv5-6D model demonstrated competitive results on the public LINEMOD dataset, achieving an average $0.1{\cdot}d$ ADD(-S) score of 96.84% and operating at a real-time speed of up to 42 FPS. These results are significantly faster compared to existing methods with comparable accuracy. Additionally, the model's performance on two new X-ray datasets —-one containing a calibration cube and the other a clinically relevant cancellous bone screw—-showcases its robustness and accuracy in the challenging X-ray domain. Specifically, the model achieves a high $0.1{\cdot}d$ ADD(-S) performance of 92.41% on the spine-phantom test set, highlighting its potential for real-world clinical applications.

The work also indicates the importance of incorporating the imaging geometry into the pose estimation process, enabling the YOLOv5-6D model to generalize across different imaging system geometries and complex environments. The proposed method offers a practical solution for real-time instrument pose estimation during minimally invasive surgeries, significantly reducing the need for postoperative CT scans and extensive external equipment.

In conclusion, the YOLOv5-6D model presents a significant advancement in the field of object pose estimation in X-ray imaging, particularly for medical applications. By addressing the technical challenges and demonstrating strong performance across different datasets, the work sets the stage for further exploration and refinement in both medical and non-medical domains. Future research can extend this approach to multi-object and multi-class pose estimation, as well as explore its applicability in other areas requiring variable imaging geometries.

Chapter 7

Chapter 8 Conclusions

8.1 Conclusions of the individual chapters

This thesis has presented significant advancements in various computer vision techniques, demonstrating enhanced detection capabilities and improved precision in guidance systems. We have introduced a PDAC detection framework that leverages external tumor-indicative features, leading to increased detection accuracy. Furthermore, the thesis details improvements in segmentation uncertainty quantification, enabling more flexible distribution modeling and incorporating full-3D information for uncertainty estimation. These methods have been applied to the segmentation of PDAC and surrounding anatomical structures, facilitating automated resectability prediction. Two novel Out-of-Distribution detection approaches have been presented: (1) wavelet-based normalizing flows for semantic OOD detection, enabling unsupervised malignant melanoma detection, and (2) covariate shift detection by modeling heteroscedastic high-frequency image components, ensuring the reliability of the images used. Finally, a novel pose estimation technique, YOLOv5-6D, has been introduced that is capable of accurately estimating the 6-Degrees-of-Freedom (6-DoF) of objects under various imaging geometries.

This section summarizes the results of the individual chapters. Section 8.2 comes back on the research questions posed in Chapter 1 and provides a detailed discussion of the answers to these questions. Section 8.3 presents an outlook on future directions and associated research themes.

Chapter 2 provides the necessary technical background information and a brief introduction to the deep learning-based image analysis algorithms applied in this thesis. The chapter introduces the technical advancements in developing deep learning-based image analysis algorithms for classification, segmentation, object detection and pose estimation. An overview of generative models, specifically Variational Autoencoders (VAEs) and Normalizings Flows (NFs) is presented. Finally, the concept of uncertainty modeling in deep learning is described.

Chapter 3 has explored the development and implementation of a Computer-Aided Detection (CADe) system for Pancreatic Ductal Adenocarcinoma (PDAC). The proposed framework integrates secondary tumor-indicative features with advanced deep learning techniques to improve the accuracy and reliability of

PDAC detection. The comprehensive evaluation has demonstrated high detection accuracy (0.99 AUROC) based on high sensitivity and specificity, indicating the robustness of the model across both internal and public datasets. The integration of external, clinically-relevant features, processed through a Residual 3D U-Net, has shown significant promise in enhancing early detection of PDAC, which is crucial for improving patient outcomes. This chapter sets the stage for subsequent discussions on segmentation and resectability assessment, laying a solid foundation for the thesis' overarching goal of improving methods and modeling for cancer detection and procedure guidance.

Chapter 4 has presented significant advancements in uncertainty quantification within medical image segmentation. By extending the capabilities of the Probabilistic U-Net (PU-Net) and incorporating normalizing flows (NFs), the research has addressed the critical need for better modeling of aleatoric uncertainty. The novel methods proposed for both 2D and 3D segmentation contexts have been rigorously evaluated, showing improvements in key metrics, like Generalized Energy Distance (GED) of 14% and Hungarian-matched Intersection over Union (IoU) of 13%. These advancements not only enhance the reliability of segmentation models, but also contribute to their practical applicability in clinical settings, where understanding and conveying interobserver and intraobserver variability is essential for informed decision-making. The chapter underscores the importance of continuous improvement in uncertainty quantification to support the safe and effective use of CADx in healthcare.

Chapter 5 has focused on predicting the resectability of pancreatic tumors, a critical aspect of PDAC treatment management. Building on the detection and segmentation techniques discussed in preceding chapters, this chapter introduces a deep learning-based framework for assessing tumor-vessel involvement, which is a key determining factor of resectability. The integration of uncertainty modeling, particularly in capturing the complexities of tumor-vessel interactions, represents a significant step forward in developing a Clinical Decision Support System (CDSS). The findings of this chapter suggest that the proposed models can provide accurate and clinically relevant predictions (0.92 scan-level sensitivity and 0.89 scan-level specificity), aiding surgeons in making better-informed decisions about the feasibility of surgical interventions. This chapter also highlights the potential of a deep learning-based image analysis method to enhance personalized treatment strategies and improve survival rates in pancreatic cancer.

Chapter 6 has addressed the critical challenge of Out-of-Distribution (OOD) detection, which is essential for ensuring the robustness and reliability of images and automated analysis models in real-world applications. The introduction of image-frequency decomposition-based Normalizing Flows (NFs) for OOD detection, particularly in the context of melanoma (semantic OOD) and covariate shift detection (covariate OOD), marks a significant innovation in the field. The methodologies developed in this chapter, including the novel CovariateFlow, indicate that different frequency bands can be more informative for OOD detection than modeling the complete image distribution. In the semantic OOD case, model-

ing the low-frequency bands with WaveletFlow provides an improved malignant melanoma detection (0.78 AUROC, up 5% over the full distribution). In the covariate case, modeling the heteroscedastic high-frequency components realizes an increase in performance of 5-10% AUROC (depending on the dataset). These methods provide a comprehensive framework for detecting and managing OOD data, thereby improving the generalization ability of machine learning systems. These contributions are crucial for advancing the safe integration of ML methods into clinical practice.

Chapter 7, the final technical chapter, has delved into the development of a novel YOLOv5-6D model for accurate 6-Degrees of Freedom (DoF) instrument pose estimation in X-ray imaging. This chapter is demonstrating that the proposed model is able to meet the stringent requirements of real-time medical applications, particularly in fluoroscopy-guided interventions for spinal screw placement. The proposed algorithm explicitly benefits from the advancements in the YOLO detection series to enable speed and keypoint prediction accuracy improvements. Additionally, excluding camera parameters from the modeling component, but including them in the pose estimation part, enables generalization across different imaging geometries. The model's robustness across different imaging geometries and its high accuracy in both laboratory (99.27% ADD, Cube dataset) and clinically-relevant setting (92.41% ADD(-S) Screw phantom test dataset) underscore its potential for real-world implementation. The chapter concludes by highlighting the broader implications of this research for both medical and non-medical domains, suggesting avenues for future exploration in multi-object and multi-class pose estimation.

8.2 Discussion on research questions

This section evaluates the proposed solutions with regards to the research questions presented in Section 1.5.

RQ1: Incorporating domain-specific knowledge in PDAC detection

RQ1a: *Can we effectively include PDAC-indicative features into a PDAC CADe system to enhance the detection performance?*

The thesis addresses this research question in Chapter 3, by discussing the critical role of secondary, PDAC-indicative features to clinicians in the conventional detection process. Key anatomical structures are identified, particularly the pancreas and bile ducts, and incorporated in a multi-stage PDAC CADe system.

The approach begins with the segmentation of these tumor-indicative features with an advanced 3D U-Net architecture. These models allow for a comprehensive view of the pancreas in CT scans and lay the groundwork for adding secondary features that often correlate with PDAC, such as bile duct dilation, pancreatic duct abnormalities, and contour irregularities. By segmenting these secondary indicators, the CADe system aligns more closely with clinical diagnostic meth-

ods, which can enhance early detection of PDAC, especially in cases where the tumor may not be prominently visible on imaging. These segmented features are then channel-wise concatenated along with the CT scan and provided to the primary detection model —-a 3D U-Net-— which processes both the tumor and surrounding anatomy, allowing for a nuanced analysis of spatial relationships critical for accurate PDAC identification. This integration of secondary indicators into the detection model has led to a marked improvement in sensitivity (1.0 ± 0.0) and specificity (0.99 ± 0.02), thereby boosting the system's ability to reliably detect early-stage PDAC. By concatenating the tumor-indicative anatomical structures along with the CT scan, the model learns to effectively extract features from the set to optimize the tumor detection performance. This approach enables a data-driven way to learning the weighting and importance of each feature towards to final task. However, it is not known whether the concatenation operation is optimal, or if other feature integration methods will yield better results.

Experimental validation within the thesis confirms the model's enhanced performance when secondary features are included. Nevertheless, the research notes in Section 3.6 that there are limitations that may inhibit the conclusions drawn from the obtained results in the context of the research question. The dataset used for training is from a single center and relatively small, which introduces further challenges. A small dataset can hinder the generalization ability of the model and reduce the statistical power of the results. Particularly related to the acquisition and integration of the secondary features, one major challenge is the variability in the appearance and prominence of these features across different patients, which makes it difficult to determine the performance of the model. The chapter also discusses the challenge of acquiring high-quality annotated data for these secondary features, as this data is essential for training and validating the models, but are extremely time-consuming to obtain. These constraints emphasize the need for larger, more diverse datasets to improve model performance and reliability.

Despite these challenges, the automated, secondary-feature enhanced CADe framework aligns well with clinical workflows, potentially reducing false negatives and enabling radiologists to interpret results with greater confidence.

RQ1b: *What is a possible setup for a complete end-to-end pancreatic CADe system?*

Chapter 3 presents a workflow consisting of several key stages for an effective and complete end-to-end PDAC CADe system. In addition, a larger, more comprehensive dataset is employed to validate this end-to-end system.

Starting with a high-resolution CT scans, the first critical component discussed is the segmentation of the pancreas from the CT volume. Localization of the pancreas focuses the subsequent analyses and enhances the execution speed of the the complete detection process. The pancreas are segmented with a U-Net on a coarse (downsampled) CT scan. The second component entails a large crop around the coarse segmentation, taken from the high-resolution scan and utilized by a fine-grained pancreas segmentation model. The same crop is exploited by the third component to segment the surrounding anatomical structures, such as the

common bile duct and pancreatic duct. From the segmentation of the pancreas (Components 1 and 2) and ducts (Component 3), the detection task is focused on the pancreatic region, while sufficient context is provided for the detection and classification of the tumor itself (Component 4). This task is carried out with an improved Residual 3D U-Net. The integration of secondary features, as discussed in response to RQ1a, plays a crucial role at this stage. The experimental results (Chapter 3) show how the combination of primary tumor detection with secondary feature analysis leads to more accurate and reliable diagnoses, reducing the number of false positives and improving the system's overall sensitivity. This process enables an AUROC score of 0.99 on the internal test set, proving the effectiveness of the approach.

Finally, these components are integrated into a unified CADe system with automated intermediate processing steps to enable a seamless workflow that allows for analysis and feedback. The development of a web-based application that brings together all these components, providing an interface that can be used by clinicians to review and analyze results. The integration of the system into clinical practice is discussed, while providing insights into how such systems can be adapted to fit the workflow of different medical institutions, ensuring that they provide clear benefit to clinicians and patients alike.

Although the proposed approach is not extensively compared against other setups, it forms a feasible solution for a complete end-to-end detection system. Additionally, it performs PDAC detection in a similar order as the clinical procedure, where the pancreas is first localized, external indicators for tumor are marked and then a detailed and focused analysis of the tumor region is carried out.

In conclusion, an effective end-to-end pancreatic CADe system consists of a pancreas localization component, followed by extracting tumor-indicative features to maximally enable PDAC segmentation. The high detection accuracy obtained is indicative of the system's capability to reliably distinguish pancreatic tumors, underscoring its potential for early detection and its value as a clinical support tool in pancreatic cancer assessment.

RQ2:Accurate ambiguity modeling for improved segmentation

RQ2a: *How can we model ambiguous ground-truths (aleatoric uncertainty) to improve the accuracy of segmentation maps?*

To address the challenge of accurate segmentation of structures under ambiguous ground truths, a probabilistic framework is introduced that captures and expresses this aleatoric uncertainty in Chapter 4. The core of this approach lies in the use of the Probabilistic U-Net (PU-Net), which is designed to model multiple plausible segmentation outcomes, rather than producing a single deterministic result. This method is particularly well-suited for medical images where ground truths can be ambiguous, due to variations in human annotation as a result of two forms of ambiguity. The first ambiguity arises from uncertainties in the data,

such as the image quality (noise, contrast) and imaging protocol (MRI vs. CT). The second from of uncertainty arises from annotation process. Experts performing the annotation have a given time (to annotate) and knowledge about the particular case. This results in variations between annotations of the same case from different annotators (interobserver), but also varying degrees of accuracy of different cases from the same annotator (intraobserver).

To address the variability introduced through the annotation process, the PU-Net utilizes a conditional Variational Autoencoder (cVAE) framework, which allows the model to learn a distribution over possible segmentation maps, thereby capturing the aleatoric uncertainty inherent in the data and annotations.

To further enhance the accuracy of segmentation under the aleatoric uncertainty conditions, Section 4.3 of the thesis proposes augmenting the posterior distribution in the PU-Net with Normalizing Flows (NFs). NFs increase the flexibility of the posterior, allowing it to capture more complex distributions that better reflect the variability and uncertainty in the ground-truth data. The thesis presents compelling evidence that when augmenting the Gaussian posterior in the PU-Net with Normalizing Flows (NFs), a substantial improvement in the quantification of aleatoric uncertainty is achieved. The standard PU-Net that assumes a Gaussian distribution for the posterior, is somewhat limited in its ability to model the true variability in the data. By incorporating NFs, the posterior distribution becomes more expressive that with a Gaussian alone, allowing it to better capture the complexities of the annotation data. This enhancement is crucial for accurately reflecting the uncertainty present in ambiguous ground-truth data.

Experimental results discussed in Chapter 4 demonstrate that the use of NFs leads to improved performance on key metrics. For instance, the GED metric, which measures the diversity of the predicted segmentation maps, shows an improvement of 14% when NFs are used. Similarly, there is a notable increase (13%) in Hungarian-matched IoU, indicating that the model's predictions are more accurate and better aligned with the ground-truth annotations. Qualitative evaluations further supported these findings, with the NF-augmented models showing greater agreement with ground truths, particularly in the central regions of segmentation structures, while still appropriately expressing uncertainty at the edges where annotators are more likely to disagree.

RQ2b: *Does aleatoric uncertainty modeling in 3D improve the accuracy, consistency and execution speed?*

Three-dimensional medical imaging data (such as CT) is is richer and more informative than the individual 2D CT slices. This is further emphasized when evaluating complex 3D structures. As a consequence, the PU-Net has to be upgraded to 3D modeling. The introduced 3D PU-Net, as described in Section 4.4.1, leverages the spatial context provided by 3D medical imaging data. Extending the PU-Net to 3D processing significantly enhances the model's ability to consistently and accurately quantify aleatoric uncertainty in the original 3D data. Qualitatively, this 3D approach enables the model to better capture the anatomical

continuity and contextual information across different slices of a volume, leading to more consistent segmentation maps and a more reliable assessment of uncertainty. Quantitatively, moving from the 2D PU-Net to 3D PU-Net leads to a 4% improvement in the 2D GED, and a low distance metric value (0.422) of the now possible GED measured over three dimensions.

Subsequently, it is demonstrated that incorporating Normalizing Flows into the 3D PU-Net further improves its performance. The 3D extension, combined with NFs, allows the model to generate more diverse and accurate segmentation maps. The experimental results with Planar Flows show a further improved 3D GED (6.8%) and Hungarian-matched 3D IoU (2%) metrics, which indicate that the 3D model not only provides more accurate segmentation but also better captures the uncertainty across the entire volume of the data giving higher consistency of the segmentation structures. Additionally, considering the computational speed of the 3D PU-Net, it is noted that although the architecture requires more computational resources, the benefits in terms of accuracy, consistency and execution speed (3.4 times faster than the 2D PU-Net) justify the increased complexity. The 3D approach is particularly effective in scenarios where understanding of the spatial relationships within the data is crucial, such as in the segmentation of lung nodules or other complex anatomical structures.

RQ3: Exploring uncertainty modeling in PDAC resectability prediction

RQ3a: *How accurate should a PDAC and relevant vasculature segmentation algorithm be to obtain a feasible automated prediction of resectability?*

Chapter 5 of the thesis discusses the implementation of a deep learning-based approach to accurately segment both pancreatic ductal adenocarcinoma (PDAC) and the relevant vasculature, which is crucial for predicting tumor resectability. The chapter details the use of advanced segmentation models, specifically focusing on U-Net architectures, enhanced with additional context information from surrounding anatomical structures. These models are trained on annotated CT scans, where both the tumor and key vascular structures, such as the superior mesenteric artery (SMA) and superior mesenteric vein (SMV), are carefully labeled to ensure precise segmentation.

The results presented in this chapter demonstrate that the models are capable of achieving a high accuracy for segmentation of both the PDAC (0.66 DSC) and surrounding vasculature (arteries 0.86 DSC, veins 0.88 DSC). This accurate segmentation is critical for assessing the extent of tumor involvement with the vasculature, which is a key determinant in deciding whether the tumor is resectable. By accurately delineating the boundaries of the tumor, and its relationships with nearby blood vessels, the model provides essential information to make informed decisions about surgical options.

From the obtained segmentation maps, a degree of tumor involvement with the vasculature can automatically be computed. The models demonstrate high

accuracy in detecting any tumor involvement, with the 3D U-Net achieving the highest sensitivity (0.88) and specificity (0.86). Furthermore, classification of the extent of the involvement is achieved with a high 0.89 accuracy with the Prob. 3D U-Net.

The provided results are based on a limited set of experiments from a small single-center dataset. With the obtained high segmentation quality, the resectability prediction accuracy is sufficiently high to align with clinical evaluations. This automated approach is shown to align closely with regular assessments made by a clinical expert, contributing to a reliable and consistent tool for resectability prediction.

RQ3b: *Is the integration of model uncertainty into prediction models applicable and sufficiently useful for resectability prediction?*

The research introduces the concept of utilizing model uncertainty to potentially enhance the clinical relevance of resectability predictions. Chapter 5 discusses the integration of uncertainty quantification into the segmentation and resectability prediction pipeline. By combining aleatoric (data-related) uncertainties and epistemic (model-related) uncertainties, the model provides a nuanced prediction that highlights areas where the model is less confident. It is quantitatively shown that through probabilistic modeling of the segmentation maps, a higher alignment (Prob. 3D U-Net R^2 0.6, up from R^2 0.54 with deterministic methods) with the ground-truth degree of involvement is achieved. Qualitative evaluation depicts the impact of the three approaches (nnU-Net, 3D U-Net OLL, Prob. 3D U-Net OLL) have on the degrees of involvement. The Prob. 3D U-Net, capable of handling the uncertainty in the CT domain, presents a more consistent resectability prediction that more closely aligns with the clinical assessment.

This information is crucial in a clinical setting, where surgeons need to understand not only the likely outcome, but also the confidence level of the predictions. For example, in cases where the model indicates high uncertainty in the segmentation of the tumor's boundary with a major vessel, clinicians can be alerted to review those areas more closely, or consider additional diagnostic imaging prior to making any surgical decision. This early approach does not offer numerical certainty in the prediction result, but it improves reliability through the collaboration between the model and a clinician. Hence, this joint approach enhances the reliability of resectability predictions and supports more informed and safer clinical decision-making.

RQ4: Density modeling for Out-of-Distribution Detection

RQ4a: *Can generative models effectively detect and quantify semantic and/or covariate shifts in natural and X-ray images?*

Generative models demonstrate significant potential in detecting and quantifying semantic images, by modeling underlying data distributions and iden-

tifying deviations indicative of anomalies or distribution shifts. Chapter 6 has investigated their application in semantic Out-of-Distribution (OOD) detection, particularly within the context of imbalanced datasets which are common in skin melanoma research. Leveraging the unsupervised capabilities of Normalizing Flows (NFs), these models are explored as tools for identifying malignant skin melanoma by training exclusively on benign cases. However, challenges emerge when generative models prioritize general features like image entropy (e.g., high-frequency elements such as hair) over truly semantic attributes, thereby limiting their ability to assign meaningful malignancy scores. The state-of-the-art GLOW NF model, designed to capture the full-image data distribution, underscores these limitations with a modest AUROC of 0.73 in distinguishing benign from malignant melanoma. These findings align with broader literature, highlighting that while generative models hold promise for semantic OOD detection, their reliance on generalized likelihoods rather than domain-specific features may constrain their effectiveness in nuanced medical imaging tasks. In conclusion, off-the-shelf generative models, such as GLOW, are limited in their ability to accurately detect OOD semantic shift in natural images.

Furthermore, Section 6.4.3 in the thesis explores the application of generative models for the detection and quantification of covariate shifts in natural images and X-ray images. This is essential for maintaining the reliability of images and of AI systems in dynamic environments. The thesis examines several generative models, including reconstruction-based evaluation for Denoising Diffusion Probabilistic Models (DDPM), likelihood-based evaluation of Variational Autoencoders (VAEs) and Normalizing Flows (NFs). However, the analysis reveals that many of these methods exhibit a predisposition towards certain types of features, which limits their effectiveness in the detection of a broad range of covariate shifts. For instance, the DDPM, VAEs and GLOW, when evaluated with log-likelihood, perform well at the detection of noise-based shifts or any shift that increases the image high-frequency components. However, the models fail to detect any change that reduces the amplitude of high-frequency and low-frequency components, such as decrease of contrast. Meanwhile, when the GLOW model is evaluated with typicality it performs in an opposite manner. The model can detect these contrast changes, but it fails to detect changes such as additional noise. Extending the application of these models to the X-ray domain further supports these findings. The modes/severity levels of covariate shift in the X-ray setting is arranged in a subjectively increasing order, with Mode 0 having the lowest noise and Mode 5 the most. In line with the observation on natural images, it is shown that all model families perform well at detecting this increasing level of noise (covariate shift).

RQ4b: *How can high-frequency heteroscedastic image components be explicitly modeled and does this lead to improved OOD covariate shift detection performance?*

Building on the limitations discussed in RQ4a, Chapter 6 further expands on generative models for covariate shift detection. In the chapter, a novel approach is proposed that explicitly models the conditional distribution of high-frequency het-

eroscedastic components within images to improve OOD detection performance, particularly for covariate shifts. Through a single Gaussian filter decomposition step, the image can be split into low-frequency and high-frequency images. The high-frequency parts are then modeled, conditioned on the low-frequency parts, by using a Normalizing Flow (NF). By explicitly modeling these components, the approach captures the In-Distribution (ID) variability in high-frequency details that are affected by covariate shift, while alleviating the complexity in modeling the complete image distribution. This complete image distribution may not be consistently represented across different datasets, making it particularly effective in distinguishing in-distribution and Out-of-Distribution (OOD) samples. This simple setup allows for modeling the heteroscedastic image components.

To effectively exploit the newly proposed modeling of image components, a new metric is introduced. The research in Chapter 6 proposes a method to unify the typicality and log-likelihood (LL) metrics for OOD detection within NFs. The independent evaluation of these methods express predisposition to specific types in distribution shift. Utilizing the strengths of each metric, the normalized score distance (NSD) is proposed as a unification metric, by simultaneously employing typicality and LL. It is on these premises that CovariateFlow is proposed as a novel method for accurate OOD covariate shift detection.

The findings from the conducted analyses validate the hypothesis that OOD covariate shifts can be effectively identified by explicitly modeling the conditional distribution between low-frequency and high-frequency components. The proposed CovariateFlow model, designed specifically to capture this distribution, surpasses other methodologies in detecting covariate shifts in natural-image datasets like CIFAR10 vs. CIFAR10-C (74.9 % AUROC) and ImageNet200 vs. ImageNet200-C (72.2 % AUROC).

In an additional series of experiments, the proposed approach is also applied to a newly collected X-ray dataset. Covariate shift detection within X-ray imaging is of critical importance to accurately identify OOD shifts to maintain the reliability and diagnostic accuracy of medical imaging systems. The proposed CovariateFlow model shows robust performance in detecting subtle covariate shifts across various imaging modes (only outperformed by GLOW with LL due to its high sensitivity to noise-based shifts). The results highlight that, while high-parameter count architectures, like GLOW, can capture a wide range of image statistics, CovariateFlow excels particularly when evaluated using the Normalized Score Distance (NSD) metric, while additionally offering a smaller model size. This superior performance, coupled with its efficiency, underscores CovariateFlow's potential as a reliable and effective tool for ensuring the consistency and safety of X-ray imaging systems, making it a valuable tool in clinical settings.

Overall, CovariateFlow with NSD proves to be a robust and generic method for OOD detection, demonstrating its effectiveness in both natural images and for X-ray imaging data.

RQ4c: *Do the decomposed frequency components of an image contain sufficient information to improve OOD detection performance?*

The thesis provides substantial evidence in Chapter 6 supporting the efficacy of decomposing images into frequency components to improve OOD detection in both semantic and covariate OOD detection.

Semantic OOD Detection: The decomposed frequency components, particularly the low-frequency components of an image, play a crucial role in improving semantic OOD detection. By utilizing wavelet-based Normalizing Flows (Wavelet Flow), the low-frequency components help capture the general structure of the image, which is essential for identifying semantic anomalies. This method is shown to be effective in detecting unseen malignant melanoma when trained on benign melanoma data alone. The model's focus on low-frequency wavelet components allows for better differentiation between ID and OOD samples based on their semantic differences, thereby improving semantic OOD detection performance and thus enhancing the reliability of OOD detection in clinical settings.

Covariate OOD Detection: In contrast, for covariate OOD detection, the decomposed high-frequency components of an image contain essential information that improves detection performance. The research introduces the CovariateFlow model, which explicitly models high-frequency heteroscedastic components conditionally on low-frequency components. This approach captures subtle changes in high-frequency details, which are often indicative of covariate shifts. These shifts, though not always visible in low-frequency data, can be detected when the conditional high-frequency components are analyzed. CovariateFlow's ability to model these high-frequency variations lead to superior performance in detecting covariate shifts, particularly in experiments involving datasets like CIFAR10 and ImageNet200. The model effectively identifies and quantifies covariate OOD shifts, further validating the importance of frequency decomposition in enhancing OOD detection.

In conclusion, the decomposed frequency components provide valuable and sufficient information for both semantic OOD and covariate shift detection. It is shown that the low-frequency components fuel improved semantic OOD detection performance, while heteroscedastic high-frequency components provide sufficient and critical information to improve covariate OOD detection. This research presents a robust solution for OOD detection in medical and natural images by leveraging frequency decomposition, ultimately advancing the accuracy of OOD detection methods.

RQ5: Instrument Pose Estimation in X-ray

RQ5a: *How to develop a general-purpose method that is both accurate and fast for 6-DoF pose estimation?*

Significant advancements in the development of a general-purpose 6-DoF pose estimation method are proposed in Chapter 7, leveraging the latest improvements

in the YOLO object detection series. The proposed YOLOv5-6D model represents a major step forward in achieving both high accuracy and real-time performance. The model's architecture is designed to efficiently predict 2D keypoints corresponding to the vertices of an object's 3D bounding box from a single X-ray image. These predictions are then used in conjunction with a Perspective-n-Point (PnP) algorithm to estimate the object's full 6-DoF pose. The use of advanced feature extraction techniques, such as CSP-Net and BiFPN, enables the model to detect objects at multiple scales and accurately localize keypoints, even in challenging X-ray images with low contrast and variable object sizes.

A key innovation of the YOLOv5-6D model is its ability to perform pose estimation in a single-shot manner, which drastically reduces the computational overhead compared to multi-stage approaches with seperate detection and pose estimation steps. This efficiency is crucial for maintaining the speed necessary for real-time applications. The model's performances on the LINEMOD dataset and the newly acquired X-ray datasets demonstrate its ability to balance accuracy with speed, achieving up to 42 FPS, while delivering competitive accuracy on standard benchmarks. This balance makes YOLOv5-6D a strong candidate for deployment in clinical settings, where the ability to quickly and accurately estimate the pose of surgical instruments can enhance the precision of minimally invasive procedures. Additionally, the model's general-purpose design allows adaptation and application in other domains beyond medical imaging.

RQ5b: *How can X-ray data and the imaging geometry be effectively incorporated into the 6-DoF pose estimation process, and what impact does this have on the model performance?*

Incorporating X-ray imaging geometry into the 6-DoF pose estimation process is crucial for maintaining the accuracy and robustness of the model across different clinical setups. Chapter 7 emphasizes the importance of accounting for the intrinsic and extrinsic parameters of the X-ray system, such as the source-image distance (SID) and detector field of view (FOV), which can vary significantly during a procedure. These parameters directly influence the appearance of the object in the X-ray image and, if not considered, can lead to substantial errors in pose estimation. The YOLOv5-6D model addresses this challenge by incorporating these imaging parameters into the PnP-based pose estimation process, allowing the model to accurately recover the 6-DoF pose of an object, despite changes in the imaging setup.

The integration of imaging geometry not only enhances the accuracy of the pose estimation, but also improves the model's ability to generalize across different X-ray systems. Since the trained model only predicts 2D keypoints, the intrinsic parameters can be dynamically adjusted during inference to obtain the object pose. This property enables the model to handle variations in the X-ray setup without requiring retraining or manual recalibration. Such capability is particularly beneficial in medical settings, where the X-ray system's configuration may need to be frequently adjusted to optimize the visualization of different

anatomical structures during surgery. The experimental results presented in the chapter demonstrate that the YOLOv5-6D model can maintain high accuracy in natural images (LINEMOD 96.84% ADD(-S)) and in X-ray settings (Cube Val. 99.27% ADD, Screw Val. 96.87% ADD-S, Screw Test 92.41% ADD-S), even under varying acquisition geometries, with a translation error as low as 0.35 mm. The obtained level of precision is critical for applications like spinal screw placement, where even small deviations can lead to suboptimal outcomes. In contrast, the results obtained with EfficientPose, a method that does not incorporate the imaging geometry, fails completely in the X-ray setting. Overall, these results highlight the robustness and generalization ability of the proposed YOLOv5-6D model.

8.3 Future directions and research challenges

This thesis has advanced the state-of-the-art in the four major topics discussed and has opened several future directions and research challenges. These avenues promise to push the boundaries of current methodologies.

8.3.1 Pancreatic cancer treatment

The development and deployment of deep learning-based Computer-Aided Detection (CADe) systems for pancreatic cancer, particularly pancreatic ductal adenocarcinoma (PDAC), face significant challenges. The scarcity of large, diverse, high-quality labeled datasets is a critical hurdle. Future research should focus on fostering collaboration and data sharing among expert centers to overcome this limitation. Moreover, the translation of these algorithms into clinical practice necessitates further exploration into improving model robustness, reliability, and transparency in evaluation metrics. Interactive AI, which combines the pattern recognition abilities of AI with the domain knowledge of clinicians, presents a promising direction to enhance the robustness of these systems in clinical settings.

8.3.2 Uncertainty quantification in medical imaging

The work presented in this thesis has focused on improving uncertainty quantification in medical image segmentation, particularly in challenging scenarios with ambiguous ground truths. This research employs methods capable of explicitly modeling the ambiguity inherently present in the data through the use of the cVAE augmented with NFs integrated in the PU-Net framework. However, other methods that implicitly model ambiguity exist. These methods, while potentially less transparent, often yield competitive performance and can be more straightforward to implement. Future research should investigate whether there is an added benefit to explicitly modeling uncertainty beyond the performance metrics, such as improved interpretation or trustworthiness in clinical applications.

Additionally, future work should explore avenues to further improve the modeling capacity of the PU-Net's latent space. While this research has demonstrated the effectiveness of combining NFs with the PU-Net to enhance uncertainty quantification, there is still potential for refining this approach.

8.3.3 Out-of-Distribution detection

Although this thesis has contributed to refining techniques in both semantic and covariate OOD detection, there is still considerable room for improvement. Future research should focus on developing methods that can disentangle and identify the types of shifts with high accuracy. Moreover, it is essential to ensure that the correct OOD detection method is employed in the appropriate context. For example, when the primary goal is early detection of anomalies or diseases, the model should prioritize semantic OOD detection to focus on identifying new or rare conditions. Conversely, when the goal is to ensure the consistency and reliability of imaging data, covariate shift detection should be emphasized.

Further exploration is needed to enhance the sensitivity and specificity of these detection methods, particularly in complex, real-world clinical environments where both semantic anomalies and covariate shifts can occur simultaneously.

8.3.4 Pose estimation

Future research in pose estimation should explore the development of model-free methods that can predict poses with high computational efficiency without requiring extensive retraining on new data.

In the longer term, pose estimation research will largely be fueled by progress in robotics and augmented reality. The work presented in this thesis on pose estimation is one of the first to bring this technology to the medical domain, where its applications will continuously grow to ultimately be integrated into medical robotics. Such systems will be employed for a variety of surgical applications where higher precision and consistent quality are essential.

8.3.5 Medical image analysis in practice

Despite the significant advancements of deep learning models for medical image analysis in laboratory settings, their clinical adoption remains limited. This can be attributed to a range of factors such as models struggling when exposed to the variability of real-world data, the need for robust technical infrastructure and governance surrounding the use of AI in clinical practice. Additionally, several nuanced and less commonly discussed factors contribute to the limited adoption of deep learning models in clinical practice.

- *Overemphasis on generalization may be overrated:* There is a prevailing belief that models must generalize across diverse populations and imaging settings to be useful. While this is an admirable goal, it may not always be practical or necessary. Instead, localized/specialized models tailored to specific institutions or demographics can yield high performance and fast adoption.
- *Clinicians may resist, not lack literacy:* It is often assumed that clinicians need more AI training for adoption to succeed. However, resistance might not stem from a lack of literacy, but more from mistrust in models that are ”black boxes”, have inconsistent performance, or disrupt existing workflows. Clinicians’ resistance may reflect generic legitimate concerns about AI systems

interfering with their decision-making autonomy and accountability.

- *Overly ambitious multi-functionality systems and complex deployment pipelines:* Industry developers frequently aim to create multi-purpose models that address a wide range of clinical tasks, from diagnostics to triage and treatment planning. While this versatility looks desirable at first glance, it significantly increases the complexity of development and regulatory approval. Additionally, this leads to complex deployment pipelines with resource-intensive infrastructure, such as cloud computing, high-performance GPUs, and real-time data streaming. Lightweight, edge models or hybrid solutions combining on-device and cloud processing could streamline adoption.

By addressing these challenges with innovative, localized solutions, fostering trust among clinicians, and advocating for adaptive regulatory frameworks, the field can unlock the full potential of deep learning in supporting clinical practice.

8.3.6 Outlook on generalist vs. domain-specific models

Much of the work in this thesis is rooted in domain-specific advancements, while the broader trend in ML research is the rise of generalist AI, which present both opportunities and challenges for the field.

The current trajectory of mainstream ML research emphasizes the creation of increasingly large, general-purpose models capable of performing a broad set of tasks across diverse domains. These models, such as foundation models in natural images and multi-modal AI systems, demonstrate impressive versatility and high general performance. However, without additional fine-tuning, they often face significant challenges in providing the level of accuracy, robustness, and system understanding required in specialized high-risk domains such as healthcare. For these domains, task-specific models remain indispensable due to their ability to incorporate domain expertise, focus on effective use of computational resources, and yield results that are both actionable and explainable. Even in a speculative future where general intelligence achieves human-like reasoning or even exceeds it, it is likely that domain-specific models will continue to be developed. From an efficiency standpoint, the creation of specialized models tailored to particular tasks allows for the optimization of resources, both computational and data-related.

The rise of generalist AI should inspire researchers to explore new horizons. However, it is essential to recognize that the coexistence of generalist and specialist approaches is not only inevitable, but also necessary. Much like the rise of the personal computer alongside centralized mainframes, the future will see large generalist foundation models coexisting with specific, embedded, and personalized models. This evolution mirrors historical patterns, demonstrating once again how technological progress balances scale with individual adaptability. History, as always, repeats itself, albeit in different instances.

Appendices

Appendix A Additional technical details

This appendix provides foundational insights into three technical areas utilized in this thesis: an overview of uncertainty quantification methods (Section A.1), the YOLO6D architecture (Section A.2) and dequantization in NFs (Section A.3).

A.1 Overview of uncertainty quantification methods

This section provides a clear, high-level summary of widely used deep learning methods for quantifying uncertainty in classification tasks, categorized by the type of uncertainty they address.

Type	Method	Brief Description
Epistemic	Bayesian Neural Networks	Introduces uncertainty in weights by treating them as distributions from which can be sampled.
	Dropout as a Bayesian approximation	Uses dropout during inference to estimate model uncertainty.
	Deep Ensembles	Trains multiple models and averages their predictions to quantify model uncertainty.
	M-Heads	Shared backbone, M output heads (often referred to as "experts").
Aleatoric	Test-Time Data Augmentation (TTA)	Applies transformations to input data to measure variability due to noise in the data.
	Heteroscedastic Neural Networks	Directly models the variance of the output to capture data uncertainty.
	Temperature Scaling	Calibrates confidence scores to better reflect aleatoric uncertainty.

Table A.1 *High-level overview of popular deep learning-based methods for uncertainty quantification in classification tasks.*

A.2 Overview of the YOLO 6D

This sections provides an overview of the YOLO 6D processing chain.

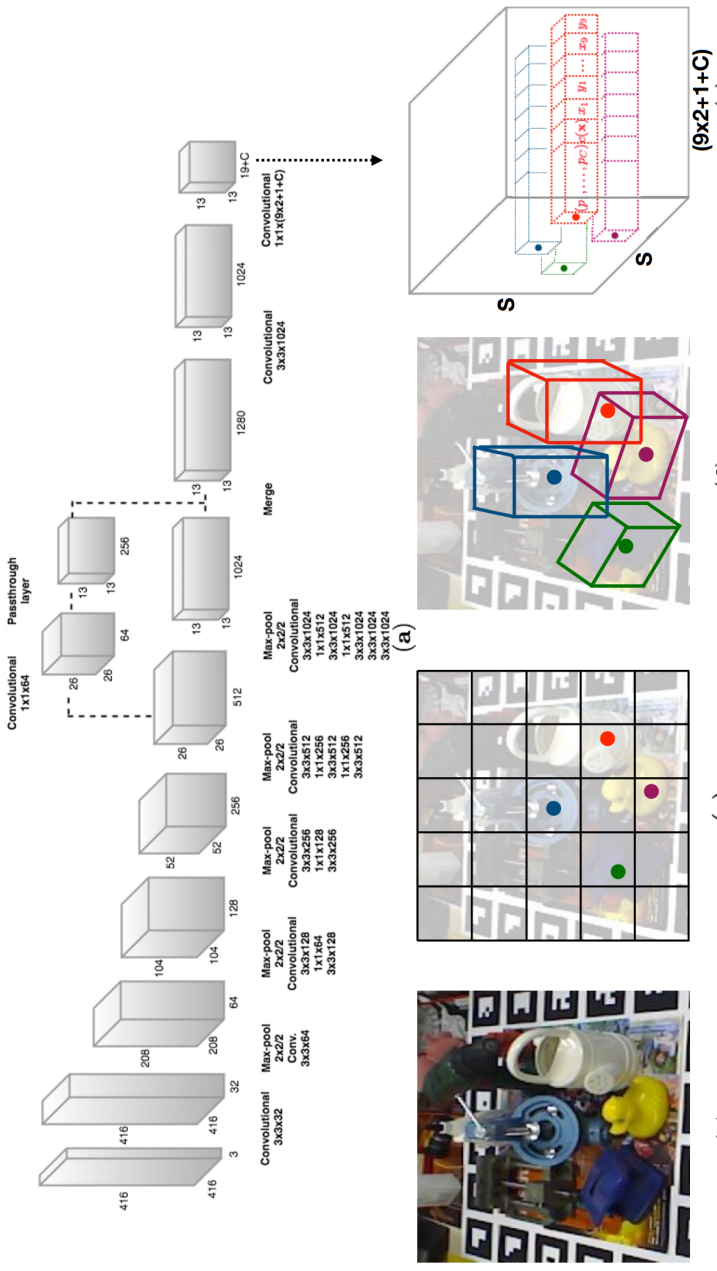


Figure A.1 *Overview of the YOLO 6D processing chain and intermediate results below. (a) The YOLO6D CNN architecture. (b) An example input image with four objects. (c) The $S \times S$ grid showing cells responsible for detecting the four objects. (d) Each cell predicts the 2D locations of the corners of the projected 3D bounding boxes in the image. (e) The 3D output tensor from the network, which represents a vector (for each cell), consisting of the 2D corner locations, the class probabilities and a confidence value associated with the prediction. Image from the original paper [301].*

A.3 Dequantization in NFs

Normalizing flows are powerful generative models that transform simple, known probability distributions into more complex ones through a series of invertible functions. They are inherently designed to model continuous probability density functions (PDFs). However, in practice, data is often discrete. Directly applying continuous models to discrete data without adjustment can lead to non-ideal solutions where the model focuses narrowly on the discrete points, thereby neglecting the overall distribution shape. Dequantization effectively converts discrete data into continuous data, allowing for the application of continuous modeling techniques. This is achieved by adding a carefully chosen noise component to each discrete data point. The continuous representation, $\tilde{\mathbf{x}}$, of a discrete variable $\mathbf{x}$ is then expressed as:

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{u} \tag{A.1}$$

where $\mathbf{u}$ is a noise variable drawn from a noise distribution $q(\mathbf{u}|\mathbf{x})$ with support in the interval $[0, 1)$. This transformation expands the domain of each discrete value $*x$ into a continuous interval. For instance, a value $x = 0$ is mapped to an interval $[0.0, 1.0)$, and $x = 1$ to $[1.0, 2.0)$, and so forth.

The probability density of the dequantized data can be modeled as:

$$p_{\mathbf{X}} = \int p_{\mathbf{X}}(\mathbf{x} + \mathbf{u})\mathrm{d}\mathbf{u} = \int \frac{q(\mathbf{u}|\mathbf{x})}{q(\mathbf{u}|\mathbf{x})} p_{\mathbf{X}}(\mathbf{x} + \mathbf{u})\mathrm{d}\mathbf{u} = \mathbb{E}_{q(\mathbf{u}|\mathbf{x})}\left[\frac{p_{\mathbf{X}}(\mathbf{x} + \mathbf{u})}{q(\mathbf{u}|\mathbf{x})}\right]. \tag{A.2}$$

This integral represents the expected value of the likelihood ratio between the probability densities of the transformed (continuous) data and the noise, averaged over the noise distribution. The choice of noise distribution $q(\mathbf{u}|\mathbf{x})$ is crucial and typically can be uniform, as it simplifies integration and ensures coverage of the entire range [0,1).

Variational dequantization [286] extends the concept of dequantization by introducing a learnable noise model instead of a fixed one. This approach leverages variational inference to optimize the noise distribution $q(\mathbf{u}|\mathbf{x})$ directly as part of the training process. The goal is to minimize the discrepancy between the empirical distribution of the dequantized data and the model's distribution. This optimization allows for more flexible and potentially more accurate modeling of complex data distributions.

In summary, dequantization is a crucial step in adapting normalizing flows for discrete data, and variational dequantization further enhances this approach by making the noise addition process adaptable for specific data characteristics.

Appendix A

Appendix B

CADe in Pancreatic Ductal Adenocarcinoma

B.1 Challenges in AI data representativeness, biases and confounders

The implementation of AI in healthcare, while promising, has encountered significant hurdles related to data representativeness, biases, and the presence of confounding factors. Despite the potential of AI to discern intricate patterns within complex datasets, its effectiveness can be compromised by these inherent limitations, which hinder widespread adoption and the generalizability of AI applications [40], [347], [348].

Generalizability remains elusive for many AI models due to the necessity for highly representative and diverse datasets that accurately mirror the target population. Variability in data collection methods and population characteristics across different healthcare settings often results in models that perform well in one setting but poorly in another. This issue was starkly highlighted in a study on the detection of abnormal chest radiographs, where the specificity of a model varied dramatically, ranging from 0.57 to 1.00 across five different datasets [349]. To address this, leveraging independent local datasets that reflect specific population characteristics as supplementary training material can potentially enhance an algorithm's adaptability prior to broader application.

Furthermore, the risk of inherent biases within training datasets poses a significant challenge. These biases may stem from a variety of factors, including incomplete data capture, insufficient sample sizes, and errors in data measurement or classification. Such biases not only undermine the reliability of model predictions, but may also perpetuate or exacerbate socioeconomic inequalities within healthcare systems [350], [351]. For example, problematic biases have been observed in non-healthcare AI applications, such as those predicting recidivism, which have demonstrated racial discrimination [352]. Similarly, in healthcare, algorithms designed to predict cardiovascular risks have shown biases against non-white populations [353]. Employing tools like the Prediction model Risk of Bias Assessment Tool (PROBAST) can aid in identifying and mitigating these risks by evaluating the bias within AI prediction models [354].

Moreover, AI systems are susceptible to forming spurious correlations or confounding relationships, deriving conclusions from coincidental or irrelevant fea-

tures present in the training data [40]. Notable examples in healthcare include algorithms misidentifying the presence of rulers or surgical markings as indicators of malignancy [355], [356]. Such misinterpretations emphasize the necessity for thorough understanding and continual refinement of the features and biases AI models learn.

In conclusion, while AI and data-driven methods holds transformative potential for medical diagnostics and treatment, ensuring the robustness, fairness, and transparency of these systems is imperative. Ongoing development and rigorous evaluation of AI technologies are essential to mitigate these challenges, enabling more reliable and equitable healthcare solutions.

B.2 Discussion PDAC resectability

Accurate assessment of tumor and vascular involvement and determining the appropriate treatment is still a growing problem with limited accurate methods for assessment. Different scoring systems have been proposed to predict vascular involvement and thus resectability status in pancreatic cancer patients [357], [358]. Given the beneficial effect of neoadjuvant treatment on cancer specific survival, a recent study developed tumor-vessel interface criteria to predict vascular involvement and resectability in borderline pancreatic cancer patients [359]. The diagnostic performance for predicting vascular involvement was evaluated between 2 readers and showed an AUROC for agreement of 0.85 - 0.88 for arterial invasion and 0.87 - 0.92 for venous invasion. In addition, CT texture analysis for predicting resectability after neoadjuvant treatment has been introduced, providing important information regarding tumor characterization by quantifying tissue heterogeneity and texture coarseness [360], [361].

This concludes the additional discussion of the work on in this thesis on data quality and PDAC resectability prediction.

Appendix C

Uncertainty quantification in medical image segmentation

C.1 Datasets for 2D uncertainty quantification

This section provides additional details on the the datasets used in the 2D PU-Net research for segmentation uncertainty quantification. Figure C.1 depicts four examples from the LIDC-IDRI [196] dataset. On the left in the figure, the 2D image slices from a CT volume containing the lesion is depicted, followed by the four labels made by four independent annotators on the right. In Figure C.2, eight examples from the Kvasir-SEG dataset [362] are depicted. The endoscopic gastrointestinal polyp images and corresponding segmentation masks can be seen.

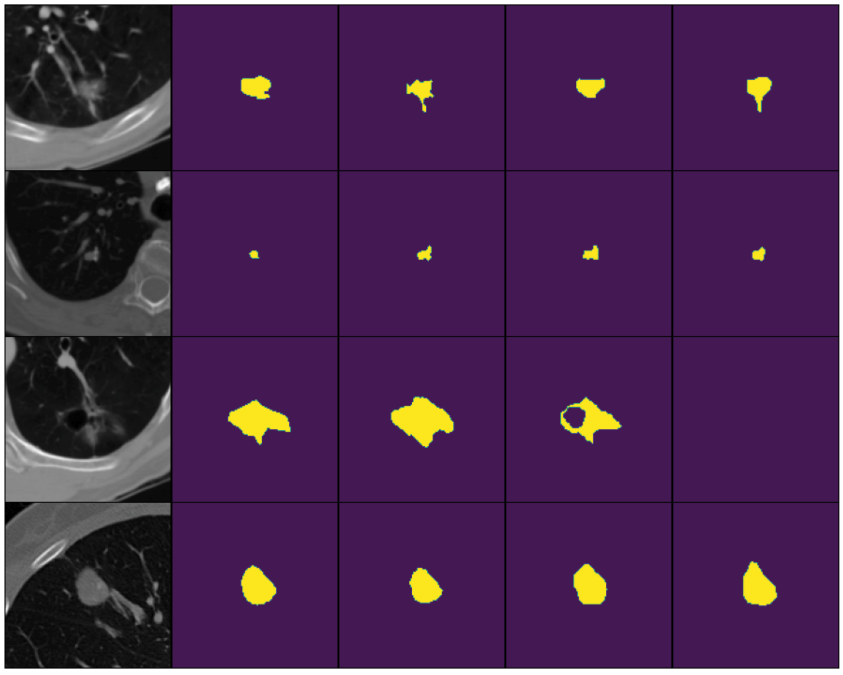


Figure C.1 *Example processed images from the LIDC-IDRI dataset.*

Figure C.2 *Example images from the Kvasir-SEG dataset.*

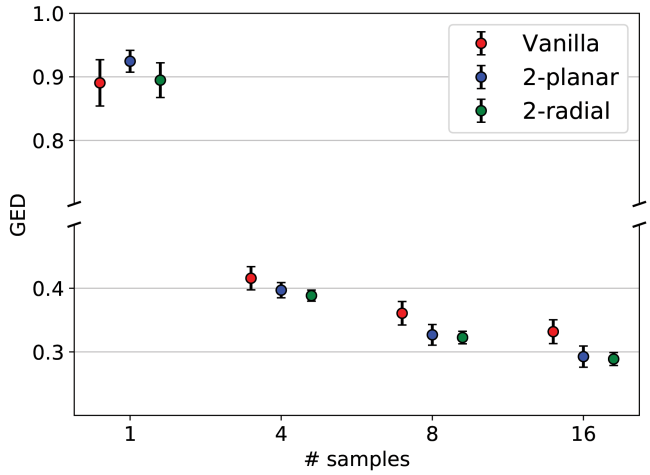
C.2 GED at different sample sizes

The GED evaluation is dependent on the number of reconstructions sampled from the prior distribution. Figure C.3 depicts this relationship for the vanilla, 2-planar and 2-radial posterior models. The error bars in the plot originate from the deviations in the results when training with ten-fold cross validation. One can observe that with increasing sample size, the GED as well as the associated uncertainty decrease. This is also the case when the posterior is augmented with a 2-planar or 2-radial flow. Particularly, the uncertainty in the GED evaluation significantly decreases.

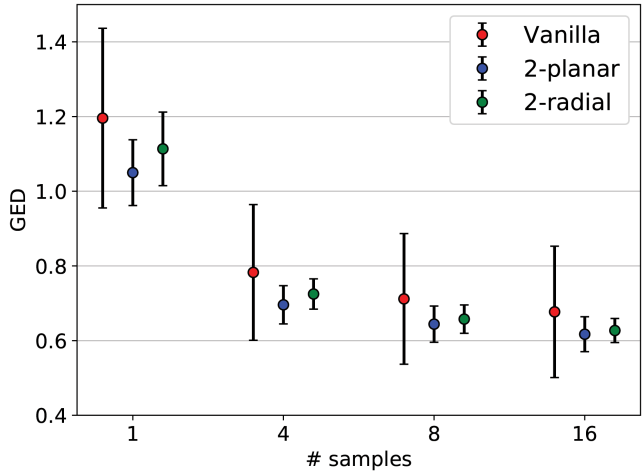
C.3 Prior distribution variance

An investigation into whether the prior distribution captures the degree of ambiguity in the input images was conducted. For every input image $\mathbb{X}$, we obtain a latent L-dimensional mean and standard deviation vector of the prior distribution $P(\boldsymbol{\mu}, \boldsymbol{\sigma}|\mathbb{X})$.

The mean of the latent prior variance vector μ_{LV}, is obtained from the input images in an attempt to quantify this uncertainty. Figure C.4 shows this for several different input images of the test set. As can be seen, the mean variance over



(a) *LIDC test set*



(b) *Kvasir-SEG test set*

Figure C.3 *The GED based on sample size evaluated on the vanilla, 2-planar and 2-radial models.*

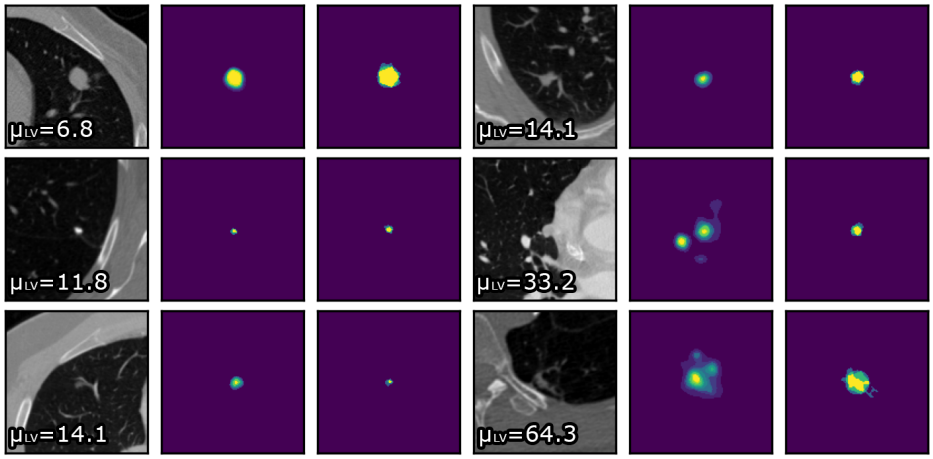


Figure C.4 *Depiction of the prior variance for various images. In the CT image, the mean of the prior distribution variance of the 2-planar model is shown. The input CT image (column 1 and 4), its average segmentation prediction from 16 samples (column 2 and 5) and ground truth from four annotators (column 3 and 6) are depicted.*

the latent prior increases along with a subjective assessment of the annotation difficulty.

C.4 Dataset for 3D uncertainty quantification

This section provides additional details on the the datasets used in the 3D PU-Net research for 3D segmentation uncertainty quantification. In line with the work on 2D segmentation uncertainty, we employ the LIDC-IDRI [196] dataset, however, the complete 3D nodule volume is extracted instead of 2D slices. We preprocess the CT scans by clustering all nodule annotations for a scan through a computation of a distance measure between the annotations. If an annotation is within one-voxel spacing of that particular CT scan from another annotation, it is grouped to belong to the same nodule. The scan is resampled to 0.5 mm along the x, y-dimensions and 1 mm along the z-dimension to obtain uniform voxel spacing between all samples. This is followed by cropping the CT scan and resulting annotations based on the center of the first assessor's mask with a dimension of 96×180×180 voxels in the z, x, y-dimensions. Finally, if the nodule does not have at least four annotations, the ground-truth (GT) masks are filled with empty annotations. Figure 4.6 depicts four ground-truth annotations of a nodule in the CT scan. Figure C.5 depicts a 3D visualization of the mean ground-truth annotation for another nodule from different angles.

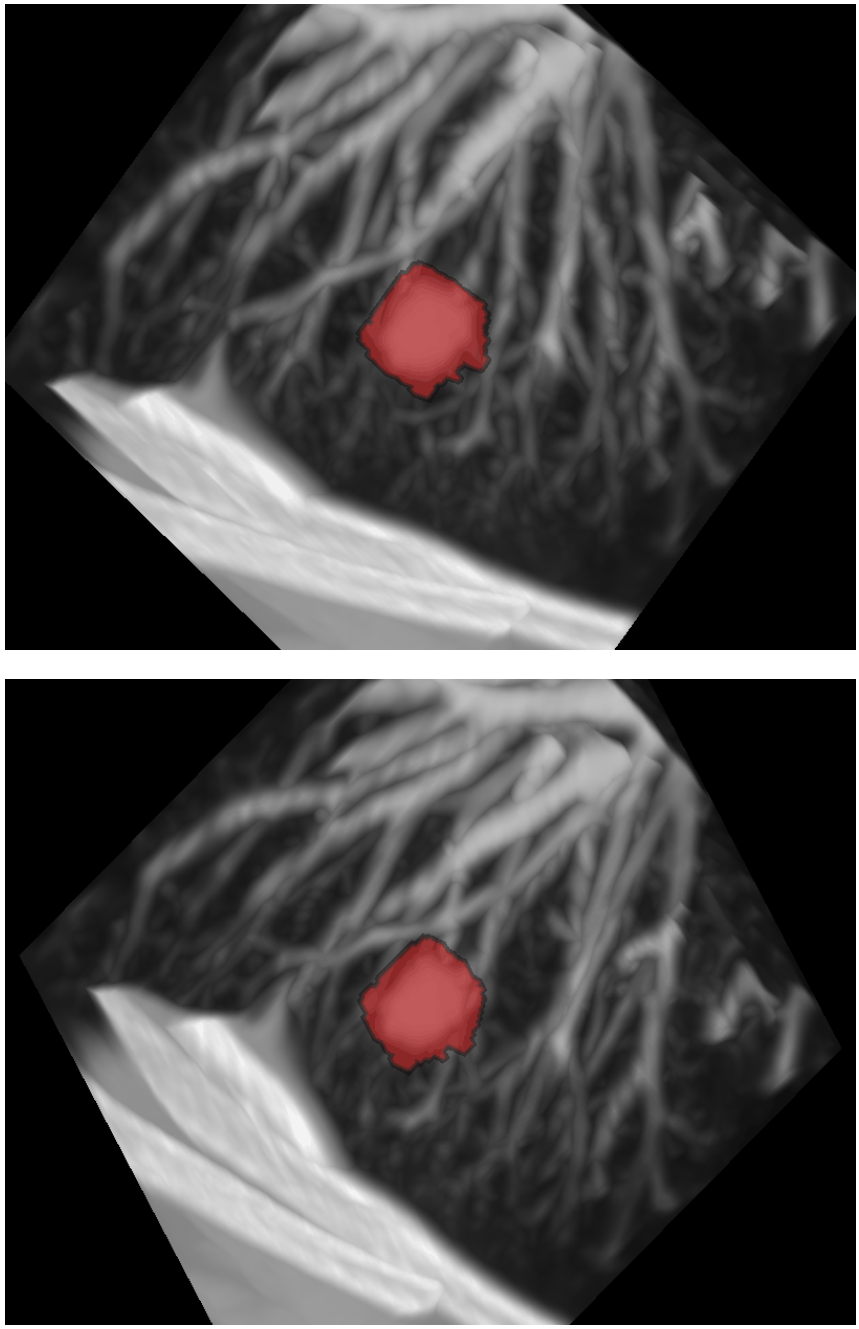


Figure C.5 *Visualization from different angles of mean 3D segmentation of a lung nodule.*

Appendix C

Appendix D OOD detection

This appendix providing more information in relation to the OOD detection work is organized as follows. Section D.1 describes the implementation details of all the models employed in this research. Section D.2 has a step-by-step rundown on how we obtain the Normalized Score Distance. Section D.3 provides detailed results on CIFAR10 and CIFAR10-C of the experiments and Section D.4 results on ImageNet200 and ImageNet200-C as described in the Experiments section of Chapter 6 on covariate shift detection. Additional information about the X-ray experiments are provided in Section D.5 and Section D.6. Finally, we provide a series of additional ablation experiments in Section D.7.

D.1 Implementation details

In this section, we detail the unsupervised training methodologies employed for five distinct baseline models and CovariateFlow aimed at OOD detection.

VAE and Adversarial VAE: The VAE is trained to minimize the standard ELBO [259] loss. Model evaluations using SSIM and KL-divergence presented the best AUROC results. The AVAE model integrates adversarial training [48] into the variational autoencoder framework to enhance its capability in generating realistic samples. For OOD detection, one can leverage the reconstruction loss (Mean Squared Error (MSE)), the KL-divergence and the discriminative loss to compute a OOD score. We adopt the implementation described in [267]. In both the VAE and AVAE we employ a 4 layer deep network with a latent dim=1024. The models were trained for 200 epochs following a cosine annealing learning rate scheduler.

VAE-FRL: The VAE with frequency-regularized learning (FRL) [291] introduces decomposition and training mechanism which incorporates high-frequency information into training and guides the model to focus on semantically relevant features. This proves effective in semantic OOD detection. We employ the pre-trained model as publicly available[1]. For the CIFAR10 experiments, the model consists of a standard 3 layer deep VAE with strided convolutional down-sampling

[1] https://github.com/mu-cai/FRL/tree/main

layer, transposed convolutional up-sampling and ReLu non-linear functions. The model has a latent dimension of 200. The OOD score is obtained by the log-likelihood (lower bound in the case of the VAE) minus the image complexity. The formulation is given as

$$S(x) = -\log p_\theta(x) - L(x),$$

where $L(x)$ is the complexity score derived from data compressors [253], such as PNG.

Denoising Diffusion Probabilistic Model: We implemented the Denoising Diffusion Probabilistic Model (DDPM) following the specifications outlined in [269] and as publicly available [2]. The method employs a time-conditioned UNet [12] architecture with a simplified training objective where the variance is set to time-dependent constants and the model is trained to directly predict the noise ϵ at each timestep t:

$$L(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t)\|^2\right]. \tag{D.1}$$

We aim to reconstruct an input x_t across multiple time steps (t), utilizing the DDPM sampling strategy which necessitates t steps for each reconstruction $\hat{x}_0, t$, with each step involving a model evaluation. To enhance efficiency, we leverage the PLMS sampler [363], a recent advancement in fast sampling for diffusion models, which significantly decreases the number of required sampling steps while preserving or enhancing the quality of samples. For evaluating the reconstructions, we employ both the mean-squared error (MSE) between the reconstructed and the input image, and the Learned Perceptual Image Patch Similarity (LPIPS) metric [364], the latter of which assesses perceptual similarity through deep feature distances. For each of the N reconstructions we compute these 2 similarity measurements. Finally we average these scores (over the two metrics and all the reconstructions) to derive an OOD score for each input, integrating both quantitative and perceptual accuracy assessments.

The model architecture is implemented exactly as described in [269]. For training, we set $T = 1000$ and employed a linear noise schedule, with β_t ranging from 0.0015 to 0.0195. The training process spanned 300 epochs, utilizing the Adam optimizer with a learning rate of $2.5e^{-5}$. During the testing, we utilized the PLMS sampler configured to 100 timesteps and, in line with AnoDDPM [261], we only test reconstructions from $T = 250$. Since we do not intend to detect semantic anomalies in this work and are more interested in high frequency image components, we focus on reconstructions later in the schedule.

Finally, we experiment with the DDPM model trained on CIFAR10 and evaluated at different reconstruction starting points. Figure D.6 depicts our results obtained with different reconstruction starting points and the average AUROC across all the degradations in CIFAR10-C.

[2] https://github.com/marksgraham/ddpm-ood

GLOW: Normalizing Flows enable OOD detection by modeling the ID data distributions with invertible transformations through a maximize the log-likelihood training objective. We employ the GLOW [206] architecture, as publicly available [3], in this study. Additionally, following the recent work in typicality (Section 6.4.5), we train our model with the *approximate mass* augmented log-likelihood objective as described in [258]. We incorporate the *approximate mass* as a component in the loss function formulation. Let $L(x;\theta) = \log(p(x;\theta))$ denote the average log-likelihood (LL) of the model, parameterized by θ, evaluated over a batch of input data x. Our revised training objective is expressed as:

$$\min_{\theta} \left(-L(x;\theta) + \alpha \left\| \frac{\partial L(x;\theta)}{\partial x} \right\| \right) \tag{D.2}$$

where $\alpha > 0$ signifies a hyperparameter that balances the trade-off between local enhancement of the likelihood and reduction of the gradient magnitude. We employ $\alpha = 2$ in the GLOW implementation. At test time, we compute the per sample LL and gradient score. These components are used to compute the NSD as described in Section 6.5.2.A.

CovariateFlow: Section 6.5.2 describes the CovariateFlow model proposed in this work. Figure 6.9 depicts the architecture and general flow of information during training and when computing the OOD scores. Figure 6.10 illustrates a detailed diagram of the low-frequency conditioned coupling steps employed in the model. Additionally, following the image decomposition through the Gaussian filtering, we encode the individual components as 16-bit depth data to avoid information loss. Our model is completely invertible and can thus also generate signal-dependent high-frequency components. The models are prepared following the typicality augmented training objective (Equation D.2). We use an Adam optimizer (starting $lr = 5e^{-4}$) with a one-cycle annealing learning rate scheduler for 300 epochs across all our experiments. The code for the model is available at https://github.com/covariateflow/CovariateFlow.

D.2 Detailed analysis of the normalized score distance (NSD)

This section details the computation of the NSD from the LL and typicality score. Figure D.5 depicts this process through the evaluation of the GLOW model applied to three different OOD covariate shifts. In Figure D.1 the LL and typicality (gradient score) of the model subject to Gaussian Noise can be seen. Following the process described in Section 6.5.2.A, column 2 depicts the standardization of both scores using validation statistics. This is followed by converting the scores to absolute distance from the expected mean in column 3. The LL distance and gradient score distance can then simply be added to obtain a unified distance

[3] https://github.com/y0ast/Glow-PyTorch

(Figure D.1). The same flow is depicted in Figure D.3 and Figure D.4 for Contrast change. Following this standardized approach, the change in each measure (LL and gradient score) w.r.t. the validation statistics are utilized and combined to provide a single and effective OOD score. All the results depicted in The Figure D.5 depicts the ID CIFAR10 test scores vs. the OOD CIFAR10-C scores.

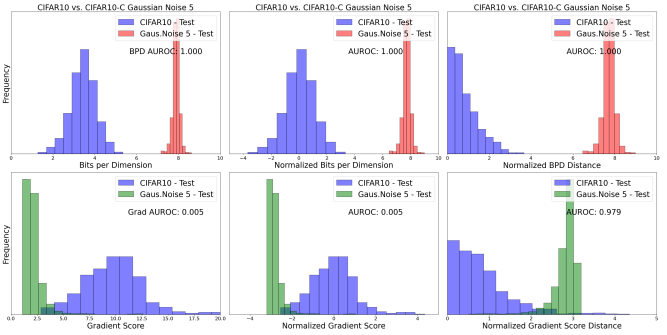


Figure D.1 *Top row:* ***Log-likelihood*** *of CIFAR10 vs. Gaussian Noise 5, the normalized LL and the absolute value of the normalized LL. Bottom row:* ***Gradient score*** *of CIFAR10 vs. CIFAR10-C Gaussian Noise 5, normalized gradient score and the absolute value of the normalized gradient score.*

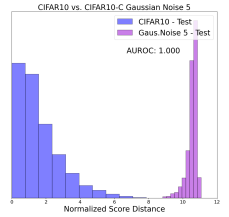


Figure D.2 *The sum of the normalized LL distance and the normalized gradient distance shown as a unified normalized score distance (NSD)*

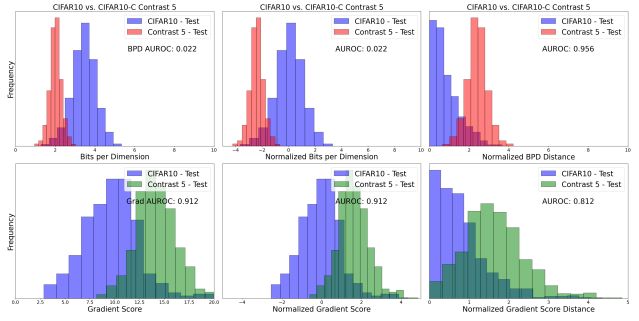


Figure D.3 *Top row:* ***Log-likelihood*** *of CIFAR10 vs. Contrast 5, the normalized LL and the absolute value of the normalized LL. Bottom row:* ***Gradient score*** *of CIFAR10 vs. Contrast 5, normalized gradient score and the absolute value of the normalized gradient score.*

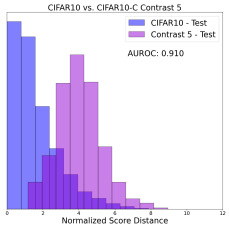


Figure D.4 *The sum of the normalized LL distance and the normalized gradient distance shown as a unified normalized score distance (NSD)*

Figure D.5 *Histograms of test results of GLOW trained on CIFAR10 and evaluated on CIFAR10-C Gaussian Noise, Gaussian Blur and Contrast. The unification between log-likelihood and typicality to compute the Normalized Score Distance (NSD) is depicted.*

D.3 Detailed results on CIFAR10 vs. CIFAR10-C

The following section presents detailed results obtained with various models on our experiments with ID CIFAR10 and CIFAR10-C as OOD.

Our analysis examines the reconstruction capabilities of the DDPM across various initial time steps, T. Figure D.6 presents the mean AUROC curve calculated for reconstructions assessed using the LPIPS, MSE, or a combination of LPIPS and MSE metrics at each time step. Notably, at larger time steps (e.g., $T = 250$), the distinction in average reconstruction error between the ID CIFAR10 test set and the OOD CIFAR10-C dataset becomes less pronounced, leading to inferior OOD detection performance. This phenomenon is attributable to the high-level image perturbations characteristic of OOD data, which are predominantly addressed in the final stages of the diffusion process. In Contrast, initial diffusion stages focus on generating lower-level image semantics, resulting in reconstructions that significantly diverge from the test image, particularly in terms of low-frequency components.

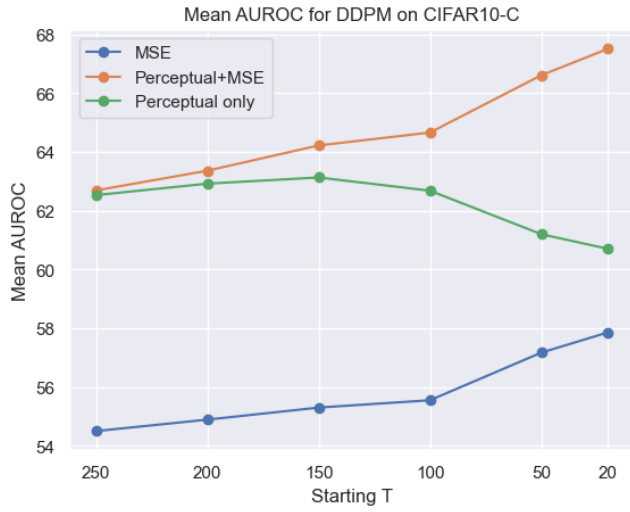


Figure D.6 *Results obtained with the DDPM on CIFAR10 and CIFAR10-C. The figure depicts mean AUROC obtained from reconstructions at different starting points, T. With covariate shift being predominantly change in high-frequency components, reconstructions starting at T=20 shows the best performance.*

Figures D.5, D.6, and Table D.1, highlight the distinct sensitivities of log-likelihood (LL) and gradient scores when applied to GLOW under severe Gaussian Noise conditions, as depicted in Figure D.1. These metrics diverge in their assessment, with LL clearly identifying distorted images as OOD, whereas gradient scores suggest such images are more typical than even the ID data. Conversely, Figure D.3 demonstrates the opposite trend for contrast changed images, where LL overestimates their likelihood relative to ID data, but gradient scores accurately

classify them as OOD. These observations corroborate Zhang *et al.*'s theoretical insights [257] about the propensity of certain model-metric combinations to misjudge the probability of natural images. To address these discrepancies, we introduce the NSD metric, which synthesizes LL and gradient movements into a unified OOD detection metric. Figures D.2 and D.4 validate the NSD metric's effectiveness in discerning OOD samples across both conditions, with extended results available in the supplementary material.

Table D.1 depicts the AUROC for 3 degradations (each severity) from CIFAR-10C that summarizes the performance of all the models employed in this work. Figure D.7 additionally depicts the average AUROC of all the models at each severity. We also present the complete performance evaluation of all the models on CIFAR10-C on all the degradtions and at every severity level. The results are depicted in order of presentation: DDPM T150-LPIPS (D.2), DDPM T20-LPIPS+MSE (D.3), VAE (Table D.4), AVAE (Table D.5), GLOW-LL (Table D.7), GLOW-Typicality (Table D.8), GLOW-NSD (Table D.9), CovariateFlow-LL (Table D.10), CovariateFlow-Typicality (Table D.11) and CovariateFlow-NSD (Table D.12).

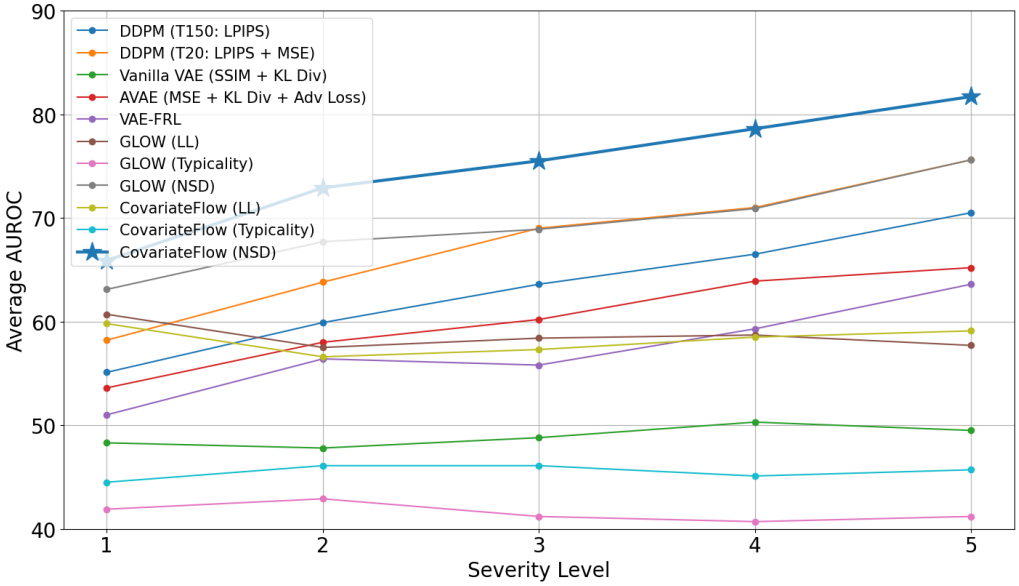


Figure D.7 *Illustration of model performance (average AUROC across all 19 degradations) per severity level.*

Method	CIFAR10 ID Gaussian Noise 1/ 2/3/4/5	CIFAR10-C OOD Gaussian Blur 1/ 2/3/4/5	 Contrast 1/ 2/3/4/5	 All Shifts Average↑ FPR95↓
Reconstruction				
DDPM [269] (T150: LPIPS)	53.7/ 59.2/ 66.3/ 70.1/ 73.5	50.7/ 68.3/ 82.4/ 92.1/ 98.8	50.1/ 50.0/ 51.2/ 50.5/ 50.2	63.1/ 83.9
DDPM [269] (T20: LPIPS + MSE)	75.6/ 91.7/ 98.2/ 99.1/ 99.6	48.7/ 58.6/ 70.6/ 82.2/ 95.1	48.2/ 48.5/ 48.3/ 46.2/ 45.0	67.5/ 75.2
Explicit Density				
Vanilla VAE [259] (SSIM + KL Div)	64.2/ 79.0/ 91.7/ 95.6/ 97.9	43.0/ 24.6/ 19.2/ 15.4/ 10.2	23.8/ 5.4/ 2.0/ 0.5/ 0.0	48.9/ 83.3
AVAE [267] (MSE + KL Div + Adv Loss)	58.4/ 68.7/ 80.6/ 86.1/ 90.6	45.5/ 34.0/ 30.7/ 28.2/ 25.7	34.1/ 38.3/ 43.5/ 48.3/ 50.3	60.2/ 73.1
GLOW [206] (LL)	100.0/ 100.0/ 100.0/ 100.0/ 100.0	44.3/ 21.3/ 14.2/ 9.8/ 5.2	39.2/ 20.9/ 14.9/ 8.8/ 2.2/	57.7/69.5
GLOW [206] (Typicality) [258]	0.0/ 0.0/ 0.2/ 0.2/ 0.47	55.4/ 65.1/ 71.1/ 76.5/ 85.53	60.4/ 66.1/ 71.5/ 77.7/ 91.2	41.6/85.8
GLOW (Normalized Distance)	100.0/ 100.0/ 100.0/ 100.0/ 100.0	48.7/ 52.7/ 60.8/ 69.2 / 82.0	49.6/ 57.7/ 64.4/ 74.0/ 91.0	69.3/65.7
CovariateFlow (LL)	100.0/ 100.0/ 100.0/ 100.0/ 100.0	42.1/ 16.4/ 10.5/ 7.4/ 4.4	31.2/ 11.0/ 6.3/ 2.8/ 0.5	58.3/63.5
CovariateFlow (Typicality) [258]	7.0/ 1.8/ 0.4/ 0.2/ 0.1	56.1/ 75.9/ 81.0/ 84.6/ 89.5	63.3/ 77.6/ 81.9/ 85.8/ 91.1	45.5/83.8
CovariateFlow (NSD)	99.5/ 99.7/ 99.8/ 99.8/ 99.8	50.1/ 69.4/ 77.3/ 82.7/ 89.4	55.3/ 76.7/ 83.9/ 90.1/ 95.9	**74.9**/**61.7**

Table D.1 *AUROC scores of various methods on detecting OOD covariate shift on CIFAR10 vs. CIFAR10-C. Note, only 3 degradations at the 5 severity levels are depicted but the average AUROC and FPR95 is computed across all degradations in the dataset.*

Severity / Metric	1 AUROC↑/ FPR95↓	2 AUROC↑/ FPR95↓	3 AUROC↑/ FPR95↓	4 AUROC↑/ FPR95↓	5 AUROC↑/ FPR95↓	Average AUROC↑/ FPR95↓
Gaussian Noise	53.74 / 95.2	59.17 / 93.3	66.26 / 89.5	70.09 / 88.6	73.54 / 84.0	64.56 / 90.12
Shot Noise	52.78 / 95.8	53.83 / 95.2	61.61 / 91.9	65.4 / 88.0	72.95 / 81.8	61.31 / 90.54
Speckle Noise	52.19 / 96.0	52.19 / 96.0	59.51 / 92.6	67.72 / 86.5	74.94 / 77.8	61.31 / 89.78
Impulse Noise	61.15 / 92.2	68.9 / 86.1	76.25 / 77.1	86.74 / 53.6	91.72 / 39.6	76.95 / 69.72
Defocus Blur	50.17 / 96.3	54.95 / 93.7	67.72 / 89.4	81.99 / 70.6	96.95 / 14.1	70.36 / 72.82
Gaussian Blur	50.71 / 96.3	68.3 / 86.5	82.39 / 68.6	92.07 / 39.8	98.81 / 5.0	78.46 / 59.24
Glass Blur	64.67 / 85.1	63.36 / 85.1	57.20 / 87.1	73.2 / 89.3	66.29 / 77.6	64.94 / 84.84
Motion Blur	60.81 / 92.1	74.48 / 77.3	83.69 / 63.2	84.0 / 66.7	90.22 / 45.9	78.64 / 69.04
Zoom Blur	71.56 / 80.9	74.32 / 76.4	80.52 / 70.9	84.12 / 61.8	89.68 / 49.0	80.04 / 67.8
Snow	51.64 / 96.3	52.8 / 95.8	51.38 / 95.2	48.4 / 96.3	46.46 / 95.3	50.14 / 95.78
Fog	56.53 / 93.9	70.66 / 80.8	81.27 / 61.4	89.63 / 39.6	94.99 / 20.7	78.62 / 59.28
Brightness	50.2 / 96.0	48.75 / 96.4	47.56 / 96.9	46.16 / 96.9	44.4 / 97.4	47.41 / 96.72
Contrast	50.12 / 95.2	49.97 / 94.7	51.18 / 95.0	50.49 / 92.8	50.19 / 95.0	50.39 / 94.54
Elastic Transform	57.55 / 93.3	58.12 / 92.7	63.66 / 88.6	60.5 / 89.1	53.99 / 93.5	58.76 / 91.44
Pixelate	52.0 / 94.4	53.44 / 93.2	54.06 / 93.3	58.04 / 89.5	64.96 / 85.9	56.5 / 91.26
JPEG Compression	54.9 / 93.9	56.57 / 93.0	57.6 / 92.7	57.65 / 93.0	60.04 / 90.2	57.35 / 92.56
Spatter	55.48 / 93.5	60.96 / 92.6	61.25 / 89.4	56.48 / 90.3	63.88 / 86.1	59.61 / 90.38
Saturate	64.44 / 90.7	73.69 / 84.1	45.51 / 97.1	40.4 / 98.0	36.85 / 98.2	52.18 / 93.62
Frost	46.57 / 97.0	47.15 / 97.0	54.29 / 94.6	56.36 / 95.0	64.77 / 91.6	53.83 / 95.04

Table D.2 *The performance of the Denoising Diffusion Probabilistic Model (DDPM) in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated with a starting* ***T=150 and using the LPIPS reconstruction*** *metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 63.2% and a False Positive Rate at 95% True Positive Rate (FPR95) of 84.0%.*

Severity / Metric	1 AUROC↑/ FPR95↓	2 AUROC↑/ FPR95↓	3 AUROC↑/ FPR95↓	4 AUROC↑/ FPR95↓	5 AUROC↑/ FPR95↓	Average AUROC↑/ FPR95↓
Gaussian Noise	75.55 / 79.1	91.74 / 30.8	98.2 / 5.0	99.14 / 2.2	99.57 / 1.3	92.84 / 23.68
Shot Noise	67.34 / 85.8	79.52 / 65.6	95.65 / 17.2	97.69 / 8.0	99.23 / 2.1	87.89 / 35.74
Speckle Noise	68.09 / 79.8	68.09 / 79.8	93.4 / 25.6	97.77 / 8.2	99.2 / 2.3	85.31 / 39.14
Impulse Noise	88.62 / 41.7	98.31 / 5.0	99.64 / 1.0	99.99 / 0.1	100.0 / 0.0	97.31 / 9.56
Defocus Blur	48.79 / 94.6	49.96 / 95.5	57.66 / 92.9	70.48 / 88.9	91.46 / 49.2	63.67 / 84.22
Gaussian Blur	48.65 / 95.4	58.56 / 93.4	70.64 / 87.6	82.24 / 76.6	95.11 / 30.3	71.04 / 76.66
Glass Blur	75.7 / 79.2	73.6 / 80.6	64.93 / 86.5	79.38 / 76.6	71.9 / 81.0	73.1 / 80.78
Motion Blur	54.68 / 93.4	64.04 / 91.5	72.44 / 86.0	72.64 / 88.1	79.69 / 81.8	68.7 / 88.16
Zoom Blur	62.21 / 91.4	65.0 / 89.7	69.71 / 85.3	73.18 / 84.2	78.89 / 79.5	69.8 / 86.02
Snow	58.47 / 93.8	66.97 / 89.0	63.95 / 91.3	59.71 / 94.4	56.71 / 96.7	61.16 / 93.04
Fog	54.46 / 92.1	62.74 / 86.0	69.58 / 77.9	77.22 / 73.8	86.11 / 54.0	70.02 / 76.76
Brightness	52.34 / 94.9	51.03 / 95.5	51.96 / 95.9	51.58 / 96.2	50.44 / 97.5	51.47 / 96.0
Contrast	48.24 / 95.4	48.47 / 95.4	48.31 / 95.6	46.2 / 95.9	44.95 / 95.9	47.23 / 95.64
Elastic Transform	52.07 / 94.0	52.03 / 93.8	55.79 / 93.8	52.77 / 93.1	50.06 / 93.7	52.54 / 93.68
Pixelate	50.41 / 95.4	54.13 / 93.9	53.29 / 93.4	57.41 / 91.5	61.82 / 91.1	55.41 / 93.06
JPEG Compression	54.49 / 93.8	55.73 / 92.4	56.5 / 92.8	56.4 / 91.5	58.97 / 90.4	56.42 / 92.18
Spatter	57.57 / 91.6	66.94 / 83.2	72.93 / 79.8	63.86 / 84.4	76.67 / 67.9	67.59 / 81.38
Saturate	53.83 / 95.9	61.16 / 94.6	51.18 / 95.4	55.59 / 92.2	61.01 / 91.5	56.55 / 93.92
Frost	51.76 / 96.5	54.24 / 97.3	61.21 / 95.4	63.83 / 94.9	70.32 / 86.5	60.27 / 94.12

Table D.3 *The performance of the Denoising Diffusion Probabilistic Model (DDPM) in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated with a starting* ***T=20 and using the MSE + LPIPS reconstruction*** *metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 67.8% and a False Positive Rate at 95% True Positive Rate (FPR95) of 75.5%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Gaussian Noise	64.2 / 82.2	79.0 / 58.6	91.7 / 26.9	95.6 / 15.0	97.9 / 7.4	85.68 / 38.02
Shot Noise	58.4 / 88.7	66.4 / 79.3	84.5 / 48.3	90.4 / 33.9	96.3 / 14.7	79.2 / 52.98
Speckle Noise	58.8 / 88.8	72.4 / 74.9	79.4 / 64.6	90.1 / 42.7	95.6 / 22.8	79.26 / 58.76
Impulse Noise	75.9 / 67.6	91.0 / 33.0	97.2 / 11.6	99.8 / 0.9	100.0 / 0.1	92.78 / 22.64
Defocus Blur	42.8 / 96.2	32.2 / 97.6	24.6 / 98.5	20.2 / 98.8	13.1 / 99.4	26.58 / 98.1
Gaussian Blur	43.0 / 96.2	24.6 / 98.5	19.2 / 98.9	15.4 / 99.3	10.2 / 99.6	22.48 / 98.5
Glass Blur	60.9 / 92.4	58.5 / 93.9	44.3 / 96.2	64.3 / 92.2	49.4 / 95.8	55.5 / 94.1
Motion Blur	31.8 / 97.7	24.0 / 98.6	18.6 / 99.1	18.6 / 99.1	14.8 / 99.4	21.56 / 98.78
Zoom Blur	27.5 / 98.2	23.9 / 98.4	21.0 / 98.6	18.5 / 98.8	15.7 / 99.1	21.32 / 98.62
Snow	58.2 / 89.4	65.8 / 81.2	71.1 / 75.6	70.2 / 75.6	65.3 / 83.7	66.12 / 81.1
Fog	31.9 / 98.1	16.6 / 99.4	10.2 / 99.6	7.0 / 99.7	4.2 / 99.7	13.98 / 99.3
Brightness	52.4 / 94.6	54.6 / 93.5	56.7 / 92.6	58.0 / 92.4	59.0 / 92.4	56.14 / 93.1
Contrast	23.8 / 98.8	5.4 / 99.8	2.0 / 99.9	0.5 / 100.0	0.0 / 100.0	6.34 / 99.7
Elastic Transform	37.3 / 97.2	33.5 / 97.7	27.6 / 98.2	27.2 / 98.4	30.4 / 98.0	31.2 / 97.9
Pixelate	49.0 / 95.3	48.1 / 95.6	47.5 / 95.8	45.9 / 96.0	43.1 / 96.3	46.72 / 95.8
JPEG Compression	49.7 / 95.1	48.5 / 95.5	47.9 / 95.6	47.4 / 95.8	46.7 / 95.7	48.04 / 95.54
Spatter	57.6 / 89.7	68.4 / 76.8	77.3 / 61.6	73.8 / 78.1	83.2 / 66.1	72.06 / 74.46
Saturate	42.4 / 96.7	42.4 / 96.9	59.1 / 91.9	69.1 / 86.3	76.3 / 82.1	57.86 / 90.78
Frost	51.8 / 93.6	53.1 / 92.7	48.1 / 93.1	44.5 / 94.2	38.4 / 95.2	47.18 / 93.76

Table D.4 *The performance of the VAE model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***MSE + KL-divergence + Adversarial*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 48.9% and a False Positive Rate at 95% True Positive Rate (FPR95) of 83.3%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Gaussian Noise	58.4 / 88.8	68.7 / 76.8	80.6 / 55.2	86.1 / 42.4	90.6 / 29.8	76.9 / 58.6
Shot Noise	54.9 / 91.8	59.8 / 87.1	73.4 / 69.2	79.7 / 58.6	88.5 / 38.5	71.3 / 69.0
Speckle Noise	55.0 / 91.8	63.7 / 83.7	69.4 / 78.1	80.5 / 63.2	89.2 / 45.0	71.6 / 72.3
Impulse Noise	66.4 / 81.3	80.3 / 60.0	89.4 / 37.4	97.6 / 9.8	99.5 / 1.8	86.7 / 38.1
Defocus Blur	45.1 / 96.1	38.5 / 96.7	33.8 / 97.4	31.2 / 97.7	27.7 / 97.9	35.2 / 97.2
Gaussian Blur	45.5 / 95.9	34.0 / 97.4	30.7 / 97.7	28.2 / 98.0	25.7 / 98.1	32.8 / 97.4
Glass Blur	57.2 / 93.0	55.10 / 94.1	46.11 / 95.6	60.0 / 92.7	50.2 / 94.6	53.7 / 94.0
Motion Blur	38.1 / 96.7	33.6 / 97.4	30.8 / 97.7	30.8 / 97.7	28.8 / 97.8	32.4 / 97.5
Zoom Blur	35.5 / 97.2	33.2 / 97.4	31.5 / 97.5	30.0 / 97.6	28.2 / 97.7	31.7 / 97.5
Snow	68.0 / 79.4	94.2 / 21.2	96.7 / 12.9	99.8 / 0.8	100.0 / 0.0	91.7 / 22.9
Fog	38.9 / 97.0	43.8 / 94.9	49.9 / 92.7	52.7 / 90.3	55.2 / 86.8	48.1 / 92.2
Brightness	62.6 / 86.2	85.6 / 47.9	97.8 / 8.7	99.8 / 0.8	100.0 / 0.0	89.2 / 28.7
Contrast	34.1 / 97.7	38.3 / 96.7	43.5 / 96.3	48.3 / 95.7	50.3 / 95.3	42.9 / 96.3
Elastic Transform	42.2 / 96.4	39.9 / 96.5	36.1 / 97.1	36.1 / 96.9	38.0 / 96.7	38.5 / 96.7
Pixelate	48.7 / 95.2	48.0 / 95.2	47.8 / 95.6	47.5 / 95.7	45.1 / 95.8	47.4 / 95.5
JPEG Compression	49.7 / 95.2	48.7 / 95.5	48.5 / 95.4	48.0 / 95.6	47.7 / 95.3	48.5 / 95.4
Spatter	55.2 / 91.4	63.2 / 82.5	74.8 / 66.1	66.9 / 84.3	75.7 / 77.0	67.2 / 80.0
Saturate	65.0 / 80.5	72.6 / 70.6	63.3 / 88.1	90.7 / 41.8	98.9 / 5.3	78.1 / 57.3
Frost	97.9 / 7.8	100.0 / 0.2	100.0 / 0.1	99.9 / 0.3	99.7 / 1.2	99.5 / 1.9

Table D.5 *The performance of the AVAE model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***MSE + KL-divergence + Adversarial*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 60.2% and a False Positive Rate at 95% True Positive Rate (FPR95) of 73.1%.*

Severity / Metric	1 AUROC↑/ FPR95↓	2 AUROC↑/ FPR95↓	3 AUROC↑/ FPR95↓	4 AUROC↑/ FPR95↓	5 AUROC↑/ FPR95↓	Average AUROC↑/ FPR95↓
Gaussian Noise	11.78 / 99.4	11.76 / 99.06	10.31 / 98.96	10.12 / 98.9	10.32 / 98.98	10.86 / 99.06
Shot Noise	13.26 / 99.62	10.25 / 99.54	12.32 / 98.8	21.2 / 96.8	80.4 / 34.66	27.49 / 85.88
Speckle Noise	14.19 / 99.54	8.76 / 99.66	7.42 / 99.74	5.68 / 99.82	5.34 / 99.82	8.28 / 99.72
Impulse Noise	82.32 / 54.02	91.32 / 26.3	94.41 / 15.58	96.6 / 8.44	97.25 / 6.08	92.38 / 22.08
Defocus Blur	54.86 / 91.56	60.7 / 88.62	66.93 / 82.18	70.74 / 75.0	78.47 / 58.0	66.34 / 79.07
Gaussian Blur	55.4 / 91.26	66.78 / 82.36	70.56 / 75.56	73.89 / 69.8	78.93 / 56.84	69.11 / 75.16
Glass Blur	29.79 / 96.54	28.62 / 97.42	22.86 / 98.56	46.11 / 89.8	26.81 / 97.32	30.84 / 95.93
Motion Blur	58.69 / 90.22	63.21 / 86.7	67.24 / 82.36	66.73 / 84.4	69.38 / 81.08	65.05 / 84.95
Zoom Blur	61.45 / 87.6	65.43 / 82.92	67.54 / 79.86	69.11 / 77.4	71.03 / 74.16	66.91 / 80.39
Snow	53.74 / 92.08	50.7 / 93.08	56.26 / 89.42	56.2 / 89.48	56.5 / 91.08	54.68 / 91.03
Fog	51.45 / 91.84	57.41 / 85.8	61.76 / 79.46	63.38 / 75.6	63.97 / 70.48	59.6 / 80.64
Brightness	49.75 / 94.9	51.09 / 94.5	53.06 / 93.46	56.62 / 91.64	63.58 / 87.32	54.82 / 92.36
Contrast	59.6 / 86.84	75.93 / 62.96	81.8 / 49.1	88.51 / 33.06	96.23 / 8.96	80.41 / 48.18
Elastic Transform	58.2 / 89.46	68.92 / 77.98	66.32 / 79.52	57.31 / 85.66	45.73 / 91.22	59.3 / 84.77
Pixelate	88.99 / 25.18	96.71 / 5.32	97.78 / 3.4	99.0 / 1.36	99.96 / 0.16	96.49 / 7.08
JPEG Compression	59.64 / 92.2	70.8 / 82.96	74.59 / 77.98	78.17 / 68.92	83.57 / 56.64	73.35 / 75.74
Spatter	54.71 / 91.9	60.38 / 88.06	60.86 / 85.94	61.72 / 86.26	67.53 / 82.96	61.04 / 87.02
Saturate	72.83 / 78.02	94.09 / 15.04	53.79 / 94.44	72.87 / 78.72	83.32 / 52.06	75.38 / 63.66
Frost	37.52 / 97.7	38.83 / 97.06	35.33 / 96.78	33.43 / 96.82	30.43 / 96.08	35.11 / 96.89

Table D.6 *The performance of the VAE FRL model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***Cross Entroy + KL-divergence - Input Complexity*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 57.2% and a False Positive Rate at 95% True Positive Rate (FPR95) of 76.3%.*

Severity / Metric	1 AUROC↑/ FPR95↓	2 AUROC↑/ FPR95↓	3 AUROC↑/ FPR95↓	4 AUROC↑/ FPR95↓	5 AUROC↑/ FPR95↓	Average AUROC↑/ FPR95↓
Gaussian Noise	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0
Shot Noise	99.9 / 0.0	99.96 / 0.0	99.99 / 0.0	100.0 / 0.0	100.0 / 0.0	99.97 / 0.0
Speckle Noise	99.73 / 0.1	99.86 / 0.0	99.89 / 0.0	99.93 / 0.0	99.95 / 0.0	99.87 / 0.02
Impulse Noise	99.46 / 2.4	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	99.89 / 0.48
Defocus Blur	44.17 / 95.7	31.83 / 97.0	20.95 / 98.1	17.69 / 98.6	9.31 / 99.0	24.79 / 97.68
Gaussian Blur	44.32 / 95.7	21.34 / 98.0	14.15 / 98.6	9.75 / 99.0	5.21 / 99.3	18.95 / 98.12
Glass Blur	87.80 / 65.4	84.70 / 77.3	79.33 / 83.9	87.56 / 69.9	82.84 / 78.1	84.45 / 74.92
Motion Blur	34.23 / 96.5	26.82 / 97.1	21.77 / 97.7	21.72 / 97.5	18.18 / 98.4	24.54 / 97.44
Zoom Blur	27.71 / 97.1	21.09 / 97.8	17.25 / 98.6	14.45 / 98.6	11.53 / 98.6	18.41 / 98.14
Snow	62.99 / 88.6	74.97 / 80.0	71.98 / 81.7	70.85 / 87.0	71.06 / 90.9	70.37 / 85.64
Fog	44.69 / 95.0	33.72 / 96.1	27.79 / 96.6	24.17 / 96.5	22.13 / 95.2	30.5 / 95.88
Brightness	57.86 / 93.6	63.6 / 91.3	67.67 / 89.8	71.17 / 88.7	73.61 / 88.2	66.78 / 90.32
Contrast	39.17 / 95.6	20.93 / 98.6	14.86 / 98.7	8.79 / 99.3	2.23 / 99.8	17.2 / 98.4
Elastic Transform	38.87 / 95.8	34.2 / 96.8	27.75 / 97.1	36.17 / 96.5	51.48 / 93.9	37.69 / 96.02
Pixelate	57.09 / 93.6	62.0 / 92.3	63.52 / 91.4	67.16 / 90.3	67.01 / 90.8	63.36 / 91.68
JPEG Compression	49.32 / 90.9	44.23 / 93.8	42.04 / 94.5	39.94 / 95.7	36.21 / 96.2	42.35 / 94.22
Spatter	68.03 / 84.3	81.11 / 64.5	88.64 / 39.6	75.23 / 70.5	88.36 / 39.6	80.27 / 59.7
Saturate	23.9 / 97.4	12.14 / 98.7	69.4 / 88.2	88.04 / 52.1	92.2 / 42.8	57.14 / 75.84
Frost	73.96 / 78.0	79.53 / 79.0	82.89 / 66.1	83.17 / 60.2	84.26 / 48.2	80.76 / 66.3

Table D.7 *The performance of the GLOW model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***log-likelihood*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 58.8% and a False Positive Rate at 95% True Positive Rate (FPR95) of 69.5%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Gaussian Noise	0.01 / 100.0	0.01 / 100.0	0.21 / 100.0	0.18 / 100.0	0.47 / 100.0	0.18 / 100.0
Shot Noise	0.45 / 100.0	0.39 / 100.0	0.33 / 100.0	0.39 / 100.0	0.73 / 100.0	0.46 / 100.0
Speckle Noise	0.74 / 100.0	0.61 / 100.0	0.58 / 100.0	0.79 / 100.0	1.77 / 100.0	0.9 / 100.0
Impulse Noise	13.31 / 99.1	10.99 / 100.0	12.43 / 100.0	19.25 / 100.0	29.25 / 99.8	17.05 / 99.78
Defocus Blur	55.71 / 92.0	60.23 / 83.9	65.64 / 73.1	72.44 / 59.5	80.08 / 45.2	66.82 / 70.74
Gaussian Blur	55.44 / 92.9	65.11 / 74.9	71.07 / 62.5	76.51 / 53.4	85.53 / 36.6	70.73 / 64.06
Glass Blur	14.78 / 100.0	17.24 / 99.8	19.79 / 99.0	14.75 / 100.0	16.81 / 99.5	16.67 / 99.66
Motion Blur	59.17 / 86.5	62.82 / 80.7	66.05 / 78.2	66.08 / 76.0	68.87 / 71.4	64.6 / 78.56
Zoom Blur	62.07 / 80.0	66.15 / 72.1	69.17 / 66.9	71.99 / 61.1	75.62 / 54.7	69.0 / 66.96
Snow	46.25 / 94.8	40.12 / 96.3	40.56 / 96.2	39.72 / 95.8	39.95 / 96.2	41.32 / 95.86
Fog	58.87 / 92.4	62.46 / 88.3	64.97 / 83.6	66.07 / 80.0	69.08 / 74.1	64.29 / 83.68
Brightness	49.3 / 95.8	44.74 / 96.7	39.64 / 97.0	35.2 / 97.7	29.03 / 98.6	39.58 / 97.16
Contrast	60.38 / 90.6	66.07 / 81.4	71.51 / 73.0	77.7 / 62.3	91.2 / 30.3	73.37 / 67.52
Elastic Transform	53.64 / 89.7	55.91 / 86.1	59.33 / 79.9	48.9 / 86.7	34.59 / 94.0	50.47 / 87.28
Pixelate	41.61 / 96.1	36.16 / 96.8	33.18 / 96.8	29.0 / 97.3	26.97 / 97.3	33.38 / 96.86
JPEG Compression	83.05 / 54.0	84.49 / 47.8	85.03 / 47.2	84.93 / 44.4	84.12 / 47.0	84.32 / 48.08
Spatter	33.34 / 96.8	20.16 / 97.6	12.59 / 98.0	26.44 / 97.3	13.98 / 98.3	21.3 / 97.6
Saturate	71.75 / 68.0	92.63 / 24.3	45.27 / 97.6	17.6 / 99.8	9.36 / 100.0	47.32 / 77.94
Frost	36.11 / 97.3	29.7 / 98.7	25.6 / 99.8	26.3 / 99.5	25.86 / 99.7	28.71 / 99.0

Table D.8 *The performance of the GLOW model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***typicality*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 41.6% and a False Positive Rate at 95% True Positive Rate (FPR95) of 85.8%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Gaussian Noise	99.99 / 0.1	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.02
Shot Noise	99.62 / 0.5	99.82 / 0.1	99.94 / 0.0	99.96 / 0.0	99.96 / 0.0	99.86 / 0.12
Speckle Noise	99.39 / 1.5	99.76 / 0.3	99.81 / 0.1	99.85 / 0.1	99.89 / 0.0	99.74 / 0.4
Impulse Noise	95.95 / 13.1	99.77 / 1.0	99.98 / 0.1	100.0 / 0.0	100.0 / 0.0	99.14 / 2.84
Defocus Blur	48.67 / 94.8	47.99 / 95.7	53.14 / 88.1	57.01 / 92.1	71.64 / 73.1	55.69 / 88.76
Gaussian Blur	48.74 / 95.2	52.7 / 88.6	60.77 / 81.1	69.22 / 72.4	81.95 / 44.6	62.68 / 76.38
Glass Blur	83.37 / 68.1	80.56 / 73.3	75.60 / 78.3	83.52 / 66.2	79.06 / 74.8	80.42 / 72.14
Motion Blur	46.99 / 95.8	49.89 / 94.2	53.37 / 89.9	53.72 / 89.9	57.48 / 87.1	52.29 / 91.38
Zoom Blur	48.54 / 92.9	52.78 / 89.9	56.63 / 86.0	60.52 / 84.1	66.06 / 79.1	56.91 / 86.4
Snow	51.08 / 92.5	60.15 / 82.8	57.14 / 87.0	58.05 / 89.9	60.55 / 85.5	57.39 / 87.54
Fog	49.6 / 94.2	50.45 / 94.8	52.6 / 93.4	53.88 / 88.3	54.22 / 90.5	52.15 / 92.24
Brightness	53.11 / 91.8	56.2 / 89.7	59.07 / 89.6	62.44 / 89.6	67.19 / 92.5	59.6 / 90.64
Contrast	49.61 / 94.3	57.67 / 84.6	64.43 / 77.4	73.96 / 65.3	90.95 / 29.6	67.32 / 70.24
Elastic Transform	46.48 / 94.9	46.47 / 94.3	47.27 / 93.5	43.97 / 92.8	50.95 / 90.2	47.03 / 93.14
Pixelate	51.88 / 95.4	55.49 / 94.3	57.67 / 94.0	61.57 / 92.1	62.39 / 91.2	57.8 / 93.4
JPEG Compression	57.41 / 65.7	57.93 / 66.3	58.43 / 68.1	58.41 / 72.2	58.23 / 75.6	58.08 / 69.58
Spatter	54.38 / 95.9	70.71 / 81.6	81.41 / 58.0	60.72 / 92.2	78.86 / 64.6	69.22 / 78.46
Saturate	54.15 / 94.0	80.13 / 62.6	60.02 / 82.5	79.14 / 75.6	87.57 / 58.7	72.2 / 74.68
Frost	60.71 / 86.2	69.18 / 80.3	71.64 / 78.7	70.37 / 78.7	69.45 / 77.3	68.27 / 80.24

Table D.9 *The performance of the GLOW model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using the* ***normalized score distance*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 69.25% and a False Positive Rate at 95% True Positive Rate (FPR95) of 65.57%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Gaussian Noise	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0
Shot Noise	99.97 / 0.03	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	99.99 / 0.01
Speckle Noise	99.87 / 0.46	99.98 / 0.0	99.99 / 0.0	100.0 / 0.0	100.0 / 0.0	99.97 / 0.09
Impulse Noise	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0	100.0 / 0.0
Defocus Blur	41.76 / 96.56	26.48 / 98.66	15.67 / 99.39	12.81 / 99.49	6.67 / 99.73	20.68 / 98.77
Gaussian Blur	42.08 / 96.55	16.35 / 99.35	10.53 / 99.64	7.43 / 99.72	4.36 / 99.79	16.15 / 99.01
Glass Blur	94.10 / 26.6	92.71 / 32.9	85.63 / 53.3	95.30 / 22.3	90.04 / 41.4	91.56 / 35.3
Motion Blur	28.51 / 98.22	20.43 / 99.04	15.41 / 99.35	15.39 / 99.34	12.06 / 99.52	18.36 / 99.09
Zoom Blur	21.57 / 99.06	15.85 / 99.3	12.66 / 99.47	10.36 / 99.54	8.13 / 99.67	13.71 / 99.41
Snow	66.81 / 76.76	78.15 / 57.68	75.76 / 62.43	74.43 / 66.1	77.02 / 62.06	74.43 / 65.01
Fog	37.97 / 96.31	22.81 / 98.66	15.71 / 99.16	11.1 / 99.39	7.29 / 99.4	18.98 / 98.58
Brightness	56.99 / 91.95	63.15 / 88.38	68.39 / 84.34	73.15 / 78.64	78.57 / 72.78	68.05 / 83.22
Contrast	31.23 / 97.77	11.0 / 99.63	6.3 / 99.73	2.84 / 99.84	0.51 / 99.98	10.38 / 99.39
Elastic Transform	34.78 / 97.95	29.44 / 98.41	22.61 / 98.92	29.91 / 98.15	46.57 / 94.68	32.66 / 97.62
Pixelate	58.57 / 91.1	63.7 / 87.87	65.79 / 86.0	70.44 / 81.05	74.14 / 75.92	66.53 / 84.39
JPEG Compression	52.06 / 90.7	48.68 / 92.9	48.04 / 93.33	46.66 / 94.63	44.78 / 95.83	48.04 / 93.48
Spatter	77.96 / 56.4	90.66 / 25.05	92.64 / 18.55	90.29 / 33.44	97.05 / 11.91	89.72 / 29.07
Saturate	22.04 / 98.92	16.43 / 99.45	71.69 / 79.1	91.61 / 33.98	96.68 / 13.23	59.69 / 64.94
Frost	70.48 / 72.15	79.48 / 58.54	82.12 / 52.39	79.68 / 56.79	78.54 / 57.5	78.06 / 59.47

Table D.10 *The performance of the CovariateFlow model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***log-likelihood*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 58.3% and a False Positive Rate at 95% True Positive Rate (FPR95) of 63.5%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Gaussian Noise	6.99 / 100.0	1.75 / 100.0	0.44 / 100.0	0.19 / 100.0	0.08 / 100.0	1.89 / 100.0
Shot Noise	13.91 / 99.99	7.24 / 100.0	1.54 / 100.0	0.86 / 100.0	0.36 / 100.0	4.78 / 100.0
Speckle Noise	14.62 / 99.99	5.91 / 100.0	3.92 / 100.0	1.79 / 100.0	0.88 / 100.0	5.42 / 100.0
Impulse Noise	1.72 / 100.0	0.25 / 100.0	0.09 / 100.0	0.03 / 100.0	0.03 / 100.0	0.42 / 100.0
Defocus Blur	56.28 / 91.03	67.29 / 77.35	75.52 / 58.82	81.35 / 48.12	87.62 / 33.84	73.61 / 61.83
Gaussian Blur	56.14 / 91.32	75.85 / 58.06	81.0 / 44.39	84.6 / 36.37	89.52 / 27.93	77.42 / 51.61
Glass Blur	31.48 / 99.64	33.31 / 99.53	40.55 / 98.72	27.37 / 99.75	35.57 / 99.28	33.66 / 99.38
Motion Blur	66.94 / 78.89	73.39 / 67.27	77.53 / 57.47	77.39 / 56.9	80.2 / 50.71	75.09 / 62.25
Zoom Blur	75.41 / 64.55	77.96 / 56.4	81.04 / 48.43	82.35 / 44.68	84.65 / 38.59	80.28 / 50.53
Snow	40.84 / 97.16	34.39 / 98.25	34.76 / 98.08	35.47 / 97.95	36.02 / 98.05	36.3 / 97.9
Fog	59.09 / 88.58	69.74 / 74.13	74.34 / 62.76	77.76 / 54.6	81.51 / 45.02	72.49 / 65.02
Brightness	46.11 / 96.49	42.79 / 97.53	39.77 / 98.37	36.64 / 98.91	32.23 / 99.42	39.51 / 98.14
Contrast	63.27 / 83.41	77.58 / 55.23	81.85 / 43.87	85.81 / 34.14	91.13 / 23.89	79.93 / 48.11
Elastic Transform	62.73 / 85.85	66.2 / 80.71	71.06 / 71.71	65.88 / 79.12	55.36 / 90.91	64.25 / 81.66
Pixelate	49.57 / 96.24	50.82 / 96.96	50.71 / 97.01	53.34 / 97.24	58.73 / 96.77	52.63 / 96.84
JPEG Compression	53.62 / 93.49	56.54 / 92.86	58.53 / 92.25	59.89 / 91.32	61.29 / 90.17	57.97 / 92.02
Spatter	40.06 / 97.67	30.67 / 98.94	27.04 / 99.36	24.72 / 99.29	16.57 / 99.89	27.81 / 99.03
Saturate	64.44 / 77.58	66.78 / 72.5	36.82 / 99.02	21.56 / 99.89	13.69 / 100.0	40.66 / 89.8
Frost	41.35 / 97.29	38.0 / 98.22	39.38 / 98.04	41.29 / 97.43	43.61 / 96.68	40.73 / 97.53

Table D.11 *The performance of the CovariateFlow model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***typicality*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 45.5% and a False Positive Rate at 95% True Positive Rate (FPR95) of 83.8%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Gaussian Noise	99.46 / 0.63	99.65 / 0.4	99.79 / 0.25	99.81 / 0.2	99.82 / 0.19	99.71 / 0.33
Shot Noise	99.19 / 1.16	99.46 / 0.71	99.69 / 0.43	99.76 / 0.39	99.81 / 0.25	99.58 / 0.59
Speckle Noise	98.96 / 2.07	99.48 / 0.78	99.59 / 0.68	99.72 / 0.47	99.79 / 0.39	99.51 / 0.88
Impulse Noise	99.68 / 0.5	99.84 / 0.19	99.88 / 0.15	99.91 / 0.12	99.92 / 0.09	99.85 / 0.21
Defocus Blur	50.33 / 94.96	58.69 / 92.74	70.05 / 84.46	75.37 / 78.97	85.71 / 52.89	68.03 / 80.8
Gaussian Blur	50.14 / 95.3	69.37 / 85.8	77.29 / 73.98	82.67 / 61.39	89.41 / 36.32	73.78 / 70.56
Glass Blur	89.41 / 45.08	87.57 / 53.53	77.47 / 74.20	91.13 / 38.94	83.31 / 64.63	85.78 / 55.28
Motion Blur	57.66 / 93.61	65.53 / 89.88	72.16 / 83.91	72.05 / 84.16	76.56 / 75.4	68.79 / 85.39
Zoom Blur	64.69 / 91.29	70.9 / 85.13	75.56 / 78.26	78.72 / 71.08	82.34 / 63.09	74.44 / 77.77
Snow	50.58 / 95.61	61.29 / 91.65	58.24 / 93.14	57.31 / 93.96	60.98 / 92.24	57.68 / 93.32
Fog	51.21 / 94.85	62.74 / 90.19	70.43 / 83.68	76.32 / 74.28	81.99 / 56.85	68.54 / 79.97
Brightness	50.58 / 95.27	53.04 / 95.21	56.92 / 94.54	61.02 / 93.45	68.04 / 91.46	57.92 / 93.99
Contrast	55.32 / 93.67	76.67 / 75.78	83.86 / 54.76	90.12 / 29.31	95.89 / 10.34	80.37 / 52.77
Elastic Transform	53.65 / 94.88	56.87 / 94.16	62.39 / 91.69	55.03 / 94.55	47.27 / 96.18	55.04 / 94.29
Pixelate	52.09 / 95.21	54.97 / 95.39	56.66 / 95.11	61.11 / 94.28	66.07 / 92.41	58.18 / 94.48
JPEG Compression	46.36 / 96.46	46.7 / 96.45	47.28 / 96.54	47.75 / 96.6	48.96 / 96.57	47.41 / 96.52
Spatter	60.21 / 93.94	81.88 / 52.79	85.73 / 38.68	80.65 / 71.55	93.5 / 23.28	80.39 / 56.05
Saturate	60.09 / 89.53	66.52 / 84.43	59.65 / 94.4	85.27 / 62.34	93.93 / 24.4	73.09 / 71.02
Frost	53.51 / 95.45	64.2 / 91.5	67.86 / 89.29	64.31 / 91.19	62.17 / 92.49	62.41 / 91.98

Table D.12 *The performance of the CovariateFlow model in detecting out-of-distribution (OOD) covariate shift between CIFAR10 and CIFAR10-C datasets is evaluated. The model is evaluated using* ***normalized score distance*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 74.9% and a False Positive Rate at 95% True Positive Rate (FPR95) of 61.7%.*

Model	VAE	AVAE	VAE	DDPM	DDPM	GLOW	GLOW	GLOW	CovFlow	CovFlow	CovFlow
Evaluation	ALL	ALL	FLR	T150	T20	LL	Typ	NSD	LL	Typ	NSD
Gaussian Noise	85.7	76.9	10.9	64.6	92.8	100.0	0.2	100.0	100.0	1.9	99.7
Shot Noise	79.2	71.3	27.5	61.3	87.9	100.0	0.5	99.9	100.0	4.8	99.6
Speckle Noise	79.3	71.6	8.2	61.3	85.3	99.9	0.9	99.7	100.0	5.4	99.5
Impulse Noise	92.8	86.6	92.4	77.0	97.3	99.9	17.0	99.1	100.0	0.4	99.8
Defocus Blur	26.6	35.3	66.3	70.4	63.7	24.8	66.8	55.7	20.7	73.6	68.0
Gaussian Blur	22.5	32.8	69.1	78.5	71.0	19.0	70.7	62.7	16.2	77.4	73.8
Glass Blur	55.5	53.7	30.8	64.9	73.1	84.4	16.7	80.4	91.6	33.7	85.8
Motion Blur	21.6	32.4	65.1	78.6	68.7	24.5	64.6	52.3	18.4	75.1	68.8
Zoom Blur	21.3	31.7	66.9	80.0	69.8	18.4	69.0	56.9	13.7	80.3	74.4
Snow	66.1	91.7	54.7	50.1	61.2	70.4	41.3	57.4	74.4	36.3	57.7
Fog	14.0	48.1	59.6	78.6	70.0	30.5	64.3	52.2	19.0	72.5	68.5
Brightness	56.1	89.2	54.8	47.4	51.5	66.8	39.6	59.6	68.0	39.5	57.9
Contrast	6.3	42.9	80.4	50.4	47.2	17.2	73.4	67.3	10.4	79.9	80.4
Elastic Transform	31.2	38.5	59.3	58.8	52.5	37.7	50.5	47.0	32.7	64.2	55.0
Pixelate	46.7	47.4	96.5	56.5	55.4	63.4	33.4	57.8	66.5	52.6	58.2
JPEG Compression	48.0	48.5	73.3	57.4	56.4	42.4	84.3	58.1	48.0	58.0	47.4
Spatter	72.1	67.2	61.0	59.6	67.6	80.3	21.3	69.2	89.7	27.8	80.4
Saturate	57.9	78.1	75.3	52.2	56.6	57.1	47.3	72.2	59.7	40.7	73.1
Frost	47.2	99.5	35.1	53.8	60.3	80.8	28.7	68.3	78.1	40.7	62.4
Average	48.9	60.2	57.2	63.1	67.5	57.7	41.6	69.3	58.3	45.5	**74.9**

Table D.13 *Comparison of the performance (AUROC) of all the employed models at detecting every CIFAR10(-C) OOD degredation type.*

D.4 Detailed results on ImageNet200 vs. ImageNet200-C

The following section depicts detailed results obtained with various models on our experiments with ID ImageNet200 and ImageNet200-C as OOD. The results are depicted in order of presentation: DDPM T20-LPIPS+MSE (D.14), GLOW-LL (Table D.15), GLOW-Typicality (Table D.16), GLOW-NSD (Table D.9), CovariateFlow-LL (Table D.18), CovariateFlow-Typicality (Table D.19) and CovariateFlow-NSD (Table D.20).

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Brightness	39.0 / 94.3	35.45 / 95.0	33.15 / 96.3	30.04 / 97.2	28.51 / 97.3	33.23 / 96.02
Contrast	59.52 / 86.7	64.78 / 85.7	73.87 / 80.5	85.18 / 63.3	89.73 / 51.5	74.62 / 73.54
Defocus Blur	57.46 / 91.8	67.45 / 87.1	82.33 / 72.9	96.75 / 16.8	98.77 / 3.7	80.55 / 54.46
Elastic Transform	49.89 / 92.1	49.37 / 93.8	54.55 / 92.1	52.63 / 92.1	50.93 / 93.6	51.47 / 92.74
Fog	59.22 / 85.7	70.98 / 79.8	81.35 / 64.7	91.6 / 43.0	95.54 / 25.1	79.74 / 59.66
Frost	34.42 / 96.2	40.97 / 96.2	48.61 / 93.0	52.6 / 92.1	57.38 / 89.6	46.8 / 93.42
Gaussian Noise	34.87 / 96.5	74.38 / 63.2	91.73 / 28.0	96.13 / 14.2	97.71 / 6.9	78.96 / 41.76
Glass Blur	56.65 / 87.2	49.07 / 92.5	50.19 / 93.2	59.12 / 90.2	82.11 / 72.6	59.43 / 87.14
Impulse Noise	49.4 / 88.2	66.25 / 75.1	89.9 / 30.0	95.79 / 14.9	98.0 / 5.4	79.87 / 42.72
JPEG Compression	43.2 / 93.9	41.98 / 94.8	47.62 / 93.6	47.03 / 92.8	52.38 / 92.1	46.44 / 93.44
Motion Blur	54.12 / 92.1	59.06 / 88.2	70.23 / 83.8	76.9 / 77.9	82.41 / 66.7	68.54 / 81.74
Pixelate	41.02 / 94.6	42.57 / 94.7	49.44 / 91.9	52.49 / 89.9	54.62 / 89.3	48.03 / 92.08
Shot Noise	40.79 / 94.2	62.07 / 81.1	80.72 / 60.5	88.17 / 48.3	94.85 / 20.5	73.32 / 60.92
Snow	45.58 / 94.6	58.46 / 90.2	44.76 / 97.2	38.28 / 97.3	39.68 / 97.2	45.35 / 95.3
Zoom Blur	64.0 / 87.5	71.22 / 81.5	77.86 / 73.2	83.17 / 62.9	87.45 / 55.1	76.74 / 72.04

Table D.14 *The performance of the DDPM model in detecting out-of-distribution (OOD) covariate shift between ImageNet200 and ImageNet200-C datasets is evaluated. The model is evaluated using* ***T20-LPIPS+MSE*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 62.87% and a False Positive Rate at 95% True Positive Rate (FPR95) of 75.80%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Brightness	38.03 / 96.5	42.45 / 95.9	45.58 / 95.6	47.64 / 95.9	48.21 / 96.5	44.38 / 96.08
Contrast	5.8 / 99.8	3.0 / 99.8	1.13 / 99.9	0.15 / 100.0	0.01 / 100.0	2.02 / 99.9
Defocus Blur	17.3 / 98.1	14.11 / 98.2	9.92 / 99.3	5.12 / 99.7	4.0 / 99.8	10.09 / 99.02
Elastic Transform	23.8 / 98.0	22.13 / 98.0	18.54 / 98.1	19.07 / 98.1	21.5 / 97.8	21.01 / 98.0
Fog	18.73 / 98.1	12.86 / 99.3	9.19 / 99.6	5.73 / 99.8	4.0 / 99.8	10.1 / 99.32
Frost	41.85 / 93.7	43.38 / 92.1	42.83 / 91.5	43.73 / 90.2	44.57 / 88.6	43.27 / 91.22
Gaussian Noise	65.57 / 65.5	97.55 / 6.9	99.71 / 0.8	99.96 / 0.2	99.99 / 0.0	92.56 / 14.68
Glass Blur	50.16 / 93.8	23.86 / 97.5	15.38 / 98.1	11.67 / 98.2	6.53 / 99.5	21.52 / 97.42
Impulse Noise	69.23 / 64.7	94.31 / 16.7	99.73 / 0.5	99.98 / 0.1	100.0 / 0.0	92.65 / 16.4
JPEG Compression	17.45 / 98.1	17.49 / 98.2	14.21 / 98.8	12.64 / 99.4	9.05 / 99.7	14.17 / 98.84
Motion Blur	21.88 / 98.0	17.4 / 98.1	14.35 / 98.2	12.19 / 98.6	10.58 / 98.9	15.28 / 98.36
Pixelate	32.68 / 97.0	32.34 / 96.9	32.99 / 96.6	30.62 / 97.0	28.46 / 97.3	31.42 / 96.96
Shot Noise	66.1 / 71.7	89.03 / 43.9	96.59 / 15.0	98.41 / 3.0	99.34 / 0.8	89.89 / 26.88
Snow	42.62 / 93.6	54.38 / 88.7	44.96 / 93.8	40.73 / 96.0	36.27 / 97.3	43.79 / 93.88
Zoom Blur	16.06 / 98.1	12.24 / 98.5	10.42 / 98.9	8.67 / 99.4	7.44 / 99.5	10.97 / 98.88

Table D.15 *The performance of the GLOW model in detecting out-of-distribution (OOD) covariate shift between ImageNet200 and ImageNet200-C datasets is evaluated. The model is evaluated using* ***log-likelihood*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 36.21% and a False Positive Rate at 95% True Positive Rate (FPR95) of 81.7%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Brightness	43.64 / 96.0	37.45 / 97.8	33.11 / 98.6	29.8 / 99.0	28.06 / 99.1	34.41 / 98.1
Contrast	81.03 / 48.7	86.93 / 36.3	92.61 / 22.5	97.37 / 7.0	99.17 / 3.4	91.42 / 23.58
Defocus Blur	59.32 / 79.3	61.55 / 74.4	66.0 / 66.8	73.11 / 53.0	75.83 / 48.4	67.16 / 64.38
Elastic Transform	56.73 / 85.4	57.43 / 83.7	59.26 / 79.8	58.04 / 80.6	56.33 / 83.0	57.56 / 82.5
Fog	65.07 / 73.1	71.91 / 62.9	76.93 / 54.3	82.61 / 45.9	86.36 / 36.5	76.58 / 54.54
Frost	44.91 / 95.0	46.15 / 93.9	48.82 / 91.4	49.75 / 90.5	50.83 / 90.1	48.09 / 92.18
Gaussian Noise	27.09 / 97.4	4.3 / 99.2	0.53 / 99.8	0.16 / 100.0	0.01 / 100.0	6.42 / 99.28
Glass Blur	45.74 / 94.6	57.55 / 80.8	61.94 / 72.1	64.63 / 67.6	70.64 / 56.9	60.1 / 74.4
Impulse Noise	22.29 / 98.1	4.99 / 99.2	0.36 / 100.0	0.07 / 100.0	0.01 / 100.0	5.54 / 99.46
JPEG Compression	64.51 / 79.8	65.75 / 80.9	69.9 / 74.4	73.9 / 70.8	81.06 / 58.7	71.02 / 72.92
Motion Blur	56.63 / 83.9	59.15 / 79.1	61.21 / 74.4	63.11 / 71.6	64.74 / 69.4	60.97 / 75.68
Pixelate	52.9 / 90.2	53.72 / 89.1	53.02 / 89.4	55.48 / 84.7	60.41 / 78.5	55.11 / 86.38
Shot Noise	27.68 / 97.4	13.79 / 99.0	5.15 / 99.6	2.85 / 99.8	1.28 / 100.0	10.15 / 99.16
Snow	51.33 / 92.2	47.83 / 94.3	54.25 / 91.4	57.88 / 90.2	63.25 / 85.8	54.91 / 90.78
Zoom Blur	60.25 / 78.2	63.37 / 71.6	65.24 / 68.4	67.39 / 63.8	69.11 / 61.2	65.07 / 68.64

Table D.16 *The performance of the GLOW model in detecting out-of-distribution (OOD) covariate shift between ImageNet200 and ImageNet200-C datasets is evaluated. The model is evaluated using* ***typicality*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 51.0% and a False Positive Rate at 95% True Positive Rate (FPR95) of 78.8%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Brightness	47.03 / 88.7	49.81 / 86.4	53.01 / 84.6	56.25 / 81.5	58.86 / 78.5	52.99 / 83.94
Contrast	79.98 / 59.4	88.76 / 41.1	95.46 / 15.6	99.2 / 4.0	99.92 / 0.4	92.66 / 24.1
Defocus Blur	55.14 / 76.4	59.14 / 70.9	65.89 / 61.7	76.95 / 46.4	80.72 / 41.3	67.57 / 59.34
Elastic Transform	49.51 / 85.9	50.52 / 84.5	53.73 / 79.7	52.89 / 79.5	50.35 / 82.4	51.4 / 82.4
Fog	54.47 / 86.2	63.48 / 80.7	71.4 / 72.8	80.79 / 59.6	86.43 / 48.1	71.31 / 69.48
Frost	44.2 / 93.3	42.28 / 95.6	40.22 / 96.2	39.61 / 96.2	39.12 / 95.6	41.09 / 95.38
Gaussian Noise	48.72 / 94.2	90.93 / 17.5	97.51 / 4.3	99.04 / 1.8	99.52 / 0.6	87.14 / 23.68
Glass Blur	47.43 / 95.6	47.72 / 87.0	56.57 / 75.6	62.44 / 66.7	73.02 / 50.5	57.44 / 75.08
Impulse Noise	54.18 / 90.5	87.27 / 26.8	97.76 / 3.9	99.29 / 1.4	99.77 / 0.4	87.65 / 24.6
JPEG Compression	57.84 / 82.5	58.69 / 83.5	63.58 / 78.9	67.34 / 77.7	76.33 / 71.9	64.76 / 78.9
Motion Blur	50.84 / 83.3	55.12 / 77.3	58.97 / 72.4	62.29 / 67.8	65.1 / 64.2	58.46 / 73.0
Pixelate	45.1 / 91.9	44.29 / 92.1	43.55 / 92.0	43.3 / 92.0	42.39 / 94.2	43.73 / 92.44
Shot Noise	51.85 / 92.6	79.07 / 67.4	91.32 / 28.1	95.42 / 12.3	98.25 / 5.0	83.18 / 41.08
Snow	41.77 / 96.6	44.85 / 96.7	42.98 / 97.0	43.99 / 97.1	46.22 / 97.0	43.96 / 96.88
Zoom Blur	56.61 / 74.6	61.92 / 67.7	65.01 / 62.7	68.36 / 58.8	71.01 / 54.6	64.58 / 63.68

Table D.17 *The performance of the GLOW model in detecting out-of-distribution (OOD) covariate shift between ImageNet200 and ImageNet200-C datasets is evaluated. The model is evaluated using* ***normalized score distance*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 64.5% and a False Positive Rate at 95% True Positive Rate (FPR95) of 65.6%.*

Severity	1	2	3	4	5	Average
Metric	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓	AUROC↑/ FPR95↓
Brightness	24.8 / 98.6	32.64 / 97.4	40.65 / 96.9	44.85 / 96.9	50.5 / 96.2	38.69 / 97.2
Contrast	0.86 / 100.0	0.35 / 100.0	0.12 / 100.0	0.03 / 100.0	0.02 / 100.0	0.28 / 100.0
Defocus Blur	9.16 / 99.8	7.8 / 99.8	4.63 / 99.9	1.71 / 100.0	1.18 / 100.0	4.9 / 99.9
Elastic Transform	12.46 / 99.7	12.18 / 99.7	9.64 / 99.7	9.73 / 99.7	10.37 / 99.7	10.88 / 99.7
Fog	6.11 / 99.8	2.82 / 100.0	1.48 / 100.0	0.77 / 100.0	0.33 / 100.0	2.3 / 99.96
Frost	23.98 / 98.1	22.32 / 97.6	17.32 / 98.5	16.03 / 98.1	15.8 / 97.9	19.09 / 98.04
Gaussian Noise	29.29 / 95.2	68.86 / 52.8	94.04 / 11.7	98.06 / 3.6	99.46 / 1.6	77.94 / 32.98
Glass Blur	26.39 / 97.8	8.77 / 99.7	4.71 / 99.8	3.11 / 99.9	1.44 / 100.0	8.88 / 99.44
Impulse Noise	42.44 / 88.1	71.39 / 60.7	96.95 / 6.9	99.42 / 1.8	99.99 / 0.1	82.04 / 31.52
JPEG Compression	12.31 / 99.5	15.34 / 99.5	11.5 / 99.7	12.1 / 99.7	10.84 / 99.8	12.42 / 99.64
Motion Blur	11.93 / 99.7	8.59 / 99.7	6.98 / 99.8	5.11 / 99.8	4.29 / 99.9	7.38 / 99.78
Pixelate	17.99 / 99.1	16.25 / 99.1	15.22 / 99.5	14.63 / 99.2	12.65 / 99.4	15.35 / 99.26
Shot Noise	29.97 / 95.2	52.29 / 82.9	80.05 / 50.3	92.4 / 20.7	98.31 / 4.5	70.6 / 50.72
Snow	23.6 / 97.1	30.26 / 94.4	28.19 / 95.5	26.6 / 96.2	23.48 / 97.4	26.43 / 96.12
Zoom Blur	8.49 / 99.8	5.74 / 99.8	4.61 / 99.8	3.86 / 99.8	2.8 / 100.0	5.1 / 99.84

Table D.18 *The performance of the CovariateFlow model in detecting out-of-distribution (OOD) covariate shift between ImageNet200 and ImageNet200-C datasets is evaluated. The model is evaluated using* ***log-likelihood*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 25.8% and a False Positive Rate at 95% True Positive Rate (FPR95) of 87.9%.*

Severity / Metric	1 AUROC↑/ FPR95↓	2 AUROC↑/ FPR95↓	3 AUROC↑/ FPR95↓	4 AUROC↑/ FPR95↓	5 AUROC↑/ FPR95↓	Average AUROC↑/ FPR95↓
Brightness	63.09 / 83.5	59.09 / 87.7	57.06 / 90.8	56.08 / 93.0	53.79 / 93.6	57.82 / 89.72
Contrast	80.11 / 52.2	82.58 / 46.6	84.46 / 43.7	86.7 / 38.5	87.91 / 38.4	84.35 / 43.88
Defocus Blur	73.57 / 65.0	76.69 / 59.2	77.32 / 56.7	79.52 / 52.1	80.14 / 51.6	77.45 / 56.92
Elastic Transform	70.95 / 76.5	71.3 / 69.2	71.14 / 68.0	71.97 / 66.7	69.81 / 68.8	71.03 / 69.84
Fog	73.8 / 64.2	77.47 / 56.5	78.13 / 54.9	79.93 / 50.5	81.97 / 46.8	78.26 / 54.58
Frost	62.49 / 81.5	61.63 / 81.4	65.72 / 76.1	64.46 / 76.5	66.97 / 72.0	64.25 / 77.5
Gaussian Noise	57.59 / 86.2	37.25 / 95.9	11.56 / 99.4	5.28 / 99.9	2.63 / 100.0	22.86 / 96.28
Glass Blur	57.5 / 84.7	70.68 / 69.2	73.88 / 62.4	75.91 / 58.8	79.3 / 52.9	71.45 / 65.6
Impulse Noise	50.98 / 92.7	37.38 / 96.1	9.52 / 99.7	2.75 / 100.0	0.58 / 100.0	20.24 / 97.7
JPEG Compression	68.64 / 72.4	66.91 / 75.4	70.38 / 70.6	69.92 / 74.2	69.95 / 71.5	69.16 / 72.82
Motion Blur	69.2 / 70.4	73.33 / 65.6	74.94 / 62.2	76.46 / 59.5	76.41 / 60.0	74.07 / 63.54
Pixelate	65.93 / 77.7	68.28 / 75.3	70.48 / 68.9	71.53 / 68.9	72.87 / 66.4	69.82 / 71.44
Shot Noise	56.38 / 83.8	46.32 / 93.2	27.58 / 97.8	15.13 / 99.4	5.62 / 100.0	30.21 / 94.84
Snow	60.07 / 81.9	56.09 / 84.8	57.87 / 83.5	59.16 / 83.5	62.24 / 79.2	59.09 / 82.58
Zoom Blur	74.45 / 61.1	75.93 / 61.7	75.98 / 59.3	76.89 / 59.0	77.85 / 57.1	76.22 / 59.64

Table D.19 *The performance of the CovariateFlow model in detecting out-of-distribution (OOD) covariate shift between ImageNet200 and ImageNet200-C datasets is evaluated. The model is evaluated using* ***typicality*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 61.8% and a False Positive Rate at 95% True Positive Rate (FPR95) of 73.1%.*

Severity / Metric	1 AUROC↑/ FPR95↓	2 AUROC↑/ FPR95↓	3 AUROC↑/ FPR95↓	4 AUROC↑/ FPR95↓	5 AUROC↑/ FPR95↓	Average AUROC↑/ FPR95↓
Brightness	58.65 / 86.1	54.19 / 93.5	53.38 / 93.2	53.62 / 94.2	54.55 / 94.1	54.88 / 92.22
Contrast	87.87 / 23.7	89.66 / 19.9	91.65 / 16.4	93.19 / 14.0	94.38 / 12.0	91.35 / 17.2
Defocus Blur	75.17 / 69.5	76.79 / 63.5	80.82 / 48.0	85.66 / 28.3	87.26 / 24.3	81.14 / 46.72
Elastic Transform	71.4 / 71.1	70.95 / 76.3	72.68 / 72.8	73.58 / 71.4	71.72 / 72.8	72.07 / 72.88
Fog	78.02 / 52.0	83.44 / 36.2	85.75 / 27.8	87.64 / 23.2	89.49 / 19.3	84.87 / 31.7
Frost	57.73 / 87.8	57.58 / 87.6	61.2 / 82.5	63.14 / 76.3	63.24 / 78.7	60.58 / 82.58
Gaussian Noise	51.71 / 91.5	42.24 / 94.7	89.85 / 18.7	95.72 / 7.2	97.91 / 3.6	75.49 / 43.14
Glass Blur	54.51 / 89.5	73.88 / 67.8	79.7 / 46.0	82.48 / 36.7	86.13 / 26.8	75.34 / 53.36
Impulse Noise	44.24 / 94.0	49.66 / 94.9	94.03 / 12.2	97.96 / 3.9	99.33 / 1.0	77.04 / 41.2
JPEG Compression	69.16 / 73.0	66.41 / 82.2	69.87 / 75.2	69.76 / 74.7	71.05 / 75.4	69.25 / 76.1
Motion Blur	70.49 / 77.0	74.58 / 66.4	76.85 / 63.0	79.64 / 53.8	80.66 / 45.2	76.44 / 61.08
Pixelate	64.6 / 82.7	65.71 / 81.4	65.98 / 81.7	67.72 / 83.3	68.87 / 77.1	66.58 / 81.24
Shot Noise	49.73 / 90.8	37.74 / 96.7	66.7 / 83.8	87.36 / 36.0	96.12 / 9.1	67.53 / 63.28
Snow	55.22 / 88.2	48.07 / 90.9	50.33 / 92.4	51.18 / 91.8	54.07 / 90.6	51.77 / 90.78
Zoom Blur	74.91 / 68.0	79.54 / 49.0	80.36 / 46.6	81.94 / 42.2	83.66 / 34.6	80.08 / 48.08

Table D.20 *The performance of the CovariateFlow model in detecting out-of-distribution (OOD) covariate shift between ImageNet200 and ImageNet200-C datasets is evaluated. The model is evaluated using* ***normalized score distance*** *as metric. The model achieves a mean Area Under the Receiver Operating Characteristic (AUROC) of 72.3% and a False Positive Rate at 95% True Positive Rate (FPR95) of 60.1%.*

D.5 X-Ray dataset details

Medical X-ray images, pivotal for diagnostic purposes, contain various noise sources that can compromise image quality and diagnostic accuracy [294]. Photon noise, stemming from the probabilistic interaction of X-ray photons with the detector, leads to pixel value variations and is more pronounced at lower doses, often modeled as signal-dependent Poisson noise. Electronic noise, including thermal, quantization, and readout noise, arises from system components, while scatter noise, due to photon scattering within the patient's body, degrades contrast and introduces artifacts. These are often modeled as signal-independent additive noise. Although manual modeling noise has been attempted [365], a data-driven modeling approach may be more suited for capturing the complexity in the noise.

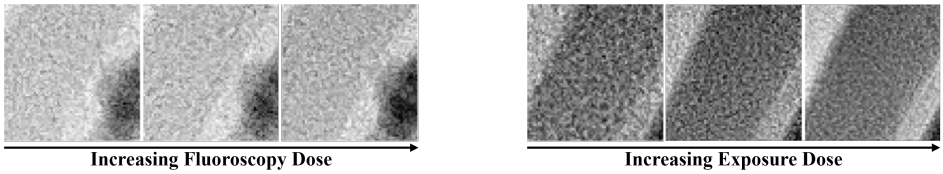


Figure D.8 *Illustration of the covariate shift introduced due to varying imaging modalities and X-Ray dose levels.*

We have collected a new dataset of X-Ray images featuring a standard test object (a clock) using the Azurion Image-Guided Therapy (IGT) system. These images, stored in Dicom format with a resolution of 12 bits per pixel, were obtained under various imaging conditions. This included: (a) two different levels of radiation dose (low and normal) and (b) the use of both pulsed fluoroscopy (subsequently referred to as fluoroscopy) and continuous radiation (subsequently referred to as exposure). These experiments were also conducted at (c) two different distances between the source and the image (SID) of 110 cm and 90 cm. The dataset includes images from six distinct modes of operation, as illustrated in Figure D.9: Mode 0 (exposure with a normal dose at 110-cm SID), Mode 1 (exposure with a low dose at 110-cm SID), Mode 2 (exposure with a normal dose at 90-cm SID), Mode 3 (exposure with a low dose at 90-cm SID), Mode 4 (fluoroscopy with a normal dose at 110-cm SID), and Mode 5 (fluoroscopy with a low dose at 90-cm SID).

Starting with Mode 0 as the reference, we anticipate significant variations in data characteristics up to Mode 5, which is expected to be the most OOD. Although the complete dataset encompasses 18 modes across two environments, this research will limit its discussion to the aforementioned six modes for clarity.

D.6 Detailed results on the X-Ray dataset

Figure D.10 depicts the performance of the DDPM model at various reconstruction starting points T. It can be seen that $T = 250$ yields the best performance,

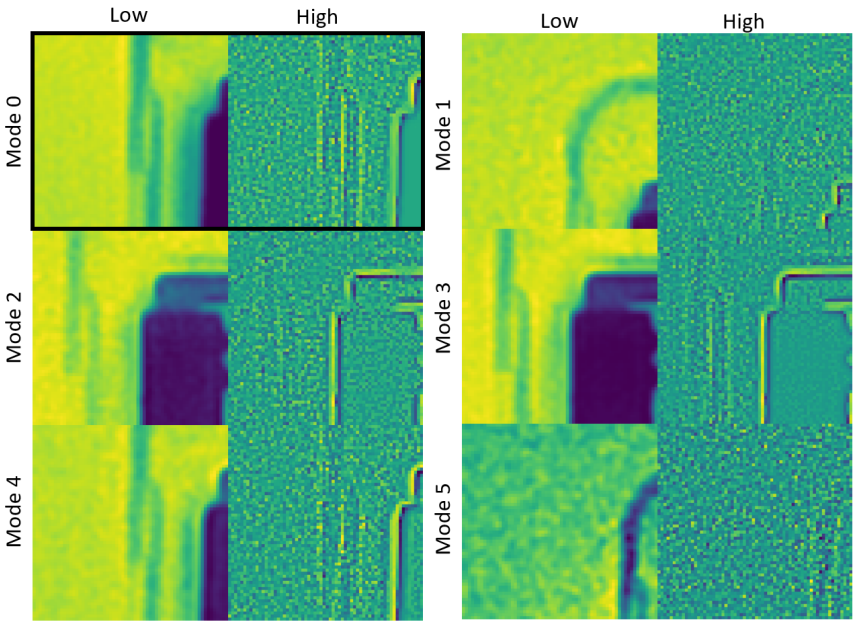


Figure D.9 *Low-frequency and high-frequency components of patches from 6 modes/settings.*

irrespective of the metric and LPIPS at $T = 250$ the overall best performance.

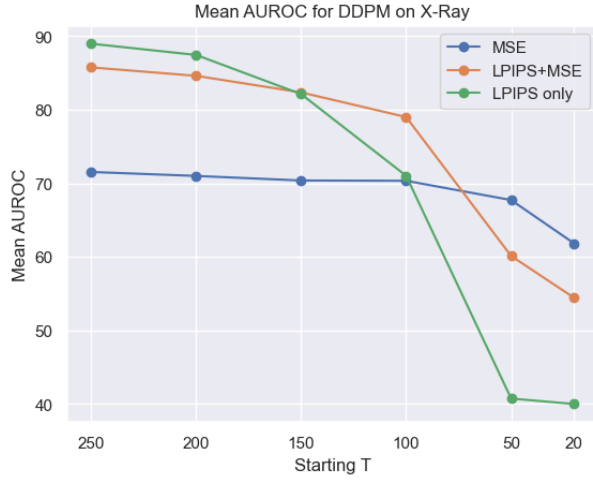


Figure D.10 *Results obtained with the DDPM on X-Ray Mode 0 vs. Mode 1-5. The figure depicts mean AUROC obtained from reconstructions at different starting points, T.*

D.7 Ablation experiments

This section details a series of ablation experiments conducted, including an analysis of the effect of the individual components in CovariateFlow on the detection performance (Table D.21), mean scores per severity of the models and resource aspects are depicted in Table D.22, model performance on a typically semantic OOD detection problem in Table D.23 and finally an example (Figure D.11) of heteroscedastic high-frequency components sampled from the fully invertible CovariateFlow.

In our ablation experiments, we test the effect of explicitly modelling the conditional distribution between the low-frequency and high-frequency signal components as described in Section 6.5.2. This is achieved by training and evaluating the CovariateFlow model in four different settings: (1) unconditional coupling flows with the full input image, (2) unconditional coupling flows subject to only the high-frequency components of the image, (3) unconditional coupling flows subject to the high-frequency components and a conditional signal-dependent layer additionally subject to the low-frequency image components and finally, (4) the high-frequency image components applied to the conditional coupling flows and a signal dependent layer subject to the low-frequency components. For each of these implementations we follow the exact same training methodologies as described in Section D.1. All the images are encoded at 16 bit depth during dequantization to ensure comparability.

Model	CIFAR10 mean BPD↓	CIFAR10-C LL↑/ Typicality↑/ NSD↑
Full Image Unconditional (1)	11.32	56.6 / 64.2 / 67.8
High Frequency & Unconditional (2)	9.85	57.9 / 40.8 / 65.5
High Frequency & Unconditional + SDL (3)	9.77	58.4 / 40.0 / 62.6
High Frequency & Conditional + SDL (4)	**5.48**	58.3 / 45.5 / **74.9**

Table D.21 *Results (Bits per dimension or AUROC) obtained from the ablation Experiments on CovariateFlow with CIFAR10(-C).*

From Table D.21 it can be seen that while model 1 is limited in modeling the complete data distribution (11.32 Bits per dimension (BPD)), it performs well on detecting OOD covariate shift with NSD (AUROC 67.8%), comparable to the performance obtained with GLOW. Only using the high-frequency image components in an unconditional setting (model 2) yields a somewhat lower OOD detection performance of 65.5% AUROC. Introducing the SDL (model 3), lowers the mean BPD and improves on LL-based OOD detection (58.4%), but adversely effects the NSD evaluation (62.6%). While the SDL layer does not show improvement in the detection performance, it significantly aided in satabalizing model training. Finally (model 4), conditioning every coupling flow in the network on the low-frequency content significantly improves modelling the high-frequency components (9.77 BPD $\rightarrow$ 5.48 BPD), indicating the value in the additional information. Modelling this conditional relation between the low-frequency and

high-frequency components also proves very effective in detecting OOD covariate shift. The model achieves a mean AUROC of 74.9% at detecting covariate factors across all variations and degredations when evaluated with NSD. Table D.22

Model	Mean distance / CIFAR10-C severity 1	2	3	4	5	Models Size (# parameters)	Inference Speed milliseconds
Vanilla VAE [259] (SSIM + KL Div)	0.0365	0.0367	0.0381	0.0408	0.0430	9,436,867	4.1ms
AVAE [267] (MSE + KL Div + Adv Loss)	0.066	0.073	0.078	0.086	0.0992	11,002,373	9.1ms
DDPM [269] (T20: LPIPS)	0.7	0.8	0.9	1.0	1.4	17,714,563	34.1ms
DDPM [269] (T20: LPIPS + MSE)	1.6	1.8	2.3	2.8	3.6	17,714,563	34.1ms
GLOW [206] (LL)	0.73	1.1	1.2	1.4	238,10.9	44,235,312	65.8ms
GLOW [258] (Typicality)	2.2	2.8	2.8	2.9	3.2	44,235,312	178.3ms
GLOW (NSD)	1.2	1.7	1.9	2.2	411,753.0	44,235,312	178.3ms
CovariateFlow (LL)	1.1	1.5	1.7	2.0	2.2	945,882	22.5ms
CovariateFlow (Typicality)	0.02	0.03	0.04	0.05	0.07	945,882	59.6ms
CovariateFlow (NSD)	2.3	2.9	3.4	3.7	4.3	945,882	59.6ms

Table D.22 *Model specific details and results. The mean distance (measured differently per model) per severity, the number of trainable parameters and the inference time are depicted. Note that the DDPM is evaluated multiple times to obtain a detection score.*

presents additional information about each of the models employed in this research. This table showcases the mean distance measurements (CIFAR10-C), taken under different evaluation criteria, across increasing severity levels of covariate shifts within the dataset. Such a detailed breakdown allows for a nuanced understanding of each model's resilience and adaptability to changes in input data distribution. Notably, the LL evaluations of GLOW at the highest severity level encountered numerical stability issues, leading to the substitution of some results with the maximum representable floating-point number. This adjustment, while necessary, underscores the challenges in maintaining computational integrity under extreme conditions and the importance of implementing robust handling mechanisms for such anomalies. It is evident from the data that there is a consistent trend of increasing mean distance scores across all models as the severity level escalates, highlighting the impact of covariate shift on model performance. This trend underscores the ability to quantify covariate shift, although only briefly evaluated here. Furthermore, the table delineates the model size, quantified by the number of trainable parameters, and the inference speed, measured in milliseconds. These metrics are critical for understanding the trade-offs between model complexity, computational efficiency, and performance.The data presented in Table D.22 not only elucidates the ability to quantify covariate shifts, but also emphasizes the importance of balancing model complexity and computational efficiency when considering the model deployment conditions. Modeling the conditional distribution between the low-frequency and high-frequency components using CovariateFlow is highly effective in detecting out-of-distribution (OOD) covariate shifts. The CIFAR10 dataset, known for its diversity, encompasses a range of in-distribution (ID) covariate conditions. When assessing CovariateFlow in the context of a semantic OOD detection problem, such as distinguishing between CIFAR10 and SVHN datasets, it is plausible that some covariate conditions in CIFAR10 overlap with those in the SVHN dataset. Despite this potential overlap,

	Models	OOD SVHN [366] AUROC ↑
CIFAR10 ID	**Reconstruction**	
	DDPM [269]	**97.9*** / 95.8
	Explicit Density	
	Vanilla VAE (SSIM + KL Div)	24.4
	AVAE (MSE + KL Div + Adv Loss)	32.0
	VAE-FRL [291]	85.4*
	GLOW-FRL [291]	91.5*
	GLOW (LL)	0.7
	GLOW (Typicality)	91.3
	GLOW (NSD)	89.9
	CovariateFlow (LL)	0.3
	CovariateFlow (Typicality)	89.9
	CovariateFlow (NSD)	90.0

Table D.23 *The performance of various models on detecting SVHN as OOD when trained on CIFAR10 as ID. * indicates values taken from the published paper.*

CovariateFlow demonstrates robust performance in identifying the OOD covariate conditions present in the SVHN dataset, as evidenced by the results shown in Table D.23. Although the DDPM (utilizing all 1000 starting points) achieves the best performance, CovariateFlow offers competitive results. This is notable given its significantly smaller size and its specific design focus on covariate conditions rather than semantic content.

Figure D.11 *High-frequency sample when conditioned on a low frequency image. The sample clearly shows the conditioning on the low-frequency image, with high-frequency components generated along the watch edges, the time and more uniformly distributed background noise on the arm.*

Appendix D

Bibliography

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[2] K. N. Fockens, M. R. Jong, J. B. Jukema, T. Boers, C. Kusters, J. van der Putten, R. Pouw, L. Duits, N. Montazeri, S. van Munster, et al. "A deep learning system for detection of early Barrett's neoplasia: a model development and validation study". In: *The Lancet Digital Health* 5.12 (2023), e905–e916.

[3] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al. "International evaluation of an AI system for breast cancer screening". In: *Nature* 577.7788 (2020), pp. 89–94.

[4] S. Wang, Y. Zha, W. Li, Q. Wu, X. Li, M. Niu, M. Wang, X. Qiu, H. Li, H. Yu, W. Gong, Y. Bai, L. Li, Y. Zhu, L. Wang, and J. Tian. "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis". en. In: *Eur. Respir. J.* 56.2 (Aug. 2020), p. 2000775.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[7] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[9] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[10] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales. "Self-Supervised Representation Learning: Introduction, advances, and challenges". In: *IEEE Signal Processing Magazine* 39.3 (May 2022), 42–62. ISSN: 1558-0792.

[11] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. *SAM 2: Segment Anything in Images and Videos*. 2024. arXiv: 2408.00714 [cs.CV].

[12] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[13] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". en. In: *Nature Methods* 18.2 (2021), pp. 203–211.

[14] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. *A Comprehensive Survey on Transfer Learning*. 2020. arXiv: 1911.02685 [cs.LG].

[15] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[16] M. L. Giger and K. Suzuki. "Computer-aided diagnosis". In: *Biomedical information technology*. Elsevier, 2008, pp. 359–XXII.

[17] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers. "Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1170–1181.

[18] C.-M. Chen, Y.-H. Chou, N. Tagawa, Y. Do, et al. "Computer-Aided Detection and Diagnosis in Medical Imaging." In: *Comput. Math. Methods Medicine* 2013 (2013), pp. 790608–1.

[19] Q. Li and R. M. Nishikawa. *Computer-aided detection and diagnosis in medical imaging*. Taylor & Francis, 2015.

[20] N. M. Hassan, S. Hamad, and K. Mahar. "Mammogram breast cancer CAD systems for mass detection and classification: a review". In: *Multimedia Tools and Applications* 81.14 (2022), pp. 20043–20075.

[21] K. V. Venkadesh, A. A. Setio, A. Schreuder, E. T. Scholten, K. Chung, M. M. W. Wille, Z. Saghir, B. van Ginneken, M. Prokop, and C. Jacobs. "Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT". In: *Radiology* 300.2 (2021), pp. 438–447.

[22] M. Ramaekers, C. G. Viviers, B. V. Janssen, T. A. Hellström, L. Ewals, K. van der Wulp, J. Nederend, I. Jacobs, J. R. Pluyter, D. Mavroeidis, et al. "Computer-aided detection for pancreatic cancer diagnosis: radiological challenges and future directions". In: *Journal of Clinical Medicine* 12.13 (2023), p. 4209.

[23] S. Nagaraj, G. N. Rao, K Koteswararao, et al. "The role of pattern recognition in computer-aided diagnosis and computer-aided detection in medical imaging: a clinical validation". In: *Int. J. Comput. Appl* 8.5 (2010), pp. 18–22.

[24] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1285–1298.

[25] L Rahib, B. D. Smith, R Aizenberg, A. B. Rosenzweig, J. M. Fleshman, and L. M. Matrisian. "Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States [published correction appears in Cancer Res". In: *Cancer Res* 74.14 (2014), pp. 2913–2921.

[26] American Cancer Society. *Facts & Figures 2019*. Atlanta, GA, USA: American Cancer Society, 2019, pp. 1–76.

[27] T. Conroy, P. Hammel, M. Hebbar, M. Ben Abdelghani, A. C. Wei, J.-L. Raoul, L. Choné, E. Francois, P. Artru, J. J. Biagi, et al. "FOLFIRINOX or gemcitabine as adjuvant therapy for pancreatic cancer". In: *New England Journal of Medicine* 379.25 (2018), pp. 2395–2406.

[28] A. E. Latenstein, L. G. van der Geest, B. A. Bonsing, B. G. Koerkamp, N. H. Mohammad, I. H. de Hingh, V. E. de Meijer, I. Q. Molenaar, H. C. van Santvoort, G. van Tienhoven, et al. "Nationwide trends in incidence, treatment and survival of pancreatic ductal adenocarcinoma". In: *European Journal of Cancer* 125 (2020), pp. 83–93.

[29] M. S. D. De La Cruz, A. P. Young, and M. T. RUFFIN IV. "Diagnosis and management of pancreatic cancer". In: *American family physician* 89.8 (2014), pp. 626–632.

[30] L. G. van der Geest, V. E. Lemmens, I. H. de Hingh, C. J. van Laarhoven, T. L. Bollen, C. Y. Nio, C. H. van Eijck, O. R. Busch, and M. Besselink. "Nationwide outcomes in patients undergoing surgical exploration without resection for pancreatic cancer". In: *Journal of British Surgery* 104.11 (2017), pp. 1568–1577.

[31] G. Gheorghe, S. Bungau, M. Ilie, T. Behl, C. M. Vesa, C. Brisc, N. Bacalbasa, V. Turi, R. S. Costache, and C. C. Diaconu. "Early diagnosis of pancreatic cancer: the key for survival". In: *Diagnostics* 10.11 (2020), p. 869.

[32] B. J. Allan, S. M. Novak, M. E. Hogg, and H. J. Zeh. "Robotic vascular resections during Whipple procedure". In: *Journal of Visualized Surgery* 4 (2018).

[33] L. Zhang, S. Sanagapalli, and A. Stoita. "Challenges in diagnosis of pancreatic cancer". In: *World journal of gastroenterology* 24.19 (2018), p. 2047.

[34] C. T. Dpcg. "Staging for Adenocarcinoma of the Pancreatic Head and Uncinate Process". 2012.

[35] M. A. Tempero, M. P. Malafa, M. Al-Hawary, S. W. Behrman, A. B. Benson, D. B. Cardin, E. G. Chiorean, V. Chung, B. Czito, and D. Chiaro. "M.; et al. Pancreatic Adenocarcinoma,

Version 2.2021, NCCN Clinical Practice Guidelines in Oncology". In: *J. Natl. Compr. Canc. Netw* 19 (2021), pp. 439–457.

[36] H. J. Asbun, A. L. Moekotte, F. L. Vissers, F. Kunzler, F. Cipriani, A. Alseidi, M. I. D'Angelica, A. Balduzzi, C. Bassi, B. Björnsson, et al. "The Miami International Evidence-Based Guidelines on Minimally Invasive Pancreas Resection". In: *Ann. Surg* 271 (2020), pp. 1–14.

[37] C. Cassinotto, A. Mouries, J.-P. Lafourcade, E. Terrebonne, G. Belleannée, J.-F. Blanc, B. Lapuyade, V. Vendrely, C. Laurent, L. Chiche, et al. "Locally Advanced Pancreatic Adenocarcinoma: Reassessment of Response with CT after Neoadjuvant Chemotherapy and Radiation Therapy". In: *Radiology* 273 (2014), pp. 108–116.

[38] R. R. White, E. K. Paulson, K. S. Freed, M. T. Keogan, H. I. Hurwitz, C. Lee, M. A. Morse, M. R. Gottfried, J. Baillie, M. S. Branch, et al. "Staging of Pancreatic Cancer before and after Neoadjuvant Chemoradiation". In: *J. Gastrointest. Surg* 5 (2001), pp. 626–633.

[39] C. Cassinotto, A. Sa-Cunha, and H. Trillaud. "Radiological Evaluation of Response to Neoadjuvant Treatment in Pancreatic Cancer". In: *Diagn. Interv. Imaging* 97 (2016), pp. 1225–1232.

[40] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. "Key Challenges for Delivering Clinical Impact with Artificial Intelligence". In: *BMC Med* 17 (2019), p. 195.

[41] L. Strohm, C. Hehakaya, E. R. Ranschaert, W. P. C. Boon, and E. H. M. Moors. "Implementation of Artificial Intelligence (AI) Applications in Radiology: Hindering and Facilitating Factors". In: *Eur. Radiol* 30 (2020), pp. 5525–5532.

[42] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, et al. "Surgical data science for next-generation interventions". In: *Nature Biomedical Engineering* 1.9 (2017), pp. 691–696.

[43] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov. "Automatic instrument segmentation in robot-assisted surgery using deep learning". In: *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2018, pp. 624–628.

[44] M. Wagner, B.-P. Müller-Stich, A. Kisilenko, D. Tran, P. Heger, L. Mündermann, D. M. Lubotsky, B. Müller, T. Davitashvili, M. Capek, A. Reinke, C. Reid, T. Yu, A. Vardazaryan, C. I. Nwoye, N. Padoy, X. Liu, E.-J. Lee, C. Disch, H. Meine, T. Xia, F. Jia, S. Kondo, W. Reiter, Y. Jin, Y. Long, M. Jiang, Q. Dou, P. A. Heng, I. Twick, K. Kirtac, E. Hosgor, J. L. Bolmgren, M. Stenzel, B. von Siemens, L. Zhao, Z. Ge, H. Sun, D. Xie, M. Guo, D. Liu, H. G. Kenngott, F. Nickel, M. von Frankenberg, F. Mathis-Ullrich, A. Kopp-Schneider, L. Maier-Hein, S. Speidel, and S. Bodenstedt. "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark". In: *Medical Image Analysis* 86 (2023), p. 102770. ISSN: 1361-8415.

[45] J. Ramalhinho, S. Yoo, T. Dowrick, B. Koo, M. Somasundaram, K. Gurusamy, D. J. Hawkes, B. Davidson, A. Blandford, and M. J. Clarkson. "The value of Augmented Reality in surgery — A usability study on laparoscopic liver surgery". In: *Medical Image Analysis* 90 (2023), p. 102943. ISSN: 1361-8415.

[46] S. Malhotra, O. Halabi, S. P. Dakua, J. Padhan, S. Paul, and W. Palliyali. "Augmented reality in surgical navigation: a review of evaluation and validation metrics". In: *Applied Sciences* 13.3 (2023), p. 1629.

[47] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: 1312.6114 [stat.ML].

[48] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[49] I. Kobyzev, S. J. Prince, and M. A. Brubaker. "Normalizing flows: An introduction and review of current methods". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 3964–3979.

[50] T. Iqbal and H. Ali. "Generative adversarial network for medical images (MI-GAN)". In: *Journal of medical systems* 42.11 (2018), p. 231.

[51] A. Kendall and Y. Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30 (2017).

[52] D. Hendrycks and K. Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016).

[53] C. G. A. Viviers, M. Ramaekers, P. H. N. de With, D. Mavroeidis, J. Nederend, M. Luyer, and F. van der Sommen. "Improved Pancreatic Tumor Detection by Utilizing Clinically-Relevant

Secondary Features". In: *Cancer Prevention Through Early Detection*. Cham: Springer Nature Switzerland, 2022, pp. 139–148. ISBN: 978-3-031-17979-2.

[54] T. A. Hellström, C. G. Viviers, M. Ramaekers, N. Tasios, J. Nederend, M. D. Luyer, F. van der Sommen, et al. "Clinical segmentation for improved pancreatic ductal adenocarcinoma detection and segmentation". In: *Medical Imaging 2023: Computer-Aided Diagnosis*. Vol. 12465. SPIE. 2023, pp. 627–633.

[55] M. Ramaekers, C. G. A. Viviers, T. A. E. Hellström, L. J. S. Ewals, N. Tasios, I. Jacobs, J. Nederend, F. v. d. Sommen, and M. D. P. Luyer. "Improved Pancreatic Cancer Detection and Localization on CT Scans: A Computer-Aided Detection Model Utilizing Secondary Features". In: *Cancers* 16.13 (2024). ISSN: 2072-6694.

[56] M. Valiuddin, C. G. Viviers, R. J. van Sloun, and F. v. d. Sommen. "Improving Aleatoric Uncertainty Quantification in Multi-annotated Medical Image Segmentation with Normalizing Flows". In: Springer, 2021, pp. 75–88.

[57] C. G. Viviers, M. A. Valiuddin, F. van der Sommen, et al. "Probabilistic 3D segmentation for aleatoric uncertainty quantification in full 3D medical data". In: *Medical Imaging 2023: Computer-Aided Diagnosis*. Vol. 12465. SPIE. 2023, pp. 341–351.

[58] C. Viviers, M. Ramaekers, A. Valiuddin, T. Hellström, N. Tasios, J. van der Ven, I. Jacobs, L. Ewals, J. Nederend, M. Luyer, et al. "Segmentation-based Assessment of Tumor-Vessel Involvement for Surgical Resectability Prediction of Pancreatic Ductal Adenocarcinoma". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2421–2431.

[59] M. A. Valiuddin, C. G. Viviers, R. J. van Sloun, P. H. de With, and F. v. der Sommen. "Efficient out-of-distribution detection of melanoma with wavelet-based normalizing flows". In: *MICCAI Workshop on Cancer Prevention through Early Detection*. Springer. 2022, pp. 99–107.

[60] C. Viviers, A. Valiuddin, F. Caetano, L. Abdi, L. Filatova, P. de With, and F. van der Sommen. *Can Your Generative Model Detect Out-of-Distribution Covariate Shift?* 2024. arXiv: 2409.03043 [cs.CV].

[61] C. G. A. Viviers, J. de Bruijn, L. Filatova, P. H. N. de With, and F. van der Sommen. "Towards real-time 6D pose estimation of objects in single-view cone-beam x-ray". In: *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by C. A. Linte and J. H. Siewerdsen. Vol. 12034. International Society for Optics and Photonics. SPIE, 2022, p. 120341V.

[62] C. G. Viviers, L. Filatova, M. Termeer, P. H. de With, and F. van der Sommen. "Advancing 6-DoF Instrument Pose Estimation in Variable X-Ray Imaging Geometries". In: *IEEE Transactions on Image Processing* 33 (2024), pp. 2462–2476.

[63] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[64] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[65] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". In: *arXiv preprint arXiv:1711.05225* (2017).

[66] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[67] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography". In: *Nature medicine* 25.6 (2019), pp. 954–961.

[68] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning". In: *Nature medicine* 24.10 (2018), pp. 1559–1567.

[69] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[70] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[71] V. Badrinarayanan, A. Kendall, and R. Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[72] A. Agarap. "Deep learning using rectified linear units (relu)". In: *arXiv preprint arXiv:1803.08375* (2018).

[73] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[74] R. Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[75] S. Ren, K. He, R. Girshick, and J. Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[76] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[77] R. Khanam and M. Hussain. *YOLOv11: An Overview of the Key Architectural Enhancements*. 2024. arXiv: 2410.17725 [cs.CV].

[78] K. Yan, X. Wang, L. Lu, and R. M. Summers. "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning". In: *Journal of medical imaging* 5.3 (2018), pp. 036501–036501.

[79] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. *Focal Loss for Dense Object Detection*. 2018. arXiv: 1708.02002 [cs.CV].

[80] X. X. Lu. "A Review of Solutions for Perspective-n-Point Problem in Camera Pose Estimation". In: *Journal of Physics: Conference Series* 1087 (2018), p. 052009.

[81] C. Payer, D. Štern, H. Bischof, and M. Urschler. "Integrating spatial configuration into heatmap regression based CNNs for landmark localization". In: *Medical image analysis* 54 (2019), pp. 207–219.

[82] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. "An introduction to MCMC for machine learning". In: *Machine learning* 50 (2003), pp. 5–43.

[83] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. "An introduction to variational methods for graphical models". In: *Machine learning* 37 (1999), pp. 183–233.

[84] J. L. W. V. Jensen. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". In: *Acta mathematica* 30.1 (1906), pp. 175–193.

[85] D. Chandler. "Introduction to modern statistical". In: *Mechanics. Oxford University Press, Oxford, UK* 5.449 (1987), p. 11.

[86] D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286.

[87] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[88] I. Kobyzev, S. Prince, and M. Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. ISSN: 1939-3539.

[89] D. Rezende and S. Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[90] L. Dinh, D. Krueger, and Y. Bengio. "Nice: Non-linear independent components estimation". In: *arXiv preprint arXiv:1410.8516* (2014).

[91] L. Dinh, J. Sohl-Dickstein, and S. Bengio. "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (2016).

[92] E. Hüllermeier and W. Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110 (2021), pp. 457–506.

[93] J. Postels, M. Segu, T. Sun, L. Sieber, L. Van Gool, F. Yu, and F. Tombari. "On the practicality of deterministic epistemic uncertainty". In: *arXiv preprint arXiv:2107.00649* (2021).

[94] D. Yoo and I. S. Kweon. "Learning loss for active learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 93–102.

[95] S. Lahlou, M. Jain, H. Nekoei, V. I. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, and Y. Bengio. "Deup: Direct epistemic uncertainty prediction". In: *arXiv preprint arXiv:2102.08501* (2021).

[96] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. "Uncertainty estimation using a single deep deterministic neural network". In: *International conference on machine learning*. PMLR. 2020, pp. 9690–9700.

[97] K. Lee, K. Lee, H. Lee, and J. Shin. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks". In: *Advances in neural information processing systems* 31 (2018).

[98] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan. "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness". In: *Advances in neural information processing systems* 33 (2020), pp. 7498–7512.

[99] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. "Laplace redux-effortless bayesian deep learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 20089–20103.

[100] M. Collier, R. Jenatton, B. Mustafa, N. Houlsby, J. Berent, and E. Kokiopoulou. "Massively scaling heteroscedastic classifiers". In: *arXiv preprint arXiv:2301.12860* (2023).

[101] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[102] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (2017).

[103] S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, and D. Batra. "Why m heads are better than one: Training a diverse ensemble of deep networks". In: *arXiv preprint arXiv:1511.06314* (2015).

[104] B. Mucsányi, M. Kirchhof, and S. J. Oh. "Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks". In: *arXiv preprint arXiv:2402.19460* (2024).

[105] S. C. Hora. "Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management". In: *Reliability Engineering & System Safety* 54.2-3 (1996), pp. 217–223.

[106] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier. "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?" In: *Uncertainty in Artificial Intelligence*. PMLR. 2023, pp. 2282–2292.

[107] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning". In: *International conference on machine learning*. PMLR. 2018, pp. 1184–1193.

[108] M. H. Shaker and E. Hüllermeier. "Ensemble-based uncertainty quantification: Bayesian versus credal inference". In: *Workshop Computational Intelligence, Proceedings 31*. Vol. 25. 2021, p. 63.

[109] D. Pfau. "A generalized bias-variance decomposition for bregman divergences". In: *Unpublished manuscript* (2013).

[110] N. Gupta, J. Smith, B. Adlam, and Z. Mariet. "Ensembling over classifiers: a bias-variance perspective". In: *arXiv preprint arXiv:2206.10566* (2022).

[111] U. Von Luxburg and B. Schölkopf. "Statistical learning theory: Models, concepts, and results". In: *Handbook of the History of Logic*. Vol. 10. Elsevier, 2011, pp. 651–706.

[112] B. Agarwal, A. M. Correa, and L. Ho. "Survival in Pancreatic Carcinoma Based on Tumor Size". In: *Pancreas* 36, e (2008), pp. 15–20.

[113] R. L. Siegel, K. D. Miller, and A. C. S. Jemal. "CA". In: *Cancer J. Clin.* 2019.69 (2019), pp. 7–34.

[114] J. C. Ardengh, G. A. de Paulo, and A. P. Ferrari. "Pancreatic Carcinomas Smaller than 3.0 Cm: Endosonography (EUS) in Diagnosis". In: *Staging and Prediction of Resectability. HPB (Oxford)* 5 (2003), pp. 226–230.

[115] K. Y. Elbanna, H.-J. Jang, and T. K. Kim. "Imaging Diagnosis and Staging of Pancreatic Ductal Adenocarcinoma: A Comprehensive Review". In: *Insights Imaging* 11 (2020), p. 58.

[116] J. D. Kang, S. E. Clarke, and A. F. Costa. "Factors Associated with Missed and Misinterpreted Cases of Pancreatic Ductal Adenocarcinoma". In: *Eur. Radiol* 31 (2021), pp. 2422–2432.

[117] S. H. Yoon, J. M. Lee, J. Y. Cho, K. B. Lee, J. E. Kim, S. K. Moon, S. J. Kim, J. H. Baek, S. H. Kim, S. H. Kim, et al. "Small". In: 20 Mm) Pancreatic Adenocarcinomas: Analysis of Enhancement Patterns and Secondary Signs with Multiphasic Multidetector CT. Radiology 259 (2011), pp. 442–452.

[118] J. C. Wong and S. Raman. "Surgical Resectability of Pancreatic Adenocarcinoma: CTA". In: *Abdom. Imaging* 35 (2010), pp. 471–480.

[119] S. Gangi, J. G. Fletcher, M. A. Nathan, J. A. Christensen, W. S. Harmsen, B. S. Crownhart, and S. T. Chari. "Time Interval between Abnormalities Seen on CT and the Clinical Diagnosis of Pancreatic Cancer: Retrospective Review of CT Scans Obtained before Diagnosis". In: *AJR. Am. J. Roentgenol* 182 (2004), pp. 897–903.

[120] K. M. Jang, S. H. Kim, Y. K. Kim, K. D. Song, S. J. Lee, and D. M. P. D. A. Choi. "Assessment of Early Imaging Findings on Prediagnostic Magnetic Resonance Imaging". In: *Eur. J. Radiol* 84 (2015), pp. 1473–1479.

[121] S. S. Ahn, M.-J. Kim, J.-Y. Choi, H.-S. Hong, Y. E. Chung, and J. S. Lim. "Indicative Findings of Pancreatic Cancer in Prediagnostic CT". In: *Eur. Radiol* 19 (2009), pp. 2448–2455.

[122] D. P. Singh, S. Sheedy, A. H. Goenka, M. Wells, N. J. Lee, J. Barlow, A. Sharma, H. Kandlakunta, S. Chandra, S. K. Garg, et al. "Computerized Tomography Scan in Pre-Diagnostic Pancreatic Ductal Adenocarcinoma: Stages of Progression and Potential Benefits of Early Intervention: A Retrospective Study". In: *Pancreatology* 20 (2020), pp. 1495–1501.

[123] M. J. A. M. Bakens, Y. R. B. M. van Gestel, M. Bongers, M. G. H. Besselink, C. H. C. Dejong, I. Q. Molenaar, O. R. C. Busch, V. E. P. P. Lemmens, and I. H. J. de Hingh. "T.; Dutch Pancreatic Cancer Group Hospital of Diagnosis and Likelihood of Surgical Treatment for Pancreatic Cancer". In: *Br. J. Surg* 102 (2015), pp. 1670–1675.

[124] J. R. Treadwell, H. M. Zafar, M. D. Mitchell, K. Tipton, U. Teitelbaum, and J. Jue. "Imaging tests for the diagnosis and staging of pancreatic adenocarcinoma: A meta-analysis". en. In: *Pancreas* 45.6 (2016), pp. 789–795.

[125] M. Hidalgo. "Pancreatic cancer". en. In: *N. Engl. J. Med.* 362.17 (2010), pp. 1605–1617.

[126] S. Park, L. C. Chu, R. H. Hruban, B. Vogelstein, K. W. Kinzler, A. L. Yuille, D. F. Fouladi, S. Shayesteh, S. Ghandili, C. L. Wolfgang, et al. "Differentiating Autoimmune Pancreatitis from Pancreatic Ductal Adenocarcinoma with CT Radiomics Features". In: *Diagn. Interv. Imaging* 101 (2020), pp. 555–564.

[127] S. Ziegelmayer, G. Kaissis, F. Harder, F. Jungmann, T. Müller, M. Makowski, and R. Braren. "Deep Convolutional Neural Network-Assisted Feature Extraction for Diagnostic Discrimination and Feature Visualization in Pancreatic Ductal Adenocarcinoma (PDAC) versus Autoimmune Pancreatitis". In: *AIP). J. Clin. Med* 9 (2020).

[128] F. Rigiroli, J. Hoye, R. Lerebours, K. J. Lafata, C. Li, M. Meyer, P. Lyu, Y. Ding, F. R. Schwartz, N. B. Mettu, et al. "CT Radiomic Features of Superior Mesenteric Artery Involvement in Pancreatic Ductal Adenocarcinoma: A Pilot Study". In: *Radiology* 301 (2021), pp. 610–622.

[129] L. C. Chu, S. Park, S. Kawamoto, Y. Wang, Y. Zhou, W. Shen, Z. Zhu, Y. Xia, L. Xie, F. Liu, et al. "Application of Deep Learning to Pancreatic Cancer Detection: Lessons Learned From Our Initial Experience". In: *J. Am. Coll. Radiol* 16 (2019), pp. 1338–1342.

[130] S.-L. Liu, S. Li, Y.-T. Guo, Y.-P. Zhou, Z.-D. Zhang, S. Li, and Y. Lu. "Establishment and Application of an Artificial Intelligence Diagnosis System for Pancreatic Cancer with a Faster Region-Based Convolutional Neural Network". In: *Chin.* Med. J. (Engl) 132, 2019, pp. 2795–2803.

[131] Z. Zhu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille. "Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer. 2019, pp. 3–12.

[132] L. C. Chu, S. Park, S. Kawamoto, D. F. Fouladi, S. Shayesteh, E. S. Zinreich, J. S. Graves, K. M. Horton, R. H. Hruban, A. L. Yuille, et al. "Utility of CT Radiomics Features in Differentiation of Pancreatic Ductal Adenocarcinoma From Normal Pancreatic Tissue". In: *AJR. Am. J. Roentgenol* 213 (2019), pp. 349–357.

[133] K.-L. Liu, T. Wu, P.-T. Chen, Y. M. Tsai, H. Roth, M.-S. Wu, W.-C. Liao, and W. Wang. "Deep Learning to Distinguish Pancreatic Cancer Tissue from Non-Cancerous Pancreatic Tissue: A Retrospective Study with Cross-Racial External Validation". In: *Lancet. Digit. Heal* 2, e303–e313 (2020).

[134] Z. Zhang, S. Li, Z. Wang, and Y. A. Lu. "Novel and Efficient Tumor Detection Framework for Pancreatic Cancer via CT Images". 2020.

[135] H. Ma, Z.-X. Liu, J.-J. Zhang, F.-T. Wu, C.-F. Xu, Z. Shen, C.-H. Yu, and Y.-M. Li. "Construction of a Convolutional Neural Network Classifier Developed by Computed Tomography Images for Pancreatic Cancer Diagnosis". In: *World J. Gastroenterol* 26 (2020), pp. 5156–5168.

[136] K. Si, Y. Xue, X. Yu, X. Zhu, Q. Li, W. Gong, T. Liang, and S. Duan. "Fully End-to-End Deep-Learning-Based Diagnosis of Pancreatic Tumors". In: *Theranostics* 11 (2021), pp. 1982–1990.

[137] J.-J. Qiu, J. Yin, W. Qian, J.-H. Liu, Z.-X. Huang, H.-P. Yu, L. Ji, and X.-X. A. Zeng. "Novel Multiresolution-Statistical Texture Analysis Architecture: Radiomics-Aided Diagnosis of PDAC Based on Plain CT Images". In: *Med*. Ed. by I. Trans. Imaging 40, 2021, pp. 12–25.

[138] S. Ebrahimian, R. Singh, A. Netaji, K. S. Madhusudhan, F. Homayounieh, A. Primak, F. Lades, S. Saini, M. K. Kalra, and S. Sharma. "Characterization of Benign and Malignant Pancreatic Lesions with DECT Quantitative Metrics and Radiomics". In: *Acad. Radiol* 29 (2022), pp. 705–713.

[139] N. Alves, M. Schuurmans, G. Litjens, J. S. Bosma, J. Hermans, and H. Huisman. "Fully Automatic Deep Learning Framework for Pancreatic Ductal Adenocarcinoma Detection on Computed Tomography". In: *Cancers (Basel)* 14 (2022).

[140] B. Kenner, S. T. Chari, D. Kelsen, D. S. Klimstra, S. J. Pandol, M. Rosenthal, A. K. Rustgi, J. A. Taylor, A. Yala, N. Abul-Husn, et al. "Artificial Intelligence and Early Detection of Pancreatic Cancer: 2020 Summative Review". In: *Pancreas* 50 (2021), pp. 251–279.

[141] G. Kaissis, S. Ziegelmayer, F. Lohöfer, H. Algül, M. Eiber, W. Weichert, R. Schmid, H. Friess, E. Rummeny, D. Ankerst, et al. "A Machine Learning Model for the Prediction of Survival and Tumor Subtype in Pancreatic Ductal Adenocarcinoma from Preoperative Diffusion-Weighted Imaging". In: *Eur. Radiol. Exp* 3 (2019), p. 41.

[142] G. Kaissis, S. Ziegelmayer, F. Lohöfer, K. Steiger, H. Algül, A. Muckenhuber, H.-Y. Yen, E. Rummeny, H. Friess, R. Schmid, et al. "A Machine Learning Algorithm Predicts Molecular Subtypes in Pancreatic Ductal Adenocarcinoma with Differential Response to Gemcitabine-Based versus FOLFIRINOX Chemotherapy". In: *PLoS One* 14, e0218642 (2019).

[143] G. A. Kaissis, S. Ziegelmayer, F. K. Lohöfer, F. N. Harder, F. Jungmann, D. Sasse, A. Muckenhuber, H.-Y. Yen, K. Steiger, J. Siveke, et al. "Image-Based Molecular Phenotyping of Pancreatic Ductal Adenocarcinoma". In: *J. Clin. Med* 9 (2020).

[144] Y. Liang, D. Schott, Y. Zhang, Z. Wang, H. Nasief, E. Paulson, W. Hall, P. Knechtges, B. Erickson, and X. A. Li. "Auto-Segmentation of Pancreatic Tumor in Multi-Parametric MRI Using Deep Convolutional Neural Networks". In: *Radiother. Oncol* 145 (2020), pp. 193–200.

[145] X. Gao and X. Wang. "Deep Learning for World Health Organization Grades of Pancreatic Neuroendocrine Tumors on Contrast-Enhanced Magnetic Resonance Images: A Preliminary Study". In: *Int. J. Comput. Assist. Radiol. Surg* 14 (2019), pp. 1981–1991.

[146] J. E. Corral, S. Hussein, P. Kandel, C. W. Bolan, U. Bagci, and M. B. Wallace. "Deep Learning to Classify Intraductal Papillary Mucinous Neoplasms Using Magnetic Resonance Imaging". In: *Pancreas* 48 (2019), pp. 805–810.

[147] X. Gao and X. Wang. "Performance of Deep Learning for Differentiating Pancreatic Diseases on Contrast-Enhanced Magnetic Resonance Imaging: A Preliminary Study". In: *Diagn. Interv. Imaging* 101 (2020), pp. 91–100.

[148] Y. Deng, B. Ming, T. Zhou, J.-L. Wu, Y. Chen, P. Liu, J. Zhang, S.-Y. Zhang, T.-W. Chen, and X.-M. Zhang. "Radiomics Model Based on MR Images to Discriminate Pancreatic Ductal Adenocarcinoma and Mass-Forming Chronic Pancreatitis Lesions". In: *Front. Oncol* 11 (2021). Article 620981.

[149] L. Zhang, S. Sanagapalli, and A. Stoita. "Challenges in diagnosis of pancreatic cancer". en. In: *World J. Gastroenterol.* 24.19 (2018), pp. 2047–2060.

[150] E. S. Lee and J. M. Lee. "Imaging diagnosis of pancreatic cancer: a state-of-the-art review". en. In: *World J. Gastroenterol.* 20.24 (2014), pp. 7864–7877.

[151] S. S. Ahn, M.-J. Kim, J.-Y. Choi, H.-S. Hong, Y. E. Chung, and J. S. Lim. "Indicative findings of pancreatic cancer in prediagnostic CT". en. In: *Eur. Radiol.* 19.10 (2009), pp. 2448–2455.

[152] M. Kriegsmann, K. Kriegsmann, G. Steinbuss, C. Zgorzelski, A. Kraft, and M. M. Gaida. "Deep learning in pancreatic tissue: Identification of anatomical structures, pancreatic intraepithelial neoplasia, and ductal adenocarcinoma". en. In: *Int. J. Mol. Sci.* 22.10 (2021), p. 5385.

[153] J. Petch, S. Di, and W. Nelson. "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology". In: *Canadian Journal of Cardiology* 38.2 (2022), pp. 204–213. ISSN: 0828-282X.

[154] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Ed. by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells. Cham: Springer International Publishing, 2016, pp. 424–432. ISBN: 978-3-319-46723-8.

[155] K.-L. Liu, T. Wu, P.-T. Chen, Y. M. Tsai, H. Roth, M.-S. Wu, W.-C. Liao, and W. Wang. "Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation". In: *The Lancet Digital Health* 2.6 (2020), e303–e313. ISSN: 2589-7500.

[156] K. Si, Y. Xue, X. Yu, X. Zhu, Q. Li, W. Gong, T. Liang, and S. Duan. "Fully end-to-end deep-learning-based diagnosis of pancreatic tumors". en. In: *Theranostics* 11.4 (2021), pp. 1982–1990.

[157] Z. Zhu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille. "Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma". In: *Lecture Notes in Computer Science*. Lecture notes in computer science. Cham: Springer International Publishing, 2019, pp. 3–12.

[158] N. Alves, M. Schuurmans, G. Litjens, J. S. Bosma, J. Hermans, and H. Huisman. "Fully Automatic Deep Learning Framework for Pancreatic Ductal Adenocarcinoma Detection on Computed Tomography". In: *Cancers* 14.2 (2022). ISSN: 2072-6694.

[159] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. Gollub, J. Golia-Pernicka, S. H. Heckers, W. R. Jarnagin, M. K. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso. *A large annotated medical image dataset for the development and evaluation of segmentation algorithms*. 2019.

[160] A. Wolny, L. Cerrone, A. Vijayan, R. Tofanelli, A. V. Barro, M. Louveaux, C. Wenzl, S. Strauss, D. Wilson-Sánchez, R. Lymbouridou, S. S. Steigleder, C. Pape, A. Bailoni, S. Duran-Nebreda, G. W. Bassel, J. U. Lohmann, M. Tsiantis, F. A. Hamprecht, K. Schneitz, A. Maizel, and A. Kreshuk. "Accurate and versatile 3D segmentation of plant tissues at cellular resolution". In: *eLife* 9 (2020). Ed. by C. S. Hardtke, D. C. Bergmann, D. C. Bergmann, and M. Graeff, e57613. ISSN: 2050-084X.

[161] H. Yang, C. Shan, A. Bouwman, A. F. Kolen, and P. H. de With. "Efficient and Robust Instrument Segmentation in 3D Ultrasound Using Patch-of-Interest-FuseNet with Hybrid Loss". In: *Medical Image Analysis* 67 (2021), p. 101842. ISSN: 1361-8415.

[162] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[163] W. Silversmith. *cc3d: Connected components on multilabel 3D & 2D images, version: 3.2.1*. 2021.

[164] G. H. Berkelmans, L. F. Fransen, A. C. Dolmans-Zwartjes, E. A. Kouwenhoven, M. J. van Det, M. Nilsson, G. A. Nieuwenhuijzen, and M. D. Luyer. *Direct oral feeding following minimally invasive esophagectomy (NUTRIENT II trial): an international, multicenter, open-label randomized controlled trial*. 2020.

[165] Y. Wu and K. He. *Group Normalization*. 2018. arXiv: 1803.08494 [cs.CV].

[166] R. Raza, U. Ijaz Bajwa, Y. Mehmood, M. Waqas Anwar, and M. Hassan Jamal. "dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI". In: *Biomedical Signal Processing and Control* 79 (2023), p. 103861. ISSN: 1746-8094.

[167] I. Loshchilov and F. Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[168] P.-T. Chen, T. Wu, P. Wang, D. Chang, K.-L. Liu, M.-S. Wu, H. R. Roth, P.-C. Lee, W.-C. Liao, and W. Wang. “Pancreatic cancer detection on CT scans with deep learning: a nationwide population-based study”. In: *Radiology* 306.1 (2023), pp. 172–182.

[169] Y. Wang, P. Tang, Y. Zhou, W. Shen, E. K. Fishman, and A. L. Yuille. “Learning inductive attention guidance for partially supervised pancreatic ductal adenocarcinoma prediction”. In: *IEEE transactions on medical imaging* 40.10 (2021), pp. 2723–2735.

[170] Z. Zhu, Y. Lu, W. Shen, E. K. Fishman, and A. L. Yuille. “Segmentation for classification of screening pancreatic neuroendocrine tumors”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3402–3408.

[171] K. Si, Y. Xue, X. Yu, X. Zhu, Q. Li, W. Gong, T. Liang, and S. Duan. “Fully end-to-end deep-learning-based diagnosis of pancreatic tumors”. In: *Theranostics* 11.4 (2021), p. 1982.

[172] M. Li, F. Lian, C. Wang, and S. Guo. “Accurate pancreas segmentation using multi-level pyramidal pooling residual U-Net with adversarial mechanism. BMC Med Imaging”. In: *21(1):168. ; PMCID:* PMC8588719 (2021). PMID: 34772359.

[173] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, and A. U-Net. “Learning Where to Look for the Pancreas”.

[174] M. Huang, C. Huang, J. Yuan, and D. A. Kong. “Semiautomated Deep Learning Approach for Pancreas Segmentation. J Healthc Eng”. In: *2021:3284493. ; PMCID:* PMC8272661 (2021). PMID: 34306587.

[175] A. Patra, K. Panagiotis, G. Suman, A. Panda, S. K. Garg, and A. Goenka. “Abstract PO084: Automated detection of pancreatic ductal adenocarcinoma (PDAC) on CT scans using artificial intelligence (AI): Impact of inclusion of automated pancreas segmentation on the accuracy of 3D-convolutional neural network”. In: *CNN). Clinical Cancer Research* 27:PO–084, . Num Pages: PO-084 (2021).

[176] S. P. Pereira et al. “Early detection of pancreatic cancer”. In: *Lancet Gastroenterol. Hepatol.* 5 (2020), pp. 698–710.

[177] M. Dbouk et al. “The multicenter Cancer of Pancreas Screening study: impact on stage and survival”. In: *J. Clin. Oncol.* 40 (2022), pp. 3257–3266.

[178] K. Blouhos, K. A. Boulas, K. Tsalis, and A. Hatzigeorgiadis. “The isoattenuating pancreatic adenocarcinoma: Review of the literature and critical analysis”. In: *Surg. Oncol* 24 (2015), pp. 322–328.

[179] J. X. Hu, C. F. Zhao, W. B. Chen, Q. C. Liu, Q. W. Li, Y. Y. Lin, and G. F. P. cancer. “A review of epidemiology, trend, and risk factors. World J Gastroenterol”. In: *27(27):4298-4321. ; PMCID:* PMC8316912 (2021). PMID: 34366606.

[180] Y. LeCun, Y. Bengio, and G. Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[181] D. van de Sande, M. E. Van Genderen, J. M. Smit, J. Huiskens, J. J. Visser, R. E. R. Veen, E. van Unen, O. H. Ba, D. Gommers, and J. v. D. Bommel. “Implementing and Governing Artificial Intelligence in Medicine: A Step-by-Step Approach to Prevent an Artificial Intelligence Winter”. In: *BMJ Heal. care informatics* 29 (2022).

[182] W. Bouwmeester, N. P. A. Zuithoff, S. Mallett, M. I. Geerlings, Y. Vergouwe, E. W. Steyerberg, D. G. Altman, and K. G. M. Moons. “Reporting and Methods in Clinical Prediction Research: A Systematic Review”. In: *PLoS Med* 9 (2012), pp. 1–12.

[183] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons. “Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis”. In: *TRIPOD): The TRIPOD Statement. Br. J. Surg* 102 (2015), pp. 148–158.

[184] G. S. Collins and K. G. M. Moons. “Reporting of Artificial Intelligence Prediction Models”. In: *Lancet (London, England)* 393 (2019), pp. 1577–1579.

[185] R. D. Riley, J. Ensor, K. I. E. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. M. Moons, G. Collins, and M. van Smeden. “Calculating the Sample Size Required for Developing a Clinical Prediction Model”. In: *BMJ* 368, m441 (2020).

[186] X. Luo, G. Wang, T. Song, J. Zhang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang. “MIDeepSeg: Minimally Interactive Segmentation of Unseen Objects from Medical Images Using Deep Learning”. In: *Med.* Ed. by I. Anal. 102102: 72, 2021.

[187] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, et al. "Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy". In: *Ophthalmology* 126 (2019), pp. 552–564.

[188] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, et al. "Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making". 2019.

[189] S. A. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. H. Maier-Hein, S. Eslami, D. J. Rezende, and O. Ronneberger. "A probabilistic u-net for segmentation of ambiguous images". In: *arXiv preprint arXiv:1806.05034* (2018).

[190] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[191] T. Nair, D. Precup, D. L. Arnold, and T. Arbel. "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation". In: *Medical image analysis* 59 (2020), p. 101557.

[192] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl. "Leveraging uncertainty information from deep neural networks for disease detection". In: *Scientific reports* 7.1 (2017), pp. 1–14.

[193] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. Fu. "A review of uncertainty estimation and its application in medical imaging". In: *Meta-Radiology* (2023), p. 100003.

[194] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[195] K. Sohn, H. Lee, and X. Yan. "Learning structured output representation using deep conditional generative models". In: *Advances in neural information processing systems* 28 (2015), pp. 3483–3491.

[196] S. Armato III, G. Mclennan, L. Bidaut, M. McNitt-Gray, C. Meyer, A. Reeves, B. Zhao, D. Aberle, C. Henschke, E. Hoffman, E. Kazerooni, H. Macmahon, E. Beek, D. Yankelevitz, A. Biancardi, P. Bland, M. Brown, R. Engelmann, G. Laderach, and L. Clarke. "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans". In: *Medical Physics* 38 (Jan. 2011), pp. 915–931.

[197] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. "The cityscapes dataset for semantic urban scene understanding". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3213–3223.

[198] A. Kendall, V. Badrinarayanan, and R. Cipolla. "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding". In: *arXiv preprint arXiv:1511.02680* (2015).

[199] R. Selvan, F. Faye, J. Middleton, and A. Pai. *Uncertainty quantification in medical image segmentation with normalizing flows*. 2020. arXiv: `2006.02683 [stat.ML]`.

[200] S. Hu, D. Worrall, S. Knegt, B. Veeling, H. Huisman, and M. Welling. *Supervised Uncertainty Quantification for Segmentation with Multiple Annotations*. 2019. arXiv: `1907.01949 [cs.LG]`.

[201] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen. *Kvasir-SEG: A Segmented Polyp Dataset*. 2019. arXiv: `1911.07069 [eess.IV]`.

[202] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. *PHiSeg: Capturing Uncertainty in Medical Image Segmentation*. 2019. arXiv: `1906.04045 [eess.IV]`.

[203] K. Pogorelov, K. Randel, C. Griwodz, T. de Lange, S. Eskeland, D. Johansen, C. Spampinato, D. T. Dang Nguyen, M. Lux, P. Schmidt, M. Riegler, and P. Halvorsen. "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection". In: June 2017.

[204] R. van den Berg, L. Hasenclever, J. M. Tomczak, and M. Welling. *Sylvester Normalizing Flows for Variational Inference*. 2019. arXiv: `1803.05649 [stat.ML]`.

[205] S. A. A. Kohl, B. Romera-Paredes, K. H. Maier-Hein, D. J. Rezende, S. M. A. Eslami, P. Kohli, A. Zisserman, and O. Ronneberger. *A Hierarchical Probabilistic U-Net for Modeling Multi-Scale Ambiguities*. 2019. arXiv: `1905.13077 [cs.CV]`.

[206] D. P. Kingma and P. Dhariwal. *Glow: Generative Flow with Invertible 1x1 Convolutions*. 2018. arXiv: 1807.03039 [stat.ML].

[207] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. "Cyclical annealing schedule: A simple approach to mitigating kl vanishing". In: *arXiv preprint arXiv:1903.10145* (2019).

[208] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. "Fixing a broken ELBO". In: *International conference on machine learning*. PMLR. 2018, pp. 159–168.

[209] L. Rahib, B. D. Smith, R. Aizenberg, A. B. Rosenzweig, J. M. Fleshman, and L. M. Matrisian. "Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States". en. In: *Cancer Res.* 74.11 (June 2014), pp. 2913–2921.

[210] L. G. de la Santa, J. A. P. Retortillo, A. C. Miguel, and L. M. Klein. "Radiology of pancreatic neoplasms: An update". en. In: *World J. Gastrointest. Oncol.* 6.9 (Sept. 2014), pp. 330–343.

[211] P. Tummala, O. Junaidi, and B. Agarwal. "Imaging of pancreatic cancer: An overview". In: *Journal of gastrointestinal oncology* 2.3 (2011), p. 168.

[212] M. A. Tempero, M. P. Malafa, M. Al-Hawary, S. W. Behrman, A. B. Benson, D. B. Cardin, E. G. Chiorean, V. Chung, B. Czito, M. Del Chiaro, et al. "Pancreatic adenocarcinoma, version 2.2021, NCCN clinical practice guidelines in oncology". In: *Journal of the National Comprehensive Cancer Network* 19.4 (2021), pp. 439–457.

[213] E. Versteijne, C. H. J. van Eijck, C. J. A. Punt, M. Suker, A. H. Zwinderman, M. A. C. Dohmen, K. B. C. Groothuis, O. R. C. Busch, M. G. H. Besselink, I. H. J. T. de Hingh, A. J. Ten Tije, G. A. Patijn, B. A. Bonsing, J. de Vos-Geelen, J. M. Klaase, S. Festen, D. Boerma, J. I. Erdmann, I. Q. Molenaar, E. van der Harst, M. B. van der Kolk, C. R. N. Rasch, G. van Tienhoven, and Dutch Pancreatic Cancer Group (DPCG). "Preoperative radiochemotherapy versus immediate surgery for resectable and borderline resectable pancreatic cancer (PREOPANC trial): study protocol for a multicentre randomized controlled trial". en. In: *Trials* 17.1 (Mar. 2016), p. 127.

[214] J. R. Treadwell, H. M. Zafar, M. D. Mitchell, K. Tipton, U. Teitelbaum, and J. Jue. "Imaging tests for the diagnosis and staging of pancreatic adenocarcinoma: A meta-analysis". en. In: *Pancreas* 45.6 (July 2016), pp. 789–795.

[215] L. G. de la Santa, J. A. P. Retortillo, A. C. Miguel, and L. M. Klein. "Radiology of pancreatic neoplasms: An update". In: *World journal of gastrointestinal oncology* 6.9 (2014), p. 330.

[216] E. Versteijne, E. Lens, A. van der Horst, A. Bel, J. Visser, C. J. A. Punt, M. Suker, C. H. J. van Eijck, and G. van Tienhoven. "Quality assurance of the PREOPANC trial (2012-003181-40) for preoperative radiochemotherapy in pancreatic cancer : The dummy run". en. In: *Strahlenther. Onkol.* 193.8 (Aug. 2017), pp. 630–638.

[217] H. J. Asbun, A. L. Moekotte, F. L. Vissers, F. Kunzler, F. Cipriani, A. Alseidi, M. I. D'Angelica, A. Balduzzi, C. Bassi, B. Björnsson, et al. "The Miami international evidence-based guidelines on minimally invasive pancreas resection". In: *Annals of surgery* 271.1 (2020), pp. 1–14.

[218] F. Alemi, F. G. Rocha, W. S. Helton, T. Biehl, and A. Alseidi. "Classification and techniques of en bloc venous reconstruction for pancreaticoduodenectomy". In: *HPB* 18.10 (2016), pp. 827–834.

[219] E. Lermite, D. Sommacale, T. Piardi, J.-P. Arnaud, A. Sauvanet, C. H. Dejong, and P. Pessaux. "Complications after pancreatic resection: diagnosis, prevention and management". In: *Clinics and research in hepatology and gastroenterology* 37.3 (2013), pp. 230–239.

[220] E. Versteijne, O. J. Gurney-Champion, A. van der Horst, E. Lens, M. W. Kolff, J. Buijsen, G. Ebrahimi, K. J. Neelis, C. Rasch, J. Stoker, et al. "Considerable interobserver variation in delineation of pancreatic cancer on 3DCT and 4DCT: a multi-institutional study". In: *Radiation Oncology* 12.1 (2017), pp. 1–10.

[221] I. Joo, J. M. Lee, E. S. Lee, J.-Y. Son, D. H. Lee, S. J. Ahn, W. Chang, S. M. Lee, H.-J. Kang, and H. K. Yang. "Preoperative CT classification of the resectability of pancreatic cancer: interobserver agreement". In: *Radiology* 293.2 (2019), pp. 343–349.

[222] H. S. Tran Cao, A. Balachandran, H. Wang, G. M. Nogueras-González, C. E. Bailey, J. E. Lee, P. W. Pisters, D. B. Evans, G. Varadhachary, C. H. Crane, et al. "Radiographic tumor-vein interface as a predictor of intraoperative, pathologic, and oncologic outcomes in resectable and borderline resectable pancreatic cancer". In: *Journal of Gastrointestinal Surgery* 18 (2014), pp. 269–278.

[223] C. G. A. Viviers, M. M. A. Valiuddin, P. H. N. de With, and F. van der Sommen. "Probabilistic 3D segmentation for aleatoric uncertainty quantification in full 3D medical data". In: *Medical*

Imaging 2023: Computer-Aided Diagnosis. Ed. by K. M. Iftekharuddin and W. Chen. Vol. 12465. International Society for Optics and Photonics. SPIE, 2023, p. 124651I.

[224] K.-L. Liu, T. Wu, P.-T. Chen, Y. M. Tsai, H. Roth, M.-S. Wu, W.-C. Liao, and W. Wang. "Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation". In: *The Lancet Digital Health* 2.6 (June 2020), e303–e313. ISSN: 2589-7500.

[225] T. Vaiyapuri, A. K. Dutta, I. S. H. Punithavathi, P. Duraipandy, S. S. Alotaibi, H. Alsolai, A. Mohamed, and H. Mahgoub. "Intelligent Deep-Learning-Enabled Decision-Making Medical System for Pancreatic Tumor Classification on CT Images". In: *Healthcare* 10.4 (2022). ISSN: 2227-9032.

[226] A. Patra, K. Panagiotis, G. Suman, A. Panda, S. K. Garg, and A. Goenka. "Abstract PO-084: Automated detection of pancreatic ductal adenocarcinoma (PDAC) on CT scans using artificial intelligence (AI): Impact of inclusion of automated pancreas segmentation on the accuracy of 3D-convolutional neural network (CNN)". In: *Clinical Cancer Research* 27.5_-Supplement (Mar. 2021). Num Pages: PO-084, PO–084. ISSN: 1078-0432.

[227] K. Si, Y. Xue, X. Yu, X. Zhu, Q. Li, W. Gong, T. Liang, and S. Duan. "Fully end-to-end deep-learning-based diagnosis of pancreatic tumors". In: *Theranostics* 11.4 (2021). Publisher: Ivyspring International Publisher, pp. 1982–1990. ISSN: 1838-7640.

[228] W. Wei, G. Jia, Z. Wu, T. Wang, H. Wang, K. Wei, C. Cheng, Z. Liu, and C. Zuo. "A multidomain fusion model of radiomics and deep learning to discriminate between PDAC and AIP based on 18F-FDG PET/CT images". In: *Japanese Journal of Radiology* 41.4 (Apr. 2023), pp. 417–427. ISSN: 1867-108X.

[229] H. Wang, Z. Wu, F. Wang, W. Wei, K. Wei, and Z. Liu. "MAFF: Multi-Scale and Self-Adaptive Attention Feature Fusion Network for Pancreatic Lesion Detection in PET / CT Images". In: *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*. EITCE '22. event-place: Xiamen, China. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1412–1419. ISBN: 978-1-4503-9714-8.

[230] S.-L. Liu, S. Li, Y.-T. Guo, Y.-P. Zhou, Z.-D. Zhang, S. Li, and Y. Lu. "Establishment and application of an artificial intelligence diagnosis system for pancreatic cancer with a faster region-based convolutional neural network". In: *Chinese Medical Journal* 132.23 (2019). ISSN: 0366-6999.

[231] Z. Zhang, S. Li, Z. Wang, and Y. Lu. "A Novel and Efficient Tumor Detection Framework for Pancreatic Cancer via CT Images". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2020, pp. 1160–1164.

[232] H. Ma, Z.-X. Liu, J.-J. Zhang, F.-T. Wu, C.-F. Xu, Z. Shen, C.-H. Yu, and Y.-M. Li. "Construction of a convolutional neural network classifier developed by computed tomography images for pancreatic cancer diagnosis". In: *World Journal of Gastroenterology* 26.34 (Sept. 2020), pp. 5156–5168. ISSN: 1007-9327.

[233] Z. Zhu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille. "Multi-scale Coarse-to-Fine Segmentation for Screening Pancreatic Ductal Adenocarcinoma". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan. Cham: Springer International Publishing, 2019, pp. 3–12. ISBN: 978-3-030-32226-7.

[234] Y. Wang, P. Tang, Y. Zhou, W. Shen, E. K. Fishman, and A. L. Yuille. "Learning Inductive Attention Guidance for Partially Supervised Pancreatic Ductal Adenocarcinoma Prediction". In: *IEEE Transactions on Medical Imaging* 40.10 (2021), pp. 2723–2735.

[235] Z. Zhu, Y. Lu, W. Shen, E. K. Fishman, and A. L. Yuille. "Segmentation for Classification of Screening Pancreatic Neuroendocrine Tumors". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. Oct. 2021, pp. 3402–3408.

[236] N. Alves, M. Schuurmans, G. Litjens, J. S. Bosma, J. Hermans, and H. Huisman. "Fully Automatic Deep Learning Framework for Pancreatic Ductal Adenocarcinoma Detection on Computed Tomography". In: *Cancers* 14.2 (2022). ISSN: 2072-6694.

[237] P.-T. Chen, T. Wu, P. Wang, D. Chang, K.-L. Liu, M.-S. Wu, H. R. Roth, P.-C. Lee, W.-C. Liao, and W. Wang. "Pancreatic Cancer Detection on CT Scans with Deep Learning: A Nationwide Population-based Study". In: *Radiology* 306.1 (2023). PMID: 36098642, pp. 172–182. eprint: https://doi.org/10.1148/radiol.220152.

[238] T. Mahmoudi, Z. M. Kouzahkanan, A. R. Radmard, R. Kafieh, A. Salehnia, A. H. Davarpanah, H. Arabalibeik, and A. Ahmadian. "Segmentation of pancreatic ductal adenocarcinoma (PDAC) and surrounding vessels in CT images using deep convolutional neural networks and texture descriptors". In: *Scientific Reports* 12.1 (2022), p. 3092.

[239] J. Yao, K. Cao, Y. Hou, J. Zhou, Y. Xia, I. Nogues, Q. Song, H. Jiang, X. Ye, J. Lu, G. Jin, H. Lu, C. Xie, R. Zhang, J. Xiao, Z. Liu, F. Gao, Y. Qi, X. Li, Y. Zheng, L. Lu, Y. Shi, and L. Zhang. "Deep learning for fully automated prediction of overall survival in patients undergoing resection for pancreatic cancer: A retrospective multicenter study". en. In: *Ann. Surg.* 278.1 (July 2023), e68–e79.

[240] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. "A survey of uncertainty in deep neural networks". In: *arXiv preprint arXiv:2107.03342* (2021).

[241] J. Zhang, Y. Dai, M. Xiang, D.-P. Fan, P. Moghadam, M. He, C. Walder, K. Zhang, M. Harandi, and N. Barnes. "Dense uncertainty estimation". In: *arXiv preprint arXiv:2110.06427* (2021).

[242] A. C. Society. *What is melanoma skin cancer?* Accessed: 2024-10-11. 2024.

[243] Melanoma Research Alliance. *Melanoma survival rates*. Accessed: 2024-10-11.

[244] Mayo Clinic. *Melanoma: Diagnosis and Treatment*. Accessed: 2024-10-11. 2024.

[245] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[246] P. Kirichenko, P. Izmailov, and A. G. Wilson. "Why normalizing flows fail to detect out-of-distribution data". In: *Advances in neural information processing systems* 33 (2020), pp. 20578–20589.

[247] J. J. Yu, K. G. Derpanis, and M. A. Brubaker. "Wavelet flow: Fast training of high resolution normalizing flows". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6184–6196.

[248] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. "Neural spline flows". In: *Advances in neural information processing systems* 32 (2019).

[249] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. "Improved variational inference with inverse autoregressive flow". In: *Advances in neural information processing systems* 29 (2016).

[250] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. "Neural autoregressive flows". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2078–2087.

[251] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31 (2018).

[252] P. Kirichenko, P. Izmailov, and A. G. Wilson. "Why normalizing flows fail to detect out-of-distribution data". In: *Advances in neural information processing systems* 33 (2020), pp. 20578–20589.

[253] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque. "Input complexity and out-of-distribution detection with likelihood-based generative models". In: *arXiv preprint arXiv:1909.11480* (2019).

[254] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)". In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.

[255] E. Nalisnick, A. Matsukawa, Y. W. Teh, and B. Lakshminarayanan. "Detecting out-of-distribution inputs to deep generative models using typicality". In: *arXiv preprint arXiv:1906.02994* (2019).

[256] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky. "Your classifier is secretly an energy based model and you should treat it like one". In: *arXiv preprint arXiv:1912.03263* (2019).

[257] L. Zhang, M. Goldstein, and R. Ranganath. "Understanding failures in out-of-distribution detection with deep generative models". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12427–12436.

[258] S. Chali, I. Kucher, M. Duranton, and J.-O. Klein. "Improving Normalizing Flows With the Approximate Mass for Out-of-Distribution Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 750–758.

[259] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[260] J. Yang, K. Zhou, Y. Li, and Z. Liu. "Generalized out-of-distribution detection: A survey". In: *arXiv preprint arXiv:2110.11334* (2021).

[261] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks. "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 650–656.

[262] J. Tian, Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira. "Exploring Covariate and Concept Shift for Out-of-Distribution Detection". In: *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*. 2021.

[263] E. Adhikarla, K. Zhang, J. Yu, L. Sun, J. Nicholson, and B. D. Davison. *Robust Computer Vision in an Ever-Changing World: A Survey of Techniques for Tackling Distribution Shifts*. 2023. arXiv: 2312.01540 [cs.CV].

[264] R. Karval and K. N. Singh. "Catching Silent Failures: A Machine Learning Model Monitoring and Explainability Survey". In: *2023 OITS International Conference on Information Technology (OCIT)*. 2023, pp. 526–532.

[265] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, and H. V. Poor. *Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey*. 2023. arXiv: 2303.06302 [cs.LG].

[266] B. Goyal, S. Agrawal, and B. Sohi. "Noise issues prevailing in various types of medical images". In: *Biomedical & Pharmacology Journal* 11.3 (2018), p. 1227.

[267] A. Plumerault, H. Le Borgne, and C. Hudelot. "Avae: Adversarial variational auto encoder". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 8687–8694.

[268] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. "Do deep generative models know what they don't know?" In: *arXiv preprint arXiv:1810.09136* (2018).

[269] M. S. Graham, W. H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, and J. Cardoso. "Denoising diffusion models for out-of-distribution detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2947–2956.

[270] A. Q. Nichol and P. Dhariwal. "Improved denoising diffusion probabilistic models". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8162–8171.

[271] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, and S. Vernekar. "Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance". In: *arXiv preprint arXiv:1812.02765* (2018).

[272] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. "A review of novelty detection". In: *Signal processing* 99 (2014), pp. 215–249.

[273] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. "Deep autoencoding gaussian mixture model for unsupervised anomaly detection". In: *International conference on learning representations*. 2018.

[274] S. B. David, T. Lu, T. Luu, and D. Pál. "Impossibility theorems for domain adaptation". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 129–136.

[275] R. Averly and W.-L. Chao. "Unified Out-Of-Distribution Detection: A Model-Specific Perspective". In: *arXiv preprint arXiv:2304.06813* (2023).

[276] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira. *Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data*. 2020. arXiv: 2002.11297 [cs.CV].

[277] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. *Moment Matching for Multi-Source Domain Adaptation*. 2019. arXiv: 1812.01754 [cs.CV].

[278] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. "Covariate shift by kernel mean matching". In: (2008).

[279] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. *Domain-Adversarial Training of Neural Networks*. 2016. arXiv: 1505.07818 [stat.ML].

[280] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe. "Guided image generation with conditional invertible neural networks". In: *arXiv preprint arXiv:1907.02392* (2019).

[281] M. Thomas and A. T. Joy. *Elements of information theory*. Wiley-Interscience, 2006.

[282] A. L. Caterini and G. Loaiza-Ganem. "Entropic issues in likelihood-based ood detection". In: *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*. PMLR. 2022, pp. 21–26.

[283] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

[284] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG].

[285] A. Abdelhamed, M. A. Brubaker, and M. S. Brown. "Noise flow: Noise modeling with conditional normalizing flows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 3165–3173.

[286] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel. "Flow++: Improving flow-based generative models with variational dequantization and architecture design". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2722–2730.

[287] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky. "Your classifier is secretly an energy based model and you should treat it like one". In: *arXiv preprint arXiv:1912.03263* (2019).

[288] A. Krizhevsky, G. Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[289] D. Hendrycks and T. G. Dietterich. "Benchmarking neural network robustness to common corruptions and surface variations". In: *arXiv preprint arXiv:1807.01697* (2018).

[290] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, K. Zhou, W. Zhang, Y. Li, Z. Liu, Y. Chen, and H. Li. "OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection". In: *arXiv preprint arXiv:2306.09301* (2023).

[291] M. Cai and Y. Li. "Out-of-distribution detection via frequency-regularized generative models". In: *Proceedings of the IEEE/CVF WACV*. 2023, pp. 5521–5530.

[292] A. Saha, S. Mishra, and A. C. Bovik. "Re-iqa: Unsupervised learning for image quality assessment in the wild". In: *Proceedings of the IEEE/CVF CVPR*. 2023, pp. 5846–5855.

[293] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. "KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4041–4056.

[294] E. Manson, V. A. Ampoh, E Fiagbedzi, J. Amuasi, J. Flether, and C Schandorf. "Image noise in radiography and tomography: Causes, effects and reduction techniques". In: *Curr. Trends Clin. Med. Imaging* 2.5 (2019), p. 555620.

[295] T. J. Learch, J. B. Massie, M. N. Pathria, B. A. Ahlgren, and S. R. Garfin. "Assessment of pedicle screw placement utilizing conventional radiography and computed tomography: a proposed systematic approach to improve accuracy of interpretation". en. In: *Spine (Phila Pa 1976)* 29.7 (2004), pp. 767–773.

[296] A. Elmi-Terander, G. Burström, R. Nachabé, M. Fagerlund, F. Ståhl, A. Charalampidis, E. Edström, and P. Gerdhem. "Augmented reality navigation with intraoperative 3D imaging vs fluoroscopy-assisted free-hand surgery for spine fixation surgery: a matched-control study comparing accuracy". In: *Scientific Reports* 10.1 (2020), p. 707. ISSN: 2045-2322.

[297] G. Burström, R. Nachabe, O. Persson, E. Edström, and A. Elmi Terander. "Augmented and Virtual Reality Instrument Tracking for Minimally Invasive Spine Surgery: A Feasibility and Accuracy Study". In: *Spine* 44.15 (2019). ISSN: 0362-2436.

[298] M. Richter, T. Mattes, and B. Cakir. "Computer-assisted posterior instrumentation of the cervical and cervico-thoracic spine". In: *European Spine Journal* 13.1 (2004), pp. 50–59. ISSN: 1432-0932.

[299] C. R. Hatt, M. A. Speidel, and A. N. Raval. "Real-time pose estimation of devices from x-ray images: Application to x-ray/echo registration for cardiac interventions". In: *Medical Image Analysis* 34 (2016). Special Issue on the 2015 Conference on Medical Image Computing and Computer Assisted Intervention, pp. 101–108. ISSN: 1361-8415.

[300] Y. Zhu, M. Li, W. Yao, and C. Chen. "A Review of 6D Object Pose Estimation". In: *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. Vol. 10. 2022, pp. 1647–1655.

[301] B. Tekin, S. N. Sinha, and P. Fua. "Real-Time Seamless Single Shot 6D Object Pose Prediction". In: *CVPR*. 2018.

[302] H. Chen, P. Wang, F. Wang, W. Tian, L. Xiong, and H. Li. "Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 2781–2790.

[303] Y. Wu, M. Zand, A. Etemad, and M. Greenspan. "Vote from the center: 6 dof pose estimation in rgb-d images by radial keypoint voting". In: *European Conference on Computer Vision*. Springer. 2022, pp. 335–352.

[304] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun. "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11632–11641.

[305] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen. "Fs6d: Few-shot 6d pose estimation of novel objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 6814–6824.

[306] B. Wen, W. Yang, J. Kautz, and S. Birchfield. "FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects". In: *arXiv preprint arXiv:2312.08344* (2023).

[307] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang. "Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images". In: *European Conference on Computer Vision*. Springer. 2022, pp. 298–315.

[308] J. Lin, Z. Wei, Y. Zhang, and K. Jia. "Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 14001–14011.

[309] R. Wang, X. Wang, T. Li, R. Yang, M. Wan, and W. Liu. "Query6DoF: Learning Sparse Queries as Implicit Shape Prior for Category-Level 6DoF Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 14055–14064.

[310] Y. Bukschat and M. Vetter. *EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach*. 2020. arXiv: `2011.04307 [cs.CV]`.

[311] M. Tan and Q. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6105–6114.

[312] M. Tan, R. Pang, and Q. V. Le. "Efficientdet: Scalable and efficient object detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10781–10790.

[313] Y. Xu, K.-Y. Lin, G. Zhang, X. Wang, and H. Li. "RNNPose: Recurrent 6-DoF Object Pose Refinement with Robust Correspondence Field Estimation and Pose Optimization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[314] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes". In: *arXiv preprint arXiv:1711.00199* (2017).

[315] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao. "Pvnet: Pixel-wise voting network for 6dof pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4561–4570.

[316] J. Redmon and A. Farhadi. "YOLO9000: Better, Faster, Stronger". In: *CVPR*. 2017, pp. 6517–6525.

[317] J. Redmon and A. Farhadi. "YOLOv3: An Incremental Improvement". In: *arXiv* (2018).

[318] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection". In: *arXiv* (2020).

[319] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. "Scaled-YOLOv4: Scaling Cross Stage Partial Network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 13029–13038.

[320] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, K. Michael, J. Fang, imyhxy, Lorna, C. Wong, Z. Yifu, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, P. Skalski, A. Hogan, M. Strobel, M. Jain, L. Mammana, and xylieong. *ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations*. Version v6.2. 2022.

[321] Z. Li, G. Wang, and X. Ji. "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 7678–7687.

[322] A. Presenti, S. Bazrafkan, J. Sijbers, and J. De Beenhouwer. *Deep learning-based 2D-3D sample pose estimation for X-ray 3DCT*. Tech. rep. 2020.

[323] A. Presenti, Z. Liang, L. F. Alves Pereira, J. Sijbers, and J. De Beenhouwer. "CNN-based Pose Estimation of Manufactured Objects During Inline X-ray Inspection". In: *2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI)*. 2021, pp. 388–393.

[324] A. Presenti, Z. Liang, L. F. A. Pereira, J. Sijbers, and J. De Beenhouwer. "Fast and accurate pose estimation of additive manufactured objects from few X-ray projections". In: *Expert Systems with Applications* 213 (2023), p. 118866. ISSN: 0957-4174.

[325] M. Bui, S. Albarqouni, M. Schrapp, N. Navab, and S. Ilic. "X-Ray PoseNet: 6 DoF Pose Estimation for Mobile X-Ray Devices". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, pp. 1036–1044.

[326] L. Kausch, S. Thomas, H. Kunze, T. Norajitra, A. Klein, L. Ayala, J. El Barbari, E. Mandelka, M. Privalov, S. Vetter, A. Mahnken, L. Maier-Hein, and K. Maier-Hein. "C-arm positioning for standard projections during spinal implant placement". In: *Medical Image Analysis* 81 (2022), p. 102557. ISSN: 1361-8415.

[327] A. Pourtaherian, H. J. Scholten, L. Kusters, S. Zinger, N. Mihajlovic, A. F. Kolen, F. Zuo, G. C. Ng, H. H. M. Korsten, and P. H. N. de With. "Medical Instrument Detection in 3-Dimensional Ultrasound Data Volumes". In: *IEEE Transactions on Medical Imaging* 36.8 (2017), pp. 1664–1675.

[328] H. Yang, C. Shan, A. F. Kolen, and P. H. N. de With. "Medical instrument detection in ultrasound: a review". In: *Artificial Intelligence Review* (2022). ISSN: 1573-7462.

[329] D. Kügler, J. Sehring, A. Stefanov, I. Stenin, J. Kristin, T. Klenzner, J. Schipper, and A. Mukhopadhyay. "i3PosNet: Instrument Pose Estimation from X-Ray in temporal bone surgery". In: *International journal of computer assisted radiology and surgery* 15.7 (2020), pp. 1137–1145. ISSN: 1861-6429.

[330] S. Liu and W. Deng. "Very deep convolutional neural network based image classification using small training sample size". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015, pp. 730–734.

[331] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes". In: *2011 International Conference on Computer Vision*. 2011, pp. 858–865.

[332] T. Hodan, P. Haluza, S. Obdrzálek, J. Matas, M. I. A. Lourakis, and X. Zabulis. "T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects". In: *CoRR* abs/1701.05498 (2017). arXiv: 1701.05498.

[333] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and M. Marín-Jiménez. "Automatic generation and detection of highly reliable fiducial markers under occlusion". In: *Pattern Recognition* 47.6 (2014), pp. 2280–2292. ISSN: 0031-3203.

[334] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[335] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. "Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes". In: *Computer Vision – ACCV 2012*. Ed. by K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 548–562. ISBN: 978-3-642-37331-2.

[336] V. Lepetit, F. Moreno-Noguer, and P. Fua. "EP n P: An accurate O (n) solution to the P n P problem". In: *International journal of computer vision* 81 (2009), pp. 155–166.

[337] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh. "CSPNet: A new backbone that can enhance learning capability of cnn". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 390–391.

[338] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision* 88.2 (2010), pp. 303–338.

[339] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe. "Synthetic occlusion augmentation with volumetric heatmaps for the 2018 eccv posetrack challenge on 3d human pose estimation". In: *arXiv preprint arXiv:1809.04987* (2018).

[340] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll'a r, and C. L. Zitnick. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312.

[341] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, et al. "Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3364–3372.

[342] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1521–1529.

[343] Y. Bukschat and M. Vetter. *EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach*. Tech. rep. 2020. arXiv: 2011.04307v2.

[344] F. Manni, A. Elmi-Terander, G. Burström, O. Persson, E. Edström, R. Holthuizen, C. Shan, S. Zinger, F. van der Sommen, and P. H. N. de With. "Towards Optical Imaging for Spine Tracking without Markers in Navigated Spine Surgery". In: *Sensors* 20.13 (2020). ISSN: 1424-8220.

[345] M. Kisantal, S. Sharma, T. H. Park, D. Izzo, M. Märtens, and S. D'Amico. "Satellite pose estimation challenge: Dataset, competition design, and results". In: *IEEE Transactions on Aerospace and Electronic Systems* 56.5 (2020), pp. 4083–4098.

[346] P. Carcagnì, M. Leo, P. Spagnolo, P. L. Mazzeo, and C. Distante. "A lightweight model for satellite pose estimation". In: *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part I*. Springer. 2022, pp. 3–14.

[347] B. van Ginneken, C. M. Schaefer-Prokop, and M. C.-A. D. Prokop. "How to Move from the Laboratory to the Clinic". In: *Radiology* 261 (2011), pp. 719–732.

[348] A. Kohli and S. Jha. "Why CAD Failed in Mammography". In: *J. Am. Coll. Radiol* 15 (2018), pp. 535–537.

[349] E. J. Hwang, S. Park, K.-N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, J. G. Cohen, et al. "Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs". In: *JAMA Netw. open* 2, e191095 (2019).

[350] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk. "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data". In: *JAMA Intern. Med* 178 (2018), pp. 1544–1547.

[351] D. S. Char, N. H. Shah, and D. Magnus. "Implementing Machine Learning in Health Care - Addressing Ethical Challenges". In: *N. Engl. J. Med* 378 (2018), pp. 981–983.

[352] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. *Larson*. Available online. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[353] C. M. Gijsberts, K. A. Groenewegen, I. E. Hoefer, M. J. C. Eijkemans, F. W. Asselbergs, T. J. Anderson, A. R. Britton, J. M. Dekker, G. Engström, G. W. Evans, et al. "Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events". In: *PLoS One* 10, e0132321 (2015).

[354] K. G. M. Moons, R. F. Wolff, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett. "PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration". In: *Ann. Intern. Med* 170 (2019), W1–W33.

[355] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks". In: *Nature* 542 (2017), pp. 115–118.

[356] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, et al. *Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition*. JAMA dermatology, 2019.

[357] T. Marinelli, A. Filippone, F. Tavano, A. Fontana, F. Pellegrini, J. Köninger, G. M. Richter, L. Bonomo, and M. W. d. Büchler. "Sebastiano". In: *P.; et al. A Tumour Score with Multidetector*

Spiral CT for Venous Infiltration in Pancreatic Cancer: Influence on Borderline Resectable. Radiol. Med 119 (2014), pp. 334–342.

[358] M. Klauss, A. Mohr, H. von Tengg-Kobligk, H. Friess, R. Singer, P. Seidensticker, H. U. Kauczor, G. M. Richter, G. W. Kauffmann, and L. A. Grenacher. "New Invasion Score for Determining the Resectability of Pancreatic Carcinomas with Contrast-Enhanced Multidetector Computed Tomography". In: *Pancreatology* 8 (2008), pp. 204–210.

[359] S. A. Ahmed, A. F. Mourad, R. A. Hassan, M. A. E. Ibrahim, A. Soliman, E. Aboeleuon, O. M. A. Elbadee, H. F. Hetta, and M. A. Jabir. "Preoperative CT Staging of Borderline Pancreatic Cancer Patients after Neoadjuvant Treatment: Accuracy in the Prediction of Vascular Invasion and Resectability". In: *Abdom. Radiol. (New York)* 46 (2021), pp. 280–289.

[360] B. R. Kim, J. H. Kim, S. J. Ahn, I. Joo, S.-Y. Choi, S. J. Park, and J. K. C. T. Han. "Prediction of Resectability and Prognosis in Patients with Pancreatic Ductal Adenocarcinoma after Neoadjuvant Treatment Using Image Findings and Texture Analysis". In: *Eur. Radiol* 29 (2019), pp. 362–372.

[361] C. Yip, D. Landau, R. Kozarski, B. Ganeshan, R. Thomas, A. Michaelidou, and V. P. E. C. Goh. "Heterogeneity as Potential Prognostic Biomarker in Patients Treated with Definitive Chemotherapy and Radiation Therapy". In: *Radiology* 270 (2014), pp. 141–148.

[362] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen. *Kvasir-SEG: A Segmented Polyp Dataset*. 2019. arXiv: 1911.07069 [eess.IV].

[363] L. Liu, Y. Ren, Z. Lin, and Z. Zhao. "Pseudo numerical methods for diffusion models on manifolds". In: *arXiv preprint arXiv:2202.09778* (2022).

[364] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[365] P. Gravel, G. Beaudoin, and J. A. De Guise. "A method for modeling noise in medical images". In: *IEEE Transactions on medical imaging* 23.10 (2004), pp. 1221–1232.

[366] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. "Reading digits in natural images with unsupervised feature learning". In: (2011).

Acronyms

AR Augmented Reality

AUROC Area Under Curve of the Receiver Operating Characteristic

CADe Computer-Aided Detection

CADx Computer-Aided Diagnosis

CAD CADx and CADe

CNN Convolutional Neural Network

FCN Fully Convolutional Network

FPR False Positive Rate

FPS Frames Per Second

AI Artificial Intelligence

DL Deep Learning

OOD Out-of-Distribution

CT Computed Tomography

SOTA State-of-the-art

PDAC Pancreatic Ductal Adenocarcinoma

DSC Dice-Sørensen coefficient

GPU Graphics Processing Unit

NIfTI Neuroimaging Informatics Technology Initiative

CZE Catharina Hospital in Eindhoven

MIS Minimally invasive surgeries

ELBO Evidence Lower Bound

OS Overall Survival

NFs Normalizing Flows

MSE Mean-Squared Error

CE Cross-Entropy

Publication List

The following conference and journal papers have been published based on the research presented in this thesis.
* denotes equal contribution.

Journal articles

[J1] M. Ramaekers*, **C. G. A. Viviers***, B. V. Janssen, T. A. Hellström, L. Ewals, K. Van der Wulp, J. Nederend, I. Jacobs, J. R. Pluyter, D. Mavroeidis, F. Van der Sommen, M. G. Besselink, M. D. P. Luyer, et al. "Computer-aided detection for pancreatic cancer diagnosis: radiological challenges and future directions". In: *Journal of Clinical Medicine* 12.13 (2023), p. 4209.

[J2] **C. G. A. Viviers**, L. Filatova, M. Termeer, P. H. N. De With, and F. Van der Sommen. "Advancing 6-DoF Instrument Pose Estimation in Variable X-Ray Imaging Geometries". In: *IEEE Transactions on Image Processing* 33 (2024), pp. 2462–2476.

[J3] A. M. M. Valiuddin, **C. G. A. Viviers,** R. J. G. Van Sloun, P. H. N. De With, and F. Van der Sommen. "Investigating and Improving Latent Density Segmentation Models for Aleatoric Uncertainty Quantification in Medical Imaging". In: *IEEE Transactions on Medical Imaging* (2024), p. 1.

[J4] M. Ramaekers*, **C. G. A. Viviers***, T. A. Hellström, L. J. Ewals, N. Tasios, I. Jacobs, J. Nederend, F. Van derSommen, and M. D. Luyer. "Improved Pancreatic Cancer Detection and Localization on CT scans: A Computer-Aided Detection model utilizing Secondary Features". In: *Cancers* 16.13 (2024), p. 2403.

International conference contributions

[C1] A. M. M. Valiuddin, **C. G. A. Viviers**, R. J. Van Sloun, P. H. N. De With, and F. Van der Sommen. "Improving aleatoric uncertainty quantification in multi-annotated medical image segmentation with normalizing flows". In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer. 2021, pp. 75–88.

[C2] **C. G. A. Viviers**, M. Ramaekers, P. H. N. De With, D. Mavroeidis, J. Nederend, M. Luyer, and F. Van der Sommen. "Improved pancreatic tumor detection by utilizing clinically-relevant secondary features". In: *MICCAI Workshop on Cancer Prevention through Early Detection*. Springer. 2022, pp. 139–148.

[C3] **C. G. A. Viviers**, A. M. M. Valiuddin, P. H. N. De With, and F. Van der Sommen. "Probabilistic 3D segmentation for aleatoric uncertainty quantification in full 3D medical data". In: *Medical Imaging 2023: Computer-Aided Diagnosis*. Vol. 12465. SPIE. 2023, pp. 341–351.

[C4] A. M. M. Valiuddin, **C. G. A. Viviers**, R. J. van Sloun, P. H. N. De With, and F. Van der Sommen. "Efficient Out-of-Distribution Detection of Melanoma with Wavelet-Based Normalizing Flows". In: *MICCAI Workshop on Cancer Prevention through Early Detection*. Springer. 2022, pp. 99–107.

[C5] **C. G. A. Viviers**, J. de Bruijn, L. Filatova, P. H. N. De With, and F. Van der Sommen. "Towards real-time 6D pose estimation of objects in single-view cone-beam x-ray". In: *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 12034. SPIE. 2022, pp. 418–423.

[C6] A. M. M. Valiuddin, **C. G. A. Viviers**, R. van Sloun, P. H. N. De With, and F. Van der Sommen. "Retaining Informative Latent Variables in Probabilistic Segmentation". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 5635–5639.

[C7] F. Mammadli*, T. A. E. Hellström*, **C. G. A. Viviers**, I. Jacobs, L. J. Ewals, N. Tasios, D. Mavroeidis, H. P. Verhees, P. H. N. De With, J. Nederend, and F. Van der Sommen. "Robustness evaluation of CAD systems for lung nodule segmentation using clinically relevant image perturbations". In: *Medical Imaging 2024: Image Processing*. Vol. 12926. SPIE. 2024, pp. 554–562.

[C8] C. H. Claessens, J. Hamm, **C. G. A. Viviers**, J. Nederend, D. Grünhagen, P. J. Tanis, P. H. N. De With, and F. Van der Sommen. "Evaluating task-specific augmentations in self-supervised pre-training for 3D medical image analysis". In: *Medical Imaging 2024: Image Processing*. Vol. 12926. SPIE. 2024, pp. 403–410.

[C9] S. E. Okel*, **C. G. A. Viviers***, M. Ramaekers, T. A. Hellström, N. Tasios, D. Mavroeidis, J. Pluyter, I. Jacobs, M. Luyer, P. H. N. De With, and F. Van der Sommen. "Advancing Abdominal Organ and PDAC Segmentation Accuracy with Task-Specific Interactive Models". In: *MICCAI International Workshop on Applications of Medical AI*. Springer. 2023, pp. 52–61.

[C10] T. A. E. Hellström*, **C. G. A. Viviers***, M. Ramaekers, N. Tasios, J. Nederend, M. D. Luyer, P. H. N. De With, and F. Van der Sommen. "Clinical segmentation for improved pancreatic ductal adenocarcinoma detection and segmentation". In: *Medical Imaging 2023: Computer-Aided Diagnosis*. Vol. 12465. SPIE. 2023, pp. 620–626.

[C11] **C. G. A. Viviers***, M. Ramaekers*, A. M. M. Valiuddin, T. Hellström, N. Tasios, J. van der Ven, I. Jacobs, L. Ewals, J. Nederend, P. H. N. De With, M. Luyer, and F. Van der Sommen. "Segmentation-based Assessment of Tumor-Vessel Involvement for Surgical Resectability Prediction of Pancreatic Ductal Adenocarcinoma". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2421–2431.

[C12] **C. G. A. Viviers**, A. M. M. Valiuddin*, F. Caetano*, L. Abdi, L. Filatova, P. H. N. De With, and F. V. der Sommen. *Can Your Generative Model Detect Out-of-Distribution Covariate Shift?* In: European Conference on Computer Vision (ECCV) 2024 Uncertainty Quantification for Computer Vision (UNCV). 2024. arXiv: 2409.03043 [cs.CV].

[C13] L. Abdi, A. M. M. Valiuddin*, **C. G.A. Viviers***, P. H. De With, and F. Van der Sommen. "Typicality Excels Likelihood for Unsupervised Out-of-Distribution Detection in Medical Imaging". In: *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. Springer. 2024, pp. 149–159.

National conference contributions

[N1] A. M. M. Valiuddin, **C. G. A. Viviers**, R. J. van Sloun, P. H. N. De With, and F. Van der Sommen. "Improving aleatoric uncertainty quantification in multi-annotated medical image segmentation with normalizing flows". In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer. 2021, pp. 75–88.

[N2] **C. G. A. Viviers**, L. Filatova, M. Termeer, P. H. N. De With, and F. Van der Sommen. "Advancing 6-DoF Instrument Pose Estimation in Variable X-Ray Imaging Geometries". In: *IEEE Transactions on Image Processing* 33 (2024), pp. 2462–2476.

Acknowledgements

If a PhD dissertation was a novel, it wouldn't be a page-turning thriller—it would be an epic saga, full of plot twists, unexpected dead ends, and the occasional moment of triumph. This thesis is the final chapter of my PhD story, but it was never a solo adventure. To my mentors, collaborators, friends, and family: thank you for being my co-authors, editors, and proofreaders, in both research and life.

They don't tell you this when you start a PhD, but instead of tying up all the loose ends in your field, you're actually creating more of them. Every experiment, every paper, every "Aha!" moment just adds fuel to the fire. Ada Lovelace summed it up perfectly: "The more I study, the more insatiable do I feel my genius for it to be." And honestly, that's how it's felt—the more I learn, the more I want to learn. But this cycle of learning is far from easy—it's incredibly humbling. With every new piece of knowledge comes the realization of how much more there is to understand, and with that comes plenty of mistakes and moments of doubt. As much as curiosity has driven this journey, the process of learning and improving has often been tough. Jake the Dog from Adventure Time said it right: "Dude, sucking at something is the first step towards being sorta good at something." This sentiment perfectly captures the humbling nature of my PhD journey. A PhD journey has its fair share of challenges, but with the help of many generous and patient people, I was able to turn those initial struggles into progress. I owe a huge debt of gratitude to everyone who supported me along the way and helped make this achievement possible.

First of all, I would like to thank my promoter, prof.dr.ir. Peter de With, co-promoter dr.ir. Fons van der Sommen and supervisor dr. Lena Filatova. Thank you for giving me the opportunity to pursue this doctoral project and for your guidance and support along this journey.

Peter, I am grateful for your diligence and patience in reading and providing feedback on my work. Through your meticulous approach, you've shown me the importance of expressing thoughts clearly and coherently. More than that, you taught me that scientific writing is not just a skill, but an art form—one that requires precision, creativity, and careful craftsmanship. Your guidance has been invaluable, and I would like to thank you for shaping both my writing and my thinking throughout this journey.

Fons, thank you for being such a great source of support and inspiration throughout this PhD journey. Our brainstorming sessions were always filled with energy and fresh ideas, and your ambition constantly pushed me to aim higher. I

also appreciate the life advice that, when progress was slow, there's nothing like a Hazy IPA to put everything in perspective. Your encouragement and drive have had a lasting impact on both my work and my mindset.

Lena, I am incredibly grateful for your unwavering and reliable support throughout this journey. You've helped me set good priorities, ensuring that I stayed focused on what truly mattered. Your trust in the work I was doing gave me the confidence to push forward, and your guidance and insight in navigating the complexities of research in industry was invaluable. Thank you!

Returning from an industry position to academia is no easy feat. For that, I would like to thank Stefan Schalk and Thomas de Laet for initiating the PhD research project and for trusting in me as a strong candidate for this opportunity. Your initiative set the foundation for this work, and your belief in my abilities allowed me to pursue this research with confidence.

I would like to extend a special thanks to all my colleagues and collaborators at Philips: Johan, Roger, Bahadir, Yeshi, Nils, Jonathan, Liam, Stefan, Mark, Klaus Jeurgen and Nico. Your expertise and support were crucial in navigating the complexities of the Philips IGT system, helping me to understand its intricacies and ensuring the progression of this research through data collection and gathering the necessary information efficiently and effectively. A big thank you to Maurice Termeer for the collaboration and insights, which led to a complete chapter in this thesis. Your knowledge has been very valuable.

I am immensely grateful to my colleagues at Catharina Ziekenhuis for their exceptional support, collaboration and for what we achieved in the research together. A special mention to prof. dr. Misha Luyer and dr. Joost Nederend, whose drive and clear research goals were pivotal in the development of the work, ensuring we investigate solutions with potential real impact. Lotte, Jon, and Igor, thank you for your hard work and camaraderie, which made the journey both productive and enjoyable. An especially heartfelt thanks to my co-author Mark Ramaekers, whose dedication and input were instrumental in bringing our work to fruition. I am truly grateful for all of your contributions.

A big thank you to all my VCA colleagues and co-authors at the university for your amazing support. The current group is extremely ambitious and talented, leading to many great and insightful conversations (whether in the lab or at Walhalla with some chupitos). Your feedback and collaboration have been a huge part of my growth and the progress of this work. Shoutout to Amaan, Terese, and Francisco—your expertise and dedication made such a difference to our projects. It's been an absolute privilege to work with such an inspiring team!

Big thanks to all my friends for making sure I didn't completely disappear into the PhD abyss! Your timely distractions and support kept me grounded (and sane) throughout this journey. Being in a new country wasn't always easy, but

these friends became my community and the people who made it all feel possible.

My opregte dank aan my familie dat hulle van kleins af 'n sin van nuuskierigheid en deursettingsvermoë in my gekweek het. Julle het my altyd aangemoedig om te verken, deur te druk, en uitdagings met moed aan te pak. Julle ondersteuning en liefde was die bestendige kragte agter hierdie prestasie. Aan my pa, baie dankie dat jy my die liefde vir wetenskap geleer het en vir die uitstekende pa wat jy in elke opsig vir my is.

Dear Katty, I owe you a special thanks for your support and encouragement, even when my PhD seemed to be my full-time hobby. Your patience, especially when I said I'd only be working 'a little longer,' has been nothing short of heroic. It's no exaggeration to say that this PhD owes much of its success to you. I am extremely grateful for you and who you are to me.

Ek dank God vir krag en wysheid gedurende hierdie PhD.

Curriculum vitae

Christiaan Viviers was born in South Africa, Pretoria on 10 January 1993. He is passionate about computer vision and artificial intelligence, especially in application areas with potential high impact, such as in the medical domain. He holds both a Bachelor's and a Master's degree in Electrical Engineering from Stellenbosch University in South Africa. His Master's thesis focused on developing novel sensing technologies for the detection of antigens, laying the groundwork for his future pursuits in technology-driven healthcare solutions.

Towards the end of his Master's program, Christiaan developed a keen interest in machine learning, which was further nurtured through his role at Philips Image Guided Therapy (IGT) systems in the Netherlands. Starting at Philips Healthcare, Best, Netherlands in 2018, he played a key role in developing various image analysis algorithms, ultimately guiding him towards exploring deep learning and modern computer vision techniques. This passion prompted him to pursue a PhD in Computer Vision and AI in Medical Imaging at Eindhoven University of Technology, Electrical Engineering in the Video Coding Architectures (VCA) research group, while continuing to assist as a software engineer at Philips Healthcare.

The research during his PhD is primarily focused on developing and advancing vision-based assistive technologies that enhance the detection and treatment of cancer, and advancing the capabilities of generative AI with applications in the medical field for improved uncertainty quantification and anomaly detection. He also has a deep interest in 3D vision, as evident from his research in surgical guidance and pose estimation.

Outside of his research, Christiaan enjoys building new and useful things, staying active through exercise, and traveling to explore new cultures and perspectives. He also enjoys the occasional AAA game, blending relaxation with a competitive edge.

Looking ahead, he will pursue a Postdoctoral position at TU/e in the European TASTI project (EUREKA) to continue leveraging generative AI and synthetic image generation in various application domains, and especially addressing healthcare challenges.