

Semantic Segmentation of Ambiguous Images



Simon Andreas Alexander Kohl

Supervisors: Prof. Rainer Stiefelhagen
Dr. Klaus H. Maier-Hein

Department of Computer Science
Karlsruhe Institute of Technology

This dissertation is submitted for the degree of
Doctor of Engineering

Semantic Segmentation of Ambiguous Images

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Simon Andreas Alexander Kohl

Tag der mündlichen Prüfung: 29.01.2020

Erster Referent: Prof. Dr.-Ing. Rainer Stiefelhagen

Zweiter Referent: PD Dr. Klaus Maier-Hein

Acknowledgements

First and foremost I would like to thank Klaus Maier-Hein for his steady guidance, the encouragement and the experience that he shared in the research and the collaborations that formed this dissertation. I am also very grateful for the flexibilities that he has provided me. This is true with respect to the direction that this work took and with respect to the significant amounts of time that he enabled me to carry out research outside of my position at the German Cancer Research Center (DKFZ) in Heidelberg. Klaus and his wife Lena Maier-Hein have created a wonderful and collaborative environment to work in, for which I cannot thank them enough.

I also owe a large debt of gratitude to Rainer Stiefelhagen, who generously took me on as an external PhD student at KIT and gave invaluable feedback in the process leading up to this thesis.

I further want to thank our radiological collaborators David Bonekamp and Patrick Schelb whose expertise and insights were indispensable and helped ground this work in clinical desiderata.

A very warm thank you goes out to Olaf Ronneberger and the (then) DeepMind Health team who I was lucky enough to intern and work with. Olaf has had a lasting impact through the mentorship and example he provided me with, for which I am truly grateful. Moreover, I am deeply thankful to DeepMind for allowing me to set aside time to finish up this dissertation after joining the team in a full-time position.

A huge thanks is due to the incredible colleagues and fellow PhD students at the DKFZ who are too numerous to list individually but made these three years thoroughly enjoyable. I do however feel obliged to shine a spotlight on the ‘deep fridge’ crew around Paul Jäger, Peter Full, Fabian Isensee, Jens Petersen, Sebastian Wirkert and Gregor Köhler. Together we not only regularly brought our Titan Xs to a boil but also enjoyed a good measure of post-lunch drone racing, Friday Mezcal tastings, alpha investments, GNTM challenge strategizing and Buffet & Bistro frenzies.

I would further like to mention the weekly paper discussion held by the Machine Learning Student Society at KIT (ML-KA), organized by Fabian Both, Jörg Franke and others, from which I profited immensely.

A resounding and heartfelt thank you shall go to my parents Andreas and Claudia and my siblings Janis, Anna and Maria who provided me with grounding, stability and joyful weekends back home whenever I visited.

Lastly, I am deeply indebted to my girlfriend Léa, who was always supportive, never (well, hardly ever) complained about long lab hours, or worse, months spent doing research abroad, and whose open ears and heart were key in getting me this far. Thank you so much for being there!

Simon Kohl
London, December 2019

Abstract (English)

Medical images can be difficult to interpret. Not only because spotting structures and potential changes therein requires experience and years of training, but also because the presented measurements are often ambiguous at heart. This is fundamentally a consequence of the fact that medical image modalities such as MRI or CT only provide indirect measurements of the underlying molecular identities. The semantics of an image therefore generally have to be inferred from the provided larger context, which is often insufficient to pin down the interpretation to a singular, unique hypothesis.

Similar scenarios exist in natural images, where the contextual information required to resolve potential ambiguities can be limited, for example due to occlusions or measurement noise. Additionally, overlapping and vague class definitions may contribute to an ill-defined or diverse solution space. The presence of such image ambiguity can hamper the training and the performance of machine learning models. Moreover, current models are mostly unable to capture complex-structured diverse outputs and instead are forced to retreat to sub-optimal singular solutions or indiscernible mixtures.

This can be particularly problematic when scaling classifiers to dense prediction tasks such as semantic segmentation. Semantic segmentation is concerned with predicting a class label for every pixel in an image. This type of detailed image interpretation also plays an important role in diagnosing and treating diseases such as cancer: Tumors are often detected from MRI or CT scans and their precise location and delineation are crucial steps in grading them, preparing potential biopsies or planning focal therapy. This clinical interpretation of images, but also the perception of our surroundings in everyday tasks such as driving, are currently performed by humans. As we move towards incorporating learning based systems in our decision making processes, it is of course vital to adequately model the tasks at hand. This involves capturing the uncertainties that exist in the models' predictions, including such that can be attributed to image ambiguities.

This thesis proposes various ways in which to deal with ambiguous image evidence. First we examine the current clinical standard of subjectively grading prostate lesions

visible on MRI, that is associated with high inter-rater variability due to the ambivalent MRI appearance of lesions. We find simple machine learning models and even just quantifying certain MRI parameters to perform better than an individual subjective expert, suggesting a promising potential to improve grading by quantifying the process.

Second, we probe the currently most successful segmentation architecture on a strongly ambiguous dataset that was collected and annotated during clinical routine. Our experiments show that the standard segmentation loss function may be sub-optimal when applied in scenarios with heavy label noise. As an alternative we learn a model of the loss with the aim of allowing for the co-existence of plausible segmentation predictions during training. We observe performance improvements when employing this training scheme to the otherwise identical deep neural network and find even more pronounced relative gains in the small data limit. Lack of data and labels, high levels of imaging and label noise and ambiguous image evidence are particularly common on medical image datasets. This part of the thesis thus exposes some of the vulnerabilities that standard machine learning techniques may face in the light of these particularities.

Current segmentation models such as the ones considered above are constrained to produce a singular prediction. This contrasts the observation that a group of graders typically produces a set of diverse but plausible annotations when given ambiguous image data. In order to lift this model constraint and allow for the appropriate probabilistic treatment of the task, we go on to develop two models that predict a distribution over plausible annotations rather than predicting just a singular deterministic one. The first of the two models combines an encoder-decoder model with the framework of variational inference and employs a global latent vector that encodes the space of possible annotations for a given image. We show that this model improves upon the performance of the considered baselines and yields well calibrated uncertainties. The second model refines the formulation of the first in that it introduces a more flexible and hierarchical latent space decomposition that allows to capture segmentation variability at different image scales. This increases the granularity of segmentation detail that the model can produce and allows to model independently varying locations and scales, which we show on the task of segmenting individual object instances. Both of these novel generative segmentation models allow to sample diverse and coherent image segmentations if admissible, which contrasts with prior work that is either deterministic, models uncertainty at the pixel level or suffers from an under-complex modelling of the appropriate diversity.

In conclusion, this thesis is concerned with machine learning applications for the interpretation of medical images: We expose the possibility to increase the standard of care in clinical practice by a quantitative use of image markers which are currently

subjectively factored into diagnoses, we show the potential utility of a new training scheme to remedy the apparent susceptibility of the standard segmentation loss formulation to strong label noise and we propose two novel probabilistic segmentation models that are able to accurately capture the distribution over admissible labels given an image. These contributions can be seen as steps towards a more quantitative, principled and uncertainty-aware analysis of medical images -an important quest as learning based systems will find increasing integration into clinical workflows.

Abstract (German)

Medizinische Bilder können schwer zu interpretieren sein. Nicht nur weil das Erkennen von Strukturen und möglichen Veränderungen Erfahrung und jahrelanges Training bedarf, sondern auch weil die dargestellten Messungen oft im Kern mehrdeutig sind. Fundamental ist dies eine Konsequenz dessen, dass medizinische Bild-Modalitäten, wie beispielsweise MRT oder CT, nur indirekte Messungen der zu Grunde liegenden molekularen Identitäten bereithalten. Die semantische Bedeutung eines Bildes kann deshalb im Allgemeinen nur gegeben einem größeren Bild-Kontext erfasst werden, welcher es oft allerdings nur unzureichend erlaubt eine eindeutige Interpretation in Form einer einzelnen Hypothese vorzunehmen.

Ähnliche Szenarien existieren in natürlichen Bildern, in welchen die Kontextinformation, die es braucht um Mehrdeutigkeiten aufzulösen, limitiert sein kann, beispielsweise aufgrund von Verdeckungen oder Rauschen in der Aufnahme. Zusätzlich können überlappende oder vage Klassen-Definitionen zu schlecht gestellten oder diversen Lösungsräumen führen. Die Präsenz solcher Mehrdeutigkeiten kann auch das Training und die Leistung von maschinellen Lernverfahren beeinträchtigen. Darüber hinaus sind aktuelle Modelle ueberwiegend unfähig komplex strukturierte und diverse Vorhersagen bereitzustellen und stattdessen dazu gezwungen sich auf sub-optimale, einzelne Lösungen oder ununterscheidbare Mixturen zu beschränken.

Dies kann besonders problematisch sein wenn Klassifikationsverfahren zu pixel-weisen Vorhersagen wie in der semantischen Segmentierung skaliert werden. Die semantische Segmentierung befasst sich damit jedem Pixel in einem Bild eine Klassen-Kategorie zuzuweisen. Diese Art des detaillierten Bild-Verständnisses spielt auch eine wichtige Rolle in der Diagnose und der Behandlung von Krankheiten wie Krebs: Tumore werden häufig in MRT oder CT Bildern entdeckt und deren präzise Lokalisierung und Segmentierung ist von grosser Bedeutung in deren Bewertung, der Vorbereitung möglicher Biopsien oder der Planung von Fokal-Therapien. Diese klinischen Bildverarbeitungen, aber auch die optische Wahrnehmung unserer Umgebung im Rahmen von täglichen Aufgaben wie dem Autofahren, werden momentan von Menschen durchgeführt. Als Teil des zunehmenden

Einbindens von maschinellen Lernverfahren in unsere Entscheidungsfindungsprozesse, ist es wichtig diese Aufgaben adequat zu modellieren. Dies schliesst Unsicherheitsabschätzungen der Modellvorhersagen mit ein, mitunter solche Unsicherheiten die den Bild-Mehrdeutigkeiten zugeschrieben werden können.

Die vorliegende Thesis schlägt mehrere Art und Weisen vor mit denen mit einer mehrdeutigen Bild-Evidenz umgegangen werden kann. Zunächst untersuchen wir den momentanen klinischen Standard der im Falle von Prostata Läsionen darin besteht, die MRT-sichtbaren Läsionen subjektiv auf ihre Aggressivität hin zu bewerten, was mit einer hohen Variabilität zwischen Bewertern einhergeht. Unseren Studien zufolge können bereits einfache machinelle Lernverfahren und sogar simple quantitative MRT-basierte Parameter besser abschneiden als ein individueller, subjektiver Experte, was ein vielversprechendes Potential der Quantifizierung des Prozesses nahelegt.

Desweiteren stellen wir die derzeit erfolgreichste Segmentierungsarchitektur auf einem stark mehrdeutigen Datensatz zur Probe der während klinischer Routine erhoben und annotiert wurde. Unsere Experimente zeigen, dass die standard Segmentierungsverlustfunktion in Szenarien mit starkem Annotationsrauschen sub-optimal sein kann. Als eine Alternative erproben wir die Möglichkeit ein Modell der Verlustfunktion zu lernen mit dem Ziel die Koexistenz von plausiblen Lösungen während des Trainings zuzulassen. Wir beobachten gesteigerte Performanz unter Verwendung dieser Trainingsmethode für ansonsten unveränderte neuronale Netzarchitekturen und finden weiter gesteigerte relative Verbesserungen im Limit weniger Daten. Mangel an Daten und Annotationen, hohe Maße an Bild- und Annotationsrauschen sowie mehrdeutige Bild-Evidenz finden sich besonders häufig in Datensätzen medizinischer Bilder wieder. Dieser Teil der Thesis exponiert daher einige der Schwächen die standard Techniken des maschinellen Lernens im Lichte dieser Besonderheiten aufweisen können.

Derzeitige Segmentierungsmodelle, wie die zuvor Herangezogenen, sind dahingehend eingeschränkt, dass sie nur eine einzige Vorhersage abgeben können. Dies kontrastiert die Beobachtung dass eine Gruppe von Annotierern, gegeben mehrdeutiger Bilddaten, typischer Weise eine Menge an diverser aber plausibler Annotationen produziert. Um die vorgenannte Modell-Einschränkung zu beheben und die angemessen probabilistische Behandlung der Aufgabe zu ermöglichen, entwickeln wir zwei Modelle, die eine Verteilung über plausible Annotationen vorhersagen statt nur einer einzigen, deterministischen Annotation. Das erste der beiden Modelle kombiniert ein ‘encoder-decoder’ Modell mit dem Verfahren der ‘variational inference’ und verwendet einen globalen ‘latent vector’, der den Raum der möglichen Annotationen für ein gegebenes Bild kodiert. Wir zeigen, dass dieses Modell deutlich besser als die Referenzmethoden abschneidet und gut kalibrierte

Unsicherheiten aufweist. Das zweite Modell verbessert diesen Ansatz indem es eine flexiblere und hierarchische Formulierung verwendet, die es erlaubt die Variabilität der Segmentierungen auf verschiedenen Skalen zu erfassen. Dies erhöht die Granularität der Segmentierungsdetails die das Modell produzieren kann und erlaubt es unabhängig variierende Bildregionen und Skalen zu modellieren. Beide dieser neuartigen generativen Segmentierungs-Modelle ermöglichen es, falls angebracht, diverse und kohärente Bild Segmentierungen zu erstellen, was im Kontrast zu früheren Arbeiten steht, welche entweder deterministisch sind, die Modellunsicherheiten auf der Pixelebene modellieren oder darunter leiden eine unangemessen geringe Diversität abzubilden.

Im Ergebnis befasst sich die vorliegende Thesis mit der Anwendung von maschinellem Lernen für die Interpretation medizinischer Bilder: Wir zeigen die Möglichkeit auf den klinischen Standard mit Hilfe einer quantitativen Verwendung von Bildparametern, die momentan nur subjektiv in Diagnosen einfließen, zu verbessern, wir zeigen den möglichen Nutzen eines neuen Trainingsverfahrens um die scheinbare Verletzlichkeit der standard Segmentierungsverlustfunktion gegenüber starkem Annotationsrauschen abzumildern und wir schlagen zwei neue probabilistische Segmentierungsmodelle vor, die die Verteilung über angemessene Annotationen akkurat erlernen können. Diese Beiträge können als Schritte hin zu einer quantitativeren, verstärkt Prinzipien-gestützten und unsicherheitsbewussten Analyse von medizinischen Bildern gesehen werden -ein wichtiges Ziel mit Blick auf die fortschreitende Integration von lernbasierten Systemen in klinischen Arbeitsabläufen.

Table of contents

1	Introduction	1
2	Medical Imaging Techniques	7
2.1	Magnetic Resonance Imaging	7
2.2	Computed Tomography Scanning	11
2.3	Histopathologic Light Microscopy	13
3	Medical Image Analysis: The Diagnosis of Prostate Cancer	15
3.1	Current Clinical Diagnosis Steps	16
3.2	Malignancy Quantification & Challenges	20
3.3	Ambiguity and Inter-rater Variability	25
3.4	Discussion	27
4	Medical Image Analysis: Algorithmic State-of-the-Art	29
4.1	From Image Classification to Semantic Segmentation	30
4.2	Predictions under Uncertainty and Noise	37
4.3	Generative Models for Images	41
5	Finding Discriminative MRI Features	45
5.1	Problem Statement	46
5.2	Prostate MRI Dataset	48
5.3	Radiomics Pipeline	49
5.4	Evaluation and Results	53
5.4.1	Lesion-based Analysis	54
5.4.2	Patient-based Analysis	56
5.4.3	Zone-based Analysis	57
5.4.4	Feature Importances	59
5.5	Discussion	60

6	Mitigating Label Noise through Adversarial Training	63
6.1	How Ambiguity interferes with CE-based Training	64
6.2	Learning to reparameterize the Loss Function	65
6.3	Dataset Details	68
6.4	Network Architecture and Training Procedure	69
6.5	Results	70
6.5.1	Improving Segmentation Performance	70
6.5.2	Increasing Robustness on Fewer Training Samples	72
6.6	Discussion	72
7	Learning Image-Global Distributions over Segmentations	75
7.1	Segmenting Ambiguous Images	76
7.2	Related Work & Baselines	77
7.3	Network Architecture and Training Procedure	79
7.3.1	Sampling	79
7.3.2	Training	80
7.4	Performance Measures and Baseline Methods	81
7.4.1	Performance Measures	81
7.4.2	Baseline Methods	81
7.5	Quantitative Results	83
7.5.1	Lung Abnormalities Segmentation	83
7.5.2	Stochastic Cityscapes Street Scene Segmentation	85
7.6	Qualitative Results	86
7.6.1	Lung Abnormalities Segmentation	86
7.6.2	Street Scene Segmentation	87
7.7	Additional Analyses	87
7.7.1	Calibration Analysis.	87
7.7.2	Ablation Analysis	90
7.7.3	Predicting Ambiguity	92
7.8	Discussion	92
8	Learning Multi-Scale Distributions over Segmentations	97
8.1	The Need for a More Flexible Model	98
8.2	Network Architecture and Learning Objective	100
8.2.1	Sampling	100
8.2.2	Training	102
8.2.3	Architecture and Training in more Detail	104

8.3	Dataset Details	105
8.4	Performance Measures	106
8.4.1	Distribution Agreement	106
8.4.2	Reconstruction Fidelity	108
8.4.3	Instance Segmentation	109
8.5	Results	109
8.5.1	LIDC: Segmentation of Ambiguous Lung Scans	110
8.5.2	SNEMI3D: Generative Instance Segmentation of Neurites	112
8.5.3	Extrapolation Task on SNEMI3D	113
8.5.4	Cityscapes Cars: Generative Instance Segmentation of Cars	113
8.5.5	Ablation Study	115
8.6	Discussion	117
9	Discussion	121
9.1	Outlook	124
9.2	A Golden Future?	127
	List of Own Publications	131
	Appendix A Finding Discriminative MRI Features	133
A.1	Systematic and Targeted MR Imaging/TRUS-Fusion Biopsies	133
A.2	Cohort Inclusion Criteria and Demographics	134
A.3	Detailed Results	136
	Appendix B The Probabilistic U-Net	139
B.1	Metrics	139
B.2	How models fit the ground truth distribution	139
B.3	Sampling LIDC masks using different models	140
B.4	Sampling Cityscapes segmentations using our model	140
B.5	Training Details	150
B.5.1	Lung Abnormalities Segmentation	150
B.5.2	Stochastic Cityscapes Street Scene Segmentation	150
	Appendix C The Hierarchical Probabilistic U-Net	153
C.1	Instance Segmentation Post-Processing	153
C.2	Training Details by Dataset	154
C.2.1	LIDC-IDRI Lung CT scans	154
C.2.2	SNEMI3D neocortex EM slices	154

C.2.3 Cityscapes Car Instances	155
C.3 GED ² on LIDC subset B	156
C.4 GED ² and Hungarian-matched IoU for Baselines on LIDC	157
C.5 LIDC, SNEMI3D and Cityscapes: Extra Examples	157
References	165
List of figures	187
List of tables	189
Acronyms	191

Chapter 1

Introduction

When looking to examine the inner state of a body part, a surgical intervention would almost always allow the best look at the state of affairs. Performing invasive procedures on the target region or taking biopsies from it is however often extremely unattractive due to associated health risks and patient discomforts. Instead, non-invasive diagnostic approaches that leave a patient's physique intact and reduce potential long term impacts, are highly sought after [Bickelhaupt et al., 2018].

A powerful toolkit in this respect is of course offered by the discipline of medical imaging. Using MRI or CT scanning technology allows a glimpse of the spatial composition of living tissue with only little risk for a patients health if any at all [Hill et al., 2016]. These modalities offer 3D images by indirectly measuring certain tissue properties in every voxel, e.g. measuring the absorption of X-rays allows to infer the tissue density [Schlegel et al., 2018]. MRI and CT scanning find wide-spread clinical application, for example in cancer diagnosis [Weinreb et al., 2016], i.e. the detection and grading of lesions, and also during cancer therapy, where a pixel-precise localization of tissue boundaries can be critical for radiation therapy [Nikolov et al., 2018] or tumor growth monitoring [Kickingreder et al., 2019].

From a naive point of view, the interpretation of medical images may appear straight forward, often akin to something like a tissue density map [Schlegel et al., 2018]. Squeezing clinically actionable information out of them however, is hard - a task that for example radiologists are only entrusted with after many years of training. Among the challenges are the often only very subtle changes in appearance that abnormal tissue displays and the fact that medical images really only indirectly measure the underlying molecular identities [Borofsky et al., 2017]. For this reason the images alone are often ambiguous and their interpretation may only be narrowed down by further evidence such as the one provided by blood tests or -if need be- biopsies [Prostate Cancer UK, 2018].

The current clinical standard relies on radiologists' giving a qualitative assessment of medical images and if required producing hand-drawn outlines of the structures of interest [Weinreb et al., 2016]. As a consequence of the subtleties and ambiguities in the images, even a set of very experienced clinicians and radiologists often shows large variability in their diagnoses [Muller et al., 2015, Pierre et al., 2018, Rosenkrantz et al., 2016b]. To give a real example, two CT scans of potential lung lesions are shown in Fig. 1.1: The 4 clinical experts asked to assess them not only produce strongly varying segmentations of the lesions but also disagree more fundamentally on whether or not the scan shows a veritable lesion in the first place [Armato et al., 2011].

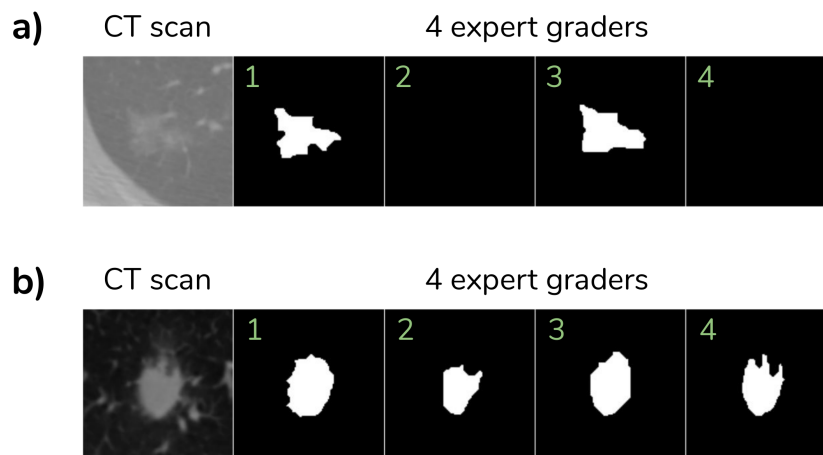


Figure 1.1 | Ambiguity in Lung CT Scans. Subfigures a) and b) show two different cases of potential lung abnormalities alongside 4 expert segmentations (enumerated in green). Two of the experts in a) disagree on whether the scan even shows an abnormality. The provided segmentations in both a) and b) vary strongly. The examples are part of the LIDC-IDRI dataset [Armato et al., 2011].

In clinical practice, grading and segmenting is typically performed by a singular reader, which is why the range of plausible interpretations is not usually known and this uncertainty is hence not taken into account in a quantitative way. This is in part due to the fact that annotating a scan by even a single reader, much less by several, may be an almost prohibitively laborious process, as manual screening and segmentation can take up to several hours per scan, which is e.g. the case when segmenting head CT scans for therapy planning [Nikolov et al., 2018].

The perhaps obvious choice is aiming to support or automate the process, e.g. by using machine learning algorithms to grade and segment the images, and thus ideally quantify the assessments and their associated uncertainties. The recent advances in computer vision algorithms such as deep convolutional neural networks (CNNs) certainly appear

ripe for applications in this domain [De Fauw et al., 2018, Isensee et al., 2019, Nikolov et al., 2018]. At a closer look however, they do not allow for a direct drop-in solution in the high-stake decision making process found in medical image analyses: CNNs excel, when trained on large-scale image data with high-quality annotations [Hénaff et al., 2019] and minimal image ambiguity [Rupprecht et al., 2017], e.g. little class confusions and visual occlusions etc. Real world data in contrast, can however suffer from measurement noise, exhibit domain shifts [De Fauw et al., 2018], come with noisy annotations [Borofsky et al., 2017, Bratan et al., 2014], be very scarce or scarcely labelled [Ross et al., 2018], and, as is often the case with medical images, may present ambiguous image evidence only [Armato et al., 2011, Kitzing et al., 2015].

Contributions All of these problems are important domains to study in order to increase robustness and reliability of deep neural networks in real world applications. In this thesis we investigate how diagnoses and in particular the segmentation of tissues or objects, can be algorithmically handled when the image evidence is (partially) inconclusive.

As a first step towards an improved quantification of the grading of lesions, we investigate which image features derived from different MRI sequences may hold discriminative power in the grading of potential tumors. The idea is, that already bare image-derived features or comparatively simple machine learning algorithms may guide clinical decisions that are currently mostly based on qualitative assessments. Such approaches could thus reduce some of the variance and lead to improved decisions. This could have direct clinical impact, in that difficult but aggressive cases could be spotted more reliably and benign cases could be spared invasive biopsies. In **Finding Discriminative MRI Features** (Chap. 5), we make the following contributions:

- We develop a simple machine learning based approach to classify the clinical significance of prostate lesions based on MRI-derived features using a dataset and biopsy reference standard that was collected in clinical practice and thus reflects realistic conditions.
- We assess the method on a held-out dataset having fixed its working point so as to reflect the radiologist’s sensitivity on the training set and observe an increased performance compared to the radiologist.
- We rank the importance of the employed features and substantiate the discriminative power of a specific MRI-derived feature largely refuting the utility of additional modalities and features as found in the literature.

- Finally, we analyze the utility of distinguishing between anatomical zones of the prostate, which are handled differently in clinical guidelines. To this end we assess the performance of separately trained models and find a combined model to perform superior.

This work relied on the availability of expert-provided lesion segmentations, which are laborious to produce manually and suffer from high inter-rater variability. Because segmentations of anatomical structures of the prostate and even more importantly the segmentation of lesions have numerous clinical applications, such as the guidance of targeted biopsies and treatments such as focal therapy [Borofsky et al., 2017], reliable machine learning models could have a large impact in supporting clinical decisions. However, as is yet to argue, MRI images tend to be particularly ambiguous in the case of the prostate [Hameed and Humphrey, 2010, Kitzing et al., 2015, Nagel et al., 2013, Sakala et al., 2017]. As a consequence, the training labels provided by radiologists can exhibit significant levels of seeming inconsistencies, acting as noisy training supervision. This can have negative bearings on the training process of deep networks, as even plausible network produced segmentations that deviate from the noisy ground truth get ‘punished’ in the training. We hypothesize that an otherwise identical segmentation network should perform better when trained with a procedure that allows for the co-existence of multiple segmentation modes. To this end we investigate the utility of a learned model of the loss in form of a separate, adversarial network. In **Mitigating Label Noise through Adversarial Training** (Chap. 6), we make the following contributions:

- We compare training of a state-of-the-art segmentation model with the standard cross-entropy loss against training it in a mini-max game against an adversarial discriminator.
- We observe increased performance when training adversarially, which we hypothesize could be attributed to reduced gradient conflicts in the noisy label setting.
- Lastly, we find further increases in relative performance when reducing the number of training examples, which seems in line with the hypothesis, that adversarial training might mitigate label noise and suggests particular utility in the small dataset regime.

This approach aims at mitigating the negative effects that a diverse ground truth can have on the training of discriminative deep models. In order to capture the uncertainty over segmentations, the models however need to learn the distribution over plausible segmentations that a given image admits. For this purpose we developed a model

that combines a state-of-the-art segmentation model with variational inference, allowing to produce image-conditional distributions over segmentations. We refer to the model presented in **Learning Image-Global Distributions over Segmentations** (Chap. 7), as the *Probabilistic U-Net* and make the following contributions:

- We propose a model that can induce complex distributions over segmentations including the occurrence of very rare modes, and is able to learn calibrated probabilities of segmentation modes.
- Our framework provides consistent segmentation maps instead of pixel-wise probabilities and can therefore give a joint likelihood of modes.
- Sampling from our model is computationally cheap.

The Probabilistic U-Net is a global latent variable model, which works well for images with singular objects and segmentation variations that are mostly global in nature. When several objects or lesions are depicted and ambiguities are present on different scales and scopes a global model can be too constraint. For these reasons we introduce a hierarchical version that is able to model complex output interdependencies on various image scales. In **Learning Multi-Scale Distributions over Segmentations** (Chap. 8), we propose the *Hierarchical Probabilistic U-Net* and make the following contributions:

- We propose a generative model for semantic segmentation able to learn complex-structured conditional distributions equipped with a latent space that scales with image size.
- Compared to prior art, strongly improved fidelity to fine structure in the models' samples and reconstructions.
- Improved modelling of distributions over segmentations including independently varying scales and locations, as demonstrated in its ability to generate instance segmentations.
- Automatic learning of factors of variations across space and scale.

Both probabilistic models allow to produce coherent samples from the distributions that they parameterize. This could be useful in various ways: A clinician could pick the most appropriate segmentation from a provided set or interact with the model's latent space to quickly produce the desired results. The samples could further be used in down-stream clinical tasks, such as disease or tumor classification and the consistency of the samples naturally lends itself to forward propagation of the captured uncertainty.

The methods for semantic segmentation developed and proposed as part of this thesis are not specific to medical images. Ambiguities similar in nature to the ones described above exist on natural images, e.g. occlusions, lens glare, resolution artifacts or vague class definitions all lead to comparable situations. Modelling and quantifying the ensuing distributions and uncertainties is also an important task with a view towards the ongoing developments of vision systems in autonomous vehicles. As demonstrated in [Chap. 7](#) and [Chap. 8](#) above techniques may also be applicable on natural images such as street scenes.

Thesis Structure The thesis is structured as follows. First, in [Chap. 2](#), we introduce the physical concepts behind relevant medical imaging techniques. Second, in [Chap. 3](#), we explain the process of clinical cancer diagnoses, following the example of prostate cancer and highlight the ambiguities and uncertainties that need to be handled in the process. In [Chap. 4](#) we detail algorithmic developments in the analysis of medical images and feature the similarity and differences with the broader field of computer vision, leading up to state-of-the-art methods including such that aim to mitigate ambiguity and model uncertainty. [Chap. 5](#) - [Chap. 8](#) describe our own contributions at length, following the order in which they were sketched out above. Finally, [Chap. 9](#) concludes this work with a review and an outlook on the many areas that remain to be worked on.

Chapter 2

Medical Imaging Techniques

2.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is based on measuring the electromagnetic signal that nuclei induce in a receiver coil, as their net magnetization after excitation by an external magnetic field resumes its initial state [Lauterbur et al., 1973]. Due to their abundance in the human body and strong response in MRI, most clinically relevant applications of MRI target to measure the effect in hydrogen nuclei [Schlegel et al., 2018]. In over-simplified terms, MRI can be thought of as measuring the local density of hydrogen, resulting in a 3D image of its density. More details on what is actually measured are given below.

Nuclei exhibit a quantum number that is referred to as spin, which can only occur in increments of $1/2$. Hydrogen nuclei for example, have a nuclear spin of $I = 1/2$. Because nuclei are charged and the spin is a vector-valued quantity analogous to a classical angular momentum, the spin is associated with a magnetic moment $\boldsymbol{\mu} = \hbar\gamma\mathbf{I}$ and can couple to an externally applied magnetic field \mathbf{B} . Here \hbar is proportional to Planck's constant and γ is the gyro-magnetic constant that is material dependent. The external field \mathbf{B} exerts a torque $\mathbf{T} = \boldsymbol{\mu} \times \mathbf{B}$ on the nucleus' magnetic moment, inducing a rotation of $\boldsymbol{\mu}$ around the field axis, called precession. This results in a potential energy of

$$E = -\boldsymbol{\mu} \cdot \mathbf{B} . \tag{2.1}$$

Because the quantum-mechanic spin and its spatial projections are quantized, so are the projections of $\boldsymbol{\mu}$ onto \mathbf{B} . Therefore the magnetic field induces only discrete and equidistant energy levels, which is known as the Zeeman effect. With the m th level energy amounting to $E = \hbar\gamma mB$, the energy levels (of which there are two in the hydrogen case)

are separated by

$$\Delta E = \hbar\gamma B = \hbar\omega_L, \quad (2.2)$$

where ω_L is the *Larmor frequency* that is associated with emitted or absorbed radiation (photons) during energy level transitions and also coincides with the frequency of precession of the nuclei. Due to this splitting, a nucleus with spin $I = 1/2$ can therefore only be in parallel or anti-parallel alignment ($I_z \pm 1/2$) with the magnetic field. Because the parallel alignment is a lower energy state, a macroscopic collection of nuclei (e.g. in the tissue volume of a voxel) in thermodynamic equilibrium has a net magnetization $\sum_i \boldsymbol{\mu}_i \neq \mathbf{0}$ parallel to the coil direction in an MRI scanner. Note that both the energy gap and the precession frequency are material- and field-dependent.

Resonance In order to excite nuclei to an energetically higher Zeeman level, additional external fields are employed. To effect transitions at *resonance*, i.e. with magnitude $\hbar\omega_L$, the employed pulses have radio-frequency (RF, in the MHz range) and are typically chosen so as to address hydrogen nuclei. Just after the application of an RF pulse, the net longitudinal magnetization M_{\parallel} (parallel to \mathbf{B}) is decreased or reversed and the spins' precessions have become synchronized. From a macroscopic view, the synchronized precession adds a rotating net transverse magnetization M_{\perp} (in the plain orthogonal to \mathbf{B}), that due to macroscopic cancellations is zero when the phases are not in sync.

Relaxation After application of the pulse, both the longitudinal and transverse magnetization relax to the initial state. The disappearance in transverse magnetization is called T1-relaxation and the restoration of the initial longitudinal magnetization is referred to as T2-relaxation. The magnetizations follow the Bloch equations as functions of time t after the pulse [McRobbie et al., 2017]:

$$M_{\parallel}(t) = M_{\parallel}^{eq} - (M_{\parallel}^{eq} - M_{\parallel}(0) \exp(-t/T_1)), \quad (2.3)$$

$$M_{\perp}(t) = M_{\perp}(0) \exp(-t/T_2), \quad (2.4)$$

where M_{\parallel}^{eq} is the equilibrium magnetization. The overall magnetization therefore follows a spiralling motion after the pulse, which induces a current in a receiver coil that is placed inside the MRI [Currie et al., 2013]. Both effects simultaneously contribute to the relaxation of $\boldsymbol{\mu}$ that is measured by the MRI. Different pulse sequences and pulse timings however are used to increase the sensitivity to either of them thus allowing to ‘up-weight’ the effect in the measurement of one versus the other. The sequences and the ensuing modalities are therefore referred to as *T1-* or *T2-weighted* (T1w or T2w)[Pooley, 2005].

Spatial coding In order to spatially resolve the signal, the external magnetic field is made dependent on space (and time) by using additional coils that induce linear field gradients G_i along the spatial axes $\hat{\mathbf{e}}_i, i \in \{x, y, z\}$ resulting in a superimposed field inside the MRI scanner of $\mathbf{B}(x, y, z, t) = \mathbf{B}_0 + G_x(t)x\hat{\mathbf{e}}_x + G_y(t)y\hat{\mathbf{e}}_y + G_z(t)z\hat{\mathbf{e}}_z$ [Rosen and Wald, 2006]. Note that simply using stationary 3D gradients can not code spatial positions as the resulting Larmor frequencies are degenerate in space. Instead, in order to reduce the problem to a 2D-measurement, typically a *slice selection* is performed first and before acquisition. This entails applying a field gradient G_z in addition to the homogeneous field $\mathbf{B}_0 = B_0\hat{\mathbf{e}}_z$, which induces a z -dependent resonance energy. This way, by employing an RF pulse with $\omega_L = \gamma(B_0 + G_z z')$, it is possible to address and select a particular tissue slice at z' , as slices with $z \neq z'$ will not be excited by the pulse. Just after the RF pulse, two additional gradients G_x and G_y can be applied. Both fields apply a torque on the macroscopic magnetization vector of each voxel and make the precession frequencies location dependent. In order to resolve both the x and y location, one of the fields, e.g. G_x , is applied for a short time only and switched off before acquisition, while the other gradient, e.g. G_y , is left constant throughout signal acquisition. The usage of a short pulse (G_x) before acquisition is referred to as *phase encoding*, because the magnetic moment component that precesses around x (in this case), picks up a location-dependent relative phase $\Delta\varphi(x) = \gamma G_x x \Delta t$. This is because magnetic moments with larger ω_L precess faster and thus ‘run off’, an effect that grows larger for larger x and longer pulse times Δt . After switching off G_x , only the phase of μ_x is altered dependent on x while the precession frequencies ω_L resume the same value for all x . Using a gradient field that remains switched on during signal acquisition on the other hand (G_y here) is called *frequency encoding* since the read-out signal will now entail a spectrum of frequencies, each of which encode a y -position through $\omega_L(y) = \gamma(G_y y + \text{const})$.

Image Reconstruction During signal acquisition, the receiver coil picks up a (scalar) signal $S(t)$ that varies with time. The challenge becomes how to reconstruct the spatial spin density from it. After slice selection as described above, this boils down to performing a 2D Fourier transform of $S(t)$, going from the frequency to the space domain. Assuming the coil equally sums up contributions from all locations, i.e. neglecting geometrical effects etc., the signal measures a superposition of all locations for the spin density $\rho(x, y)$ times the modulation by its phase $\gamma G_x x t + \gamma G_y y \Delta t$:

$$S(t) = \iint_{\text{tissue}} \rho(x, y) e^{i\gamma G_x x t + i\gamma G_y y \Delta t} dx dy = \iint_{\text{tissue}} \rho(x, y) e^{-ik_x x - ik_y y} dx dy, \quad (2.5)$$

where we introduced k -space variables that are conjugate to x and y . To solve for the spin-density, we can apply the inverse Fourier transform (FT⁻¹) [Rosen and Wald, 2006]:

$$\rho(x, y) = \text{FT}^{-1} [S(k_x, k_y)] = \iint_{k\text{-space}} S(k_x, k_y) e^{ik_x x + ik_y y} dk_x dk_y, \quad (2.6)$$

In practice the k -space integral is approximated with discrete steps in k constructing a binned matrix $S(k_x, k_y)$ that is filled using repeated measurements. The matrix is typically sampled one row at a time since the pulse durations Δt in $k_y = -\gamma G_y y \Delta t$ can only be varied using a new RF-pulse, whereas various time samples and thus samples of $k_x = -\gamma G_x x t$ can be obtained after a single resonance [Schlegel et al., 2018].

Diffusion Weighting Given an appropriate MRI sequence, it turns out something like the local water mobility can be measured: **Diffusion-weighted Imaging** can be performed by employing an MRI pulse sequence ('contrast') that is sensitive to the Brownian motion of water molecules, referred to as diffusion. Diffusion is measured in terms of the average squared distance that a molecule travels in some time t through a d dimensional space, $D(t) = \langle (\mathbf{x}_1 - \mathbf{x}_2)^2 \rangle / (2dt)$ [Schlegel et al., 2018]. The diffusion of water in vivo is affected by cellular parameters such as cell dimensionality, compartmentation and transport processes [Posse et al., 1993]. Due to changes in cell morphology, diffusivity can be reduced in cancerous tissue [Chatterjee et al., 2015] making **DWI** a contrast with frequent use in oncology.

The general principle to measure diffusion relies on the observation that protons moving along some path $\mathbf{r}(t)$ for a time τ in a gradient field $G(r)$ pick up a phase [Posse et al., 1993]:

$$\Delta\varphi(\tau) = \gamma \int_0^\tau G(r) r(t) dt, \quad (2.7)$$

which affects the signal intensity as a phase factor under the integral in Eq. 2.6 and results in a signal attenuation:

$$\rho_b(\mathbf{x}) = \rho_0(\mathbf{x}) e^{-D(\mathbf{x}) \cdot b}, \quad (2.8)$$

where ρ_0 is the signal intensity without diffusion, D is the diffusion coefficient in each voxel and b is the gradient factor. b is $\propto (\gamma G \tau)^2 \Delta$, where Δ is the time between gradient pulses (see below), and thus depends on the chosen sequence [Schlegel et al., 2018]. Reduced Brownian motion of water molecules, translates to smaller D and thus increased intensity in the **DWI** image ρ_b .

As suggested by [Stejskal and Tanner, 1965], a trick can be used to measure this attenuation: First a gradient pulse is used to dephase the initially coherent precession (similar to the phase encoding above). Then, after a time Δ , a second gradient pulse that reverses the precession axis is applied. In a characteristic point in time the spins of immobile nuclei again co-incide in-phase and induce a signal, called the *spin echo*, whereas nuclei that moved in the mean-time will have picked up a phase as in Eq. 2.7, resulting in a reduced signal.

In clinical scenarios the diffusion D cannot be measured without interference from non-linear effects involving the localization gradient fields [Le Bihan and Breton, 1985], such that in practice a combined measure, referred to as the *Apparent Diffusion Coefficient* (ADC), is captured.

The ADC-map is commonly obtained from two or multiple b-value images, e.g. using the relation

$$ADC(\mathbf{x}) = \frac{1}{b_1 - b_2} \ln(\rho_{b_2}(\mathbf{x})/\rho_{b_1}(\mathbf{x})) . \quad (2.9)$$

For reduced noise, several different b-value images can be taken and an exponential fit for each voxel is performed, e.g. combining images with gradient factors $b = 40, 400$ and 800 s mm^{-2} [Mark Hammer, 2013].

Both T2 relaxation and diffusion are measured using phase shifts. Individual b-value images ρ_b are therefore also sensitive to T2 relaxation, which is known as *T2 shine-through*, as e.g. both lesions with restricted diffusion and long T2 relaxation will appear very bright in DWI. ADC-maps on the other hand provide a modality that is cleared from this effect [Mark Hammer, 2013]. Note that ADC maps invert the relation to diffusivity compared to b-value images ρ_b : On ADC low intensity voxels correspond to lower Brownian motion, which can be indicative of lesions.

2.2 Computed Tomography Scanning

Computed Tomography (CT) scanning employs X-rays to produce 3D images of the local tissue density (to a good approximation). It is one of the most common medical imaging modalities to date owing its popularity to its high resolution, speed of acquisition and reconstruction, high availability, and importantly its quantitative nature.

This section largely follows [Schlegel et al., 2018]. The principle behind CT is the attenuation of X-rays through absorption (‘photo effect’) or scattering (‘Compton effect’) of photons as they traverse matter. According to the Lambert-Beer’s law, this results in a reduction of the initial photon count (X-ray intensity) N_0 to N as a collection of photons

passes through a material with location dependent absorption coefficient $\mu(\mathbf{x})$. The rays' paths can be parameterized in terms of an origin \mathbf{s} and a unit vector $\boldsymbol{\theta}$ resulting in a signal attenuation of:

$$N = N_0 e^{-\int d\lambda \mu(\mathbf{s} + \lambda\boldsymbol{\theta})} . \quad (2.10)$$

These line integrals of X-ray absorption in a probe, e.g. a human body, are successively captured from many different angles. In order to obtain a 3D scan, CT scanners are built with the x-ray source and the detector rotating around the subject, while the tray carrying the subject can often move through the ring-shaped scanner in order to capture the 3rd dimension.

Once transformed into Cartesian coordinates (see below), CT scans depict the spatial distribution of the absorption coefficient $\mu(\mathbf{x})$. To a good degree of approximation $\mu(\mathbf{x})$ is proportional to the local tissue density, allowing it to be interpreted as a density map. Importantly, CT is a quantitative measure that is quantified in terms of *Hounsfield Units* (HU). Diagnostic CT systems are calibrated using the HU value of water:

$$CT(\mathbf{x}) = \frac{\mu(\mathbf{x}) - \mu_{\text{water}}}{\mu_{\text{Water}}} \cdot 1000 \text{ HU} . \quad (2.11)$$

Using the conventional unsigned 12 bit encoding, CT values range from -1024 HU to 3071 HU. This is suitable to cover the human body's density range, with air typically ranging at -1000 HU and bone at up to 2000 HU on above scale.

Despite CT's many favorable characteristics, there are downsides and reasons to prefer MRI over it: For one, although decreasing due to technical advancements, CT deposits a non-negligible amount of radiation dose in a subject's tissue, which could be harmful. For another, in certain use cases, such as prostate imaging, CT is less informative than MRI [American Cancer Society, 2016b] and is known to result in 'notoriously inaccurate [segmentations]' [Moghanaki et al., 2017].

Image Reconstruction In order to obtain $\mu(\mathbf{x})$, the measured line integrals need to be inverted. Under specific geometric conditions the inversion can be performed analytically, which is why such conditions are adopted in clinical CT set-ups. Analytical signal inversions are for example possible for parallel, fan or cone beam geometries.

In parallel beam tomography, the employed machinery, called *filtered back projection*, can perhaps be sketched out most easily. Parallel beam tomography entails scanning a probe with parallel X-rays that are detected on an array situated behind the probe and perpendicular to the beam. The scans are repeated under different angles θ which overall results in the scan of a 2D slice of a probe. For each angle θ and spatial distance ξ from

the origin (defining the 1D spatial coordinate on the detector), the detector array thus measures the forward transform, called Radon transform [Radon, 1986]:

$$p(\theta, \xi) = \int d\lambda \mu(\mathbf{s} + \lambda\boldsymbol{\theta}), \quad (2.12)$$

whose inversion, i.e. transform to Cartesian space, can be shown to equal a convolution with a ‘Ramp kernel’ k , followed by an integration of the acquisition angle [Schlegel et al., 2018]:

$$\mu(x, y) = \int_0^\pi d\theta p(\theta, \xi) * k(\xi) \Big|_{\xi=x \cos \theta + y \sin \theta}, \quad (2.13)$$

where $*$ denotes a convolution and the ramp kernel’s functional form can be derived in ξ ’s Fourier conjugate space (here $k \propto -1/\xi^2$).

In practice this corresponds to filling a CT scan’s pixels (x, y) with signal contributions one angle after the other: For each θ , first the detector signal is *filtered* with the ramp kernel, yielding $\hat{p}(\theta, \xi) = p(\theta, \xi) * k(\xi)$, which is followed by smearing back (*back projecting*) \hat{p} along the beam’s ray leading through each detector position ξ (a line given by $\xi = x \cos \theta + y \sin \theta$). For this reason the inverse transform is referred to as *filtered backprojection*.

2.3 Histopathologic Light Microscopy

In contrast to the non-invasive imaging techniques covered above, histopathologic examinations require the surgical extraction of a tissue sample. In order to enable the interpretation of the sample under a microscope, a number of steps need to be carried out, which we elaborate upon below, following [Peckham et al., 2013].

First, to prevent tissue decay, the probes require chemical fixation, e.g. by application of formalin, which inactivates enzymes, kills bacteria etc. Additionally, in order to enable the processing of the sample into thin slices, its mechanical stability during slicing needs to be increased. This is either done by embedding the probe in paraffin (wax) or by freezing it using a cryostat during slicing, a process referred to as frozen section processing.

The most commonly employed technique however is the former, wax-embedding, and it requires a dehydration of the specimen by subjecting it to increasing concentrations of ethyl alcohol. Subsequently, the probe is embedded in warm paraffin, which fills voids formerly occupied by water, and after cooling yields a hardened sample ready to be cut by a slicer, called a microtome. Typical slice thicknesses rang at $\sim 5 \mu\text{m}$.

After chemical fixation and dehydration, most cells are colourless and nearly transparent. In order to increase contrast, dyes are applied. Most commonly a dye is chosen, that stains some of the cell components in a bright colour, together with a counter-stain that stains the rest of the cells in a contrasting colour. The most popular dye of this type is Haemotoxylin and Eosin (H&E) staining, which results in cell nuclei appearing purple and most other cell components pink in tone, e.g. see [Fig. 3.2c](#)). Before staining, the specimen are mounted on the microscope glass slide. In order to apply the dye, the fixation wax needs to be dissolved and removed, upon which the probe is re-hydrated, essentially reversing the steps required for section slicing [[Anderson, 2019](#)].

The stained and mounted specimen are then inspected through conventional light microscopy, allowing resolutions of down to ~ 200 nm (which refers to the smallest spacing between two point objects that is still distinguishable) [[Peckham et al., 2013](#)].

Whole Slide Imaging Comparatively simple, man-operated and -interpreted light microscopy is still the most common tool in histopathology [[Barghaan, 2015](#)]. Increasingly the field however turns towards a technology, referred to as *whole slide imaging* (WSI). Here scanners are employed to automatically create a digital image of the whole slide of a mounted specimen [[Farahani et al., 2015](#)]. Under the hood, a light microscope is combined with a digital camera, mechanisms to position the slide and a software that stitches individual images together. The advantages of WSI lie in a simplified sharing and archiving of histopathologic images and the possibility of an automated image processing and interpretation. The images are typically taken at several magnifications, just like a regular microscope would allow for, and saved alongside one another in what is called an image pyramid. This results in very large files (several gigabytes per image), with the highest resolution images exhibiting gigapixel sizes.

Chapter 3

Medical Image Analysis: The Diagnosis of Prostate Cancer

Diagnosing cancer is typically a process that involves several successive steps, which are set in motion upon an initial suspicion. Each such step involves taking a measure, such as an [MRI](#) scan and its interpretation, and is chosen so as to bring information to the table, that the previous step might not have been able to produce. For example because a scan and its interpretation may have come out ambiguous. In that manner, the space of hypotheses can be narrowed down successively, as gradually more powerful prognostic tools are employed.

While this sounds fairly straight-forward on paper, it is a very complex process in reality. A large part of the reason is that the decision making process needs to carefully weigh the health risk, discomfort, financial cost and the uncertainties that are associated with each diagnosis step. Traversing the chain of steps, it is in principle often possible to fully disambiguate the diagnosis. In practice however, the personal cost associated with that may often be unacceptably high. For example fully removing a suspicious prostate that may or may not harbor aggressive cancer through what is called radical prostatectomy would give the clearest diagnostic picture of what is (was) going on, but would have a strong negative impact on a patient's quality of life.

In this chapter we take a closer look at this diagnosis process and highlight the steps and their abilities as well as their limitations as we follow along the example of prostate cancer diagnosis. Besides arguably making for a particularly complex case, staying close to the example of prostate cancer has the added benefit of directly introducing relevant concepts for [Chap. 5](#) and [Chap. 6](#). After describing the pitfalls and ambiguities involved in the interpretation of [MRIs](#) and histo-pathological images of biopsy probes, we discuss studies on the ensuing inter- and intra-rater variance on those modalities.

3.1 Current Clinical Diagnosis Steps

To exemplify the procedure for diagnosing cancer, we present it following the process for a very frequent type: [Prostate Cancer \(PCa\)](#).

The prostate is a gland that is part of the male urinary and reproductive system responsible for producing and secreting a fluid that constitutes $\sim 20-30\%$ of the ejaculate. It is embedded deep inside a man's pelvis, below the bladder and in front of the rectum. The gland wraps around the upper part of the urethra, a tube that carries urine from the bladder out of the body. The prostate of men in their 20s and 30s is about the size of a walnut but will often transform to the size of a peach later in life as part of a benign growth. [Fig. 3.1](#) gives more details on the anatomy of a 'normal prostate' and its distinct anatomical zones.

According to the American Cancer Society about 1 man in 9 will be diagnosed with prostate cancer during his lifetime. About 6 cases in 10 are diagnosed in men aged 65 or older whereas it is rare before age 40 [[American Cancer Society, 2016a](#)]. The average age at the time of diagnosis is about 66. It is estimated that PCa is the most frequent cancer type among new male cancer cases in 2019 (20% of all new cases in the United States) [[American Cancer Society, 2019](#)].

Although PCa is also the most deadly in absolute numbers (10% of all cancer-related deaths in the US), only about 1 in 49 diagnosed men will die from the disease. If detected at a local stage, the 5-year survival chance is approaching 100%, when detected late (after infiltrating beyond the prostate gland) the chances of survival decrease to 30% [[American Cancer Society, 2019](#)]. Therefore, as is true for cancer in general, an early detection is key. These numbers, with the good chances of long-term survival, however also highlight that there may be a unique chance to offer milder forms of treatment that enable a high quality of life while the disease is on-going and monitored.

Progressively Narrowing down the Diagnosis Suspicious early symptoms for PCa may be detected as part of routine screenings or surface in the form of certain urinary conditions that a subject may experience and bring to the attention of a physician. Upon an initial suspicion the first clinical steps involve what is called a [Digital Rectal Examination \(DRE\)](#), where a clinician uses their finger to feel for whether the prostate's size is enlarged and whether it appears hard or lumpy, see [Fig. 3.2a](#)) for a schematic of the process. This is naturally a subjective measure with limited accuracy and reproducibility [[Smith and Catalona, 1995](#)], but it is fast, immediate, cheap and simple and may help detect very unambiguous cases. Another early test that is often done before or in addition to a DRE is a blood-test that screens for an elevated [Prostate-specific Antigen \(PSA\)](#)

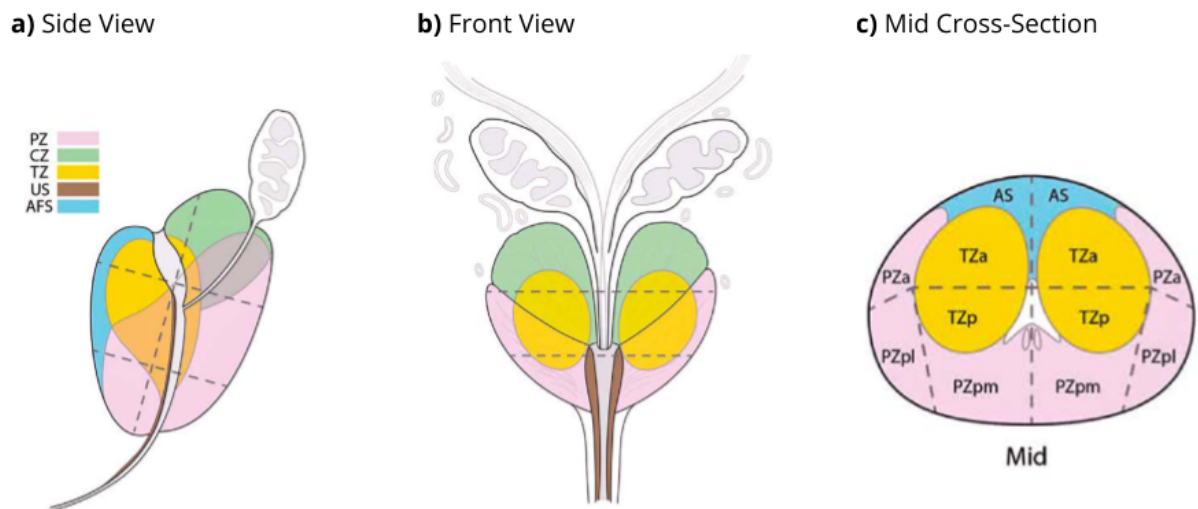
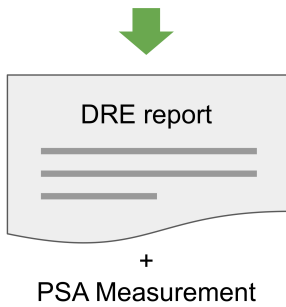
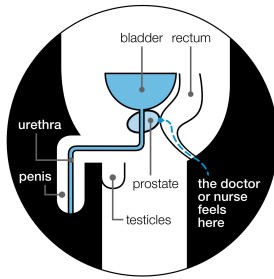


Figure 3.1 | Anatomy of the Prostate. The prostate gland's *transitional zone* (TZ, in yellow), *central zone* (CZ, in green) and the *anterior fibromuscular stroma* (AFS, in blue) are partially wrapped by its *peripheral zone* (PZ, in pink). The *urethral sphincter* (US, in brown) follows along the urethra that leads through the prostate. The illustrations show an idealized 'normal prostate' from a) a side, b) a frontal and c) a cross-sectional view (in which further sectorial distinctions are shown). Image credit: [Weinreb et al., 2016].

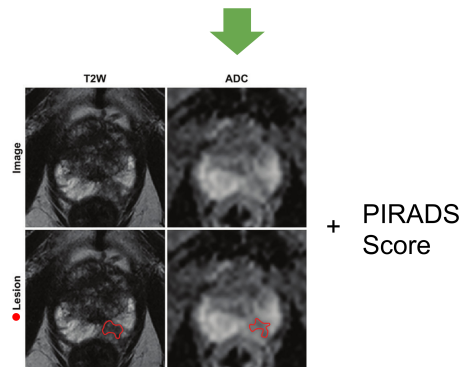
level. A PSA level above 4 ng mL^{-1} is indicative of higher chances for PCa, although a significant number of actual PCa cases may have a PSA test come out below 4 ng mL^{-1} [Prostate Cancer UK, 2018] and high numbers can also be caused by non-cancerous diseases, highlighting the test's fair but in isolation insufficient diagnostic value for prostate cancer.

In case the suspicions hold up in the DRE or the PSA measurement, further diagnosis steps are warranted. In the past this meant a prospective patient directly underwent surgical biopsy, see Fig. 3.2c). Aside from discomfort during the procedure and the days or weeks after, surgical biopsy involves the risk of infections of the prostate gland. Furthermore, without knowledge of the locations of potential lesions within the prostate, a grid of needle positions over the prostate needs to be scanned across in order to get the necessary coverage, which further increases health risks and might negatively affect potential imaging of the prostate at a later stage. For these reasons, suspicious subjects increasingly undergo **Multi-parametric MRI (mpMRI)** scanning of their pelvis region including the prostate gland - carried out before or ideally even instead of biopsy. Typically **T1-weighted/T2-weighted** as well as **DWI** protocols (in order to calculate **ADC** maps) are run. The resulting scans are then interpreted by a radiologist who produces a report encompassing locations of prospective lesions, often delineating them on the

a) Digital Rectal Exam (DRE)



b) MRI Scans + Radiologist Segmentation



c) Trans-rectal Biopsy + Pathologic Assessment

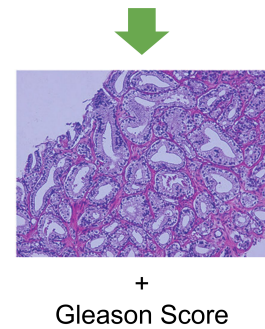
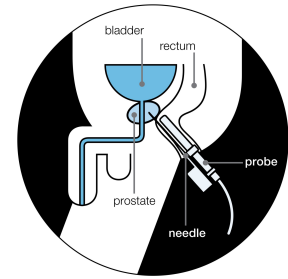


Figure 3.2 | Current Clinical Prostate Cancer Diagnosis. a) After initial suspicions or during routine screening a **Digital Rectal Examination (DRE)** may be conducted along with a blood test to measure the subject's **Prostate-specific Antigen (PSA)** level. b) Multiparametric MRI scans are taken and interpreted by a radiologist who produces a report entailing segmentations of potential lesions and a **PIRADS** score for each lesion that assesses its malignancy based on MRI appearance. c) If deemed appropriate, a targeted and/or systematic (raster) **Trans-rectal Ultra-Sound Guided Biopsy (TRUS-biopsy)** is performed by a urologist. Using biopsy needles, probes containing tissue are extracted, stain-dyed and examined by a pathologist under a microscope at cell-level resolution, who assigns a **Gleason Score** to assess the malignancy of the tissue in each probe. The images are borrowed from various sources: the schematics showing patients are taken from [Prostate Cancer UK, 2018], the MRI scanner is a Magnetom Prisma by Siemens [Siemens Healthineers, 2019], the mpMRI images are taken from [Bonekamp et al., 2018] and the histopathology slide is from [Chen and Zhou, 2016].

image (akin to a semantic segmentation of the lesion), and a score that expresses the radiologist's level of suspicion for malignancy for each prospectively detected lesion, called the **PIRADS** score. This process is sketched in Fig. 3.2b).

Delineating and grading lesions from mpMRI images serves several purposes. For one, it can provide image guidance for targeted biopsy and surgery and thus allows to reduce the number of sampled probes or may even permit to avoid the invasive procedure of a biopsy altogether. Instead, a subject may be declared cancer-free from the mpMRI evidence alone or border cases meeting certain criteria could be admitted into *active surveillance*, which encompasses regular monitoring without aggressive treatments. The circumstances under which it is warranted to make this distinction and to forgo biopsy are currently under research, e.g. see Chap. 5. Part of this quest is the development of diagnosis strategies that are not only very sensitive to PCa but also very specific, thus reducing the amounts of false positives and allowing to skip invasive surgery or treatments prone to side effects.

ADC images in particular are known to provide discriminative evidence for radiologists experienced in the grading of prostate cancer from MRI. As is discussed below in more detail, there however often still remain significant ambiguities in this modality. For this reason, if the ambivalency persists, further diagnostic steps to decrease the uncertainty on the malignancy grade of the apparent lesions are required and surgical biopsy may be unavoidable. The biopsy is most commonly carried out using **Trans-rectal Ultra-Sound Guided Biopsy (TRUS-biopsy)**, see Fig. 3.2c), in which a urologist takes several tissue probes using needles that are inserted through the wall of the back passage into the prostate. This process is guided by ultra-sound which offers a good degree of needle localization and allows to target lesions seen on mpMRI (*targeted biopsy*) as well as sampling across a grid of locations (*systematic biopsy*). After extraction the tissue probes are stain-dyed by a pathologist and viewed under a microscope of cell-level resolution, see Sec. 2.3. The pathologist then assigns a score to each biopsy sample that is dependent on the two predominant cells morphologies, the **Gleason Score (GS)**, see below. This score is often the best available ground truth and therefore treated as gold standard, but because it is a human-assigned subjective measure, it also suffers from inter-rater variances, as is discussed below.

By design, a prospective grid of biopsy samples does not cover the entire prostate and there is a chance that areas are under-sampled. Similarly, target locations may be missed. This introduces an additional degree of uncertainty. Unfortunately the most reliable diagnosis involves removing the prostate gland altogether in a process called *radical prostatectomy*. Because of risks such as the potential for incontinence and erectile

dysfunction, it is very desirable to avoid this measure and only used as a last resort in more advanced cases. Cases that had undergone TRUS-biopsy and later on radical prostatectomy allow to assess the sensitivity of TRUS-biopsy, which was shown to find up to 97% of the most severe lesions that were found after post vivo biopsy of the prostate [Radtke et al., 2016], therefore promising strong sensitivity while however still leaving some patients misdiagnosed.

3.2 Malignancy Quantification & Challenges

The core question after a suspicion for cancer is whether or not an abnormal finding is benign or malignant. Benign lesions are typically defined as tissue alterations that do not spread, which is in contrast to malignant lesions, that are characterized by aggressive growth and spreading. Unfortunately, most of the times the distinction cannot be made univocally and the grading instead aims at expressing a (subjective) likelihood for malignancy given the evidence. This is true for the interpretation of MRIs and, perhaps surprisingly, it is also true for the interpretation of biopsy probes, as is discussed in the following.

Assessing MRI Evidence: The PIRADS System In an attempt to quantify their malignancy assessment, radiologists currently employ a 5-grade system, called **The Prostate Imaging Reporting and Data System (PIRADS)**. A PIRADS score of 4 or 5 is considered as clinically significant tumor. The system aims at standardizing the interpretation and reporting of prostate mpMRI examinations. It specifies rules for how to grade and detect prostate cancer given different mpMRI modalities. If applicable these rules are also specific to different anatomical parts of the prostate, such as the prostate's *peripheral* and *transitional zone* (PZ and TZ, see Fig. 3.1), as tumors can exhibit distinct MRI appearance in different anatomical regions. PIRADS v2, the second version of the protocol, is adopted as its current standard [Weinreb et al., 2016].

As discussed in Sec. 2.1, MRI images indirectly measure tissue properties such as the diffusivity of water (DWI scans and ADC maps) and the tissue's proton density and proton spin relaxation times (T1w- and T2w-sequences). The difficulty in assessing the tumor grade from MRI lies in the often subtle difference in appearance of lesions with different malignancy, given those indirect measurements [Borofsky et al., 2017]. This is particularly problematic in prostate MRI, as the prostate gland can exhibit tissue density that is naturally heterogeneous and additionally suffer from 'benign mimickers' that can exhibit the same appearance as tumors, but are non-cancerous conditions. This

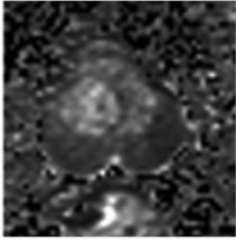

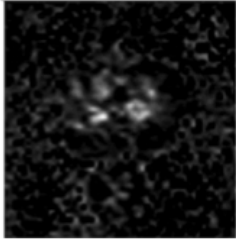

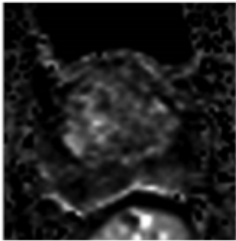
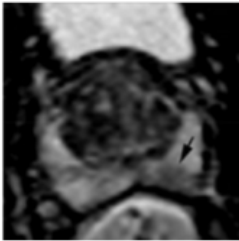
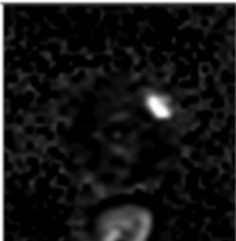
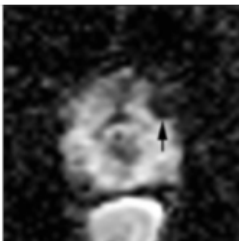
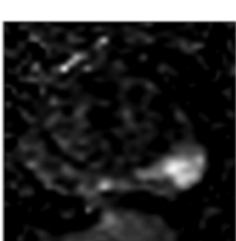
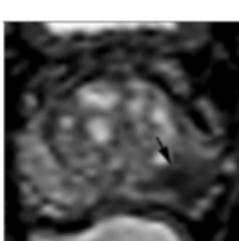
PIRADS	High-b DWI map	ADC map	Assessment
1			No abnormality (i.e. normal) on ADC and high b-value DWI.
2			Indistinct hypointense (lower intensity) on ADC (see arrow).
3			Focal mildly/moderately hypointense on ADC (arrow) and isointense/mildly hyperintense on high b-value DWI.
4			Focal markedly hypointense on ADC (arrow) and markedly hyperintense on high b-value DWI; Lesion size < 1.5 cm on axial images.
5			Same as 4 but ≥ 1.5 cm in greatest dimension (arrow) or definite extraprostatic extension / invasive behavior

Figure 3.3 | Exemplary PIRADS Assessments on DWI images. This figure presents examples for each PIRADS score (1 through 5) that a radiologist assigned based on the provided high b-value DWI and ADC scan for the five different cases. Note that the images are shown in an axial view, akin to a mid cross-section as in Fig. 3.1c). The rightmost column gives an assessment based on PIRADS guidelines for the prostate's peripheral zone. Images and assessments are borrowed from [Weinreb et al., 2016].

possibly ambiguous appearance can also make it difficult to pin down a precise location or segmentation of lesions [Puech et al., 2012].

The PIRADS system recommends usage of both high b-value DWI- and ADC-maps as well as T1w/T2w scans in order to narrow down the diagnosis by exploiting their complementary measurements. The 5-grade scale gives empirical categorizations of lesions, where a higher grade is more severe. Lesions are ‘manually’ distinguished based on MRI intensity, focality, definiteness of margins, shape and size.

Because DWI measurements are non-quantitative, it is recommended to compare a prospective lesion’s ADC-intensity to normal appearing tissue in the respective scan. On DWI, lesions in both the TZ and PZ appear hyperintense (elevated intensity) on high b-value images and hypointense (decreased intensity) on ADC-maps. Fig. 3.3 gives example cases for all five PIRADS grades alongside the grading rules for PZ lesions on ADC. For a comprehensive set of PIRADS grading rules, we refer to [Weinreb et al., 2016].

On T2w images, clinically significant cancers in the PZ usually appear as round or ill-defined hypointense (decreased intensity), focal lesions. This appearance is however not specific and there are again many benign conditions or other non-cancerous diseases that appear similar (‘benign mimickers’) [Weinreb et al., 2016]. TZ lesions on T2w may appear as ‘non-circumscribed homogeneous, moderately hypointense lesions (‘erased charcoal’ or ‘smudgy fingerprint’ appearance) [and have] spiculated margins [...]’ [Weinreb et al., 2016]. The more of these features are present, the higher is the likelihood of clinically significant prostate cancer. An instance of the ‘erased charcoal signal’ is illustrated in Fig. 3.4b), which shows a case where tumor tissue is very difficult to discern from normal or benign tissue.

Some benign conditions or non-cancerous diseases may have a similar signature to tumor on both T2w and ADC (signal hypointensity). Fig. 3.4a) shows a case of this type, where prostatitis, a non-cancerous inflammation, ‘mimics’ the appearance of a lesion, as can be seen in the displayed comparison to a (low-grade) prostate lesion. In some cases high b-value images can break the tie and allow a distinction since on them, non-cancerous cases can appear with lower intensity as compared to tumor. *Benign Prostate Hyperplasia* (BPH) however, a benign condition very common among older men which is associated with prostate growth later in life, unfortunately can be difficult to distinguish from cancer based on DWI. Aside from the uncertainty on the grade, ambiguous appearance on ADC often leads to an underestimation of tumor sizes compared to the true volumes found through prostatectomy [Bratan et al., 2014], indicating the difficulty of faithfully segmenting lesions on MRI.

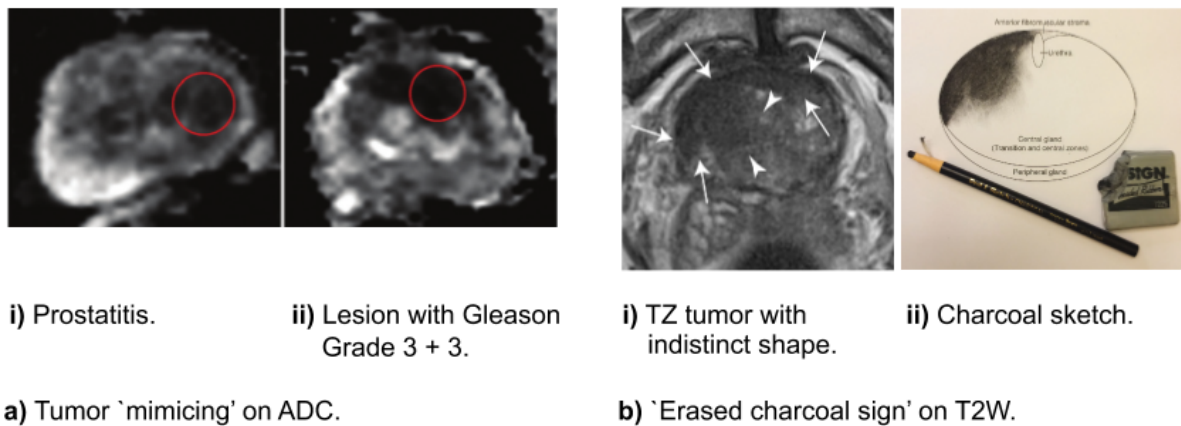


Figure 3.4 | Prostate MRI Ambiguities. Figure a) shows two very similar appearing ADC-maps, with abnormal appearing tissue in red circles. Histopathology however revealed that i) showed prostatitis (a non-cancerous inflammation) versus ii) harbored a benign prostate lesion. In Subfigure i) of the right hand side (Figure b)) a case of TZ prostate tumor that is embedded in BPH-tissue and shows very indistinct, smudged boundaries, that are difficult if not impossible to tell apart from the benign BPH tissue. This is akin to a charcoal sketch, see subfigure ii), and thus referred to as an 'erased charcoal sign'. Figures a) are borrowed from [Nagel et al., 2013] and Figures in b) from [Sakala et al., 2017].

The different MRI sequences are also believed to hold different discriminative power depending on the prostate's anatomical zone. For abnormalities in the TZ, T1w-/T2w-MRI is considered the primary determining sequence, whereas PZ foci are recommended to be graded based on DWI-images according to the PIRADS guidelines [Weinreb et al., 2016]. Due to the increased tissue homogeneity, the detection and characterization of clinically significant tumor is generally regarded more reliable in the PZ than in the TZ.

Assessing Biopsy Evidence: The Gleason System After surgical biopsy, the extracted tissue samples are prepared on microscope slides and stain-dyed so as to increase contrast giving a characteristic pink color signature, see Fig. 3.2c) and details in Sec. 2.3. A pathologist then carries out the diagnosis of each probe using a microscope to assess the morphology of the extracted cells. The morphology patterns are characterized based on how well differentiated the cells appear. Normal cells have very regular, well differentiated appearance and are closely packed whereas cancerous tissue shows cells of increasingly indistinct, heterogeneous morphology with more loose spread, called anaplasia. Fig. 3.5 gives a definition of the 5 different Gleason patterns [Gleason, 1966] that are employed and schematically illustrates the associated cell morphologies.

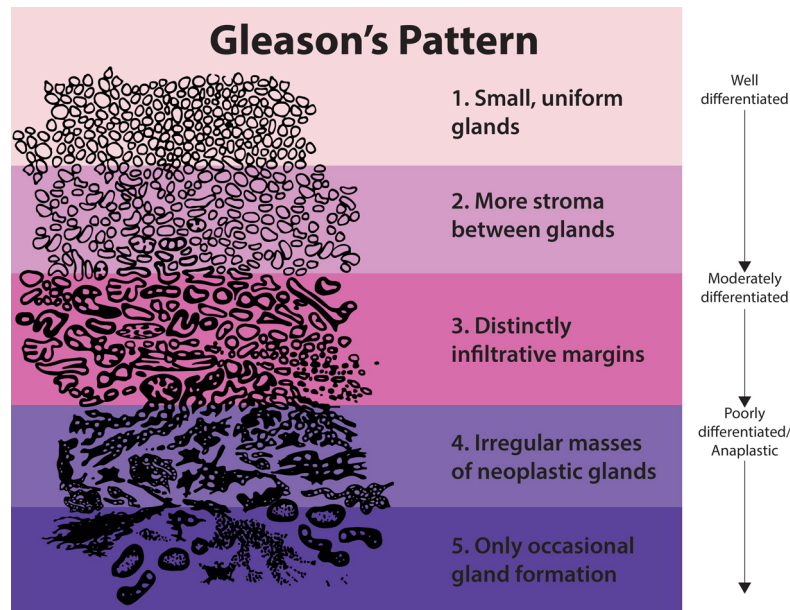


Figure 3.5 | Gleason Pattern. A schematic illustration of different cell morphologies as found in histologic tissue samples of the prostate alongside the 5 different Gleason patterns that are employed to characterize them. Image borrowed from [National Cancer Institute, 2019].

Tissue probes very commonly exhibit more than one of the distinguished cell morphologies, which is why pathologists summarize their findings in a score that combines the found Gleason patterns, called the **Gleason Score (GS)**. The score is produced by adding the two most frequent Gleason patterns, such that it technically results in 9 different grades, from $1 + 1 = 2$ to $5 + 5 = 10$. Because significant differences in patient hazards have been found depending on which of the Gleason patterns dominate the probes [Epstein et al., 2016], further distinctions are commonly made for when the probe shows Gleason patterns above 3. In these cases the predominant cell type (comprising more than 50 % of the probe) is put first and the less frequent one (less than 50 % but more than 5 % of the probe) is reported second, resulting in **GS** grades that account for this majority-minority order by defining e.g. $3 + 4 =: 7a$ and $4 + 3 =: 7b$. In practice it is exceedingly rare for pathologists to diagnose Gleason patterns below 3, which is why the **GS** effectively starts at $3 + 3 = 6$. Taking this into account alongside statistically determined clinical significance of the groups, a simplification to a 5-grade system called *Gleason Grade Group* was recently proposed and is being increasingly applied [Epstein et al., 2016]. Table 3.1 specifies the ensuing mappings between Gleason patterns, Gleason Score and Gleason Grade Group.

Table 3.1 | Gleason Score Categories. The table gives the mapping between the Gleason pattern combinations and the Gleason Score and Grade Group.

Gleason Patterns	3+3	3+4	4+3	4+4	4+5	5+4	5+5
Gleason Score	6	7a	7b	8	9a	9b	10
Gleason Grade Group	1	2	3	4	5		

While there is some debate about which **GS** threshold should be used to distinguish benign from aggressive and thus malignant tumors, the consensus view appears to define aggressive tumors as such with $GS \geq 7a$ [Carter et al., 2012, Epstein et al., 2016, Loeb et al., 2016].

The fact that the **Gleason Score**'s attempt at quantifying abnormal cell morphology holds prognostic value for the aggressiveness of prostate cancer [Epstein et al., 2016] supports the view that **Diffusion-weighted Imaging** can give a handle on assessing tumor aggressiveness in vivo, as the diffusivity of water correlates with changes in cellular morphology [Chatterjee et al., 2015].

However, even the more direct examination of cellular morphology under a microscope and the assessment in terms of the **Gleason Score** has considerable limitations, since the qualitative nature of the assessment as well as the aggregation of only two predominant patterns without further quantification, naturally leads to inter-rater variability and potential inaccuracies. A large part of the problem may also be identified in the Gleason pattern scheme itself, which bins the continuous cell alterations. Moreover, just like on **mpMRI**, there are other, non-cancerous conditions that are known to mimic the histopathologic appearance of prostate cancer [Hameed and Humphrey, 2010]. For these reasons the **Gleason Score** should be viewed as the best available gold standard, but not as an infallible ground truth.

3.3 Ambiguity and Inter-rater Variability

A range of studies have examined the variability in the interpretations of the respective diagnosis steps. They largely confirm that each step's reliance on an ambiguous measure, results in pronounced inter-rater variability. Before diving into individual findings, let us briefly review two popular statistics that measure the inter-grader agreement on categorical assessments. In order to compare agreement between two graders, *Cohen's Kappa* (κ_C) [Cohen, 1960] is often the statistic of choice, it is however limited to exactly

two graders. For a fixed but arbitrarily large number of graders *Fleiss's Kappa* (κ_F) [Fleiss, 1971] is often used instead. Both statistics correct for chance agreement such that they come out 0 when agreement is attributable to chance only and reach a maximum value of 1 for complete agreement between raters. [Landis and Koch, 1977] suggest a categorization of intervals on κ -values for a more readily available interpretation: $\kappa < 0$ - poor agreement, $0 < \kappa \leq 0.2$ - slight agreement, $0.2 < \kappa \leq 0.4$ - fair agreement, $0.4 < \kappa \leq 0.6$ - moderate agreement, $0.6 < \kappa \leq 0.8$ - substantial agreement and $0.8 < \kappa \leq 1$ - almost perfect agreement.

DRE [Smith and Catalona, 1995] compared the inter-rater variability for **Digital Rectal Examination** to distinguish between malignant and benign cases using a cohort of 116 subjects. They found a $\kappa_C = 0.22$, which was statistically significantly above chance agreement, but according to [Landis and Koch, 1977] may be described as no more than a *fair* agreement, therefore highlighting a reduced reproducibility and reliability of the (urologist performed) **DRE** examination in diagnosing prostate cancer.

PIRADS Several independent studies have assessed the inter-rater variability of the **PIRADS v2** interpretation of **mpMRI** images by radiologists. [Muller et al., 2015] have examined the inter-rater variability of 5 radiologists given 162 lesions in 94 patients and found a multirater $\kappa_C = 0.46$, indicating *moderate* agreement. [Rosenkrantz et al., 2016b] asked 6 radiologists to give a **PIRADS v2** assessment for 40 cases, upon which all of them received extra training and discussed prior results. In a second session, all of them were asked to grade another 80 cases. The inter-rater variability for **PIRADS v2** ≥ 4 on **PZ** lesions amounted to $\kappa_C = 0.59$ and $\kappa_C = 0.51$ on **TZ** lesions, in the first session. The inter-rater agreement did not change significantly after the extra training session and discussions. Overall this study thus found a *moderate* agreement on **PIRADS v2** ≥ 4 . [Pierre et al., 2018] found the inter-rater agreement on 92 **PZ** lesions in 74 patients for two radiologists to be only *fair* ($\kappa_C = 0.39$).

Aside from image ambiguities on **mpMRI**, there is a known learning curve, such that some of the variability may be reduced with more experience [Latchamsetty et al., 2007].

TRUS-biopsy As mentioned above, using **Trans-rectal Ultra-Sound Guided Biopsy** can be error prone due to under-sampling or missing target locations. Comparing cases that subsequently underwent radical prostatectomy allows to assess the accuracy of **TRUS-biopsy**. [Epstein et al., 2012] evaluated the **GS** assessment of almost 8000 **TRUS-biopsy** cores against the later prostatectomy result and found that about a fourth of the cores

that had been graded $GS \leq 6$ were undergraded and were really clinically significant cancer cores of grade $GS \geq 7a$. On a patient level, i.e. combining the assessments of all biopsy cores for the respective patient, the agreement may be more robust, as is suggested in [Radtke et al., 2016], who found that the agreement with respect to the overall and thus worst GS score per patient matched the one determined by radical prostatectomy in 97% of the times (evaluated on 120 patients).

Lesion Segmentation [Bratan et al., 2014] analyzed lesion segmentations of two radiologists, provided for mpMRI images of 202 patients. Both radiologists significantly undersegmented tumor on both T2W- and ADC-scans, as measured in tumor volume agreement. [Borofsky et al., 2017] similarly found two radiologists to undersegment lesion volume in 8% of the cases as observed on a dataset of 162 lesions from 100 patients.

Gleason Score [Melia et al., 2006] examined the inter-rater variability of 9 pathologists. Each of them was asked to give an assessment for 81 slides of cancer-diagnosed histopathology slides in terms of the Gleason score groups 2–4, 5–6, 7, 8–10. A statistic of $\kappa_F = 0.54$ indicated only *moderate agreement* between them.

3.4 Discussion

The process of narrowing down the diagnosis for prostate cancer successively reduces ignorance and ambiguity of the cancer grade. Informally speaking, this ignorance is gradually lowered along the chain of potential steps given by DRE + PSA-measurement, mpMRI scan + interpretation, TRUS-biopsy and, as a last resort, radical prostatectomy.

While the perhaps ultimate measure, radical prostatectomy, faithfully reveals the full picture, most of the times it is not an acceptable option. Instead, less invasive measures are preferred, ideally even shunning the need for biopsy altogether. This paradigm comes at the cost of the remaining limitations of the individual measures, which are mostly linked to an irreducible amount of uncertainty.

As documented above, there are considerable amounts of inter-rater variability present in almost all diagnosis measures. Sometimes this variability can be partially reduced, e.g. when it is due to different amounts of reader experience. For this reason it might be helpful to develop algorithms that perform (at least) at the level of an average radiologist and provide an algorithmic second opinion, that could supply clinical decision support.

Other times the variability cannot be reduced given the available evidence, due to inherent ambiguity. In order to still enable the most informed decision possible, it

is important to be aware of this ambiguity, quantify it and if admissible take further diagnostic steps to reduce it.

An important part of the diagnostic pipeline is the delineation, i.e. segmentation, of prostate lesions. A precise localization in this form offers many benefits, e.g. it enables image-guided biopsy, surgery and treatments such as focal therapy and radiation therapy. Monitoring tumor growth is yet another important application, especially in prostate cancer, where active surveillance is a desirable way to forgo aggressive treatment and invasive surgery for the sake of an improved quality of a patient's life. All these applications could benefit largely from well-calibrated uncertainty that allows an understanding of the image ambiguities.

Chapter 4

Medical Image Analysis: Algorithmic State-of-the-Art

By virtue of processing image-type data, medical image analyses naturally bear much similarity to other vision problems on natural images. Medical image analysis, in the past, has however required a lot of expert knowledge and posed barriers such as data and annotation scarcity, which often lead to very tailored solutions. As more data is becoming available and with natural image algorithms becoming increasingly domain agnostic, medical imaging analysis techniques are more and more in sync with those for natural images.

For this reason, medical image analyses have profited largely from the progress in the field of computer vision that came over the last decade. [Convolutional Neural Networks \(CNNs\)](#) were put (back) on the map with the sweeping win of 2012's ImageNet classification challenge [[Deng et al., 2009](#)] by AlexNet and have went on to dominate virtually all current (natural) image understanding challenges (CIFAR10/100 [[Krizhevsky et al., 2009](#)], MSCOCO [[Lin et al., 2014](#)], PascalVOC [[Everingham et al., 2010](#)] and Cityscapes [[Cordts et al., 2016](#)]).

Aside from whole-image classification (the task of assigning a single label per image) the field has increasingly considered more fine-grained image understanding tasks, such as *semantic segmentation*, where a label for every pixel in an image is sought after. This task also happens to be of large clinical value, as dense annotations allow to plan biopsies or radiation therapy, monitor tumor growth or quantify the heart volume over time, etc. For this reason, it is unsurprising, that after recording an [MRI](#) or a [CT](#) scan, radiologists often produce a pixelwise annotation of structures of interest, thereby enabling subsequent clinical steps that depend on the location and the semantics of things visible in the scan.

Given enough and -if possible- clean annotations, the task of semantic segmentation is now also handled very successfully by deep CNNs.

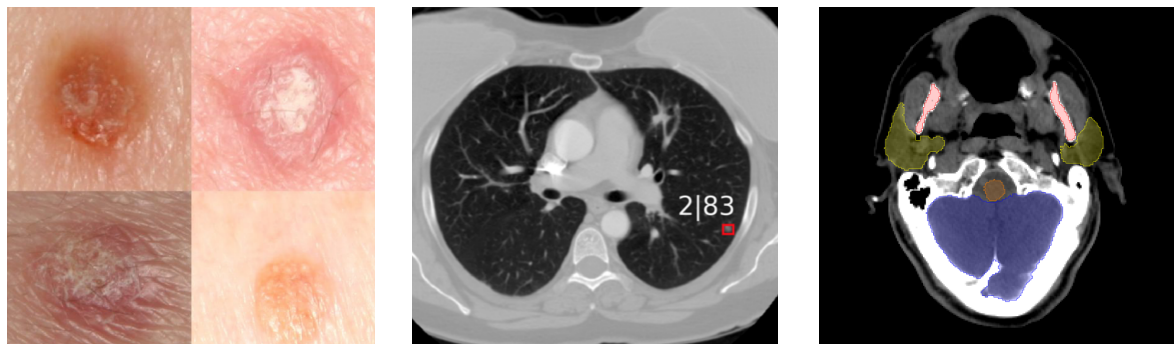
One of the foundations of this thesis however, is the observation that in certain cases, producing a pixel-perfect segmentation may be difficult if not impossible. This class of cases happens to be particularly common in medical images: *ambiguous* images, which do not provide sufficient evidence to nail down a unique hypothesis. For examples thereof, revisit [Chap. 3](#).

The present chapter lays out the background that may underpin the discussions of our contributions towards improvements in this scenario. In the following, models for both image- and pixel-level classification, their relationship and developments up to the current [State-of-the-Art \(SotA\)](#) are elaborated upon first. Then current tools that aim at dealing with uncertainty and label noise in these *discriminative models* are presented. The last part of this chapter is dedicated to current *generative models* and provides context for [Chap. 6](#) - [Chap. 8](#).

4.1 From Image Classification to Semantic Segmentation

A large area of computer vision is concerned with the interpretation and thus the classification of images. The interpretation of an image can be performed on different levels of granularity and the appropriate level is naturally chosen dependent on the task of interest. On medical images for example, we might be interested in whether a photographed skin lesion is benign, thus asking for an image global classification [[Esteva et al., 2017](#)]. When screening for the necessity of biopsies, we might want to know the location along with a malignancy estimate for a potential lesion in a lung CT scan [[Jaeger et al., 2018](#)], therefore requiring an object-level classification. And conversely, when planning for radiation treatment, we would like to have a pixelwise classification of *organs at risk* in order to be able to minimize the radiation impact on healthy tissue [[Nikolov et al., 2018](#)]. For an illustration of these examples see [Fig. 4.1](#).

Historically, the interpretation of images was carried out at a local level, mostly ignoring global context. Then came CNNs, allowing to learn a complex mapping from the full image to a class label. Today's deep semantic segmentation models allow to make dense predictions, but this time around incorporating different scales of context in a principled way rather than literally classifying every pixel independently. This development is unfolded below in a (mostly) chronological way, leading up to the respective [State-of-the-Art \(SotA\)](#).



(a) Image-Level: Images of skin lesions to be classified. **(b) Object-Level:** Detection of a malignant lesion on a lung CT. **(c) Pixel-Level:** Semantic segmentation of a brain MRI.

Figure 4.1 | Classification of Medical Images at different Levels of Granularity. Subfigure (a) shows images of skin lesions whose malignancy is to be classified [Esteva et al., 2017], (b) depicts a CT scan with an overlaid bounding-box and malignancy confidence for a detected lesion [Jaeger et al., 2018] and (c) shows an MRI slice with overlaid multi-class pixelwise annotations made by a deep neural network [Nikolov et al., 2018].

Hand-crafted Feature Classifiers Largely ignoring the pioneering work on *CNNs* of the 1980s, other early efforts towards the algorithmic interpretation of images were trying to deduce the semantic content of an image from local (dense), hand-crafted features such as edge detectors [Gavrila and Philomin, 1999], Haar wavelets [Viola et al., 2001], texture features [Tieu and Viola, 2004] or intensity gradients [Dalal and Triggs, 2005]. It was quite apparent that modeling complex relationships, both between objects and also in terms of the interplay of local and global image cues, should help in further improving algorithmic image understanding. To this end many works proposed different handcrafted hierarchies of local features, hierarchies of parts-templates or cascades of weak classifiers [Fleuret and Geman, 2001, Serre et al., 2006, Ullman et al., 2002]. One of the most widely known approaches in this spirit is SIFT (‘scale-invariant feature transform’) [Lowe, 2004], which constructs a hierarchy of features produced by difference of (fixed) Gaussian functions.

Using local handcrafted features has several obvious pitfalls. For one, the set of features may neither be optimal for, nor adapt to the intended application domain (e.g. by learning). For another, producing classifications at the intended level of coarse-to-fine granularity, requires either pre- or post-processing steps, or hand-crafting of a feature or classifier cascade, that is prone to be brittle and suffer from modeling inadequacies.

Today such systems have by far and large been replaced by learning-based, deep neural networks. In medical imaging however, they have remained a fairly popular tool. The reason is two-fold: the hunger for labelled data of *CNNs* as well as the difficulty

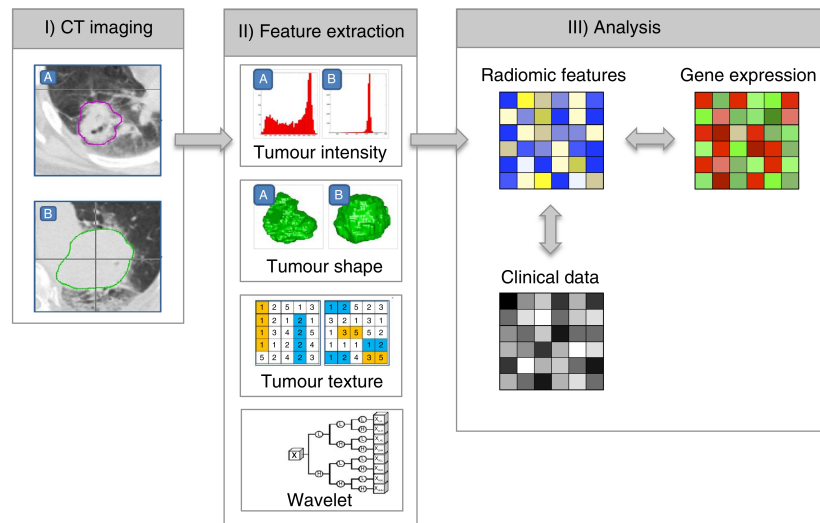


Figure 4.2 | Radiomic Feature Extraction Pipeline. Starting with step I), typically the manual segmentation of structures of interest, step II) and III) entail the extraction of handcrafted features and their analysis, for example by any type of classifier. Image borrowed from [Aerts et al., 2014].

of interpreting individual classifications of *CNNs*. The shortage of labelled data often prohibits the use of deep-learning approaches as they require a large number of labeled examples in order to learn discriminative features without over-fitting the data. For example see [Hénaff et al., 2019] who document the pronounced performance dependence of a strongly performing ResNet [He et al., 2016b] on the number of examples per class on ImageNet [Deng et al., 2009].

Radiomics One way to get by with less data, is to reduce the expressiveness of a classifier (e.g. by using decision trees or kernel-based methods such as support vector machines) and indeed train it on hand-crafted image-derived features as opposed to learning them. This intuitive and empirical solution is also backed by theory, e.g. following the argument in [Friedman et al., 2001] that models with increased expressiveness trade off their bias for increased variance (a model’s tendency to explain the training data by simpler versus more complex functions). Adding to that advantage, such simpler models, based on handcrafted features, may allow an increased degree of interpretability, in the sense that they often permit to determine the importance of individual features. This scheme and the class of simpler *ML*-models has been popular in the analysis of medical images and is referred to as *radiomics* in the medical imaging literature [Aerts et al., 2014, Gillies et al., 2015].

The reduced expressiveness of such models however comes at the obvious cost of a reduced ability to model highly complex relationships between images and class labels. Furthermore additional manual (or automated) steps as compared to end-to-end learning approaches may be required. This is because the typically highly local radiomic features do not readily allow to find the regions of interest (RoIs) that are discriminative (for an example see Chap. 5). Instead they are often calculated on RoIs that are pre-segmented by clinicians [Aerts et al., 2014, Bonekamp et al., 2018], see Fig. 4.2 and Fig. 5.1 for an illustration of a typical pipeline.

CNN Classifiers Given enough data of sufficient annotation quality, the classification of images is currently best handled by deep Convolutional Neural Network (CNN)s. The task of assigning a global label to an image largely assumes that it depicts a single object, most of the times centered and visible in full in the foreground. On natural images, the prime example of this sort of task is the ImageNet dataset [Deng et al., 2009] whose training set roughly holds 1000 unique classes with 1000 examples each and has been used excessively to benchmark and develop deep CNN architectures.

The general building blocks of CNNs have not changed much since their conception in the 1980s [Fukushima, 1980, LeCun et al., 1989, Waibel, 1989]: They use stacks of convolutional filters with shared weights followed by non-linear activation functions and down-sampling operations. Applied in a chain, these building blocks progressively reduce the spatial resolution of the produced activations, while using an increasing number of kernels, until typically an image global representation is reached [Krizhevsky et al., 2012, LeCun et al., 1998].

More recent advances have primarily focused on increasing the expressiveness of the models by means of two main levers: the number of layers (‘network depth’) [Simonyan and Zisserman, 2014, Szegedy et al., 2015] and the number of convolutional kernels per layer (‘network width’) [Wu et al., 2016b, Zagoruyko and Komodakis, 2016]. In order to reach even larger network depths without suffering from vanishing gradients in the early network layers, new connectivity patterns between the convolutional layers were conceived such as *residual blocks* [He et al., 2016a,b] and *dense blocks* [Huang et al., 2017]. Additionally different normalization schemes such as *batch-*, *instance-*, *layer-* and *group-norm* [Ioffe and Szegedy, 2015, Lei Ba et al., 2016, Ulyanov et al., 2016, Wu and He, 2018] have found wide-spread use, for the same and largely empirical reason. In the last few years this toolkit has brought about large improvements in classification accuracy on ImageNet and beyond. The task of finding improvements in architectures is

increasingly automated, e.g. by evolution strategies, and indeed has led to the latest **SotA** results on ImageNet [Real et al., 2018, Tan and Le, 2019].

Common across architectures is their feed-forward formulation that produces a categorical distribution over labels Y . This distribution is typically parameterized as a categorical softmax probability $P(Y|X)$ which is trained in fully supervised fashion, i.e. each training example input X has a unique and known (‘one-hot’) target Y . The training itself proceeds iteratively in that mini-batches $(X, Y) \sim P_{\text{data}}$, are sampled for which the classifier is set up to minimize a **Cross Entropy (CE)** loss \mathcal{L}_{CE} :

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}_{(X,Y) \sim P_{\text{data}}} [Y \log P(Y|X)], \quad (4.1)$$

where a mean-aggregation across batch instances and an additional mean aggregation across pixels is implied in the case when the output Y has spatial dimensions, such as in semantic segmentation.

CNN Classifiers in Medical Image Analyses Whole image classification is an important task in the analysis of medical images, as many clinical diagnoses could be reformulated as a global assessments of the presented image evidence. In the past two years large **CNNs** have been successfully trained to classify melanoma [Esteva et al., 2017], mammographic lesions [Kooi et al., 2017] and retinal diseases from optical coherence tomography OCT images [De Fauw et al., 2018]. Remarkably, a range of works report physician-level classification performance, e.g. in the task of identifying moles from melanomas, diabetic retinopathy, cardiovascular risk, referrals from fundus and OCT images of the eye, breast lesion detection in mammograms, and spinal analysis with magnetic resonance imaging [Esteva et al., 2019].

Pixel-level Semantic Segmentation The task of semantic segmentation is another instantiation of the canonical classification problem. In this task one aims for an image understanding at the finest resolution the image itself allows for, i.e. the pixel-level, which has been considered a difficult computer vision problem for a long time.

Before the era of deep learning was ushered in, many approaches attempted to model segmentation at either mostly local or mostly global resolutions. The line of local models built on handcrafted features, very similar to the ones discussed above, e.g. leveraging SIFT features [Lowe, 2004, Suga et al., 2008]. Other works tried to model the local correlations between pixels and neighboring segmentation labels using random fields over pixels [Boykov and Jolly, 2001, Rother et al., 2004] or random fields expressed over features aggregated over local image areas (‘superpixels’) [Fulkerson et al., 2009, He et al.,

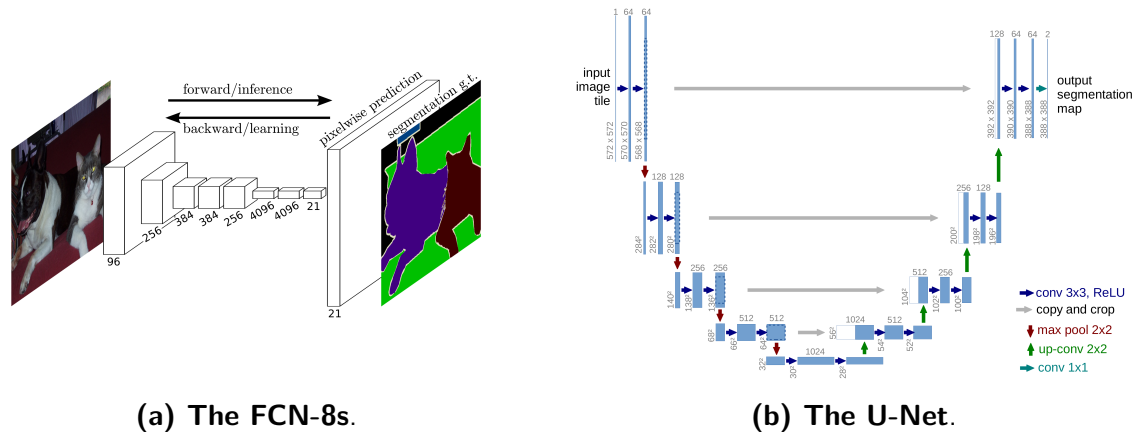


Figure 4.3 | Deep CNN Architectures for Semantic Segmentation. (a) shows a CNN for ImageNet classification, adapted for dense semantic segmentation [Long et al., 2015]. (b) shows the U-Net architecture with a symmetric encoder and decoder, linked by skip connections [Ronneberger et al., 2015].

2006]. The line of more global approaches on the other hand was largely template driven: by modeling the objects' outline, e.g. using statistical shape models [Cootes et al., 1995, Winn and Jojic, 2005] or by modeling the internal composition of objects, e.g. in terms of parts-based [Kumar et al., 2005] or fragment-based models [Borenstein and Ullman, 2008]. A more comprehensive discussion of these pre deep learning segmentation methods can be found in [Eslami, 2014].

As laid out above, CNNs instead promise to learn a hierarchy of non-linear features, encoding location and semantics in a local to global cascade. Because their original instantiations involved a contraction in resolution leading to global features, it was only gradually worked out how to successfully apply them to dense, i.e. full resolution, prediction tasks such as semantic segmentation. The simplest but also most wasteful approach conceivable is to literally, pixel by pixel, slide a CNN across an image and classify each pixel individually, which indeed was one of the early deep learning based proposals [Ciresan et al., 2012]. [Farabet et al., 2012] on the other hand, used several fully convolutional networks to produce dense segmentations for different image resolutions. The outputs were upsampled to matching resolutions and concatenated and then post-processed through random fields and superpixel-based classifiers, therefore constituting a model made up of components that were not jointly trained.

The first fully convolutional networks (FCNs) that were trained end-to-end were presented in [Hariharan et al., 2015, Long et al., 2015]. Here the first steps to fuse coarse semantic and fine appearance features were undertaken. [Long et al., 2015] proceeded by pretraining the best known classification networks of the time (AlexNet [Krizhevsky et al.,

2012], VGG-Net [Simonyan and Zisserman, 2014] and GoogLeNet [Szegedy et al., 2015]) on ImageNet and then replacing their final dense layers by 1×1 -convolutions thereby turning them into **fully convolutional networks** able to produce coarse, but nonetheless spatially resolved, output predictions (downsampled by a factor of 32 w.r.t the input), see Fig. 4.3a. In order to improve localization they added segmentation heads at two coarser locations (at stride 16 and 8) and used deconvolutions to learn an upsampling for each of those heads. The upsampled outputs of each head were combined by addition and then fine-tuned end-to-end on PascalVOC, yielding a new segmentation **SotA** by a large margin at the time. This re-use of parts of the network at a later stage is referred to as skip-connections.

Shortly after, the U-Net [Ronneberger et al., 2015] was introduced, which presented a principled way of combining the high-level semantic information of later layers with the fine-grained localization of earlier layers. The main idea is to not only add upsampled segmentation outputs of earlier layers, but to successively combine coarser features with more localized features in a *learned* fashion by means of convolutions. To this end the U-Net forms an *encoder-decoder* architecture by adding a decoder made up of convolutional and upsampling layers that mirrors its encoder (the network part that resembles a whole-image classification network) in reverse order, see Fig. 4.3b. Each processing scale of the decoder receives as input the concatenation of the upsampled decoder features from the resolution scale below with the encoder features of the same resolution. Allowing to successively recombine features of different scales by convolutions and skip-connections at all resolutions turns out to be a very powerful approach, leading to very successful semantic segmentation applications.

Another popular strand of work, referred to as DeepLab [Chen et al., 2017a], initially refrained from using a learned decoder. Instead so-called *à-trous* convolutions, which increase a convolution kernel’s field of view without increasing its number of parameters, are used to aggregate context across the image in the model’s encoder. In its first proposition DeepLab employed a conditional random field to post-process the produced segmentation, this was dropped in a second incarnation [Chen et al., 2017b] and finally a decoder with skip-connections and upsampling were introduced, called ‘DeepLabv3+’ [Chen et al., 2018], effectively assimilating the model to the U-Net. DeepLabv3+ holds the current **SotA** for semantic segmentation on Cityscapes and PascalVOC.

The encoder-decoder scheme proposed through the U-Net is the most widely and successfully used archetype for semantic segmentation models today. Especially in medical image segmentation it enjoys large popularity, with many successful adaptations in clinical applications [De Fauw et al., 2018, Nikolov et al., 2018] and **SotA** entries in segmentation

challenges like the Medical Segmentation Decathlon (10 distinct medical datasets), ACDC (heart MRI) and LiTS (liver CT) among many others, see [Isensee et al., 2019].

Object-detection & Instance Segmentation Semantic segmentation models predict a class label for every pixel in an image, but do not distinguish between instances of a class. In certain scenarios it might be of interest to distinguish instances, for example one might want to count objects or assign scores to individual objects, such as a malignancy prediction for detected lesions, see Fig. 4.1b. The task of joint localization and classification is referred to as *object detection*. In this task, it is not required to find a pixelwise segmentation of the instance and usually bounding box coordinates are regressed. The SotA model on this task is Mask R-CNN [He et al., 2017], an evolution of Faster R-CNN [Ren et al., 2015]. Both employ a U-Net like encoder-decoder as the ‘backbone’ from which they branch off additional classification and regression heads.

The task of additionally finding pixelwise masks of each object instance is referred to as *instance segmentation*, for which Mask R-CNN is also the SotA model by employing a separate mask head.

4.2 Predictions under Uncertainty and Noise

Deep classification and semantic segmentation models as described above are discriminative models. That means they are trained to find a complex, non-linear decision boundary between classes and behave deterministically, i.e. they produce a singular prediction for a given image. This prediction is made in the form of a softmax probability, which is often falsely interpreted as a model’s confidence in its prediction, when really it is a relative measure of how far the data point is away from the learned decision boundary. In fact a model can be uncertain in its prediction even when predicting with high softmax probability [Blundell et al., 2015, Gal and Ghahramani, 2016].

Uncertainty can originate from various sources, among them the uncertainty from an ambiguous mapping of $X \rightarrow Y$ and such that stems from not having found the right model parameters due to limited access to the data distribution (not enough labelled data). Given the high stakes nature of clinical applications, correctly handling ambiguity and uncertainty can be of particular importance in medical image analyses. Here, two important desiderata are sought after: 1) negative effects of diverse labels on model performance should be reduced where possible, and 2) uncertainty should be meaningful, i.e. calibrated as well as interpretable. Both goals are subject to ongoing research, which is traced out below.

Image Ambiguity and Label Noise Medical images in particular often only present ambiguous image evidence for target measures of interest, like class labels or pixelwise semantic segmentations. This is because they only indirectly measure the molecular identity of the tissue within each voxel. This similarly holds true on natural images, where the image evidence may be compromised due to limited resolution, measurement noise and occlusions. Sometimes even the label space itself is ambiguous as a consequence of an arbitrary quantification of a continuous space, e.g. think of *cat* vs. *kitty*, a concept called ‘implicit class confusion’ [Lee et al., 2016]. In all cases, this ambiguity leads to sets of plausible interpretations $Y \in (y_0, \dots, y_n)$ for a given image X , which is well documented on clinical tasks (see Sec. 3.2) and also on annotations for natural images [Gurari et al., 2018]. Because of the aforementioned lack of image evidence, the uncertainty with respect to the labels is irreducible, even in the limit of infinite amounts of labelled data. This type of uncertainty can be referred to as *aleatoric* uncertainty [Kendall and Gal, 2017].

Mitigating Label Noise The fact that there can be an irreducible amount of label uncertainty for a given image does not mean that there is no unique ground-truth, only that it cannot be pinned down beyond a certain degree from the image alone (a biopsy however could for example resolve the ambiguity). Arguing that there can therefore be labels that are preferable over others, with some labels even hurting a classifier’s performance, there is a line of work seeking to mitigate the negative effects of label ambiguity, which is *label noise* in this view.

In semantic segmentation, the perhaps simplest approach to reduce label noise is to mask ambiguous image regions in the loss calculation. Such masks are often available on datasets such as Cityscapes [Cordts et al., 2016], which however is neither principled nor does it generalize well. For this reason a line of work is concerned with finding loss-functions that are inherently more robust to label noise than the standard CE-loss (Eq. 4.1). [Ghosh et al., 2017] for example propose to use a mean squared error (L2) loss for multi-class classification under label noise and [Zhang and Sabuncu, 2018] show an increased classification performance when employing a parametric relaxation between a mean absolute error (L1) loss and a CE-loss under label noise.

The latter proposal hinges on the observation that the CE-loss gradient is inversely proportional to the classifier’s softmax probability for the ground truth label. While this is desirable on ‘clean’ labels or unambiguous classification tasks as a form of implicit hard-negative mining, it might hurt on ambiguous ones. This is because putting a lot more emphasis on difficult, therefore potentially ambiguous cases, may cause the classifier

to overfit to ambiguous labels or to perform poorly as a result of large, confusing loss gradients.

An extension towards a learned reduction of the impact from aleatoric label noise is presented in [Kendall and Gal, 2017, Lakshminarayanan et al., 2017]. Here, a deep segmentation network is equipped with an additional regression head that enables the network to attenuate the loss function by down-weighting difficult pixels or examples. This attenuation is learned indirectly, i.e. without uncertainty targets, and may be interpreted as a pixelwise aleatoric uncertainty. Other works seek to ‘clean-up’ the labels by excluding such training instances that are below some softmax probability threshold [Northcutt et al., 2017] or deemed noisy by a separate network that is trained on known clean labels [Veit et al., 2017]. Alternatively training and finetuning on self-generated labels has been proposed [Tanaka et al., 2018] or learning to construct a training curriculum that prioritizes data points with lower estimated noise [Jiang et al., 2017].

Capturing Label Distributions Another line of work seeks to model the distributions of labels or moments of these distributions, rather than mitigating noisy training signals.

Making use of the *Expectation Maximization* (EM) algorithm, [Khetan et al., 2017] estimate the label posterior distribution by explicitly modeling the annotation quality of individual annotators and [Vahdat, 2017] infer a graphical model with clean labels, noisy labels and images as nodes that allows to produce a probability distribution over clean labels. Similar approaches exist for semantic segmentation, where the EM algorithm is used to estimate the true underlying segmentation from a set of segmentations by modeling the quality of individual annotators [Warfield et al., 2004], which however relies on the availability of multiple annotations and only indirectly depends on the underlying image itself.

Other works modify the architecture of deep nets, e.g. by adding an extra layer that adapts the network outputs to match the (image global) label distribution by modelling the confusion matrix between prospective true and observed labels [Goldberger and Ben-Reuven, 2016, Sukhbaatar et al., 2014]. Relatedly, [Guan et al., 2018] model individual annotators on retinopathy images by branching off classification heads at the end of a core CNN, each of which is solely trained to match the diagnoses given by their corresponding expert. In a second step they learn weights for each of the heads by matching the ground truth label distribution and show that a respective weighted average across predictions performs better than a naive arithmetic mean of the predictions (compared against the arithmetic mean of the annotators). Again having access to multiple independent diagnoses per retinopathy image, [Raghu et al., 2019] propose to directly predict the rater

variance per image and show an improved performance in predicting rater disagreement as compared to derivative uncertainty measures such as the entropy of the softmax probabilities.

Modelling diverse outputs for semantic segmentation has also been considered in the literature. For example by using an ensemble of models [De Fauw et al., 2018, Lee et al., 2015, 2016] or by using multiple distinct output layers in combination with a stochastic loss that encourages diverse specialization of the heads [Ilg et al., 2018, Rupprecht et al., 2017]. Finally a *cVAE* for semantic segmentation was proposed in [Sohn et al., 2015a], which was applied to small scale natural images and predates powerful encoder-decoder segmentation architectures. More details on this line of work can also be found in [Sec. 7.2](#).

Capturing Model Distributions Aside from uncertainty due to irreducible ambiguity in particular observations, a model can be uncertain due to a lack of observed data. This type of uncertainty is referred to as *epistemic* uncertainty and can be reduced by providing data for unseen regions of the ‘data space’. In machine learning, epistemic uncertainty unfolds in the uncertainty about which model is appropriate given all of the seen data. As neural networks are parameterized by learned weights \mathbf{w} , this can be viewed as the uncertainty expressed by a distribution over the network weights $Q(\mathbf{w}|\mathcal{D})$, having observed labelled data \mathcal{D} . Intuitively, as the number of examples in \mathcal{D} increases, $Q(\mathbf{w}|\mathcal{D})$ should become sharper.

The exact Bayesian inference of $Q(\mathbf{w}|\mathcal{D})$ is intractable for deep neural networks and so a range of approximations have been proposed. Several works use variational approximation techniques, explicitly parameterizing weight distributions [Blundell et al., 2015, Graves, 2011, Hinton and Van Camp, 1993]. Other approaches, popular for their ease of implementation, re-interpret a regularization scheme that stochastically sets individual network weights to zero (known as ‘dropout’ [Srivastava et al., 2014]) as a variational Bernoulli distribution over the weights [Gal and Ghahramani, 2016, Gal et al., 2017a, Kendall et al., 2015] and apply dropout at test time, called *MC-dropout*. Interestingly, the *MC-dropout* technique may be viewed as subsampling smaller networks and ensembling them. And indeed [Lakshminarayanan et al., 2017] showed that training actual model ensembles can capture uncertainty and yield comparable (or improved) performance to *MC-dropout* in terms of uncertainty calibration.

Distinguishing Aleatoric and Epistemic Uncertainty While the distinction between *aleatoric* and *epistemic* uncertainty is well grounded from a theoretical perspective, it can be a difficult one to make on real world, high dimensional data. [Kendall and Gal, 2017]

for example report that the separate machinery they put in place to capture the two, MC-dropout over the weights and a learned loss attenuation (see above), can in practice also cover the respective other uncertainty type. [Smith and Gal, 2018] propose different uncertainty measures to distinguish aleatoric from epistemic uncertainty and given MC-dropout samples from a CNN classifier show that the distinction may be possible on a simple task such as classifying digits on MNIST, but even here the separation appears far from definite.

Informally, there appears to be a more fundamental problem: In real world vision applications, such as classifying a small lesion on a CT scan, it cannot be known whether the presented image evidence is truly ambiguous (giving rise to aleatoric uncertainty) or whether our model has simply not learned the discriminative image cues yet, due to not having observed enough data (causing epistemic uncertainty). This problem intensifies on high-dimensional data such as CT or MRI scans. While there may be good reasons to nonetheless strive to distinguish the two, e.g. for the detection of adversarial examples [Smith and Gal, 2018] or for active learning [Siddhant and Lipton, 2018], the distinction may be less relevant when diagnosing or segmenting medical images, where a combined uncertainty assessment, the *predictive uncertainty*, can be sufficient for practical intents.

4.3 Generative Models for Images

As described in Sec. 4.1, whole-image classification and semantic segmentation are usually approached with *discriminative models*. Although producing a categorical softmax distribution $P(Y|X)$, they are deterministic models trained to discriminate labels Y based on learned boundaries and given the observed image evidence X . While conditional *generative models* formally also model $P(Y|X)$, they induce a complex distribution over the possible values of Y , having observed X . The distinction is well apparent in applications such as dense prediction, i.e. pixelwise outputs, as can be observed in the later Chap. 7 - Chap. 8. Here, discriminative models produce a pixelwise predictive distribution that is not meant to be sampled from, as it would result in independent and thus incoherent samples, whereas generative models attempt to produce coherent output samples.

This chapter provides the background for different types of *generative models* that are used or referred to further down the line in the context of aiming at modeling complex interdependencies for $Y|X$.

Variational Autoencoders Variational Auto-Encoders (VAEs) aim to explicitly model the likelihood of data-points, e.g. images X or labels given an image $Y|X$. They do so by assuming that there is an intermediate, called *latent*, variable that maps to X and is typically assumed much lower dimensional than X . From this point of view, finding the likelihood of X becomes the marginalisation of the latent variable:

$$P(X) = \int_{\mathbf{z}} P(X|\mathbf{z})P(\mathbf{z})d\mathbf{z}. \quad (4.2)$$

For convenience $P(\mathbf{z})$ is usually chosen to be a spherical Gaussian and $P(X|\mathbf{z})$ is parametrized as a neural network, e.g. a CNN. In order to maximize $P(X)$ for observed X during training, we would like to adjust this net's weights accordingly. Computing $P(X)$ turns out to be difficult though, as the marginalization of $P(X|\mathbf{z})$ is analytically intractable and approximations like MC-integration are inefficient as in practice only small volumes of \mathbf{z} are expected to contribute to particular mappings to X .

The solution lies in exploiting this intuition and assuming that it should be possible to infer \mathbf{z} that are likely to have produced X [Doersch, 2016]. For this purpose a distribution $Q(\mathbf{z}|X)$, parameterized as a neural net, is introduced. During training this distribution shall become as close as possible to the true (but unknown) posterior $P(\mathbf{z}|X)$, which can be measured by means of a Kullback-Leibler divergence (KL) divergence, D_{KL} . Using Bayes theorem, this gives:

$$D_{\text{KL}}(Q(\mathbf{z}|X)||P(\mathbf{z}|X)) := \mathbb{E}_{z \sim Q}[\log Q(\mathbf{z}|X) - \log P(\mathbf{z}|X)] \quad (4.3)$$

$$= \mathbb{E}_{z \sim Q}[\log Q(\mathbf{z}|X) - \log P(X|\mathbf{z}) - \log P(\mathbf{z}) + \log P(X)], \quad (4.4)$$

regrouping and using $D_{\text{KL}} \geq 0$ then yields:

$$\log P(X) = \mathbb{E}_{z \sim Q}[\log P(X|\mathbf{z})] - D_{\text{KL}}(Q(\mathbf{z}|X)||P(\mathbf{z})) + D_{\text{KL}}(Q(\mathbf{z}|X)||P(\mathbf{z}|X)) \quad (4.5)$$

$$\geq \mathbb{E}_{z \sim Q}[\log P(X|\mathbf{z})] - D_{\text{KL}}(Q(\mathbf{z}|X)||P(\mathbf{z})) := -\mathcal{L}_{\text{ELBO}}. \quad (4.6)$$

The Evidence Lower Bound (ELBO) objective $\mathcal{L}_{\text{ELBO}}$ is a tight lower bound assuming that our distribution $Q(\mathbf{z}|X)$ is sufficiently close to the true posterior. From this it becomes apparent, that indeed the log-likelihood of a data point (Eq. 4.5) can be approximated (lower-bounded) by means of samples from the posterior $Q(\mathbf{z}|X)$, plus the evaluation of a divergence term that is commonly analytically tractable. In practice often only a single posterior sample suffices for training.

For images, a contracting CNN is set up, that regresses the parameters for $Q(\mathbf{z}|X)$ (commonly a Gaussian), which is referred to as the *encoder*, *inference* or *posterior network*. $P(X|\mathbf{z})$ on the other hand is setup as a CNN that gradually increases its resolution, having as input a sample \mathbf{z} and outputting an image resolution distribution $P(X|\mathbf{z})$, referred to as the *decoder*. The distribution in the output space is a design choice, for natural images typically a pixel-wise Gaussian distribution, which turns $-\mathbb{E}_{\mathbf{z}\sim Q}[\log P(X|\mathbf{z})]$ into a pixelwise $L2$ -loss.

The derivation of the training objective for **conditional Variational Auto-Encoders** (cVAEs) in order to model $P(Y|X)$ follows the same lines as Eq. 4.3 - Eq. 4.6 and is detailed in [Doersch, 2016].

Autoregressive Models & Flows Two other classes of explicit likelihood models are popular as generative models for images. *Autoregressive models* for one, use the chain rule of probabilities to factorize the image likelihood in terms of pixelwise conditional probabilities [Oord et al., 2016, Van den Oord et al., 2016]. These models are typically trained to maximize the pixelwise likelihoods directly in rgb-space and while they are very slow to sample from, yield state of the art likelihoods. Their applicability to large scale image generation was not shown until very recently [De Fauw et al., 2019, Razavi et al., 2019] with interestingly both works generating images in a hierarchical fashion.

Flows [Dinh et al., 2016, Kingma and Dhariwal, 2018], like VAEs, are latent variable models. They allow for an exact likelihood maximization, by careful construction of the neural network that maps $f(X) = \mathbf{z}$, such that f is bijective. This allows maximizing the likelihood $P(X)$ by maximizing the likelihood of the latents $P(\mathbf{z})$ under a simple distribution, such as a spherical Gaussian, using the change of variables theorem. The bijectivity requirement however dictates the total number of dimensions after each layer to remain the same, which means that overall, \mathbf{z} must have the same dimensionality as X and the qualifying functions are reduced in expressiveness.

To the best of our knowledge, neither autoregressive models nor Flows have been explored in the context of image-conditional generative modeling ($P(Y|X)$) to date.

Generative Adversarial Networks (GANs) allow the training of (deep) generative models without maximizing an explicit likelihood of the data. The GAN-framework involves two networks that are trained simultaneously, a *generator* filling the role of the generative model and a *discriminator*, which is required only during training [Goodfellow et al., 2014]. The generator G receives as input latent variables \mathbf{z} and maps them to an output X , e.g. an image, just like the *decoder* in the VAE-framework. Unlike the VAE

case, G is however not trained by reconstructing the image X , that an *inference* network has encoded in the form of \mathbf{z} . Instead it learns the mapping $G(\mathbf{z}) = X$ simply by trying to receive a high classification score for the generated X from the discriminator D . Since D is parameterized as a neural net, it is possible to backpropagate gradients through it, and training G can thus be formulated as minimizing:

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z} \sim P_{\mathbf{z}}} [\log D(G(\mathbf{z}))], \quad (4.7)$$

see [Goodfellow et al., 2014], where the latent \mathbf{z} is sampled from a fixed distribution $P_{\mathbf{z}}$, such as typically a spherical Gaussian. The discriminator is set up so as to answer the question whether the input it receives may plausibly be part of the real data distribution P_{data} . This means D is trained to tell apart real data instances $X \sim P_{\text{data}}$ (target label 1) from those generated by G (target label 0), by means of a simple binary classification with loss:

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{z} \sim P_{\mathbf{z}}} [\log (1 - D(G(\mathbf{z})))] - \mathbb{E}_{X \sim P_{\text{data}}} [\log D(X)], \quad (4.8)$$

The two networks play an *adversarial* mini-max game against one another, in which both gradually get better, G at counterfeiting fakes and D at telling them apart from real ones. This game is notoriously brittle, in that sudden and irrecoverable degradations of performance can occur, and much research has dedicated effort towards understanding and mitigating these issues [Arjovsky et al., 2017, Brock et al., 2018, Lucic et al., 2017, Miyato et al., 2018].

Making GANs image-conditional (cGANs), i.e. again modelling $P(Y|X)$, for the purpose of image-to-image translation tasks of all sorts, has equally been difficult to achieve. Many authors observed the latents \mathbf{z} being largely ignored in favor of the image to condition on [Isola et al., 2017, Luc et al., 2016]. Ways of constraining the architectures so as to prevent this from happening have been found. For example [Zhu et al., 2017a] use a chain of GANs to go back and forth between $X \leftrightarrow Y$, essentially adding in a reconstruction constraint. [Zhu et al., 2017b] similarly use a reconstruction constraint and have combined cGANs with VAEs in order to infer the semantics of the employed latent space.

Chapter 5

Finding Discriminative MRI Features

Invasive diagnostic steps such as biopsies, have the ability to improve the grading of prostate lesions as they can disambiguate the findings from non-invasive techniques, see the more detailed discussion [Chap. 3](#). However, because of the implied patient discomfort and health risks, there is a quest to forgo surgical biopsy and instead rely on image evidence, where admissible.

In current clinical practice, [Multi-parametric MRI \(mpMRI\)](#) images are qualitatively interpreted by radiologists using rule-based grading systems ([PIRADS](#)). Due to the subjective nature of the assessment, differences in rater experience as well as ambiguous image evidence, these diagnoses can be sub-optimal and are known to suffer from high inter-rater variability, see [Sec. 3.3](#).

In this chapter we explore whether it is possible to find simple [mpMRI](#)-derived features, that allow a quantitative and reproducible assessment of the image evidence while matching or surpassing the radiologists' diagnostic performance. We make the following contributions:

- We develop a simple machine learning based approach to classify the clinical significance of prostate lesions based on [mpMRI](#)-derived features using a dataset and biopsy reference standard that was collected in clinical practice and thus reflects realistic conditions.
- We assess the method on a held-out dataset having fixed its working point so as to reflect the radiologist's sensitivity on the training set and observe an increased performance compared to the radiologist.
- We rank the importance of the employed features and substantiate the discriminative power of a specific [MRI](#)-derived feature largely refuting the utility of additional modalities and features as found in the literature.

- Finally, we analyze the utility of distinguishing between anatomical zones of the prostate, which are handled differently in clinical guidelines. To this end we assess the performance of separately trained models and find a combined model to perform superior.

With kind permission by the Radiological Society of North America (RSNA), this chapter reproduces many parts of the following publication:

David Bonekamp*, Simon Kohl*, Manuel Wiesenfarth, Patrick Schelb, Jan Philipp Radtke, Michael Götz, Philipp Kickingereeder, Kaneschka Yaqubi, Bertram Hitthaler, Nils Gählert, Tristan Anselm Kuder, Fenja Deister, Martin Freitag, Markus Hohenfellner, Boris A Hadaschik, Heinz-Peter Schlemmer, Klaus H Maier-Hein. “Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC values.” *Radiology* 289, no. 1 (2018): 128-137,

where * indicates equal contributions by David Bonekamp (radiologist and senior physician at the German Cancer Research Center) and the author of this thesis, who devised the machine learning models, carried out their training, evaluation and analysis and co-wrote the manuscript. In the following the publication is cited as [Bonekamp et al., 2018].

5.1 Problem Statement

Interpretation of mpMRI according to the [The Prostate Imaging Reporting and Data System \(PIRADS\)](#) and its recent update to version 2.0 [Barentsz et al., 2012, Vargas et al., 2016, Weinreb et al., 2016], in combination with [Trans-rectal Ultra-Sound Guided Biopsy \(TRUS-biopsy\)](#), has shown promise in detecting clinically significant cancer, with sensitivities reaching 97% [Moldovan et al., 2017, Radtke et al., 2016].

However, identifying patients with high sensitivity also comes at the cost of possible overdiagnosis, and further improvement in the differentiation of non-aggressive and aggressive prostate cancer is necessary [Cooperberg and Carroll, 2015, Donati et al., 2013, 2014, Woo et al., 2017]. While PIRADS scoring is based on the qualitative assessment of DWI and T2w images, see [Sec. 3.2](#), the ability of quantitative [Apparent Diffusion Coefficient \(ADC\)](#) measurements to improve inter-reader concordance has recently been pointed out [Hansen et al., 2017, Pierre et al., 2018].

DWI, and specifically the ADC, can be considered the current best mono-parametric component of prostate MRI assessment, resulting from its ability to probe the micro-

environment of neoplastic tissues efficiently and detect alterations in compartmental volumes and cellularity [Chatterjee et al., 2015], see Sec. 2.1 for technical background.

The potential of *radiomics* to improve diagnostic accuracy, by extracting a large number of hand-crafted, quantitative features from these radiologic images, has recently received significant attention in clinical literature [Fehr et al., 2015, Vignati et al., 2015, Wang et al., 2017, Wibmer et al., 2015].

Initial studies have reported promising performance of radiomics, with and without the use of machine learning, in the prediction of the prostate cancer **Gleason Score (GS)** [Vignati et al., 2015, Wang et al., 2017, Wibmer et al., 2015]. Accuracy however varies depending on the machine learning approach used [Wibmer et al., 2015]. Additionally a magnitude of features has been reported as discriminative, e.g. various textural features have been associated with prostate cancer aggressiveness and the pathologic index lesion: Homogeneity gray-level co-occurrence matrix texture features from **T2w** images and **ADC** maps have been suggested to be superior to first order **ADC** statistics such as **mADC** [Vignati et al., 2015], and radiomics has been reported advantageous compared with and in combination with retrospective **PIRADS** assessment [Wang et al., 2017].

The purpose of our study was to further examine multi-parametric quantitative models, with the use of an independent, comparatively large test set and with direct comparison to established mono-parameters (**mADC**) and clinical assessment (**PIRADS**). For a definition of the **PIRADS** system please refer to Sec. 3.2 and the **mADC** denotes the mean of the Apparent Diffusion Coefficient (see Sec. 2.1) calculated across the pixels or voxels of a given segmentation mask.

The characterization of **MRI**-detected lesions was formulated as a binary classification between clinically significant (aggressive) lesions and such with lower significance (non-aggressive lesions). The binary targets reflect histopathological findings, such that lesions with a Gleason Grade Group of 2 or larger (equivalent to $GS \geq 7a$, see Sec. 3.2 and specifically Table 3.1), were regarded as clinically significant ground truth [Carter et al., 2012, Loeb et al., 2016].

We compared radiomics predictions and the **mean Apparent Diffusion Coefficient (mADC)** (calculated across the respective radiologist provided segmentation masks) under this prostate lesion classification task. For comparison on the independent test set, working points corresponding to the sensitivity threshold of a radiologist’s clinical reporting (employing a threshold on the given **PIRADS** scoring) were considered.

5.2 Prostate MRI Dataset

This retrospective analysis was performed in a cohort of men undergoing MRI and Transrectal Ultra-Sound Guided Biopsy (TRUS-biopsy), that was collected during routine clinical practice at the German Cancer Research Center and the University of Heidelberg Medical Center. This dataset is not currently publicly available and therefore requires specification, which is provided in the following.

Cohort Inclusion Criteria The institutional and governmental ethics committee approved the study and waived informed consent. All patients had a clinical indication leading to prostate biopsy (details on the biopsy protocol are given in Sec. A.1) which was based on Prostate-specific Antigen (PSA) elevation, suspicious DRE results and MRI examination or participation in the University of Heidelberg Medical Center’s active surveillance program. MRI data of 316 consecutive patients (median age = 64 years; interquartile range (IQR) = 58–71 years) examined with a single 3 T MRI system in 2015-2016 were included in the analysis. 183 patients examined from May 2015 until January 2016 were included in the training cohort for training and validation (median age = 64.5 years; IQR = 59–71 years), while 133 patients, examined between January 2016 and September 2016, comprise the independent test cohort (median age = 63 years; IQR = 58–71 years). Inclusion criteria were (a) imaging performed on our main institutional 3 T MRI system and (b) extended systematic and targeted TRUS-biopsy performed after MRI, see Fig. 3.2c). Exclusion criteria were (a) history of treatment for prostate cancer (antihormonal therapy, radiation therapy, focal therapy, prostatectomy); (b) biopsy within the past 6 months prior to the MRI examination; and (c) incomplete sequences or severe artifacts on MRI images. More details on the inclusion and exclusion criteria can be found in Fig. A.1 and Fig. A.2 gives specifics on the demographics and the distributions of Gleason and PIRADS scores in the cohorts.

MRI Acquisition and Interpretation MR images at 3 T were acquired prior to biopsy according to the European Society of Urogenital Radiology, or ESUR, guidelines (Magnetom Prisma, Siemens Healthcare, Erlangen, Germany, see Fig. 3.2c)). T2-weighted and DWI MR images were acquired according to the institutional prostate MRI protocol (with b-values images at $b = 50, 500, 1000$ and 1500 s mm^{-2}). Interpretation of multiparametric MRI images was performed by board-certified radiologists during clinical routine; eight radiologists, seven of them with at least 3 years of experience in prostate MR image interpretation, read 311 (98%) studies, and one younger colleague who joined the team after completing a departmental training period interpreted five (2%) studies. All exami-

nations were reviewed again in an interdisciplinary conference prior to biopsy for quality assurance and all radiologists participated in regular retrospective review of MRI reports and biopsy results. Clinical reports included PIRADS assessment for each detected lesion and a pictogram indicating lesion location. For subsequent quantitative analysis, ADC images, DWI images with b value of 1500 s mm^{-2} (in the following referred to as B1500), and T2w images were extracted and upsampled to 0.25 mm in-plane resolution and 3 mm section thickness by using the medical imaging toolkit¹ (MITK [Nolden et al., 2013]).

5.3 Radiomics Pipeline

Fig. 5.1 sketches the employed radiomics pipeline: MRI acquisition was followed by the radiologist-performed lesion segmentation, an image normalization and the extraction and selection of radiomics features which finally fed into the training and evaluation of ML classification models. The individual steps are described in more detail below.

MRI Lesion Segmentation 3D volumes of interest (VOIs) of clinical lesions were segmented by one investigator (with 6 months of experience in prostate MRI), using the afore-generated clinical reports (MRI images and location pictograms) in consensus with and under supervision of a board certified radiologist with 8 years of experience in prostate MRI using MITK [Nolden et al., 2013], and performed separately on T2w images and ADC images. Due to the natural co-registration of ADC maps to the source b-value images, the ADC segmentations carry over to B1500 image on which the segmentation process therefore does not need to be repeated. VOIs were drawn on consecutive axial sections by using a polygon tool, encompassing the whole lesion while trying to avoid areas of partial volume effects at the border and in regions of diffuse tumor infiltration. A total of 462 lesions were segmented. Aside from the lesions, the whole gland as well as the PZ were segmented. In addition, with the aim of normalizing the non-quantitative modalities, normal appearing Peripheral Zone (PZ) was segmented, excluding any lesion and minimizing diffuse signal changes while encompassing at least 50 voxels on at least three adjacent sections. Fig. 5.2 depicts example segmentations of PZ, prostate boundary, and a PIRADS 5 lesion in the anterior Transitional Zone (TZ) in a representative patient with volume renderings.

Image Normalization T2w and B1500 images were normalized by dividing voxel intensities with the mean value of background PZ tissue, which was delineated as described

¹www.mitk.org

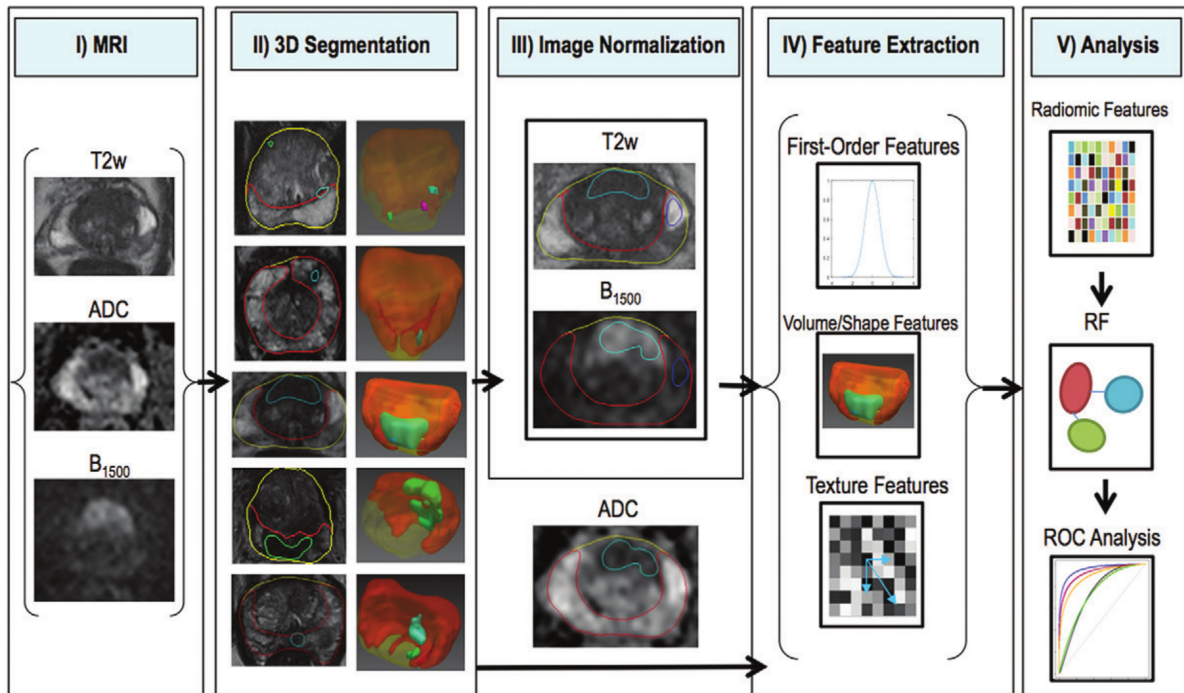


Figure 5.1 | Radiomics Workflow. From left to right: I) T2w and DWI (ADC and B₁₅₀₀) are extracted from the multiparametric MRI protocol; II) 3D segmentations of lesions, peripheral zone, and prostate are drawn by radiologists, which is shown in five representative patients overlaid on representative axial T2w images and using volume rendering; III) regions of normal-appearing peripheral zone (dark blue) is used to normalize T2w and B₁₅₀₀ images; IV) radiomic features are extracted, including first-order, volume, shape, and texture features; V) the radiomic features are combined with clinical information and entered into machine learning analysis (Random Forest (RF)). Performance is assessed by using receiver operating characteristics (ROC) analysis.

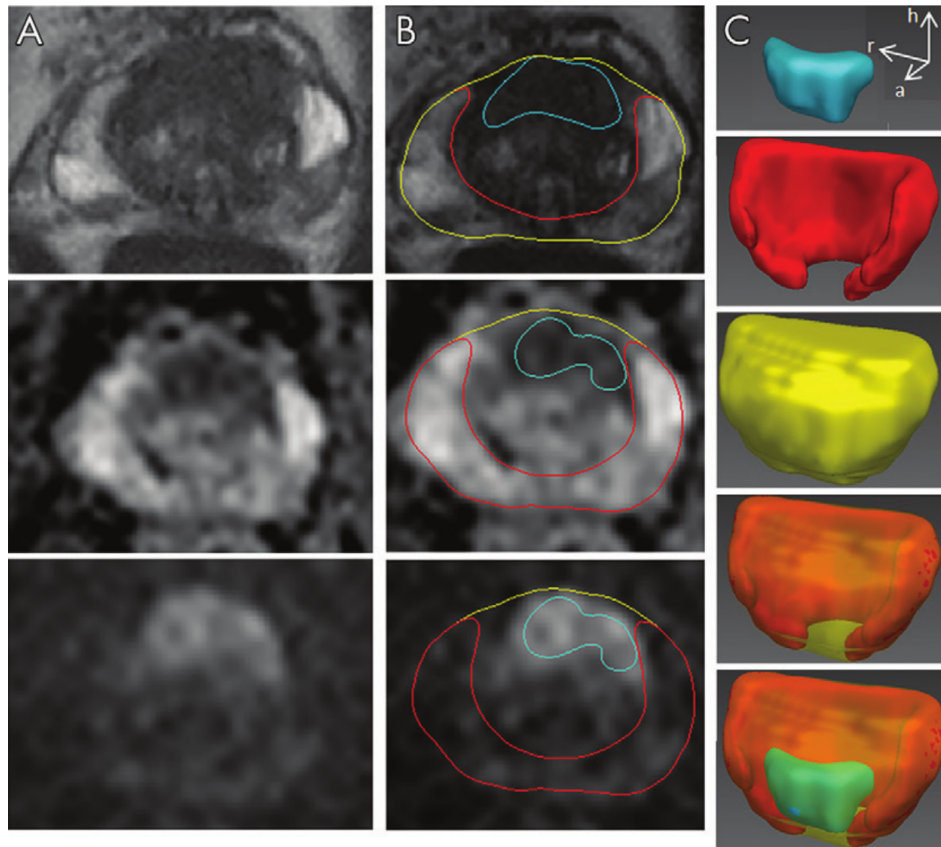


Figure 5.2 | Radiologist-Performed Example Segmentations. Segmentations in prostate MRI taken for a 56-year-old patient with initial PSA level of 7 ng ml^{-1} . Highly suspicious lesion in the anterior transition zone and anterior stroma (PIRADS category 5). A) Top: T2-weighted image demonstrates erased charcoal sign and ill-defined margins; Middle: ADC image shows moderate diffusion restriction; Bottom: B1500 shows moderate diffusion restriction. B) Segmentations of prostate (yellow), peripheral zone (red), and suspicious lesion (cyan) overlaid on corresponding images. C) Three-dimensional renderings from top to bottom of lesion, peripheral zone, prostate, prostate combined with peripheral zone, and all volumes of interest combined. Targeted biopsy revealed prostate cancer with a Gleason Grade Group of 2 in 95% of cores. h = head, a = anterior, r = right.

above. **ADC**, being a quantitative measurement, was not normalized. The normal appearing tissue was segmented on **ADC** and the resulting segmentation map then employed in the normalization of the **B1500** image.

Extraction and Selection of Features Radiomic feature calculations were performed by using the pyradiomics package² [Van Griethuysen et al., 2017], resulting in a vector of features for each pre-segmented lesion. Within each **VOI**, (a) 19 first-order features, (b) 16 volume and shape features, and (c) 59 texture features were calculated, leading to 94 features per **VOI**. Because these features were calculated separately on the available **ADC** maps, the **T2w** and **B1500** images, a total of 282 radiomics features were available for each lesion.

First order features depend on image intensities and comprise statistics such as the intensity mean (therefore also including the **mADC**), minimum or variance of the image within the lesion **VOI**. The shape features rely entirely on the binary **VOI** masks and aim at quantifying their shapes by means of measures such as the respective diameter, sphericity or surface-to-volume ratio. The texture features are based on co-occurrence, run-length and size-zone based features which are calculated from the image **VOI** in question from discretized image intensities. Co-occurrence features assess spatial relationships between adjacent voxels while run-length and size-zone features quantify the extent to which pixels of a given gray-value appear in succession.

The number of radiomic features was reduced by univariate feature selection with the remaining features serving as input into **Random Forests** (RFs), an approach well-established in radiomics [Parmar et al., 2015]. We employed the Wilcoxon rank-sum test [Wilcoxon, 1992] for feature selection and reduced the set of 282 features per **VOI** to 150 features, a choice that was selected in cross-validation. We integrated the feature selection into the classifier bagging procedure [Breiman, 1996], thus the feature selection was performed for each training data sub-split (fold), see below.

Model Training and Hyper-parameter Optimization We employed **Random Forests** (RFs) [Breiman, 2001] to perform a binary classification for the aggressiveness of prostate lesions and used the **RF** implementation available in scikit-learn³ [Pedregosa et al., 2011]. In a nutshell, random forests learn to classify by finding successive cuts on features that optimally separate the classes. Each succession of cuts makes a tree and a number of trees executed in parallel forms a forest.

²<https://github.com/Radiomics/pyradiomics>

³<http://scikit-learn.org/>

Random forests require comparatively little parameter tuning, e.g. allowing a limited-scope hyper-parameter search over their number of trees, maximum depth, minimum number of samples for a node-split and number of features (ranked according to the Wilcoxon rank-sum test). This search was integrated into a bagging procedure, i.e. a subject-stratified, repeated nested cross-validation, with the aim of preventing over-fitting and reducing classifier variance [Cawley and Talbot, 2010].

As mentioned before, we grouped 243 lesions (183 patients) in a training cohort and left the subsequent 219 lesions (133 patients) as an independent test cohort. The training cohort was employed to pin down the RF hyper-parameters and the number of selected features. For each bootstrap on the training cohort there exist held-out test folds, which, across fold rotations and bootstraps, enabled to assess the ensemble’s performance on the entire training cohort, see below. The final ensemble was additionally validated on the independent test cohort of 133 patients.

During cross-validation on the training cohort, in each of 100 bootstraps a random ten-fold split of the subjects was sampled, with one fold withheld for later testing. Training proceeded on six of ten folds and validated the hyper-parameter search on three folds. Six folds picked from the nine available folds were permuted in a nested inner loop. Class balances in the training split were addressed by re-weighting samples by the inverse class frequencies during determination of node cuts.

The best hyper-parameters were found to be: 500 trees per RF, each grown using a maximum depth of 7 splits (corresponding to 7 features), a minimum number of four samples to split a node and a reduction to 150 best features according to the feature-selection. Employing these hyper-parameters, we retrained a bagged RF ensemble, this time using 9 of the 10 folds across 100 bootstraps. This final RF ensemble thus comprises 1000 forests. Because the random bootstrap splits were seeded to be equivalent to the ones during cross-validation, we can aggregate the performance of the ensemble members on their respective held-out test sub splits (which they were neither trained nor validated on during cross-validation). Additionally, we evaluated the entire RF ensemble on the 133 patient large held-out test cohort.

Note that the bagged ensemble provides an additional source of bagging on top of both feature and sample bagging inherent to random forests [Breiman, 1996].

5.4 Evaluation and Results

As mentioned above, the binary target was set to distinguish between aggressive and non-aggressive lesions according to the Gleason Grade Group (defined as 2 or higher).

We analyzed the bagged RF ensemble (also referred to as Radiomic Machine Learning (RML)) as well as the mADC, in terms of their area under the ROC curve (ROC-AUC). The ROC curve is created by plotting the true positive rate (TPR, equivalent to the sensitivity) against the false positive rate (FPR, equivalent to $1 - \text{specificity}$) at various classifier thresholds. The AUC thereof serves as an aggregate measure of performance across all possible thresholds.

ROC curves were generated for mADC and RML and compared using the Delong test [DeLong et al., 1988] in both the training and test cohort, see Fig. 5.3. Thresholds for mADC and RML were selected to match the sensitivity of clinical assessment in the training cohort (corresponding to PIRADS ≥ 4) in order to construct working points of the models that maintain clinically achieved detection rates for significant prostate cancer. In order to evaluate how the model's performance translates to clinical findings, their sensitivity and specificity was compared on a per-lesion as well as on a per-patient basis. Specifically, the models' performance was assessed against the radiologist's performance based on the reduction of false-positive (FP) lesions or patients with FP lesions and by expressing this reduction as a ratio with the number of observations.

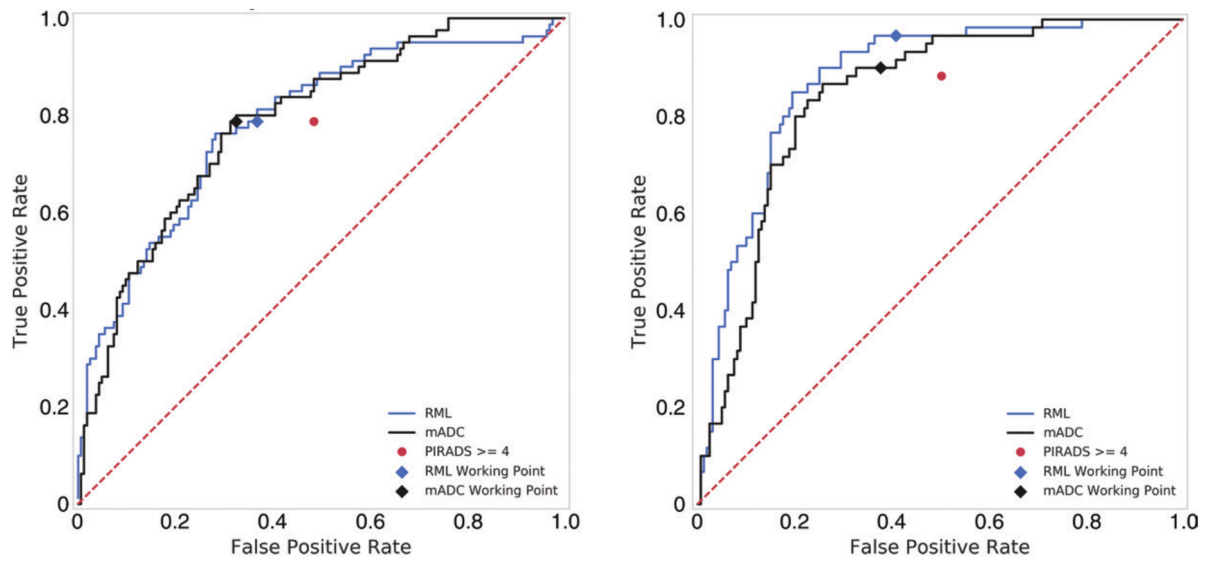
In order to test for statistically significant classification improvements as compared to the radiologist, the McNemar test [McNemar, 1947] was employed. The test assesses whether the disagreement between two paired sets of binary classifications is significantly larger than chance and is frequently applied in the medical sciences, e.g. to compare the sensitivity and specificity of two diagnostic tests on the same patient group [Hawass, 1997]. Correction for multiple comparisons was performed by using the Holm method [Holm, 1979].

38 patients had no MRI-detected lesions and did not contribute to the lesion-based analysis, which focused entirely on the task of lesion classification. Patient-based analysis metrics were calculated for the entire cohort to assess the overall combined radiologist lesion detection and model-based lesion classification performance.

A global radiomics model was trained on all lesions independent of lesion location (i.e., TZ and PZ). For zone-specific performance assessment, previously proposed as advantageous [Ginsburg et al., 2017], separate radiomics models were independently trained on TZ and PZ lesions and their predictions combined to obtain a performance assessment in the entire cohorts.

5.4.1 Lesion-based Analysis

Training Cohort The cross-validated analysis on the training cohort included 243 MRI-detected lesions, 33% of which were positive for clinically significant prostate cancer



(a) Training Cohort: Cross-validated ROC.

(b) Test Cohort: Independent test ROC.

Figure 5.3 | ROC Curves for Prostate Lesion Classification. The ROC curves for the RF ensemble (RML) and the mono-parameter mADC are shown in blue and black respectively. The red point corresponds to the radiologist’s clinical performance. The blue and black diamonds depict the model working points, which are chosen to match the radiologist’s true positive rate on the training cohort, see subfigure a).

(found in 157 of 183 patients in the training cohort, see Fig. A.2). The classifier ROC-AUC) was not found significantly different between mADC (0.79) and the RML (0.78) (see Fig. 5.3a)).

At the fixed radiologist sensitivity of 79% (63 of 80 positive lesions), the specificity of mADC (67% [110 of 163 negative lesions]) and RML (63% [103 of 163 negative lesions]) was improved, compared with 52% (84 of 163 negative lesions) for lesion classification by radiologists (see Fig. A.3). The corresponding model working points were determined to be $732 \text{ mm}^2/\text{s}$ for mADC (values below indicate positive predictions) and 0.28 for the RML model (values above indicate positive predictions). In comparison to clinical interpretation by radiologists, measurement of the mADC reduced False Positive (FP) lesions by 26 (10.7%) and RML reduced FP lesions by 19 (7.8%), both leaving False Negative (FN) lesions unchanged. At their working points, both models performed significantly better than the radiologist according to the McNemar test, see Fig. A.3.

Test Cohort This analysis included 219 MRI-detected lesions, 27% of which were positive for prostate cancer (found in 121 of 133 patients of the independent test cohort, see Fig. A.2). The classifier’s AUC was not significantly different for the mADC (0.84)

versus RML (0.88) ($p = .176$, DeLong test) (see Fig. 5.3b). The radiologist interpretation of mpMRI had a per-lesion sensitivity of 88% (53 of 60 positive lesions) and specificity of 50% (79 of 159 negative lesions). In comparison, measurement of the mADC reduced the number of FP lesions from 80 to 60 (specificity, 62% [99 of 159 negative lesions]) and the number of FN lesions from seven to six (sensitivity = 90% [54 of 60 positive lesions]; $p = .048$, significant according to McNemar test). RML reduced the number of FP lesions from 80 to 66 (specificity, 58% [93 of 159 negative lesions]) and the number of FN lesions from seven to two (sensitivity = 97% [58 of 60 positive lesions]; $p = .176$, insignificant according to McNemar test) (see Fig. A.3).

Fig. 5.4 and Fig. 5.5 show the mpMRI images of two test cohort cases, along with the respective lesion segmentations. Biopsy revealed insignificant or no prostate cancer, while both cases were assessed as suffering from aggressive PCa by the respective radiologist. The captions give more details on the clinical assessments as well as the corresponding model predictions.

5.4.2 Patient-based Analysis

By design and as described above, the RF ensemble (RML) and the mono-parameter mADC were set up to classify individual lesions. However there is also large clinical relevance in correctly classifying the patient as a whole. For this reason we further evaluated above models on a patient basis by pooling the lesions of a patient and assigning the patient both the most severe ground truth label and the most severe model prediction in the pool (where labels and predictions are still binary).

Training Cohort The cross-validated per-patient specificity of the mADC (67% [80 of 120 negative lesions]) and RML classifier (62% [74 of 120 negative patients]) was higher compared to that of the radiologist interpretation (57% [68 of 120 negative patients]). Compared to radiologist interpretation, measurement of the mADC reduced the number of FP patients by 12 and did not reduce the number of FN patients. RML showed a lower reduction in FP patients (by six) and an increase by two in the number of FN patients (see Fig. A.4).

Test Cohort Radiologist interpretation provided a per-patient sensitivity of 89% (40 of 45 positive patients) and specificity of 43% (38 of 88 negative patients). In comparison, measurement of the mADC reduced the number of FP patients from 50 to 43 (specificity = 51% [45 of 88 negative patients]) and the number of FN patients from five to three (sensitivity = 93% [42 of 45 positive patients]) ($p = .496$, insignificant according to

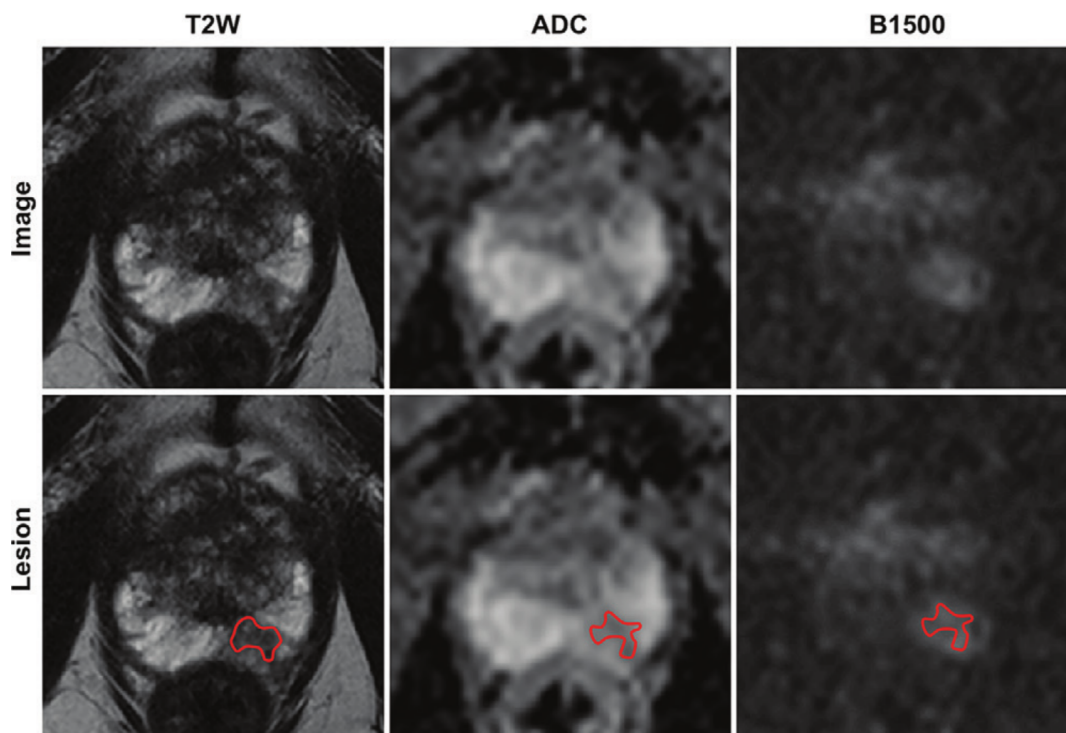


Figure 5.4 | Test Cohort Example Case 1. Images in a 74-year-old patient with mildly elevated PSA level of 6.2 ng mL^{-1} and negative DRE . $T2w$, ADC and $B1500$ images are shown in columns, and images without and with the superimposed, outlined segmented lesion (in red) are shown in rows. A lesion in the left mid **Peripheral Zone** is shown, which was read as **PIRADS** category 4 ('aggressive'). This lesion was rated negative according to a $mADC$ value of $895 \text{ mm}^2/\text{sec}$ (above the $732 \text{ mm}^2/\text{sec}$ cut-off), and it was also negative, as in below the radiomic machine learning (RML) cut-off, according to RML with a score of 0.12 (cut-off 0.28). Targeted biopsy revealed no cancer at this location. Gleason Grade 1 prostate cancer was found in systematic cores and a targeted biopsy from the **MRI** index lesion in the left mid anterior transition zone (not shown).

McNemar test). The use of **RML** reduced the number of **FP** patients from 50 to 43 (specificity = 51% [45 of 88 negative patients]) and the number of **FN** patients from five to two (sensitivity = 96% [43 of 45 positive patients]) ($p = .496$, insignificant according to McNemar test) (Fig. A.4).

5.4.3 Zone-based Analysis

As described in [Sec. 3.2](#), prostate tumors can exhibit different **MRI** appearance depending on whether they are located in the **PZ** or **TZ**. This is reflected in the zone-specific **PIRADS** guidelines and was also found a useful distinction in related radiomics studies [[Ginsburg et al., 2017](#)].

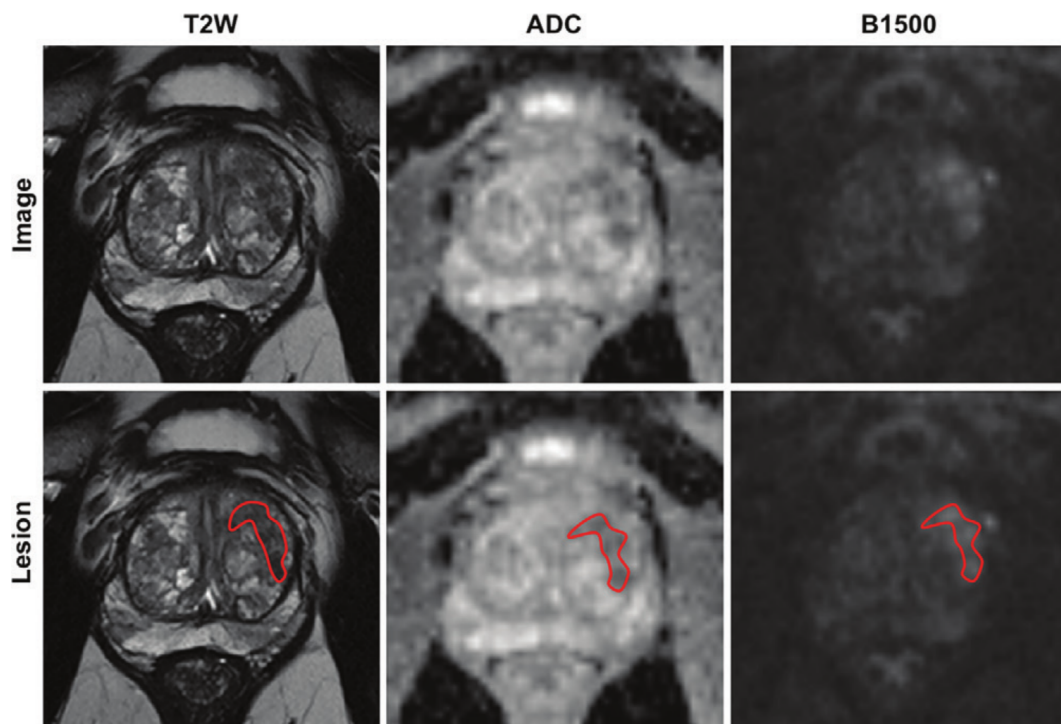


Figure 5.5 | Test Cohort Example Case 2. Images in a 65-year-old patient with strongly elevated PSA of 19.5 ng mL^{-1} and negative DRE. T2w, ADC and B1500 images are shown in columns, and images without and with the superimposed, outlined segmented lesion (in red) are shown in rows. A lesion in the left mid anterior transition zone is shown, which was read as PIRADS category 4 ('aggressive'). This lesion was rated negative according to a mADC value of $756 \text{ mm}^2/\text{sec}$ (below the $732 \text{ mm}^2/\text{sec}$ cut-off), although it was found positive, as in above the radiomic machine learning (RML) threshold, by RML with a score of 0.50. Targeted biopsy revealed no cancer at this location. There was no presence of any cancer in any of the systematic and targeted biopsy cores in this patient.

For this reason we trained zone-specific RF ensembles (RML) for both the PZ and TZ and equally distinguished two mADC 'models'. For each zone-specific model the employed procedure was identical to the zone-agnostic models described above.

Zone-specific RML performance is shown in Fig. A.3 on a per lesion basis and in Fig. A.4 on a per-patient basis. In both cases, the performance was lower than that for the zone-agnostic ('global') model. For example, on a per-lesion basis in the test set, sensitivity was 92% (55 of 60 positive lesions) and specificity was 53% (84 of 159 negative lesions), compared with 97% (58 of 60 positive lesions) and 58% (93 of 159 negative lesions), respectively, for the global model. The AUC for RML was 0.84 compared with 0.83 for mADC in the PZ ($p = .822$, insignificant according to McNemar test), and

0.89 compared with 0.87, respectively, in the TZ ($p = .493$, insignificant according to McNemar test).

5.4.4 Feature Importances

As mentioned above, the training of decision trees involves the finding of optimal cuts on feature values. During training, this optimality can be determined by measuring the decrease in Gini index [Gini, 1936] between the parent and the child nodes. The Gini index measures the impurity of the data points in the respective node with respect to the target classes. The higher the Gini index, the lower is the node’s ability to separate, i.e. discriminate, the target classes. A decrease in Gini index after a split on a variable thus indicates the respective variable’s ability to discriminate between the target labels.

Assessing this reduction in impurity after training can produce an empirical feature importance, which can be produced in various ways. The decrease in Gini index by a feature can for example be weighted by the probability with which the feature is cut (i.e. employed in a node) and aggregated across a forest’s trees [Ronaghan, 2018], which is the method we employed. After training the final zone-agnostic RF ensemble, we summed these feature importances for each feature across the ensemble members. The resulting top-10 most important features are reported in Table 5.1.

Table 5.1 | Top 10 Most Important Features. Ranking according to decrease in node impurity across ensemble members and splits as measured by reduction in Gini index.

Rank	Feature Name	$ \Delta\text{Gini index} $
1	ADC1500_original_firstorder_Maximum	41.72
2	ADC1500_original_firstorder_RootMeanSquared	39.41
3	ADC1500_original_firstorder_Median	36.84
4	ADC1500_original_firstorder_Mean	33.24
5	ADC1500_original_firstorder_90Percentile	30.63
6	ADC1500_original_firstorder_10Percentile	29.54
7	T2_original_shape_Maximum2DDiameterColumn	14.26
8	ADC1500_original_firstorder_Minimum	13.73
9	T2_original_shape_MinorAxis	13.61
10	T2_original_shape_SurfaceArea	12.37

The top most important features all relate to first order statistics of the [Apparent Diffusion Coefficient](#), including its mean value ([mADC](#) which is found the 4th most important feature), with only few [T2w](#)-derived shape features ranking low within the

top-10 ranking. This finding quantitatively substantiates the discriminative nature of [ADC](#) features, which already play an important role in the subjective lesion grading protocol that reflects current clinical standard ([PIRADS](#), see [Sec. 3.2](#))

5.5 Discussion

Prostate [MRI](#) may have the potential to spare patients the discomfort and potential morbidity of biopsies [[Ahmed et al., 2017](#)], but limitations in the detection of clinically significant prostate cancer with prostate [MRI](#) are well known [[Borofsky et al., 2017](#)]. In particular, diagnostic accuracy varies based on the individual radiologist [[Greer et al., 2017](#), [Rosenkrantz et al., 2016a, 2017](#)].

Machine learning techniques, radiomics and quantitative assessment however could have potential for decision support [[Fehr et al., 2015](#), [Sonn et al., 2017](#), [Vignati et al., 2015](#), [Wang et al., 2017](#)] and our results provide an assessment of their ability to aid [MRI](#) interpretation. In the present study, quantitative measurement of the [mADC](#), when compared with clinical assessment, is able to significantly reduce the misclassification of [MRI](#)-detected lesions. This is in agreement with recent analyses for the characterization of [TZ](#) lesions [[Pierre et al., 2018](#)]. We found improved lesion classification over [PIRADS](#) in both the [PZ](#) and [TZ](#), confirming the high value of [mADC](#) for whole gland assessment. This result is quite remarkable, given that the clinical value of [ADC](#) is already widely known and reflected in the significance that current clinical guidelines place upon them. The problem with current clinical standards may lie in their reliance on visual inspection of [ADC](#) maps and subjective comparison of prospective lesions in relation to other prostate regions. Our results suggest that a move to include quantitative measures, such as simply calculating first-order [ADC](#) statistics across delineated lesions and comparing them to an absolute threshold, may significantly improve clinical grading of lesions.

We further investigated the performance of radiomic-feature based machine learning to assess if there is added value over measurement of the [mADC](#) alone. We found the [RF](#) models ([RML](#)) did not perform better than simple [mADC](#), as assessed by [ROC](#). In the [RF](#) model, all highly ranked features were in fact closely related to first-order [ADC](#) features rather than to textural or morphologic information. Our results thus refute findings of prior studies [[Nketiah et al., 2017](#), [Wibmer et al., 2015](#)] of an added value of radiomic features and lesion morphology derived from [T2w](#) images.

Interestingly, the performance of our [ML](#) model, the [mADC](#) and that of the [PIRADS](#) assessment, all increased on the held-out test cohort. This may reflect the radiologists' learning curve since the introduction of the [PIRADS](#) version 2 system and, upon review

of the clinical images, may also be attributable to a larger number of patients with small solitary PZ lesions in the training cohort, which could be more difficult to classify in general.

The size of our training cohort and the presence and size of the test cohort exceeded that of other radiomics studies for prostate cancer [Ginsburg et al., 2017, Khalvati et al., 2015, Nketiah et al., 2017, Wang et al., 2017]. In fact, to the best of our knowledge, this dataset forms the largest collection reported to date in a radiomic analysis involving consecutive patients under suspicion for PCa that underwent a homogeneous protocol on a single MRI unit, followed by mpMRI and TRUS-biopsy examinations.

Limitations of our study further include the use of radiomics for lesion characterization but not for lesion detection, thus not examining if mADC or RML are better than radiologists at cancer detection. Lesions were segmented manually, a time consuming task that has to be regarded as prohibitive when very large databases are evaluated in the future, requiring the development of automated segmentation techniques. Furthermore, a histopathologic assessment based on MRI and TRUS-biopsy rather than radical prostatectomy specimens was used. However, our biopsy approach has been tested against radical prostatectomy as the reference standard and showed a sensitivity of 97% for significant prostate cancer at the final histopathologic examination [Radtke et al., 2016].

In conclusion, this study compared the use of mADC and radiomics with machine learning for the characterization of lesions that were prospectively detected during routine clinical interpretation. Quantitative assessment of the mADC was more accurate than qualitative PIRADS assessment in classifying a lesion as clinically significant prostate cancer. Radiomics provided additional evidence that ADC-based features (including mADC) were more discriminative than other MRI features. In fact, at the current cohort size, no added benefit of the radiomic approach was found, and mADC is suggested as the best choice for quantitative prostate assessment.

Using quantitative image features aids in reducing inter-rater variability as it reduces the subjectivity inherent in current clinical grading. The presented techniques however produce singular and deterministic assessments and are thus blind to plausible variations due to ambiguities with the classifier scores and mADC values potentially presenting a false sense of confidence. Given much larger datasets, model performance, robustness and utility could however likely be taken much further e.g. by i) learning hierarchical feature representations end-to-end using CNNs, ii) learning semantic segmentation models end-to-end using FCNs (see Chap. 6) and iii) by employing methods that yield calibrated uncertainty estimates in order to better guide down-stream decision making when facing ambiguity.

Chapter 6

Mitigating Label Noise through Adversarial Training

The interpretation of medical images suffers from large inter-rater variability. This annotation variability is largely due to ambiguous image evidence and affects the segmentation of anatomical structures and critically that of tumor lesions, see [Chap. 3](#). The ensuing diversity in the ground truth labels, can be seen as noise in the training process of [ML](#) algorithms.

In this chapter we discuss how the de facto standard training procedure for deep semantic segmentation approaches can be sub-optimal when learning from noisy ground truth annotations. As a possible solution that is empirically found more robust to such noise, we propose an adversarial training scheme borrowed from the framework of [Generative Adversarial Network \(GAN\)](#) methods (see [Sec. 4.3](#)), while leaving the deep segmentation network architecture itself unchanged. Our core contributions are as follows:

- We compare training a U-Net with the standard pixelwise [Cross Entropy \(CE\)](#) loss for semantic segmentation to training it in a mini-max game against an adversarial discriminator.
- We observe increased performance when training adversarially, which we hypothesize could be attributed to reduced gradient conflicts in the noisy label setting
- Lastly, we find further increases in relative performance when reducing the number of training examples, which seems in line with the hypothesis, that adversarial training might mitigate label noise and suggests particular utility in the small dataset regime.

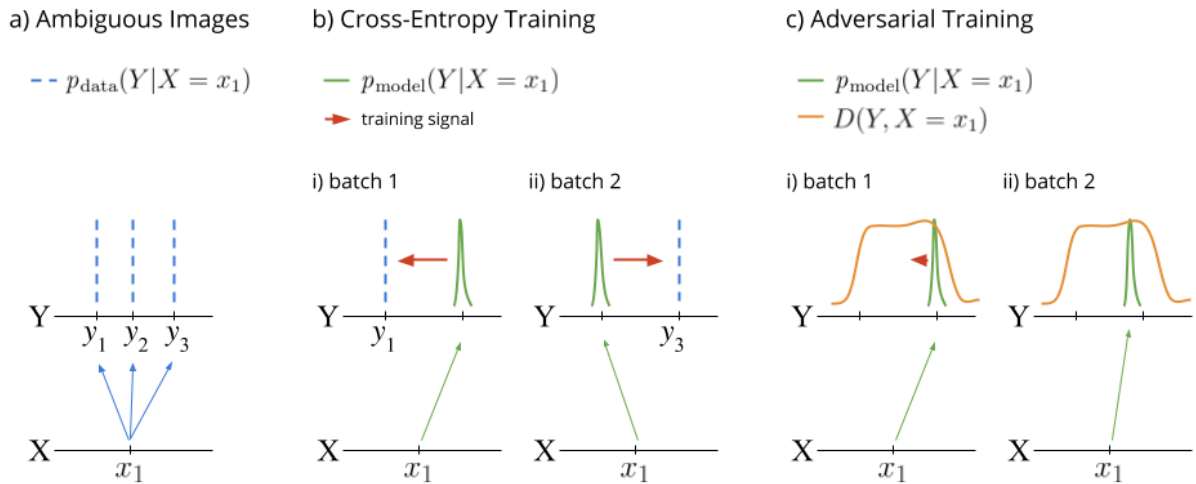


Figure 6.1 | Training under Ambiguous Images. a) shows a diverse mapping from $X \rightarrow Y$ in terms of an empirical data distribution between simplified image and segmentation manifolds. b) Training with **Stochastic Gradient Descent (SGD)** and a **CE-loss** requires sampling of $(x, y) \sim P_{\text{data}}$ and may lead to high variance if not contradicting training signals across batches, see scenario i) and ii). c) Using adversarial training and a discriminator network D to distinguish between plausible and implausible segmentations is akin to learning a model of the ground truth and holds the potential to better accommodate for co-existing modes. Note that D does not induce a likelihood $P(Y|X)$, but instead constitutes a discriminative model telling apart plausible from implausible (X, Y) -pairs.

This chapter closely follows our prior publications [Kohl et al., 2017a,b] one of which was presented at the *Machine Learning for Health Workshop* at the *Advances in Neural Information Processing Systems Conference (NeurIPS) 2017*¹.

The main idea of this work is rooted in the observation that when faced with noisy labels, stochastic gradient descent on a **CE-loss** produces conflicting training signals with possible negative effects on segmentation performance. If however we were able to formulate or learn a loss function that accommodates for sets of plausible annotations, we could reduce the negative effects of noisy or conflicting labels. We elaborate on this idea and its evaluation in more detail below.

6.1 How Ambiguity interferes with **CE-based Training**

As is standard in multiclass classification tasks, the learning target in semantic segmentation is formulated as a static and deterministic cross-entropy loss between the produced softmax probabilities $P_{\text{model}}(Y|X)$ and the one-hot ground truth segmentations Y . The

¹See <https://ml4health.github.io/2017/>.

given image and segmentation pairs (X, Y) of a dataset form an empirical distribution P_{data} from which during training mini-batches are sampled to compute parameter updates via [Stochastic Gradient Descent \(SGD\)](#). Repeating [Eq. 4.1](#), the loss thus reads

$$\mathcal{L}_{CE} = -\mathbb{E}_{(X,Y) \sim P_{data}} [Y \log P(Y|X)], \quad (6.1)$$

where here $P_{data}(X, Y)$ is a mixture of Dirac-delta functions each centered on an image-segmentation-pair. As elaborated upon in [Sec. 4.2](#), ambiguous images X may lead to sets of plausible ground truth segmentations $Y \in (y_1, y_2, \dots)$, corresponding to multi-modal conditional distributions $P(Y|X)$. This is illustrated in [Fig. 6.1a](#)), which presents a schematic for a diverse mapping of an image to the segmentation manifold. Under these circumstances the optimal solution for the model is to assign equal probability to all ground truth modes $Y \in (y_1, y_2, \dots)$, i.e. predict a mixture. This can be difficult to learn because when using mini-batch [SGD](#), the sampling of (X, Y) -pairs may lead to conflicting training signals across batches. A caricature of this is shown in [Fig. 6.1b](#)), where scenarios i) and ii) show the sampling of different segmentation targets y_i for a given $X = x_1$ into different mini-batches, therefore providing high-variance and potentially even contradicting gradients across training iterations. Additionally it might be more useful to find a single plausible mode than an implausible mixture of all of them. For this reason we develop a way of reparameterizing the multi-modal ground-truth, that holds the potential to allow for the co-existence of plausible modes under the loss, even across mini-batches. The hope is for this scheme to mitigate potentially conflicting gradients and to discourage implausible mixtures of modes.

6.2 Learning to reparameterize the Loss Function

The generation of high-fidelity images has profited largely from the introduction of [Generative Adversarial Networks \(GANs\)](#) [[Goodfellow et al., 2014](#)]. The GAN model is really a model tandem in which one of them, the discriminator D , learns to distinguish between plausible images, e.g. real natural images, and implausible images, i.e. the ones generated by its tandem partner, the generator G . As D forms the decision boundary between what is plausible and implausible, G can take steps towards directing its output towards plausible regions in the output manifold, see [Sec. 4.3](#). In this scenario D provides the training loss that G optimizes for and may be viewed as a learned higher-order loss [[Isola et al., 2017](#), [Luc et al., 2016](#)], that distills what makes a good output in

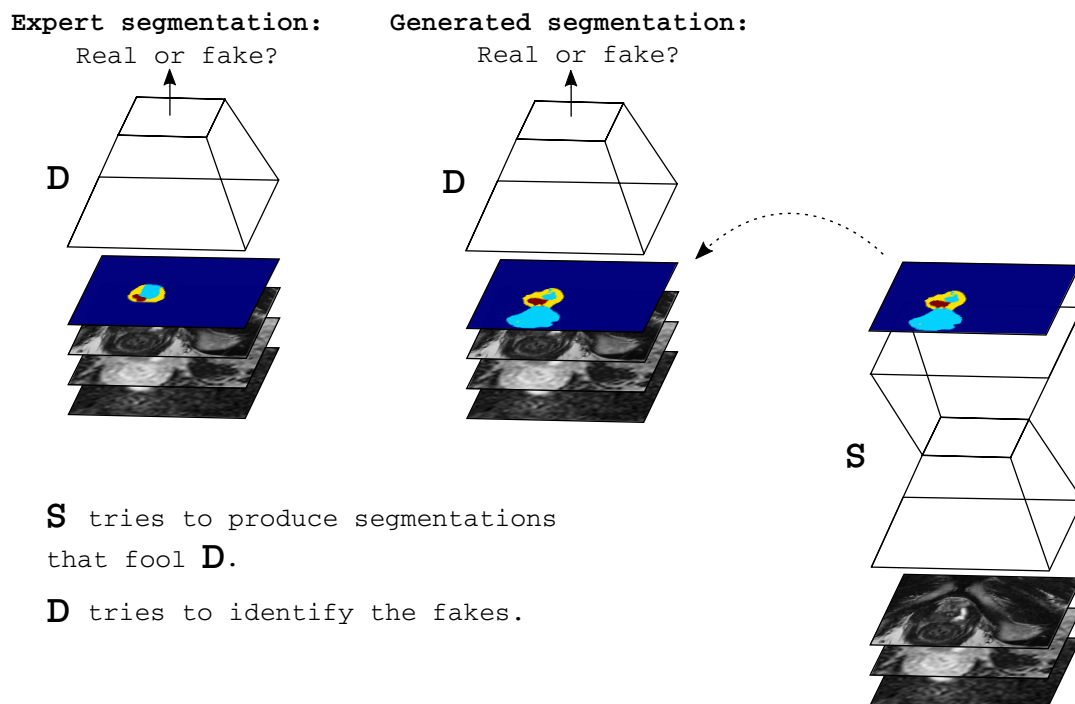


Figure 6.2 | Schematic of Adversarial Training for Semantic Segmentation. From the left to right: the discriminator D is shown expert annotations alongside a stack of corresponding MRI images. To the right thereof, D is illustrated for the case of receiving the segmentations net's (S 's) output alongside the stack of MRI images. When run deterministically G is equivalent to S .

form of a parametric model. In other words: the discriminator D constitutes a learned reparameterization for the loss.

The work presented below hinges on the idea that this learned loss could provide a training signal that allows for the co-existence of segmentation modes without encouraging to mix them. As stated above, the optimal prediction in the presence of equally plausible solutions is a uniform one over plausible labels. In the proposed scheme, the discriminator is set up to learn the plausible data region (Y, X) , therefore taking the burden from the segmentation network to produce a mixture of all plausible segmentations. This is sketched out in Fig. 6.1c). Instead, D is satisfied with a single mode for as long as it is plausible and thus is expected to result in a loss without high-variance or conflicting gradients. As is explained below, G is set up as a deterministic segmentation model and not expected to produce diverse outputs, as is otherwise common in GANs and other generative models.

The aim of our approach falls within the type of works that seek to mitigate negative effects on model training under noisy labels, which was discussed in Sec. 4.2. To the best of our knowledge none of them have considered reparameterizing the loss for this purpose. To make the GAN-framework amenable for the training of deterministic semantic segmentation networks it needs to be extended to the image-conditional case: G takes on the form of a semantic segmentation network S , that produces a pixelwise softmax output $S(X) = P_{\text{model}}(Y|X)$. Note that we do not condition S on noise and instead opt to use a strong deterministic model, the U-Net [Ronneberger et al., 2015]. D is a binary classification network that outputs a single softmax node $D(Y, X)$, given both images and segmentations. The training objective for S and D then becomes:

$$\mathcal{L}_S = -\mathbb{E}_{X \sim p_{\text{data}}} [\log D(S(X), X)], \quad (6.2)$$

$$\mathcal{L}_D = -\mathbb{E}_{X \sim p_{\text{data}}} [\log (1 - D(S(X), X))] - \mathbb{E}_{(X, Y) \sim p_{\text{data}}} [\log D(Y, X)]. \quad (6.3)$$

The adversarial training scheme between S and D that these losses provide for is further illustrated in Fig. 6.2. S and D are updated in alternating fashion. Optimal training requires for D to be near its optimal solution at all times. For this purpose, D can be trained using k mini-batch gradient descent steps for each such step performed on S [Goodfellow et al., 2014]. For semantic segmentation [Luc et al., 2016] further propose a hybrid loss term for the segmentation net S in form of a weighted sum of the discriminator-derived loss and the standard cross-entropy loss (Eq. 4.1): $\mathcal{L}'_S = \mathcal{L}_S + \lambda \mathcal{L}_{CE}$, which we also compare against below.

Note that the problem of image ambiguity and the diverse labels it induces exists also when only a single annotation is available per image, rather than a set of annotations $Y \in (y_1, y_2, \dots)$. In this case multi-modal distributions over labels manifest across images, e.g. lesions on different images with very similar appearance might be annotated very differently as a consequence of intra-grader variance or in the case for when different parts of the data are annotated by different graders. The proposed adversarial training does not rely on the availability of multiple annotations and may provide mitigation of noise even when only a single annotation is available per image, as is the case here.

6.3 Dataset Details

We base our analysis on an internal prostate MRI dataset, a clinical case that can be particularly ambiguous [Hameed and Humphrey, 2010, Kitzing et al., 2015, Nagel et al., 2013, Sakala et al., 2017]. The employed dataset contains 152 patients with MRI acquired using a Siemens Prisma 3 T machine at the National Center for Tumor Diseases (NCT) in Heidelberg, Germany. This dataset is a subset of the dataset employed in the analysis described in the previous chapter (Sec. 5.2), which had been extended after the analysis of the present chapter was carried out. All patients had a suspicious screening result and a TRUS-biopsy yielding pathological classification, i.e. Gleason Score (GS) [Gleason, 1966]. Image analysis was based on a T2-weighted (T2w), an Apparent Diffusion Coefficient (ADC) map and a high b-value diffusion weighted image at $b = 1500 \text{ s mm}^{-2}$ (B1500). The T2w images have an in-plane resolution of 0.25 mm, the other two modalities were upsampled accordingly. The prostate’s anatomical details as well as lesions were segmented independently on both the T2w and the ADC-map by an experienced radiologist. The segmentation annotations comprise -if present- four classes: tumor lesion, Peripheral Zone, Transitional Zone and other (i.e. non-prostatic, lesion-free tissue). The TZ segmentation was obtained as the complement of the PZ within the whole gland segmentation. Example cases can be found in Fig. 5.2. Image registration was performed using rigid translation maximizing the overlap between the masks of the peripheral zone, as they are most informative for the relative alignment of the images.

After registration, the two independent segmentations of all classes were fused by a hierarchical label consensus. In that process first the intersection of the lesions was found and fixed, then the intersection of the peripheral zone and then that of the transitional zone, thus ensuring anatomically plausible results. We define aggressive lesions as such with a biopsy-determined Gleason-assessment of $GS \geq 7a$ in line with Chap. 5. Lesions

found to exhibit $GS < 7a$ are regarded free of aggressive tumor and the respective lesion’s segmentations are removed from the ground truth annotations thus falling back to normal appearing tissue.

6.4 Network Architecture and Training Procedure

Architecture We use an identical 2D U-Net-type architecture for the segmentation network in each experiment, see Fig. 4.3b. We follow [Isola et al., 2017] and introduce InstanceNorm layers after each convolutional layer, thus opting against BatchNorm, conjecturing that it avoids harmful stochasticity, introduced by small batch-sizes. Let $C(I)N$ denote a Convolution-(InstanceNorm-)ReLU layer with N feature maps each. Then the U-Net’s encoder takes on the following form: C64-CI128-CI256-CI512-CI1024, while the decoder can be represented as: CI512-CI256-CI128-CI64-C4. The architecture used for the discriminator in large parts mirrors that of the U-Net’s encoder: C64-CI128-CI256-CI512-CI512-CI1024-GP1, where GP1 denotes a global average pooling layer followed by a dense layer with one output node. InstanceNorm is neither applied to the first nor the last layer in S and D . Convolutional layers employ 3×3 -filters, except for the last one in S ’s decoder which uses 1×1 -filters. D takes $7 \times 416 \times 416$ inputs, featuring one channel for each of the three MRI modalities and four channels encoding the class labels. Accordingly S receives inputs of shape $3 \times 416 \times 416$.

Training To provide meaningful comparison, the training protocol is the same for all evaluated schemes. We use a set of 55 patients (\mathcal{S}_{agg}) comprising 188 2D-slices with biopsy-confirmed aggressive tumor lesions of $GS \geq 7a$ and 97 patients (\mathcal{S}_{free}) with 475 2D-slices that were diagnosed lesion free (slice size $3 \times 416 \times 416$). The experiments are performed using four-fold cross-validation on \mathcal{S}_{agg} with mutually exclusive subject allocation to the folds, while \mathcal{S}_{free} is used during training only. In each cross-validation permutation, 2 folds are employed for training the model, one fold for model selection according to the DSC metric for tumor (see below, Sec. 6.5), and one held-out fold for validation. All segmentation models are trained for 225 epochs, with 80 randomly sampled batches each, using an initial learning rate (LR) of 10^{-5} , that is halved every 75 epochs. During the adversarial training scheme we train the discriminator D on 3 batches for each batch the segmentor is trained on while using fixed $LR = 10^{-5}$ for D . For parameter optimizations we use *Adam* [Kingma and Ba, 2014]. The training data is augmented by in-plane rotations with angle $\phi \sim \mathcal{U}[-\pi/8, \pi/8]$, crops with a mask shifted by $(\Delta x, \Delta y) \sim (\mathcal{U}[-50, 50], \mathcal{U}[-50, 50])$ and random left-right mirroring. We use

a batch-size of 5 with importance sampling, averaging to 3.5 samples from \mathcal{S}_{agg} in each batch.

6.5 Results

Because we are interested in evaluating whether it is hurtful to use a CE-loss for the task of finding the semantic segmentation of ambiguous tissue such as aggressive lesions, we exclude lesion free images in the evaluation, allowing to focus on the segmentation performance rather than assessing the detection performance. As performance metrics we employ the Sørensen–Dice Coefficient (DSC) [Dice, 1945, Sørensen, 1948], which is defined in terms of the number of True Positive (TP), the False Positive (FP) and the False Negative (FN) pixels given a binary ground truth label map Y and a predicted map Y' :

$$\text{DSC}(Y, Y') = \frac{2|Y \cap Y'|}{|Y| + |Y'|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \quad (6.4)$$

DSC is a popular metric in medical image segmentation and bears close resemblance to the Intersection over Union (IoU) metric that is in turn popular as a segmentation metric on natural images, see Eq. 7.5 for a definition. Additionally we report the pixelwise sensitivity and specificity for aggressive lesions (tumor).

6.5.1 Improving Segmentation Performance

The adversarial approach scored better for tumor segmentation both in the Sørensen–Dice Coefficient (DSC) as well as the sensitivity, see Table 6.1, which reports the inner-loop test set results on \mathcal{S}_{agg} . The improvement in Sørensen–Dice Coefficient (DSC) and sensitivity were significant as determined by means of the Wilcoxon signed-rank test [Wilcoxon, 1945] (p-value < 0.001). The specificities between the approaches were equal. Using a hybrid loss that adds both the CE-loss and the adversarial loss with the same weighting as [Luc et al., 2016], i.e. $\lambda = 0.5$, results in improvements over the CE-loss based training but does not reach the performance of the proposed scheme of adversarial training only. This is further evidence that the CE-loss might be a suboptimal choice when training on ambiguous images with noisy labels. In Fig. 6.3 we show a comparison between the ground truth and the U-Net predictions when training with either exclusively CE or adversarial losses for a range of different patients.

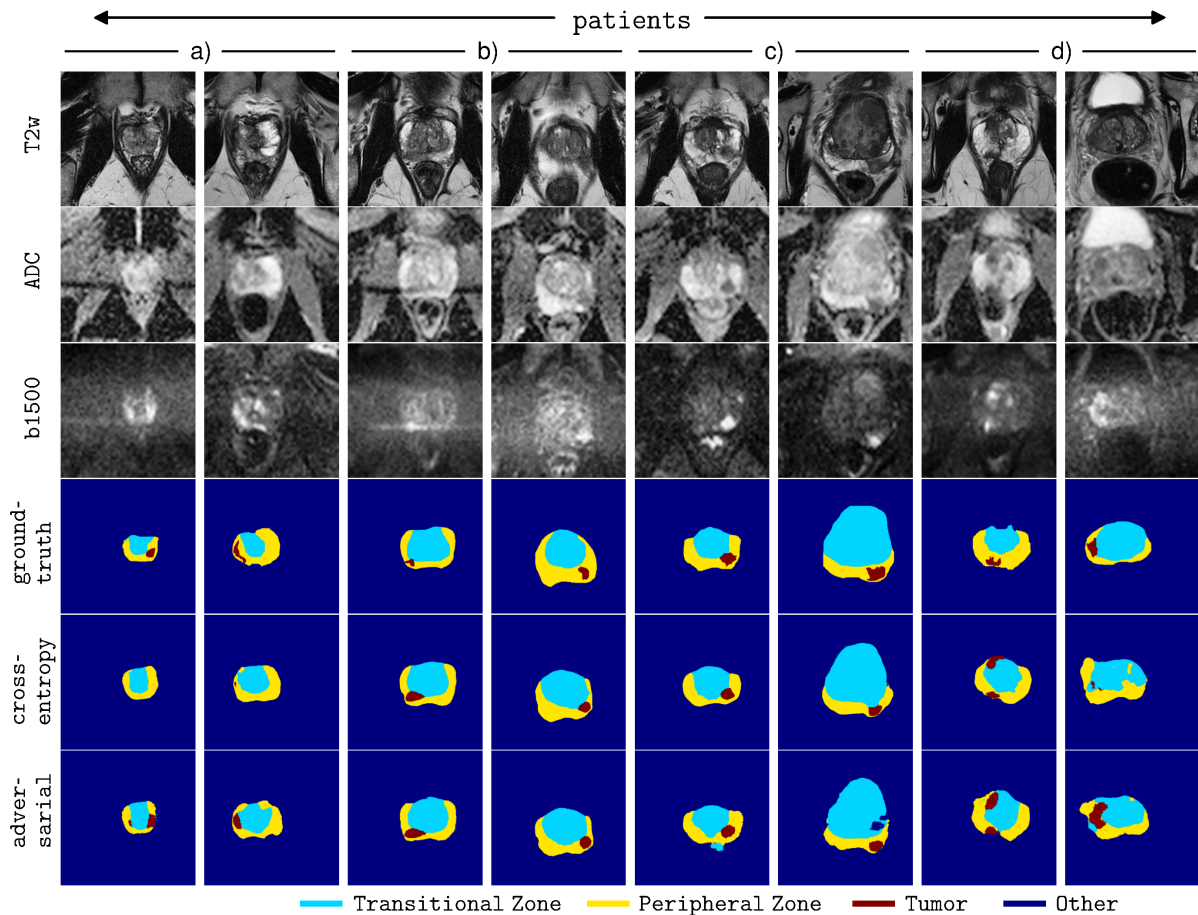


Figure 6.3 | Prostate MRI Example Cases. For each depicted test set example the three MRI modalities, the expert annotation as well as the segmentations produced by training the U-Net S with different loss schemes are shown (in that order starting from the top). The first two columns from the left, i.e. columns a), depict examples in which the adversarial is clearly more sensitive to aggressive tumor than the cross-entropy training. Columns b) show examples for which the methods are on par. Columns c) feature examples for which the adversarial method yields partially defective label maps. Columns d) exhibit examples for which both methods deviate considerably from the ground-truth, the first of which likely shows tumor detection by both methods, missed by the expert.

Table 6.1 | Quantitative Results for Prostate Tumor Segmentation. Results are obtained from an inner-loop test set of a four-fold cross-validation for $GS \geq 7$ tumor.

Training Scheme Loss	Cross-Entropy \mathcal{L}_{mce}	Adversarial $\mathcal{L}_S \& \mathcal{L}_D$	Hybrid $\mathcal{L}_{mce}/2 + \mathcal{L}_S \& \mathcal{L}_D$
Tumor DSC	0.35 ± 0.29	0.41 ± 0.28	0.39 ± 0.29
Tumor Sensitivity	0.37 ± 0.33	0.55 ± 0.36	0.49 ± 0.35
Tumor Specificity	0.98 ± 0.14	0.98 ± 0.14	0.98 ± 0.14

6.5.2 Increasing Robustness on Fewer Training Samples

Having only few examples to learn from, makes it hard to tell apart different classes, which gives rise to an uncertainty even when the classes could be disambiguated given more data (epistemic uncertainty, see Sec. 4.2). In the small data limit, the effect on the model may be the same as in the case for when classes are inherently ambiguous (aleatoric uncertainty, Sec. 4.2), since it receives training signals that it can not reconcile, in addition to potentially inherent ambiguities.

For this reason we seek to evaluate how the training schemes compare on progressively smaller datasets. This is interesting as it allows to further probe our hypothesis that adversarial training may mitigate conflicting training signals in addition to analyzing a relevant case for medical image analyses, since they typically deal with small amounts of labelled data. To this end we successively take away positive training samples, i.e. examples from \mathcal{S}_{agg} , from the fold that both schemes coincided to perform best on. We then train from scratch in the exact same manner as described above and evaluate on the same held-out fold from before.

The results for tumor DSC and sensitivity are reported in Fig. 6.4a) and b) respectively. As expected, decreasing the number of positive patients in the training set results in performance decreases for both training schemes. The adversarially trained U-Net however increases in DSC and sensitivity relative to the CE-trained U-Net (reported in percent in the lower panels of Fig. 6.4a) and b)) and thus exhibits a markedly more robust segmentation performance when dealing with increasingly noisy training signals.

6.6 Discussion

Ambiguous images and associated noisy labels are very common in the medical disciplines. Despite this, current semantic segmentation models are almost exclusively framed and trained in a deterministic fashion using the Cross Entropy (CE)-loss. In this work we

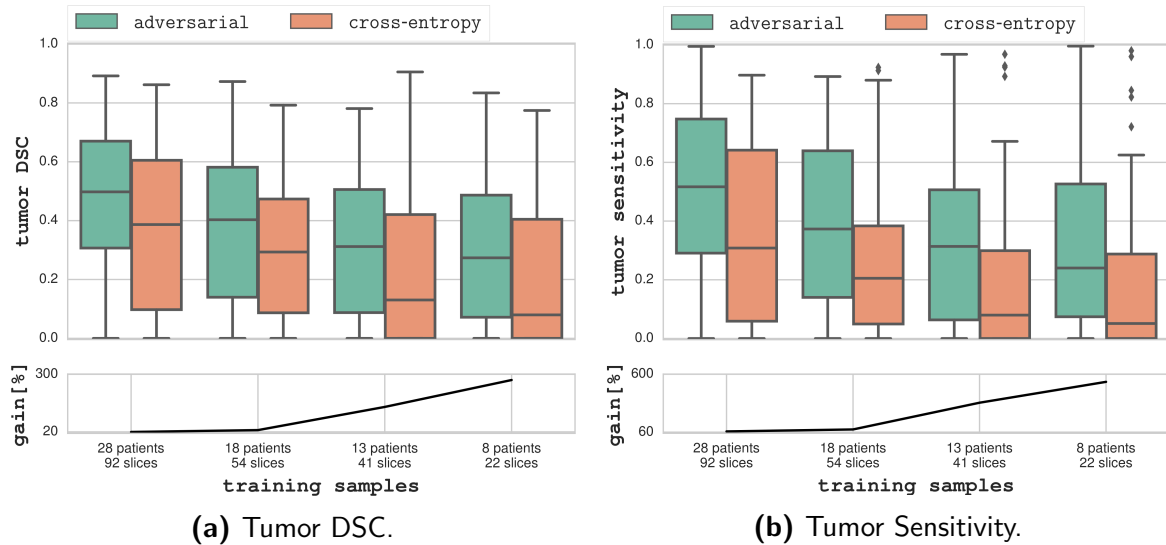


Figure 6.4 | Performance Comparison in the small Dataset Limit. Performance measured in terms of (a) tumor DSC and (b) sensitivity between the adversarial and cross-entropy training when successively taking away positive training data. The upper panels illustrate the respective performance distributions for the two schemes in the form of box-plots. The lower panels show the relative gain in median of the adversarial over the cross-entropy training, from which particularly pronounced gains are visible in the small dataset limit. Specificity (not shown) was around 0.98 in all experiments.

show that the [CE-loss](#) can exhibit sub-optimal performance when segmenting highly ambiguous [MRI](#) images of the prostate.

Starting with the observation that the [CE-loss](#) encourages implausible mixtures of segmentation modes when applied to multi-modal problems, we hypothesized that mini-batch gradient descent may lead to conflicting gradients that are detrimental to model performance. With the aim to ameliorate the gradient dynamics, we then turned to an adversarial training scheme in which a model of the loss is learned. This model (the discriminator D) distinguishes between real and fake segmentations and thus acknowledges collective subspaces of plausible segmentation modes. Training an otherwise identical segmentation network under both schemes empirically showed that the adversarial training may be advantageous under label noise. In conjunction with strong results in the small data limit these findings appear to affirm the utility of a learned reparameterization of the loss when faced with multi-modal labels.

From a theoretical stand-point reducing the number of training data induces a different type of uncertainty than the presence of ambiguous image evidence, see the discussion in [Sec. 4.2](#). From a practical point of view however, both scenarios may have comparable bearing on the training dynamics, as both little training data and data ambiguity may

result in training signals that are irreconcilable to the model. Having limited access to labelled data is very common for medical images and the empirical advantages of the adversarial training scheme over the regular CE-training suggests a possible avenue of future research towards more data efficiency.

Unfortunately adversarial training in the GAN-framework comes with its own caveats and may thus not constitute a drop-in solution for the simpler CE-based training: For one, it is well documented in the literature that the GAN setup can result in unstable training behavior such as sudden and irrecoverable deterioration of model performance (see Sec. 4.3). For another, the necessity of co-training a separate, potentially large discriminator, may limit the model and batch size that can be employed for the segmentation network and additionally significantly reduce training speed.

The extent to which the documented performance gains can be attributed to our intuition that the adversarial scheme allows for the co-existence of modes, which we expect to be violated in CE training, requires further scrutiny. To this end the losses for multiple annotations of a single image (and the statistics of the ensuing gradients) could be compared between the adversarial and the Cross Entropy training. Such an analysis was unfortunately not readily possible given the single annotations of our dataset.

Another possibility to ascertain the mechanism behind the empirical advantages would be a comparison to other training schemes designed to better cope with label noise such as [Ghosh et al., 2017, Zhang and Sabuncu, 2018]. These methods however were not published at the time this work was carried out.

As sufficiently described above and with even more detail in Chap. 3, prostate mpMRI images are often ambiguous with respect to lesion segmentations. Their outline and location may thus appear different on T2w and ADC, for example see Fig. 5.4 and Fig. 5.5. Another limitation of this work therefore lies in the approach of concurrently using several mpMRI modalities and fusing their respective annotations, as the fusion process itself may introduce additional label noise.

Lastly, training a *discriminative* model such as the U-Net with a discriminator D results in the deterministic prediction of only singular modes. A deterministic model however can not capture the admissible segmentation modes and the associated uncertainties across them. One way to extend the framework in this direction is to make it *generative* with the aim of producing multiple plausible modes given an image. Ways of doing so by means of variational models rather than GANs are proposed and evaluated in the following.

Chapter 7

Learning Image-Global Distributions over Segmentations

In this chapter we discuss a novel conditional generative model, the Probabilistic U-Net. This model is based on the observation of prior chapters, that given a medical scan alone, a single unique ground truth for the depicted tissues can often not be determined. Instead of trying to learn a deterministic mapping to segmentations, we propose to model plausible distributions over semantic segmentations for a given image in the presence of such ambiguity. This is approached by combining a U-Net with a [cVAE](#), samples of which can be decoded to unique interpretations of a scan. We treat image observations as evidence that can be used to narrow down the space of interpretations for the image. The main contributions of this work are:

- Our framework provides consistent segmentation maps instead of pixel-wise probabilities and can therefore give a joint likelihood of modes.
- Our model can induce arbitrarily complex output distributions including the occurrence of very rare modes, and is able to learn calibrated probabilities of segmentation modes.
- Sampling from our model is computationally cheap.
- In contrast to many existing applications of generative models that can only be qualitatively evaluated, our application and datasets allow quantitative performance evaluation including penalization of missing modes.

This work has been published in the Proceedings of *Advances in Neural Information Processing Systems*, see [\[Kohl et al., 2018\]](#), where it was also presented as a spotlight

talk. We provide an open-source re-implementation of our approach at https://github.com/SimonKohl/probabilistic_unet.

7.1 Segmenting Ambiguous Images

The semantic segmentation task assigns a class label to each pixel in an image. While in many cases the context in the image provides sufficient information to resolve the ambiguities in this mapping, there exists an important class of images where even the full image context is not sufficient to resolve all ambiguities. Such ambiguities are common in medical imaging applications, e.g. in lung abnormalities segmentation from CT images. A lesion might be clearly visible, but the information about whether it is cancer tissue or not might not be available from this image alone. Similar ambiguities are also present in photos. E.g. a part of fur visible under the sofa might belong to a cat or a dog, but it is not possible from the image alone to resolve this ambiguity. Most existing segmentation algorithms either provide only one likely consistent hypothesis (e.g., ‘all pixels belong to a cat’) or a pixel-wise probability (e.g., ‘each pixel is 50% cat and 50% dog’). Especially in medical applications where a subsequent diagnosis or a treatment depends on the segmentation map, an algorithm that only provides the most likely hypothesis might lead to misdiagnoses and sub-optimal treatment. Providing only pixel-wise probabilities ignores all co-variances between the pixels, which makes a subsequent analysis much more difficult if not impossible.

Here we present a segmentation framework that provides multiple segmentation hypotheses for ambiguous images (Fig. 7.1a). Our framework combines a **conditional Variational Auto-Encoder (cVAE)** [Jimenez Rezende et al., 2014, Kingma and Welling, 2013, Kingma et al., 2014, Sohn et al., 2015b] which can model complex distributions, with a U-Net [Ronneberger et al., 2015] which delivers state-of-the-art segmentations in many medical application domains. A low-dimensional latent space encodes the possible segmentation variants. A random sample from this space is injected into the U-Net to produce the corresponding segmentation map (see Fig. 7.1). One key feature of this architecture is the ability to model the joint probability of all pixels in the segmentation map. This results in multiple segmentation maps, where each of them provides a consistent interpretation of the whole image. Furthermore our framework is able to also learn hypotheses that have a low probability and to predict them with the corresponding frequency. We demonstrate these features on a lung abnormalities segmentation task [Armato et al., 2015, 2011, Clark et al., 2013], where each lesion has

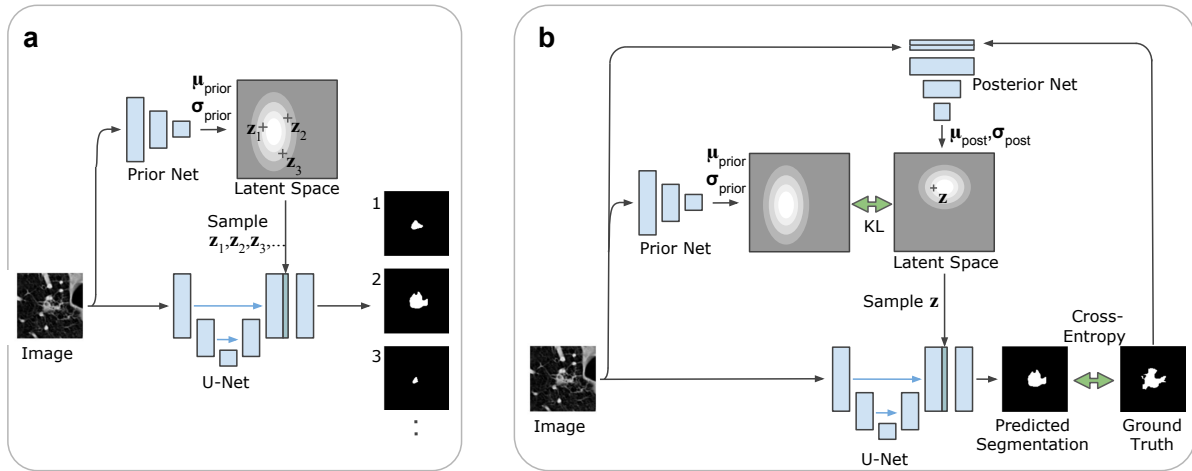


Figure 7.1 | The Probabilistic U-Net. (a) Sampling process. Arrows: flow of operations; blue blocks: feature maps. The heatmap represents the probability distribution in the low-dimensional latent space \mathbb{R}^N (e.g., $N = 6$ in our experiments). For each execution of the network, one sample $z \in \mathbb{R}^N$ is drawn to predict one segmentation mask. Green block: N -channel feature map from broadcasting sample z . The number of feature map blocks shown is reduced for clarity of presentation. (b) Training process illustrated for one training example. Green arrows: loss functions.

been segmented independently by four experts, and on the Cityscapes dataset [Cordts et al., 2016], where we artificially flip labels with a certain frequency during training.

Depending on the down-stream task, it may be beneficial or even required to have self consistent segmentation samples rather than e.g. pixel-wise samples at hand. For example a classifier can be trained to map segmentations to diagnoses, such as in [De Fauw et al., 2018], which naturally lends itself to propagate segmentation ambiguity to classifier uncertainty, if and when multiple consistent hypotheses are available. Equally, multiple plausible segmentations are arguably more readily interpreted by a clinician as opposed to pixel-wise uncertainty estimates. They could be used to suggest further diagnostic tests to resolve the ambiguities, or, when additional non-imaging information is available, the appropriate one(s) can be selected for subsequent steps such as further diagnosis or treatment planning and monitoring.

7.2 Related Work & Baselines

A body of work with different approaches towards probabilistic and multi-modal segmentation exists. An overview is given in Sec. 4.2 and we discuss the closely relate work in more detail below.

The most common approaches provide independent pixel-wise probabilities [Kendall and Gal, 2017, Kendall et al., 2015]. These models induce a probability distribution by using dropout over spatial, network-internal feature maps. While this strategy fulfills this line of work’s objective of quantifying the pixel-wise uncertainty, it produces inconsistent outputs. A simple way to produce plausible hypotheses is to learn an ensemble of (deep) models [Lakshminarayanan et al., 2017]. While the outputs produced by ensembles are consistent, they are not necessarily diverse and ensembles are typically not able to learn the rare variants as their members are trained independently. In order to overcome this, several approaches train models jointly using the oracle set loss [Guzman-Rivera et al., 2012], i.e. a loss that only accounts for the closest prediction to the ground truth. This has been explored in [Lee et al., 2015] and [Lee et al., 2016] using an ensemble of deep networks, and in [Rupprecht et al., 2017] and [Ilg et al., 2018] using one common deep network with M heads. While multi-head approaches may have the capacity to capture a diverse set of variants, they are not equipped to learn the occurrence frequencies of individual variants. Two common disadvantages of both ensembles and M heads models are their ungraceful scaling to large numbers of hypotheses, and their requirement of fixing the number of allowed hypotheses at training time. Another set of approaches to produce multiple diverse solutions relies on graphical models, such as junction chains [Chen et al., 2013], and more generally Markov Random Fields [Batra et al., 2012, Kirillov et al., 2015a,b, 2016]. While many of the previous approaches are guaranteed to find the best diverse solutions, these are confined to structured problems whose dependencies can be described by tractable graphical models.

The task of image-to-image translation [Isola et al., 2017] tackles a very similar problem: an under-constrained domain transfer of images needs to be learned. Many of the recent approaches employ GANs which are known to suffer from challenges such as ‘mode-collapse’ [Goodfellow, 2016]. In an attempt to solve the mode-collapse problem, the ‘bicycleGAN’ [Zhu et al., 2017b] involves a component that is similar in architecture to ours. In contrast to our proposed architecture, their model encompasses a fixed prior distribution and during training their posterior distribution is only conditioned on the output image. Recent work on generating appearances given a shape encoding [Esser et al., 2018] also combines a U-Net with a VAE, and was developed concurrently to ours. In contrast to our proposal, their training requires an additional pretrained VGG-net that is employed as a reconstruction loss. Finally, [Bouchacourt et al., 2016] proposed a probabilistic model for structured outputs based on optimizing the dissimilarity coefficient [Rao, 1982] between the ground truth and predicted distributions. The resultant approach is assessed on the task of hand pose estimation, that is, predicting the location of 14

joints, arguably a simpler space compared to the space of segmentations we consider here. Similarly to the approach presented below, they inject latent variables at a later stage of the network architecture.

7.3 Network Architecture and Training Procedure

Our proposed network architecture is a combination of a **conditional Variational Auto-Encoder** [Jimenez Rezende et al., 2014, Kingma and Welling, 2013, Kingma et al., 2014, Sohn et al., 2015b] with a U-Net [Ronneberger et al., 2015], with the objective of learning a conditional density model over segmentations, conditioned on the image.

7.3.1 Sampling

The central component of our architecture (Fig. 7.1a) is a low-dimensional latent space \mathbb{R}^N (e.g., $N = 6$, which performed best in our experiments). Each position in this space encodes a segmentation variant. The ‘prior net’, parametrized by weights ω , estimates the probability of these variants for a given input image X . This prior probability distribution (called P in the following) is modelled as an axis-aligned Gaussian with mean $\boldsymbol{\mu}_{\text{prior}}(X; \omega) \in \mathbb{R}^N$ and variance $\boldsymbol{\sigma}_{\text{prior}}(X; \omega) \in \mathbb{R}^N$. To predict a set of m segmentations we apply the network m times to the same input image (only a small part of the network needs to be re-evaluated in each iteration, see below). In each iteration $i \in \{1, \dots, m\}$, we draw a random sample $\mathbf{z}_i \in \mathbb{R}^N$ from P

$$\mathbf{z}_i \sim P(\cdot|X) = \mathcal{N}(\boldsymbol{\mu}_{\text{prior}}(X; \omega), \text{diag}(\boldsymbol{\sigma}_{\text{prior}}(X; \omega))) , \quad (7.1)$$

broadcast the sample to an N -channel feature map with the same shape as the segmentation map, and concatenate this feature map to the last activation map of a U-Net (the U-Net is parameterized by weights θ). A function $f_{\text{comb.}}$ composed of three subsequent 1×1 convolutions (ψ being the set of their weights) combines the information and maps it to the desired number of classes. The output, S_i , is the segmentation map corresponding to point \mathbf{z}_i in the latent space:

$$S_i = f_{\text{comb.}}(f_{\text{U-Net}}(X; \theta), \mathbf{z}_i; \psi) . \quad (7.2)$$

Notice that when drawing m samples for the same input image, we can reuse the output of the prior net and the feature activations of the U-Net. Only the function $f_{\text{comb.}}$ needs

to be re-evaluated m times.

7.3.2 Training

The networks are trained with the standard training procedure for conditional VAEs (Fig. 7.1b), i.e. by minimizing the **Evidence Lower Bound (ELBO)** (Eq. 7.4). The main difference with respect to training a deterministic segmentation model, is that the training process additionally needs to find a useful embedding of the segmentation variants in the latent space. This is solved by introducing a ‘posterior net’, parametrized by weights ν , that learns to recognize a segmentation variant (given the raw image X and the ground truth segmentation Y) and to map this to a position $\boldsymbol{\mu}_{\text{post}}(X, Y; \nu) \in \mathbb{R}^N$ with some uncertainty $\boldsymbol{\sigma}_{\text{post}}(X, Y; \nu) \in \mathbb{R}^N$ in the latent space. The output is denoted as posterior distribution Q . A sample \mathbf{z} from this distribution,

$$\mathbf{z} \sim Q(\cdot|X, Y) = \mathcal{N}(\boldsymbol{\mu}_{\text{post}}(X, Y; \nu), \text{diag}(\boldsymbol{\sigma}_{\text{post}}(X, Y; \nu))), \quad (7.3)$$

combined with the activation map of the U-Net ($f_{\text{U-Net}}$) must result in a predicted segmentation S (Eq. 7.2) identical to the ground truth segmentation Y provided in the training example. A cross-entropy loss penalizes differences between S and Y (the cross-entropy loss arises from treating the output S as the parameterization of a pixel-wise categorical distribution P_c). Additionally there is a **Kullback-Leibler divergence (KL)** $D_{\text{KL}}(Q||P) = \mathbb{E}_{z \sim Q} [\log Q - \log P]$ which penalizes differences between the posterior distribution Q and the prior distribution P . Both losses are combined as a weighted sum with a weighting factor β , as e.g. done in [Higgins et al., 2017]:

$$\mathcal{L}_{\text{ELBO}}(Y, X) = \mathbb{E}_{z \sim Q(\cdot|Y, X)} [-\log P_c(Y|S(X, z))] + \beta \cdot D_{\text{KL}}(Q(z|Y, X)||P(z|X)). \quad (7.4)$$

The training is done from scratch with randomly initialized weights. During training, this **KL** loss ‘pulls’ the posterior distribution (which encodes a segmentation variant) and the prior distribution towards each other. On average (over multiple training examples) the prior distribution will be modified in a way such that it ‘covers’ the space of all presented segmentation variants for a specific input image.

7.4 Performance Measures and Baseline Methods

In this section we first present the metric used to assess the performance of all approaches, and then describe each competitor approach used in the comparisons.

7.4.1 Performance Measures

As it is common in the semantic segmentation literature, we employ the [Intersection over Union \(IoU\)](#) as a measure to compare a pair of segmentations:

$$IoU(Y, Y') = \frac{|Y \cap Y'|}{|Y \cup Y'|} = \frac{TP}{TP + FP + FN}, \quad (7.5)$$

where [TP](#), [FP](#) and [FN](#) denote the number of [True Positive](#), [False Positive](#) and [False Negative](#) predictions between a predicted segmentation Y' and a ground truth segmentation Y .

In the present case however we not only want to compare a deterministic prediction with a unique ground truth, but rather we are interested in comparing distributions of segmentations. To do so, we use the [Generalized Energy Distance \(GED\)](#) [[Bellemare et al., 2017](#), [Salimans et al., 2018](#), [Székely and Rizzo, 2013](#)], which leverages distances between observations:

$$D_{\text{GED}}^2(P_{\text{gt}}, P_{\text{out}}) = 2\mathbb{E}[d(S, Y)] - \mathbb{E}[d(S, S')] - \mathbb{E}[d(Y, Y')], \quad (7.6)$$

where d is a distance measure, Y and Y' are independent samples from the ground truth distribution P_{gt} , and similarly, S and S' are independent samples from the predicted distribution P_{out} . The energy distance D_{GED} is a metric as long as d is also a metric [[Klebanov et al., 2005](#)]. In our case we choose $d(x, y) = 1 - \text{IoU}(x, y)$, which as proved in [[Kosub, 2016](#), [Lipkus, 1999](#)], is a metric. In practice, we only have access to samples from the distributions that models induce, so we rely on statistics of [Eq. 7.6](#), \hat{D}_{GED}^2 . The details about its computation for each experiment are presented in [Sec. B.1](#).

7.4.2 Baseline Methods

With the aim of providing context for the performance of our proposed approach we compare against a range of baselines. To the best of our knowledge there exists no other work that has considered capturing a distribution over multi-modal segmentations and has measured the agreement with such a distribution. For fair comparison, we train the baseline models whose architectures are depicted in [Fig. 7.2](#) in the exact same manner

as we train ours. The baseline methods all involve the same U-Net architecture, i.e. they share the same core component and thus employ comparable numbers of learnable parameters in the segmentation tasks.

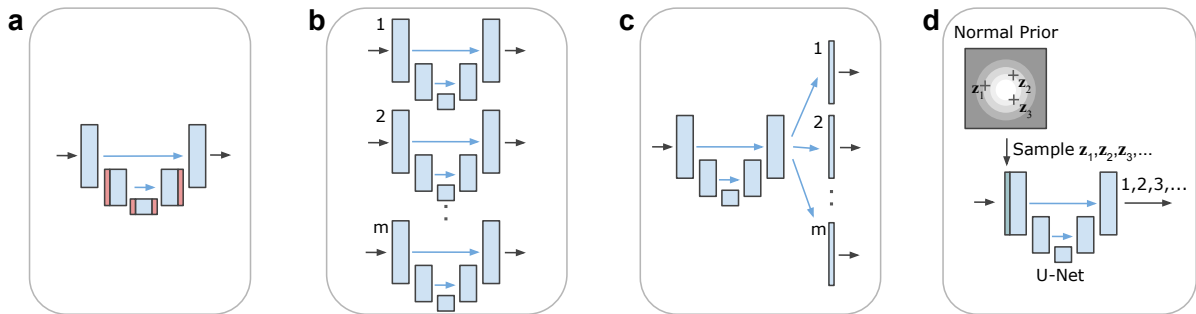


Figure 7.2 | Baseline architectures. Arrows: flow of operations; blue blocks: feature maps; red blocks: feature maps with dropout; green block broadcasted latents. Note that the number of feature map blocks shown is reduced for clarity of presentation. (a) Dropout U-Net. (b) U-Net Ensemble. (c) M-Heads. (d) Image2Image VAE.

Dropout U-Net (Fig. 7.2a). Our ‘Dropout U-Net’ baselines follow the Bayesian segnet’s [Kendall et al., 2015] proposition: we dropout the activations of the respective incoming layers of the three inner-most encoder and decoder blocks with a dropout probability of $p = 0.5$ during training as well as when sampling.

U-Net Ensemble (Fig. 7.2b). We report results for ensembles with the number of members matching the required number of samples (referred to as ‘U-Net Ensemble’). The original deterministic variant of the U-Net is the 1-sample corner case of an ensemble.

M-Heads (Fig. 7.2c). Aiming for diverse semantic segmentation outputs, the works of [Rupprecht et al., 2017] and [Ilg et al., 2018] propose to branch off M heads after the last layer of a deep net each of which contributes one output variant. An adjusted cross-entropy loss that adaptively assigns heads to ground-truth hypotheses is employed to promote diversity while reducing the risk of idle heads: the loss of the best performing head is weighted with a factor of $1 - \epsilon$, while the remaining heads each contribute with a weight of $\epsilon/(M - 1)$ to the loss. For our ‘M-Heads’ baselines we again employ a U-Net core and set $\epsilon = 0.05$ as proposed by [Rupprecht et al., 2017]. In order to allow for the evaluation of 4, 8 and 16 samples, we train M-Heads models with the corresponding number of heads.

Image2Image VAE (Fig. 7.2d). In [Zhu et al., 2017b] the authors propose a U-Net VAE-GAN hybrid for multi-modal image-to-image translation, that owes its stochasticity to normal distributed latents that are broadcasted and fed into the encoder path of the U-Net. In order to deal with the complex solution space in image-to-image translation

tasks, they employ an adversarial discriminator as additional supervision alongside a reconstruction loss. In the fully supervised setting of semantic segmentation such an additional learning signal is however not necessary and we therefore train with a cross-entropy loss only. In contrast to our proposition, this baseline, which we refer to as the ‘Image2Image VAE’, employs a prior that is not conditioned on the input image (a fixed normal distribution) and a posterior net that is not conditioned on the input either.

In all cases we examine the models’ performance when drawing a different number of samples (1, 4, 8 and 16) from each of them.

7.5 Quantitative Results

A quantitative evaluation of multiple segmentation predictions per image requires annotations from multiple labelers. Here we consider two datasets: The [LIDC-IDRI](#) dataset [[Armato et al., 2015, 2011](#), [Clark et al., 2013](#)] which contains 4 annotations per input, and the Cityscapes dataset [[Cordts et al., 2016](#)], which we artificially modify by adding *synonymous classes* to introduce uncertainty in the way concepts are labelled.

7.5.1 Lung Abnormalities Segmentation

The [Lung Image Database Consortium \(LIDC\)](#) and [Image Database Resource Initiative \(IDRI\)](#) dataset [[Armato et al., 2015, 2011](#), [Clark et al., 2013](#)] contains 1018 lung CT scans from 1010 lung patients with manual lesion segmentations from four experts. This dataset is a good representation of the typical ambiguities that appear in CT scans. For each scan, 4 radiologists (from a total of 12) provided annotation masks for lesions that they independently detected and considered to be abnormal. We use the masks resulting from a second reading in which the radiologists were shown the anonymized annotations of the others and were allowed to make adjustments to their own masks.

For our experiments we split this dataset into a training set composed of 722 patients, a validation set composed of 144 patients, and a test set composed of the remaining 144 patients. We then resampled the CT scans to $0.5 \text{ mm} \times 0.5 \text{ mm}$ in-plane resolution (the original resolution is between 0.461 mm and 0.977 mm, 0.688 mm on average) and cropped 2D images (180×180 pixels) centered at the lesion positions. The lesion positions are those where at least one of the experts segmented a lesion. By cropping the scans, the resultant task is in isolation not directly clinically relevant. However, this allows us to ignore the vast areas in which all labelers agree, in order to focus on those where there is uncertainty. This resulted in 8882 images in the training set, 1996 images in the

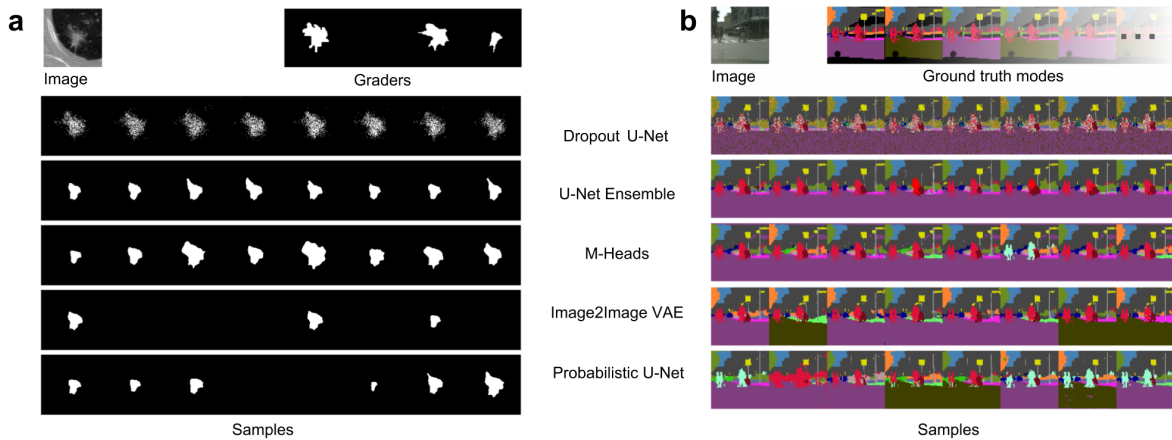


Figure 7.3 | Qualitative results. The first row shows the input image and the ground truth segmentations. The following rows show results from the baselines and from our proposed method. **(a)** lung CT scan from the LIDC test set. Ground truth: 4 graders. **(b)** Cityscapes. Images cropped to squares for ease of presentation. Ground truth: 32 artificial modes. Best viewed in colour.

validation set and 1992 images in the test set. Because the experts can disagree whether the lesion is abnormal tissue, up to 3 masks per image can be empty. Fig. 7.3a shows an example of such lesion-centered images and the masks provided by 4 graders.

As all models share the same U-Net core component and for fairness and ease of comparability, we let all models undergo the same training schedule, which is detailed in Subsec. B.5.1.

In order to grasp some intuition about the kind of samples produced by each model, we show in Fig. 7.3a, as well as in Sec. B.3, representative results for the baseline methods and our proposed Probabilistic U-Net. Fig. 7.4a shows the squared generalized energy distance \hat{D}_{GED}^2 for all models as a function of the number of samples. The data accumulations visible as horizontal stripes are owed to the existence of empty ground-truth masks. The energy distance on the 1992 images large lung abnormalities test set, decreases for all models as more samples are drawn indicating an improved matching of the ground-truth distribution as well as enhanced sample diversity. Our proposed Probabilistic U-Net outperforms all baselines when sampling 4, 8 and 16 times (numerical results can be found in Table B.1). The performance at 16 samples is found significantly higher than that of the baselines (p -value $\sim \mathcal{O}(10^{-13})$), according to the Wilcoxon signed-rank test. Finally, in Subsec. 7.7.3 we show the results of an experiment regarding the capacity different models have to distinguish between unambiguous and ambiguous instances (i.e. instances where graders disagree on the presence of a lesion).

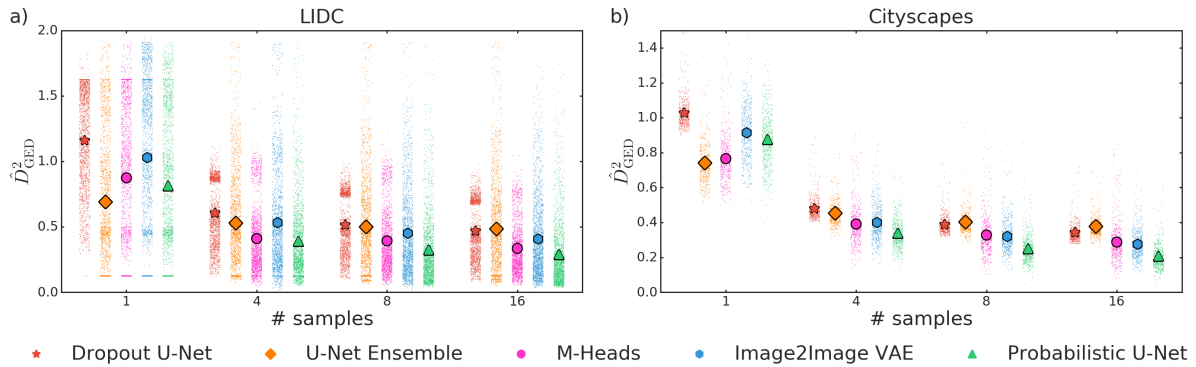


Figure 7.4 | Comparison of approaches using the GED. Lower energy distances correspond to better agreement between predicted distributions and ground truth distribution of segmentations. The symbols that overlay the distributions of data points mark the mean performance. **(a)** Performance on lung abnormalities segmentation on our LIDC-IDRI test-set. **(b)** Performance on the official Cityscapes validation set (our test set).

7.5.2 Stochastic Cityscapes Street Scene Segmentation

As a second dataset we use the Cityscapes dataset [Cordts et al., 2016]. It contains images of street scenes taken from a car with corresponding semantic segmentation maps. A total of 19 different semantic classes are labelled. Based on this dataset we designed a task that allows full control of the ambiguities: we create ambiguities by artificial random flips of five classes to newly introduced classes. We flip ‘sidewalk’ to ‘sidewalk 2’ with a probability of $8/17$, ‘person’ to ‘person 2’ with a probability of $7/17$, ‘car’ to ‘car 2’ with $6/17$, ‘vegetation’ to ‘vegetation 2’ with $5/17$ and ‘road’ to ‘road 2’ with probability $4/17$. This choice yields distinct probabilities for the ensuing $2^5 = 32$ discrete modes with probabilities ranging from 10.9% (all unflipped) down to 0.5% (all flipped). The official training dataset with fine-grained annotation labels comprises 2975 images and the validation dataset contains 500 images. We employ this official validation set as a test set to report results on, and split off 274 images (corresponding to the 3 cities of Darmstadt, Mönchengladbach and Ulm) from the official training set as our internal validation set. As in the previous experiment, in this task we use a similar setting for the training processes of all approaches, which we present in detail in Subsec. B.5.2.

Fig. 7.3b shows samples of each approach in the comparison given one input image. In Sec. B.4 we show further samples of other images, produced by our approach. Fig. 7.4b shows that the Probabilistic U-Net on the Cityscapes task outperforms the baseline methods when sampling 4, 8 and 16 times in terms of the energy distance (numerical results can be found in Table B.2). This edge in segmentation performance at 16 samples is highly significant according to the Wilcoxon signed-rank test [Wilcoxon, 1945] (p -value

$\sim \mathcal{O}(10^{-77})$). We have also conducted ablation experiments in order to explore which elements of our architecture contribute to its performance. These were (1) Fixing the prior, (2) Fixing the prior, and not using the context in the posterior and (3) Injecting the latent features at the beginning of the U-Net. Each of these variations resulted in a lower performance. Detailed results can be found in [Subsec. 7.7.2](#).

7.6 Qualitative Results

The embedding of the segmentation variants in a low-dimensional latent space allows a qualitative analysis of the internal representation of our model. For a 2D or 3D latent space we can directly visualize where the segmentation variants get assigned, see [Fig. 7.5](#) and [7.6](#).

The segmentation variants from the proposed Probabilistic U-Net correspond to latent space samples from the learned prior distribution. [Fig. 7.5](#) and [Fig. 7.6](#) below show samples from the Probabilistic U-Net for an LIDC-IDRI and a Cityscapes example respectively. The samples are arranged so as to represent their corresponding position in a 2D-plane of the respective latent space. This allows to interpret how the model ends up structuring the space to solve the given tasks.

7.6.1 Lung Abnormalities Segmentation

In the LIDC-IDRI case the z_0 -component of the prior happens to roughly encode lesion size including a transition to complete lesion absence. The probability mass allocated to absence is relatively small in the particular example, which arguably is in tune with the fact that 1 of the 4 graders assessed the image as lesion free. The z_1 -component on the other hand appears to encode shape variations. In the training, the posterior and the prior distribution are tied by means of the KL-divergence. As a consequence they ‘live’ in the same space and the graders (alongside the image to condition on) can be projected into the same latent space. [Fig. 7.5](#) shows the grader’s position in the form of green dots. All four graders map into the 3-sigma interval of the prior. One of the segmentations that indicates lesion presence and arguably delineates only the most salient region (grading number 3) is highly likely, i.e. within the 1-sigma isoprobability contour, under the prior. [Fig. 7.3](#) gives more LIDC-IDRI examples with their corresponding grader masks and 16 random samples of the Probabilistic U-Net. It appears that our model agrees very well with cases for which there is inter-grader disagreement on lesion presence. For cases where the graders agree on presence, our model at times apparently shows an

under-conservative prior, in the sense that uncertainty on presence can be elevated. The shape variations however are covered to a very good degree as attested by quantitative experiments above.

7.6.2 Street Scene Segmentation

In the Cityscapes task we employ a latent space with more dimensions than on the lung abnormalities task in order to equip the prior with sufficient capacity to encode the grader modes. The best performing model used a 6D latent space, however, for ease of presentation the following discusses the latent structure of a 3D latent space version. Fig. 7.6 shows a z_0 - z_1 plane of the latent space in which we again map corresponding segmentation samples, this time for a Cityscapes example. The precisely defined grader modes in the Cityscapes task can be identified with coherent regions in the latent space. As the space is 3D, not all 32 modes are fully manifest in the shown z_2 -slice. The location of the modes is shown via white mode numbers and the degree of transparency indicated the proximity in z_2 relative to the shown slice. As this particular task involves discrete modes, the semantically different regions are coherent and well confined as hoped for. There however inevitably are transitions between those latent space regions that will translate to mixtures of the grader modes that cross over. Ideally these transitions are as sharp as possible relative to the order of magnitude of the prior variance, which is arguably the case. Fig. B.9 shows Cityscapes examples with their corresponding grader masks and 16 random samples of the Probabilistic U-Net. The shown samples exhibit largely coherent variants alongside occasional variant mixtures that correspond to semantic cross overs in the latent space. As alluded to quantitatively before, the samples also appear to respect the grader variant frequencies, which are captured by structuring the latent-space under the prior in such fashion that the correct probability mass is allocated to the respective mode. In the upper boundary region of Fig. 7.6 improper samples are found that show miss-segmentations (although those are unlikely under the prior). The erroneously encoded modes found here are presumably attributable to the presence of inherent ambiguities in the dataset.

7.7 Additional Analyses

7.7.1 Calibration Analysis.

In the Cityscapes segmentation task, we can provide further analysis by leveraging our knowledge of the underlying conditional distribution that we have set by design. In

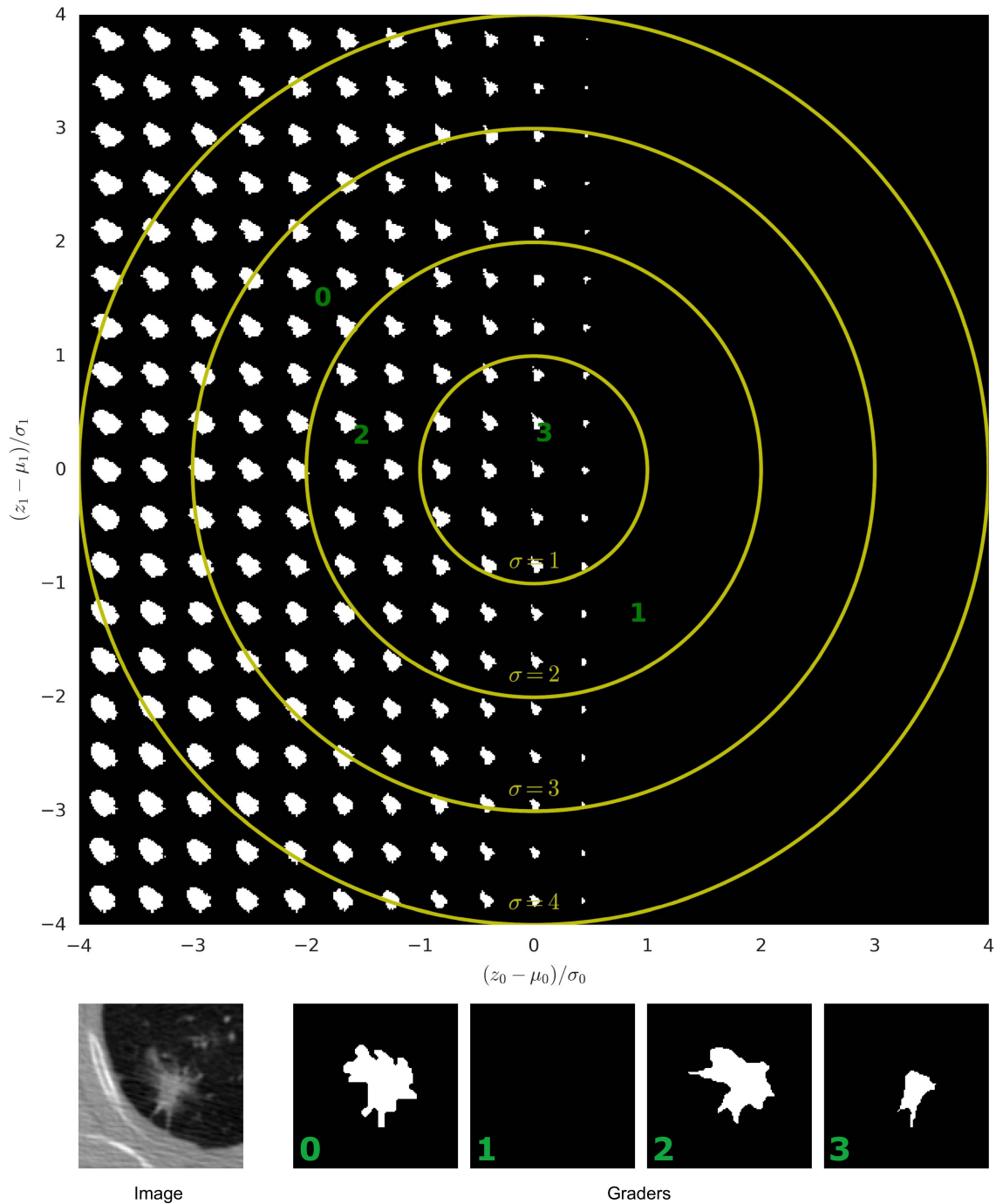


Figure 7.5 | Visualization of the latent space for the LIDC segmentation. 19×19 samples for a LIDC-IDRI test set example mapped to their prior latent-space position, using our model trained with a latent space of only 2 dimensions. For ease of presentation, the latent space is re-scaled so that the prior likelihood is a spherical unit-Gaussian. The isoprobable yellow circles denote deviations from the mean in sigma. The ground-truth grader masks' posterior position in this latent space is indicated by green numbers. The input image is shown in the lower left, to the right of it, the 4 grader masks are shown.

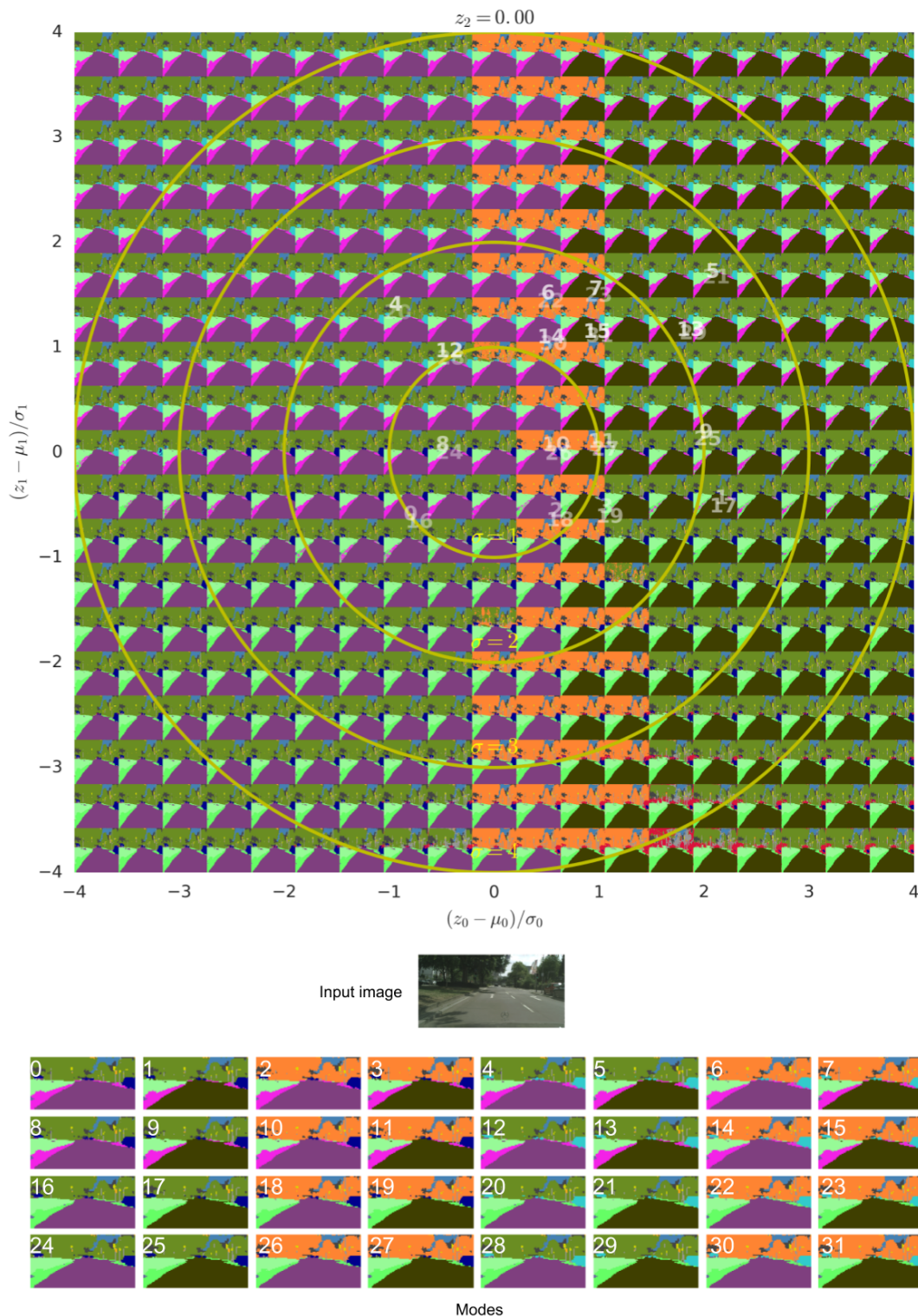


Figure 7.6 | Visualization of the latent space on Stochastic Cityscapes. 19×19 samples of a Cityscapes validation set example, mapped here to their latent-space position in the z_0 - z_1 plane ($z_2 = 0$) of the learned prior, using our model trained with a latent space of only 3 dimensions. For ease of presentation, the samples are squeezed to rectangles and the latent space is re-scaled so that the prior likelihood is a spherical unit-Gaussian. The isoprobable yellow circles denote deviations from the mean in sigma. The ground-truth grader masks' posterior position in this latent space is indicated by white numbers. (color-map as in Fig. B.9).

particular we compare the frequency with which every model predicts each mode, to the corresponding ground truth probability of that mode. To compute the frequency of each mode by each model, we draw 16 samples from that model for all images in the test set. Then we count the number of those samples that have that mode as the closest (using 1-IoU as the distance function).

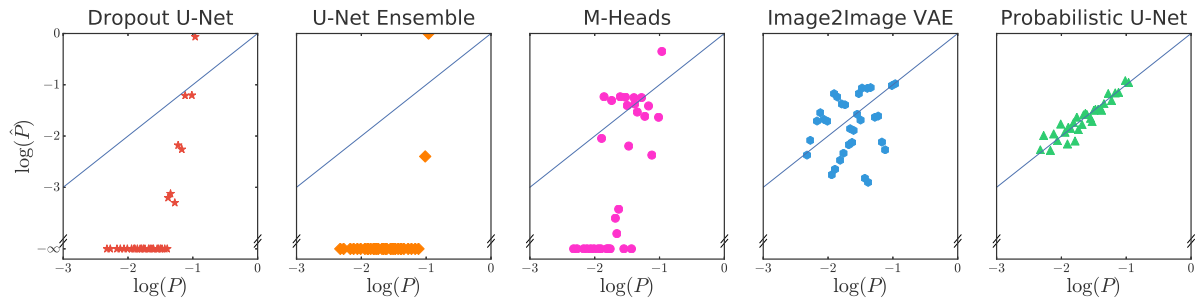


Figure 7.7 | Calibration of Mode Frequencies of the Probabilistic U-Net on Stochastic Cityscapes. The artificial flipping of 5 classes results in 32 modes with different ground truth probability (x-axis). The y-axis shows the frequency of how often the model predicted this variant in the whole test set. Agreement with the bisector line indicates calibration quality.

In Fig. 7.7 (and Figs. B.1, B.2, B.3 in Sec. B.2) we report the mode-wise frequencies for all 32 modes in the Cityscape task and show that the Probabilistic U-Net is the only model in this comparison that is able to closely capture the frequencies of a large combinatorial space of hypotheses including very rare modes, thus supplying calibrated likelihoods of modes. The Image2Image VAE is the only model among competitors that picks up on all variants, but the frequencies are far off as can be seen in its deviation from the bisector line in blue. The other baselines perform worse still in that all of them fail to represent modes and the modes they do capture do not match the expected frequencies.

7.7.2 Ablation Analysis

In this section we explore variations in the architecture of our approach, in order to understand how each design decision affects the performance. We have tried three variations over the original approach, these are:

Fixing the prior: Instead of making the prior a function of the context, here we fix it to be a standard Gaussian distribution.

Fixing the prior, and not using the context (input image) in the posterior: In addition to fixing the prior to be Gaussian, we also make the posterior a function of the ground truth mask only, ignoring the context.

Injecting the latent features at the beginning of the U-Net: Starting from our original model, we change the position in which the latent variables are used. Specifically here we concatenate them to the context (input image) and propagate that through the U-Net.

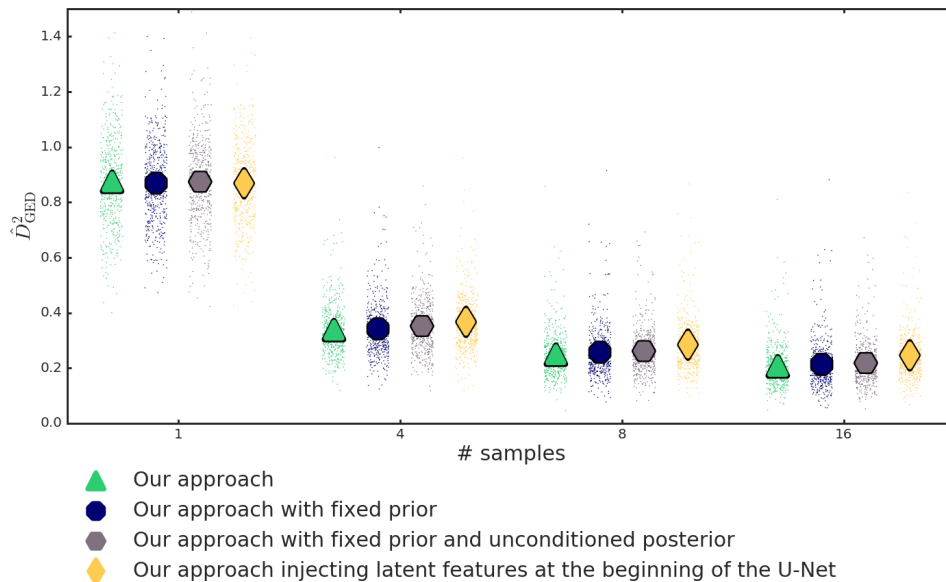


Figure 7.8 | Ablation analysis on Stochastic Cityscapes. Comparison of architectural variations of our approach using the energy distance. Lower energy distances correspond to better agreement between predicted distributions and ground truth distribution of segmentations. The symbols that overlay the distributions of data points mark the mean performance.

In Fig. 7.8 we can observe that our approach is better than the other variations. As the mechanisms that induce the distributions over segmentations during sampling and training are blinded towards the context image, the performance in terms of the IoU-based energy distance decreases. In particular, our model is much better than the variation that injects latent samples at the beginning. This is a pleasant finding, given that our decision of injecting the latent variables at the end of the U-Net was motivated by efficiency reasons when sampling. Here we find that we do not lose performance by doing so, but instead observe an improved matching of the samples with the ground-truth distribution. We hypothesize that injecting the latent variables at the final stage of the pipeline makes it easier for the model to account for different segmentations given the same input. This hypothesis is supported by the slightly better performance shown by the alternative architecture when sampling only once, and how this advantage is lost, and actually reversed, when sampling several times.

7.7.3 Predicting Ambiguity

In this section we assess the capacity of different models trained on LIDC for distinguishing between unambiguous and ambiguous instances. Specifically we define an instance to be ambiguous if 1 or more graders disagree on the presence of abnormal tissue. To do so, for each model we draw 16 samples per instance (as in all other experiments in the paper) and count the number of lesion presences out of the 16. This lesion presence is binned in two histograms with $[0, 16]$ bins, one for ambiguous and one for unambiguous instances (they are shown in Fig. 7.9). Finally we evaluate the discriminatory power of such histograms by computing the best threshold that separates ambiguous and unambiguous instances on the validation set. We present the accuracy scores on the test set in Table 7.1, which shows the advantage that our approach has over the competitors in this regard.

Table 7.1 | Predicting ambiguity for the presence of abnormalities. This table gives the accuracy for the prediction of whether graders disagree on the presence of an abnormality (evaluated on the test set). This prediction is made using a threshold on the number of non-empty samples which is determined on the validation set, see Fig. 7.9.

Dropout U-Net	U-Net Ensemble	M-Heads	Image2Image VAE	Probabilistic U-Net
0.328	0.699	0.678	0.678	0.736

7.8 Discussion

Our first set of experiments demonstrates that our proposed architecture provides consistent segmentation maps that closely match the multi-modal ground-truth distributions given by the expert graders in the lung abnormalities task and by the combinatorial ground-truth segmentation modes in the Cityscapes task. The employed IoU-based energy distance measures whether the models’ individual samples are both coherent as well as whether they are produced with the expected frequencies. It not only penalizes predicted segmentation variants that are far away from the ground truth, but also penalizes missing variants. On this task the Probabilistic U-Net is able to significantly outperform the considered baselines, indicating its capability to model the joint likelihood of segmentation variants.

The second type of experiments demonstrates that our model scales to complex output distributions including the occurrence of very rare modes. With 32 discrete modes of largely differing occurrence likelihoods (0.5% to 10.9%), the Cityscapes task requires

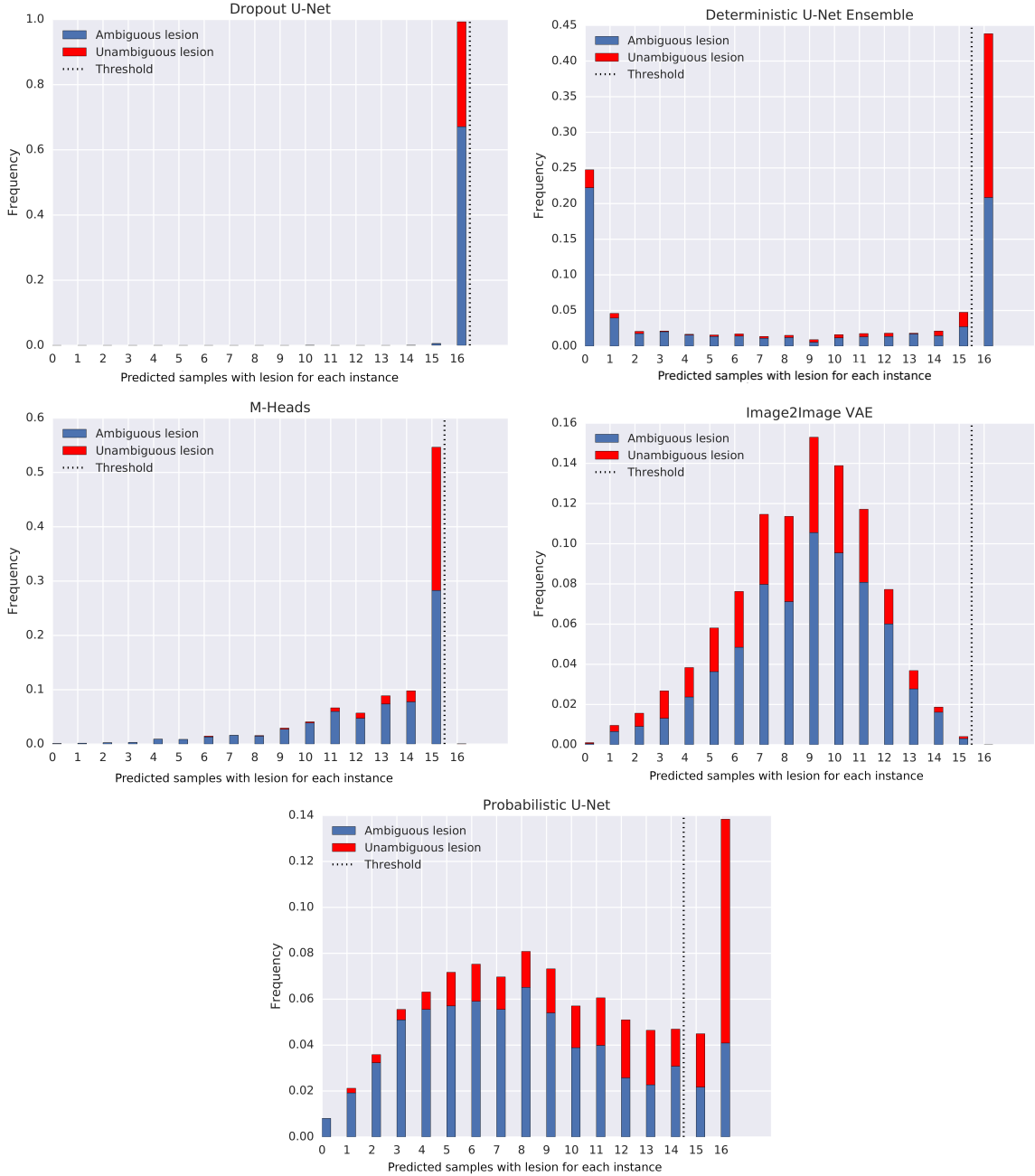


Figure 7.9 | Histograms of non-empty samples. The histograms result from 16 samples for each model and are evaluated on the validation set. For each model two histograms are produced, one for cases that are ambiguous (blue) and one for such that are non-ambiguous (red) with respect to abnormality presence, according to the set of ground truth annotations.

the ability to closely match complex data distributions. Here too our model performs best and picks the segmentation modes very close to the expected frequencies, all the way into the regime of very unlikely modes, thus defying mode-collapse and exhibiting excellent probability calibration. As an additional advantage our model scales to such large numbers of modes without requiring any prior assumptions on the number of modes or hypotheses.

The lower performance of the baseline models relative to our proposition can be attributed to design choices of these models. While the *Dropout U-Net* successfully models the pixel-wise data distribution (Fig. B.1a bottom right, in the Appendix), such pixel-wise mixtures of variants can not be valid hypotheses in themselves (see Fig. 7.3). The *U-Net Ensemble*'s members are trained independently and each of them can only learn the most likely segmentation variant as attested to by Fig. B.1b. In contrast to that the closely related *M-Heads* model can pick up on multiple discrete segmentation modes, due to the joint training procedure that enables diversity. The training does however not allow to correctly represent frequencies and requires knowledge of the number of present variants (see Fig. B.2a, in the Appendix). Furthermore neither the U-Net Ensemble, nor the M-Heads can deal with the combinatorial explosion of segmentation variants when multiple aspects vary independently of each other.

The *Image2Image VAE* shares similarities with our model, but as its prior is fixed and not conditioned on the input image, it can not learn to capture variant frequencies by allocating corresponding probability mass to the respective latent space regions. Fig. B.8 in the Appendix shows a severe miss-calibration of variant likelihoods on the lung abnormalities task that is also reflected in its corresponding energy distance. Furthermore, in this architecture, the latent samples are fed into the U-Net's encoder path, while we feed in the samples just after the decoder path. This design choice in the Image2Image VAE requires the model to carry the latent information all the way through the U-Net core, while simultaneously performing the recognition required for segmentation, which might additionally complicate training (see analysis in Subsec. 7.7.2).

Beside that, our design choice of late injection has the additional advantage that we can produce a large set of samples for a given image at a very low computational cost: for each new sample from the latent space only the network part after the injection needs to be re-executed to produce the corresponding segmentation map (this bears similarity to the approach taken in [Bouchacourt et al., 2016], where a generative model is employed to model hand pose estimation).

Aside from the ability to capture arbitrary modes with their corresponding probability conditioned on the input, our proposed *Probabilistic U-Net* allows to inspect its latent

space. This is because as opposed to e.g. GAN-based approaches, VAE-like models explicitly parameterize distributions, a characteristic that grants direct access to the corresponding likelihood landscape. Sec. 7.6 discusses how the Probabilistic U-Net chooses to structure its latent spaces.

Compared to aforementioned concurrent work for image-to-image tasks [Esser et al., 2018], our model disentangles the prior and the segmentation net. This can be of particular relevance in medical imaging, where processing 3D scans is common. In this case it is desirable to condition on the entire scan, while retaining the possibility to process the scan tile by tile in order to be able to process large volumes with large models with a limited amount of GPU memory.

On a more general note, we would like to remark that current image-to-image translation tasks only allow subjective (and expensive) performance evaluations, as it is typically intractable to assess the entire solution space. For this reason surrogate metrics such as the inception score based on the evaluation via a separately trained deep net are employed [Salimans et al., 2016]. The task of multi-modal semantic segmentation, which we consider here, allows for a direct and thus perhaps more meaningful manner of performance evaluation and could help guide the design of future generative architectures.

All in all we see a large field where our proposed Probabilistic U-Net can replace the currently applied deterministic U-Nets. Especially in the medical domain, with its often ambiguous images and highly critical decisions that depend on the correct interpretation of the image, our model's segmentation hypotheses and their likelihoods could 1) inform diagnosis/classification probabilities or 2) guide steps to resolve ambiguities. Our method could prove useful beyond explicitly multi-modal tasks, as the inspectability of the Probabilistic U-Net's latent space could yield insights for many segmentation tasks that are currently treated as a uni-modal problem.

Chapter 8

Learning Multi-Scale Distributions over Segmentations

In [Chap. 7](#) we proposed a generative model that can induce a global distribution over semantic segmentations. This image-global approach however has limitations in that it can not efficiently model more local factors of variation. A common example for such a case may be the presence of multiple lesions in a scan. Speaking more generally, medical images can exhibit regions whose ambiguity may inter-depend in complex ways and across various scales, spanning all the way from the pixel to the image level.

To improve upon our earlier model, a more flexible generative model is required, such as the [Hierarchical Probabilistic U-Net \(HPU-Net\)](#), which we describe in this chapter. The main contributions of this work are:

- A generative model for semantic segmentation able to learn complex-structured conditional distributions equipped with a latent space that scales with image size.
- Compared to prior art, strongly improved fidelity to fine structure in the models' samples and reconstructions.
- Improved modelling of distributions over segmentations including independently varying scales and locations, as demonstrated in its ability to generate instance segmentations.
- Automatic learning of factors of variations across space and scale.

This work was presented as an oral at the Med-NeurIPS Workshop at *Advances in Neural Information Processing Systems* and is available under [[Kohl et al., 2019](#)], which this chapter follows closely. We provide an open-source re-implementation of our

approach at https://github.com/deepmind/deepmind-research/tree/master/hierarchical_probabilistic_unet.

The key idea presented here, is the usage of a hierarchical latent space, that is weaved into the U-Net’s decoder structure. This way there is a top-down inter-dependency between latent samples that allows for a very flexible but coordinated generation of appropriate semantic segmentation maps.

8.1 The Need for a More Flexible Model

As discussed in [Sec. 7.2](#) several algorithms have been proposed that provide samples from the output distribution (here: consistent segmentation maps instead of pixel-wise samples). They are based on ensembles [[Lakshminarayanan et al., 2017](#)], networks with multiple heads [[Batra et al., 2012](#), [Lee et al., 2015, 2016](#), [Rupprecht et al., 2017](#)], or image-conditional generative models [[Isola et al., 2017](#), [Liu et al., 2017](#), [Park et al., 2019](#), [Zhu et al., 2017a,b](#)] such as cVAEs [[Esser et al., 2018](#), [Jimenez Rezende et al., 2014](#), [Kingma and Welling, 2013](#), [Kingma et al., 2014](#), [Sohn et al., 2015a](#)].

The Probabilistic U-Net introduced a number of significant improvements over these approaches (see contributions in [Chap. 7](#)). In practice it however shares a shortcoming with the aforementioned prior work: Depending on the model, they (can) work well for a single object in the image or for other global variations (like different segmentation styles, e.g. more narrow or more inclusive outlining), but do not scale to images containing multiple objects with uncorrelated variations.

In the case of the Probabilistic U-Net, this is because the image global latent space that it employs does not have any explicit spatial correspondence for the structure(s) and the segment(s) that it models, e.g. see [Fig. 7.1](#). In theory it should be able to learn to parameterize such mappings, in practice however a global latent vector proves to be a strong constraint on the model, even when allotting many dimensions to the used latent space, as we show below. This is undesirable, as in a more general case, the interpretations of different regions in a medical image can vary in complex ways including conditional independence, e.g. a [CT](#) scan may show several lesions with independent possible interpretations. There may further be conditional top-down dependencies, e.g. a patient’s genetic predisposition for a certain disease may alter the interpretation of the scan as a whole with effects for more local annotation decisions, based on the presence of indiscernible tissue types, or fuzzy borders at different scales.

In order to model complex high-dimensional data such as images, expressive models with the power to model complex interactions between elements are required. One way

to model rich interactions is by means of hierarchical generative models. The idea of employing a hierarchical approach for the unconditional generation of images has entered all main strands of generative models by now: GANs, employed for large scale image synthesis, have been endowed with the ability to decompose and employ a conditional input such that it modulates layers at different depths and scales of the generator [Brock et al., 2018, Karras et al., 2018, Park et al., 2019]. Autoregressive models such as PixelCNN [Van den Oord et al., 2016] have been stacked across different image scales to form hierarchical generative models with improved log-likelihood performance and sample coherence [De Fauw et al., 2019, Menick and Kalchbrenner, 2018, Salimans et al., 2017]. Various examples of VAEs [Jimenez Rezende et al., 2014, Kingma and Welling, 2013, Kingma et al., 2014] equipped with hierarchical latent spaces are reported in the literature: ConvDRAW [Gregor et al., 2016] performs an iterative inference using two layers of spatially arranged Gaussian latents, the Ladder VAE [Sønderby et al., 2016] employs up to 5 latent scales and both the Inverse Autoregressive Flow VAE [Kingma et al., 2016] and BIVA [Maaløe et al., 2019] make use of bi-directional inference respectively using up to 8 and 15 latent scales.

With the exception of one piece of concurrent work (see below), hierarchical generative models have not been adapted to image-conditional tasks such as image-to-image translation or semantic segmentation before.

We propose the **Hierarchical Probabilistic U-Net** and benchmark it against what we refer to as the **standard Probabilistic U-Net (sPU-Net)** from hereon (to distinguish it from the hierarchical model) and demonstrate the improved quality of the segmentations on LIDC, e.g. the sPU-Net often produces only ‘blobby’ segmentation samples (low segmentation fidelity). Furthermore we show the ability of the model to learn highly complex probability distributions, by presenting an instance segmentation task, where we ask the model to label (‘colorize’) each instance consistently with a random instance id. We test this ability on neuronal structures in **Electron Microscopy (EM)** images (a dataset called **SNEMI3D** [Kasthuri et al., 2015]) as well as on car instances in natural images (Cityscapes [Cordts et al., 2016]). Finally we show that the model is also capable of predicting consistent segmentations with corresponding uncertainties in a blacked out region of the image. In a medical application this could be used to predict disease progression by applying a 4D version of the proposed network to time series (3 spatial axes and 1 time axis), where the blacked out part corresponds to the unknown future development of the disease.

8.2 Network Architecture and Learning Objective

The sPU-Net models segmentation ambiguities using a low-dimensional, image global latent vector, that is sampled from a separate ‘prior net’ and is combined with U-Net features by means of a shallow network of 1×1 -convolutions [Kohl et al., 2018]. As we show below, this image-global latent space heavily constrains the granularity at which the output space can be modelled. While our proposed architecture also combines a U-Net with a cVAE, it instead employs a hierarchical latent space that resides in the U-Net’s decoder, as illustrated in Fig. 8.1. A hierarchical decomposition yields a much more flexible generative model that can further easily model top-down dependencies. E.g. the global part can model the patient’s genetic predisposition for a certain disease, while the local parts can model indiscernible tissue types, or fuzzy borders at different scales. The spatial arrangement of the latent variables further enables the network to easily model local independent variations (like multiple lesions). Due to the fully-convolutional architecture, it can also generalize from few to many lesions at arbitrary locations. Beside these fundamental extensions, we additionally removed the separate prior net and instead use U-Net internal features to predict the parameters of the prior distributions (as in [Esser et al., 2018]), which results in parameter and run-time savings. For the network to employ the full hierarchy, we further found it crucial to minimize obstructions between latent scales by introducing (pre-activated) residual blocks [He et al., 2016b] (as discussed in Sec. C.2 and in line with [Kingma et al., 2016, Maaløe et al., 2019]).

8.2.1 Sampling

The architecture’s main feature is its highly flexible parameterization of the conditional prior that it employs. This prior is composed of a) a deterministic feature extractor that computes features at spatial resolutions up to scale L (counted with ascending resolution) for the given input image X and b) a cascade of distributions interleaved with the U-Net’s decoder, that allows to hierarchically sample latents. In a conventional U-Net, the U-Net decoder’s features of every resolution are up-sampled and then concatenated with the features of the U-Net’s encoder from the respective resolution above [Ronneberger et al., 2015]. In our proposed architecture there is one additional step at each scale of the latent hierarchy: Conditioned on the decoder features of each scale $i \leq L$, we sample a spatial grid of latents \mathbf{z}_i and concatenate it with the input decoder features, before the usual up-sampling and concatenation with encoder features from above takes place, see Fig. 8.1a. The latents of each scale i thus depend on the input image X and on all latents of scales $i' < i$ that have already been sampled lower in the hierarchy, which we

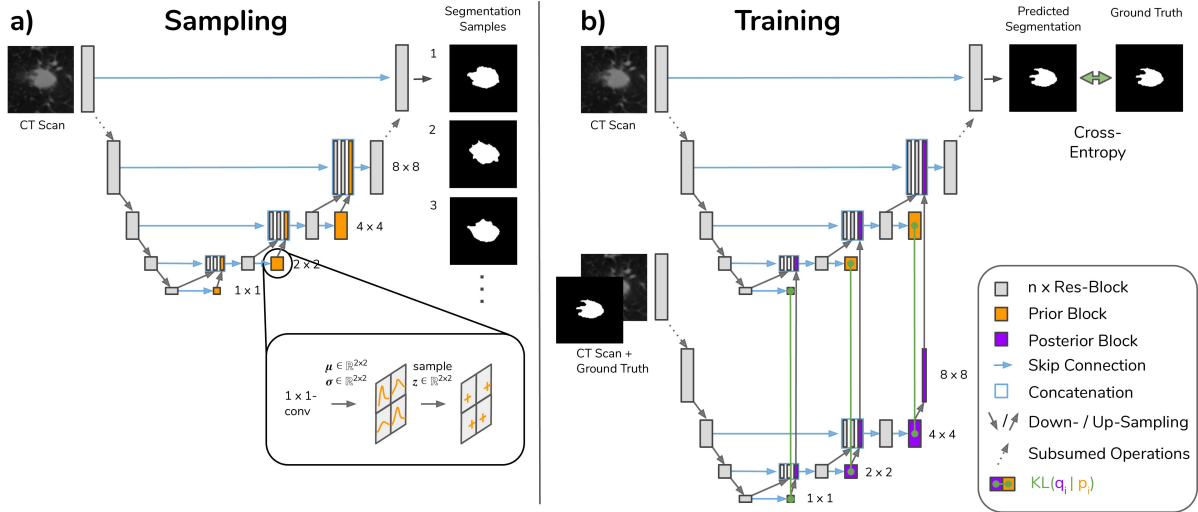


Figure 8.1 | The Hierarchical Probabilistic U-Net. The model is based on a U-Net and adds a hierarchy of spatially arranged Gaussian distributions that is interleaved with the U-Net’s decoder. (a) Sampling process: For each iteration of the network, latents \mathbf{z}_i at scale i (slim orange blocks) are successively sampled from the prior when going up the hierarchy towards increasing resolutions. (b) Training process illustrated for one training example: During training samples \mathbf{z}_i from the posterior (slim purple blocks) are injected into the U-Net’s decoder and used to reconstruct a given segmentation. Green connections: loss functions. For more details see [Sec. 8.2](#) and [Sec. C.2](#).

collectively denote as $\mathbf{z}_{<i} := (\mathbf{z}_{i-1}, \dots, \mathbf{z}_0)$. At each scale with spatial dimensions $H_i \times W_i$ the model uses conditional Gaussian distributions with mean $\boldsymbol{\mu}_i^{\text{prior}} \in \mathbb{R}^{H_i \times W_i}$ and variance $\boldsymbol{\sigma}_i^{\text{prior}} \in \mathbb{R}^{H_i \times W_i}$. The means and variances are predicted by 1×1 -convolutions for each spatial position of that scale. Sampling from the corresponding Gaussian distribution results in the spatial latents $\mathbf{z}_i \in \mathbb{R}^{H_i \times W_i}$:

$$\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i^{\text{prior}}(\mathbf{z}_{<i}, X), \boldsymbol{\sigma}_i^{\text{prior}}(\mathbf{z}_{<i}, X)) =: p(\mathbf{z}_i | \mathbf{z}_{<i}, X). \quad (8.1)$$

Our experiments did not benefit from going beyond scalar latents at each spatial location, which however is a choice that one might want to make depending on the application. The hierarchical (ancestral) sampling results in a joint distribution for the prior that decomposes as follows:

$$P(\mathbf{z}_0, \dots, \mathbf{z}_L | X) = p(\mathbf{z}_L | \mathbf{z}_{<L}, X) \cdot \dots \cdot p(\mathbf{z}_0 | X). \quad (8.2)$$

Every run of the network yields a segmentation hypothesis $Y' = S(X, \mathbf{z})$ for the given image (where $\mathbf{z} = (\mathbf{z}_L, \dots, \mathbf{z}_0)$ and S stands for the segmentation network), which is illustrated in [Fig. 8.1a](#). Note that only the U-Net’s decoder (including the hierarchical

sampling) needs to be rerun to produce the next segmentation samples for the same image. The number of latent scales L is a hyper-parameter and typically chosen smaller than the full number of scales of the U-Net; our models use $L = 3$ (4 scales).

8.2.2 Training

As is standard practice for VAEs, the training procedure aims at maximizing the so-called evidence lower bound (ELBO) on the likelihood $p(Y|X)$, where in our case Y is a segmentation and X is an image, as was the case for the sPU-Net, see Eq. 7.4. This requires to model a variational posterior $Q(\cdot|X, Y)$ that depends on both X and Y . By choice, the structure matches with that of the prior:

$$\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i^{\text{post}}(\mathbf{z}_{<i}, X, Y), \boldsymbol{\sigma}_i^{\text{post}}(\mathbf{z}_{<i}, X, Y)) =: q(\mathbf{z}_i|\mathbf{z}_{<i}, X, Y), \quad (8.3)$$

$$Q(\mathbf{z}_0, \dots, \mathbf{z}_L|X, Y) = q(\mathbf{z}_L|\mathbf{z}_{<L}, X, Y) \cdot \dots \cdot q(\mathbf{z}_0|X, Y). \quad (8.4)$$

The posterior Q is modeled in form of a separate network with the same hierarchical topology in which for each scale $i \leq L$, we compute conditional Gaussian distributions with mean $\boldsymbol{\mu}_i^{\text{post}} \in \mathbb{R}^{H_i \times W_i}$ and variance $\boldsymbol{\sigma}_i^{\text{post}} \in \mathbb{R}^{H_i \times W_i}$. During training, samples $\mathbf{z} \sim Q$ are fed into the U-Net's decoder (as illustrated in the bottom half of Fig. 8.1b) with the aim of learning to reconstruct the given input segmentation Y . The reconstruction objective (\mathcal{L}_{rec}) is formulated as a cross-entropy loss between the prediction Y' and the target Y (below formulated as a pixel-wise categorical distribution P_c). Additionally there is a Kullback-Leibler divergence $D_{\text{KL}}(Q||P) = \mathbb{E}_{\mathbf{z} \sim Q} [\log Q - \log P]$, that assimilates P and Q , just as for the sPU-Net, see Eq. 7.4. The KL-divergence terms for our choice of posterior and prior topology come about as follows:

$$D_{\text{KL}}(Q||P) = \mathbb{E}_{\mathbf{z} \sim Q} [\log Q - \log P] \quad (8.5)$$

$$= \int_{\mathbf{z}_0, \dots, \mathbf{z}_L} \prod_{j=0}^L q(\mathbf{z}_j|\mathbf{z}_{<j}) \sum_{i=0}^L [\log q(\mathbf{z}_i|\mathbf{z}_{<i}) - \log p(\mathbf{z}_i|\mathbf{z}_{<i})] d\mathbf{z}_0 \dots d\mathbf{z}_L, \quad (8.6)$$

$$\text{using } \int \phi(\mathbf{z}_i) \prod_{j=0}^L q(\mathbf{z}_j|\mathbf{z}_{<j}) d\mathbf{z}_0 \dots d\mathbf{z}_L = \int \phi(\mathbf{z}_i) \prod_{j=0}^i q(\mathbf{z}_j|\mathbf{z}_{<j}) d\mathbf{z}_0 \dots d\mathbf{z}_i \quad (8.7)$$

$$= \sum_{i=0}^L \int_{\mathbf{z}_0, \dots, \mathbf{z}_i} \prod_{j=0}^i q(\mathbf{z}_j|\mathbf{z}_{<j}) [\log q(\mathbf{z}_i|\mathbf{z}_{<i}) - \log p(\mathbf{z}_i|\mathbf{z}_{<i})] d\mathbf{z}_0 \dots d\mathbf{z}_i \quad (8.8)$$

$$= \sum_{i=0}^L \int_{\mathbf{z}_0, \dots, \mathbf{z}_i} \prod_{j=0}^{i-1} q(\mathbf{z}_j | \mathbf{z}_{<j}) q(\mathbf{z}_i | \mathbf{z}_{<i}) [\log q(\mathbf{z}_i | \mathbf{z}_{<i}) - \log p(\mathbf{z}_i | \mathbf{z}_{<i})] d\mathbf{z}_0 \dots d\mathbf{z}_i \quad (8.9)$$

$$= \sum_{i=0}^L \mathbb{E}_{\mathbf{z}_{<i} \sim Q} D_{\text{KL}}(q(\mathbf{z}_i | \mathbf{z}_{<i}) || p(\mathbf{z}_i | \mathbf{z}_{<i})), \quad (8.10)$$

where for improved clarity we omit X and X, Y as conditional arguments to p and q in the derivation above. For brevity our notation additionally subsumes $q(\mathbf{z}_0) := q(\mathbf{z}_0 | \mathbf{z}_{-1})$ and similar for $p(\mathbf{z}_0)$. For our choice of posterior and prior distribution (see Eq. 8.1 and 8.3) the KL-terms above can be evaluated analytically. The expectations in Eq. 8.11 and 8.12 using samples $\mathbf{z} \sim Q$ are performed with a single sampling pass. This yields the following ELBO objective:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{\mathbf{z} \sim Q} [-\log P_c(Y | S(X, \mathbf{z}))] + \beta \cdot \sum_{i=0}^L \mathbb{E}_{\mathbf{z}_{<i} \sim Q} D_{\text{KL}}(q_i(\mathbf{z}_i | \mathbf{z}_{<i}, X, Y) || p_i(\mathbf{z}_i | \mathbf{z}_{<i}, X)). \quad (8.11)$$

Minimizing $\mathcal{L}_{\text{ELBO}}$ leads to sub-optimally converged priors in our experiments. For this reason we make use of the recently proposed *GECCO*-objective [Rezende and Viola, 2018] that adds in a constraint on the reconstruction term and thus dynamically balances it with the KL terms from above:

$$\mathcal{L}_{\text{GECCO}} = \lambda \cdot \left(\mathbb{E}_{\mathbf{z} \sim Q} [-\log P_c(Y | S(X, \mathbf{z}))] - \kappa \right) + \sum_{i=0}^L \mathbb{E}_{\mathbf{z}_{<i} \sim Q} D_{\text{KL}}(q_i(\mathbf{z}_i | \mathbf{z}_{<i}, X, Y) || p_i(\mathbf{z}_i | \mathbf{z}_{<i}, X)), \quad (8.12)$$

where κ is chosen as the desired reconstruction loss value and the Lagrange multiplier λ is updated as a function of the exponential moving average of the reconstruction constraint. This formulation initially puts high pressure on the reconstruction and once the desired κ is reached it increasingly moves the pressure over on the KL-term. For more details we refer to Sec. C.2 and the literature [Rezende and Viola, 2018].

We additionally perform an online hard-negative mining, specifically, we only back-propagate the gradient of the k th percentile of the worst pixels of the batch [Wu et al., 2016a], $\mathcal{L}_{\text{rec}} \rightarrow \text{top_k_mask}(\mathcal{L}_{\text{rec}})$. We chose $k = 0.02$ (the worst 2% pixels) in all experiments of the *HPU-Net* and stochastically pick the k th percentile [Nikolov et al., 2018] (we sample from a Gumbel-Softmax distribution [Jang et al., 2016] over \mathcal{L}_{rec} per pixel).

8.2.3 Architecture and Training in more Detail

Architecture The **HPU-Net** implementations employed for the different tasks differ in their number of processing scales, the depth of each such scale and the number of latent scales employed. At each processing scale of the **HPU-Net** we employ a stack of n pre-activated residual blocks [He et al., 2016b] (grey blocks in Fig. 7.1), where n determines the depth of that scale. For both the LIDC and SNEMI3D experiments we use $n = 3$ residual blocks and for the Cityscapes experiment we use $n = 2$ residual blocks at each processing scale of the U-Net’s encoder and decoder respectively. Similar to [Kingma et al., 2016, Maaløe et al., 2019], we find the use of unobstructed connections (in our case residual blocks) between latent scales of the hierarchy to be crucial for the lower scales to be employed by the generative model. Without the use of residual blocks the KL-terms between distributions (indicated by green connecting lines in Fig. 8.1) at the beginning of the hierarchy often become ~ 0 early on in the training, essentially resulting in uninformative and thus unused latents.

In each residual block the residual feature map is calculated by means of a series of three 3×3 -convolutions, the first of which always halves the number of the feature maps employed at the present scale, such that the residual representations live on a lower dimensional manifold. At the end of the residual branch a single (un-activated) 1×1 -convolution projects the features back to the number of features of the given scale. The resulting residual is then added to the skipped feature map, which is skipped forward (i.e. left untouched) unless the number of feature maps is set to change, in which case it is projected by a 1×1 -convolution. This happens only at transitions that change the feature map resolutions. For down-sampling of feature maps we use average pooling and upsample by using nearest neighbour interpolation. As described in Sec. 8.2, the spatial grid of latent variables is sampled at the end of each U-Net decoder scale that is part of the hierarchy and concatenated to the final feature map produced at this scale, before both are up-sampled.

The number of latent scales is chosen empirically such as to allow for a sufficiently granular effect of the latent hierarchy. For the tasks and image resolutions considered here, we found 3 - 5 latent scales to work well. The number of processing scales is chosen such that a smallest possible spatial resolution is achieved in the bottom of the U-Net. For the square images in LIDC and SNEMI3D this means a resolution of 1×1 and for the Cityscapes task the minimum resolution is 1×2 (in this case we however employ 2×4 , which is detailed below). The employed separate posterior mirrors the number of scales and the number of feature maps of the corresponding components in the U-Net, see the bottom part of Fig. 8.1b. Its only architectural difference is its first

convolutional layer, which processes the input image concatenated with the corresponding one-hot segmentation along the channel axis. All weights of all models are initialized with orthogonal initialization having the gain (multiplicative factor) set to 1, and the bias terms are initialized by sampling from a truncated normal with $\sigma = 0.001$.

Training The HPU-Net is trained using the GECO-objective (Eq. 8.12) and a stochastic top-k reconstruction loss. As described in Sec. 8.2, the k th percentile employed for the top-k objective is fixed across tasks to 2% of each batch’s pixels. The GECO-objective aims at matching a reconstruction target value κ . For each experiment we chose κ sufficiently low so as to correspond to a strong reconstruction performance while resulting in a training schedule that is not dominated by the reconstruction term over the entire course of the training (e.g. if κ is chosen too high, the Lagrange multiplier λ , and thus the learning pressure it exerts, mounts and remains on the reconstruction term rather than moving over on the KL terms). The desired behavior of the reconstruction objective \mathcal{L}_{rec} and the Lagrange multiplier λ can be observed in Fig. C.1 and Fig. C.2, where λ rises until \mathcal{L}_{rec} matches κ , after which λ drops and the pressure on the KL-terms increases.

In contrast to the regular cross-entropy employed in semantic segmentation, the reconstruction error in Eq. 7.4, 8.11 and 8.12 is not averaged but summed over individual pixels (before being averaged across batch instances). This is because the likelihood is assumed to factorize over the pixels of an image and so their log-likelihood is summed over. For comparability we however report \mathcal{L}_{rec} and κ per pixel (e.g. in Fig. C.1, Fig. C.2 and in Table 8.2).

The precise training setups for each of the tasks and models are reported in Sec. C.2. Note that the training objectives for all models encompass an additional weight-decay term that is weighted by a factor of $1e^{-5}$.

8.3 Dataset Details

LIDC-IDRI We again use the LIDC-LIDC dataset [Armato et al., 2015, 2011, Clark et al., 2013] in exactly the same fashion as described in Subsec. 7.5.1 and Subsec. B.5.1.

LIDC-IDRI subset B For ‘Subset B’ we consider only those test set cases, which have annotations by all 4 graders, i.e. all graders agree on the presence of an abnormality. This results in 638 images, so close to a third of the full test set.

SNEMI3D As a second dataset we use the [SNEMI3D challenge](#)¹ dataset that is comprised of a fully annotated 3D block of a sub-volume of mouse neocortex, imaged slice by slice with an electron microscope [Kasthuri et al., 2015]. This stack is $1024 \times 1024 \times 100$ voxels large, comes at a voxel size of $6 \times 6 \times 29 \text{ nm}^3$ and contains a total of 400 fully annotated neurite instance annotations. We use the first 80 z-slices as our training dataset, the adjacent 10 slices as a validation set and the remaining 10 slices as a test set to report results on. We crop non-overlapping patches of size $256 \times 256 \times 1$ resulting in 1280 images for training, 160 for validation and 160 for testing. During training we randomly map the instance identifiers (ids) of the cells to one of 15 labels, thereby treating the instance id of the cells as latent information that the networks need to model. Because the number of individual cells per image can surmount this number, the training task does not necessitate a unique predicted instance id for every cell. This means that in order to obtain a predicted instance segmentation at test time, we need to aggregate a number of samples for a given image. For this purpose we propose a greedy Hamming distance [Hamming, 1950] based clustering across n samples followed by a light-weight post-processing, detailed in [Algorithm 1](#) below and [Sec. C.1](#) (we chose $n = 16$ and Hamming distance threshold $\alpha = 16$).

Cityscapes As a third dataset we use the Cityscapes street scene dataset that comes with both dense category segmentations, as well as with instance segmentations for a number of categories. General information on the dataset can be found in [Subsec. 7.5.2](#), where it was used as an experiment for the [sPU-Net](#). We again employ the official validation set of 500 images as a test set to report results on, and split off 274 images (corresponding to the 3 cities of Darmstadt, Mönchengladbach and Ulm) from the official training set as an internal validation set. As opposed to the Experiments in [Chap. 7](#) we do not randomly flip semantic segmentation classes. Instead, we randomly sample car instance segmentation classes while keeping the remaining 18 semantic segmentation classes unaltered, which is described in more detail below. At test time, we cluster 32 samples per image (see [Algorithm 1](#)), using a threshold of $\alpha = 32$.

8.4 Performance Measures

8.4.1 Distribution Agreement

As introduced before (see [Eq. 7.6](#)) we report how well the distribution produced by the respective generative model and the given ground-truth distribution agree by means of

¹<http://brainiac2.mit.edu/SNEMI3D/home>

Algorithm 1 | Hamming Distance based greedy Clustering. The clustering makes use of the assumption that pixels of the same object vary together across samples. Pseudo-code used to get instance segmentations from segmentation samples \mathbf{Y}_i for a given image X of size $H \times W$. The employed algorithm assigns pixels to clusters based on the Hamming distance between a cluster's prototype and the pixel's vector representation. The Hamming distance is a simple count of element-wise mismatches. Both vectors consist of the respective pixels' sampled class labels in one-hot form, i.e. for n samples and C classes, they have length nC . The algorithm proceeds in a greedy manner, i.e. once no more matches satisfying an upper bound on the distance to the current prototype are found, a new prototype is randomly picked from the remaining unassigned pixels. Sampling at random rather than picking the next available pixel minimizes the clustering run-time (which is $\mathcal{O} \leq (HW)^2$) and the likelihood of picking cluster prototypes from object boundaries. The algorithm starts with assigning pixels to a provided background class label. This assures that cluster $c = 0$ always corresponds to the background class, but is not strictly necessary, alternatively the algorithm can omit the case distinction in line 10ff.

Result: Instance Segmentation $\mathbf{I} \in \mathbb{Z}^{H \times W}$.

Parameters: n : number of samples, α : threshold.

```

1 Retrieve  $n$  sample segmentations  $\mathbf{Y}_i^{\text{prob}} \in \mathbb{R}^{H \times W \times C}$ ;  $\mathbf{Y}_i^{\text{prob}} \leftarrow S(X, \mathbf{z}_i)$ ,  $\mathbf{z}_i \sim P(\cdot | X)$ ;
2 Transform samples to one-hot  $\mathbf{Y}_i \in \mathbb{Z}^{H \times W \times C}$ ,  $\mathbf{Y}_i \leftarrow \text{one\_hot}(\text{argmax}(\mathbf{Y}_i^{\text{prob}}))$ ;
3 Concatenate samples over channels  $\mathbf{Y} \in \mathbb{Z}^{H \times W \times nC}$ ;  $\mathbf{Y} \leftarrow \text{concat}([\mathbf{Y}_0, \dots, \mathbf{Y}_n])$ ;
4 Initialize Instance Segmentation  $\mathbf{I} \in \mathbb{Z}^{H \times W}$ ;  $\mathbf{I} \leftarrow [[-1, \dots], \dots]$ ;
5 Initialize set of unassigned pixels  $\mathcal{U} = \text{where}(\mathbf{I} == -1)$ ;
6 Initialize background one-hot vector  $\mathbf{b} \in \mathbb{Z}^{C \times 1}$ ;  $\mathbf{b} \leftarrow \text{one\_hot}(\text{background label})$ ;
7 Initialize prototype  $\mathbf{p} \in \mathbb{Z}^{nC \times 1}$  with the prototype of the background class
   $\mathbf{p} \leftarrow \text{concat}([\mathbf{b}, \mathbf{b}, \dots])$ ;
8 Initialize cluster id  $c = 0$ ;
9 while  $|\mathcal{U}| > 0$  do
10   if  $c == 0$  then
11     | Do nothing, as  $\mathbf{p}$  is initially assigned to background class prototype;
12   else
13     | Draw a random pixel from the set of unassigned pixels  $i \sim \mathcal{U}$ ;
14     | Use the one-hot sample vector of this pixel as the  $c$ th cluster's prototype
15     |  $\mathbf{p} \leftarrow \mathbf{Y}[i]$ ;
16     |  $\mathbf{I}[i] \leftarrow c$ ;
17     | Drop  $i$  from set of unassigned pixels  $\mathcal{U} \leftarrow \mathcal{U} \setminus \{i\}$ ;
18   end
19   foreach  $j \in \mathcal{U}$  do
20     | Retrieve one-hot sample vector of the pixel  $j$  as  $\mathbf{v} \leftarrow \mathbf{Y}[j]$ ;
21     | Calculate Hamming distance  $d = \text{hamming\_distance}(\mathbf{v}, \mathbf{p})$ ;
22     | if  $d \leq \alpha$  then
23     | |  $\mathbf{I}[j] \leftarrow c$ ;
24     | |  $\mathcal{U} \leftarrow \mathcal{U} \setminus \{j\}$ ;
25     | end
26   end
27    $c \leftarrow c + 1$ ;
28 end

```

the **Generalized Energy Distance** (GED) with distance kernels of the form $1 - \text{IoU}(Y, Y')$. In such cases where the models' samples only poorly match the ground truth samples, we found this measure inadequate since the metric then unduly rewards sample diversity, regardless of the samples' adequacy. As an alternative that appears less vulnerable to such pathological cases, we propose to use the Hungarian algorithm [Kuhn, 1955, Munkres, 1957] to match samples of the model and the ground-truth. The Hungarian algorithm finds the optimal 1:1-matching between the objects of two sets, for which we use $\text{IoU}(Y, Y')$ to determine the similarity of two samples. We report the match as the **Hungarian-matched IoU**, i.e. the average **IoU** of all matched pairs and duplicate both sets so that their number of elements matches their least common multiple. As empty segmentations can be valid gradings in the LIDC dataset we need to define how the **IoU** enters the distribution metrics for the case of correctly predicted absences, which is detailed below. As an additional benefit, the Hungarian-matched **IoU** may be more readily interpreted by those familiar with semantic segmentation metrics.

8.4.2 Reconstruction Fidelity

The reconstruction fidelity is an upper bound to the fidelity of the conditional samples². In order to assess this upper bound on the fidelity of the produced segmentations we measure how well the models' posteriors are able to reconstruct a given segmentation in terms of the **IoU** metric, i.e. we report the **reconstruction IoU**, $\text{IoU}_{\text{rec}}(Y, Y')$ where $Y' = S(X, \mu^{\text{post}}(X, Y))$. Whenever we employ the **IoU**-metric, i.e. also when it enters the measures for distribution agreement, we calculate it with respect to the stochastic foreground classes only (as done for the **sPU-Net**). We further do not calculate it globally across all the test set pixels (i.e. across all the pixels of all images, as is regularly done in semantic segmentation challenges, e.g. in Cityscapes [Cordts et al., 2016]), but calculate it across the pixels of each image and then average across all test set images. The reason for this is that evaluating the predicted distribution over segmentations is only meaningful on the image level. While the reconstructions could be evaluated across all test set pixels, we stick with the image level evaluation for consistency between the metrics. As a consequence the question arises how to deal with a correctly predicted absence of a class in an image, a case for which the **IoU** metric is undefined (the denominator would be 0). For the LIDC dataset, empty ground-truth segmentations can be a valid grading which is why we proceed as for the **sPU-Net** (see Sec. B.1) and define a correctly predicted

²The term reconstruction refers to decoding posterior latents, implying a conditioning on the ground truth that is to be reconstructed. The term sampling on the other hand denotes the decoding of prior latents.

absence as $\text{IoU} = 1$. In the [SNEMI3D](#) and Cityscapes instance segmentation tasks we do not want to evaluate whether a model correctly predicts a class' absence, which is why correct class absences are simply excluded from the mean IoU of an image, while wrongly predicted absences are penalized (in practice we perform a 'NAN-mean' over the classes of interest). In the Cityscapes case we additionally make use of the provided ignore-masks, keeping unlabeled pixels out of the evaluations.

8.4.3 Instance Segmentation

In order to score how well the predicted instance segmentations (the instance clusters) agree with the ground truth, we calculate the **Rand Error**. This measure is defined as $1 - F$ -score, where the precision and recall values that enter the F -score are determined from whether pixel pairs between the ground truth clustering and a predicted clustering belong to the same segment (positive class) or different segments (negative class) [[Arganda-Carreras et al., 2015](#), [Rand, 1971](#)]. We use the foreground-restricted version as employed in the [SNEMI3D](#) challenge³.

On Cityscapes instance segmentation we additionally report the **Average Precision** (AP). It is based on object level scoring and defined as the area under the precision recall curve for all predicted object detections. To span the precision recall curve, an object level score that quantifies a model's confidence in the 'objectness' of its prediction is required. For our car instance segmentation experiments we employ the Cityscapes evaluation scheme⁴, reporting AP_{50} and AP, the average precision when requiring predictions to match above a thresholded **IoU**, $\text{IoU}_{\text{thres}} > 0.50$, and when averaging across multiple such thresholds (10 different overlaps ranging from 0.5 to 0.95 in steps of 0.05), respectively. To artificially obtain object-level scores we average the softmax scores of all stochastic classes across samples and pixels of a predicted instance mask [[Kulikov et al., 2018](#)].

8.5 Results

The [sPU-Net](#) has established significant performance advantages over other approaches in segmenting ambiguous images, see [Fig. 7.4](#) and [7.7](#). With the [HPU-Net](#) we aim at improving on the flexibility of the [sPU-Net](#) to model complex output interdependencies as well as segmentation fidelity. To this end we compare the two models' performance on the segmentation task of CT scans showing potential lung abnormalities annotated by four expert graders, called **LIDC** which we considered before (for a discussion of those

³Available as `adapted_rand_error` in the python package [gala](#) [[Nunez-Iglesias et al., 2014](#)].

⁴Official evaluation code can be found [here](#).

experiments and results see [Subsec. 7.5.1](#)). Samples and reconstructions of both models are shown in [Fig. 8.2](#). We further consider the task of segmenting individual instances, i.e. inferring both a latent id and a mask for each object in an image, and use it to assess the models’ ability to capture correlated pixel-uncertainty. We use the [EM](#) dataset of the [SNEMI3D](#) challenge (published in [[Kasthuri et al., 2015](#)]), which contains instance segmentations of neuronal cells (examples are shown in [Fig. 8.4](#)) and further probe our model’s performance on the segmentation of car instances on natural street scenes from Cityscapes (see [Fig. 8.6](#)). For training details in the respective tasks we refer to [Sec. C.2](#).

Table 8.1 | Test Set Results. Mean and standard deviation for the [HPU-Net](#) and [sPU-Net](#) are calculated from results of 10 random model initializations and 1000 bootstraps with replacement. Data splits are defined in [Sec. 8.3](#).

i) LIDC	GED ²	Hung.-m. IoU	Hung.-m. IoU (subset B)	IoU _{rec}
HPU-Net	0.27 ± 0.01	0.53 ± 0.01	0.47 ± 0.01	0.97 ± 0.00
sPU-Net	0.32 ± 0.03	0.50 ± 0.03	0.37 ± 0.07	0.75 ± 0.04
ii) SNEMI3D	Rand Error	IoU _{rec}		
HPU-Net	0.06 ± 0.00	0.60 ± 0.00		
sPU-Net	0.52 ± 0.10	0.13 ± 0.03		
iii) Cityscapes Car Instances	Rand Error	AP ₅₀	IoU _{rec}	
HPU-Net	0.13	46.8	0.62	

8.5.1 LIDC: Segmentation of Ambiguous Lung Scans

The LIDC results are reported in [Table 8.1a](#). The [HPU-Net](#) performs better in terms of the Hungarian-matched IoU (and in terms of $\text{GED}^2 = 0.27 \pm 0.01$), while showing a largely improved reconstruction fidelity, that amounts to a near perfect posterior reconstruction of $\text{IoU}_{\text{rec}} = 0.97$. Retraining the [sPU-Net](#) with an identical training set-up as in [[Kohl et al., 2018](#)], we obtain an unsatisfactorily low value of 0.75 for the foreground-restricted reconstruction IoU (IoU_{rec}) and recapture [[Kohl et al., 2018](#)]’s GED^2 of 0.29 (re-implementation: $\text{GED}^2 = 0.32 \pm 0.03$). We additionally evaluate the models on the test subset of samples for which all 4 graders agree on the presence of an abnormality (‘subset B’, see [Sec. 8.3](#)), exposing the [HPU-Net](#)’s significantly improved ability to capture shape variations (see also [Sec. C.3](#)). For the sake of completeness we also report the GED^2 and Hungarian-matched IoU of the baselines described in [Subsec. 7.4.2](#) on the full LIDC test set, see [Table C.1](#).

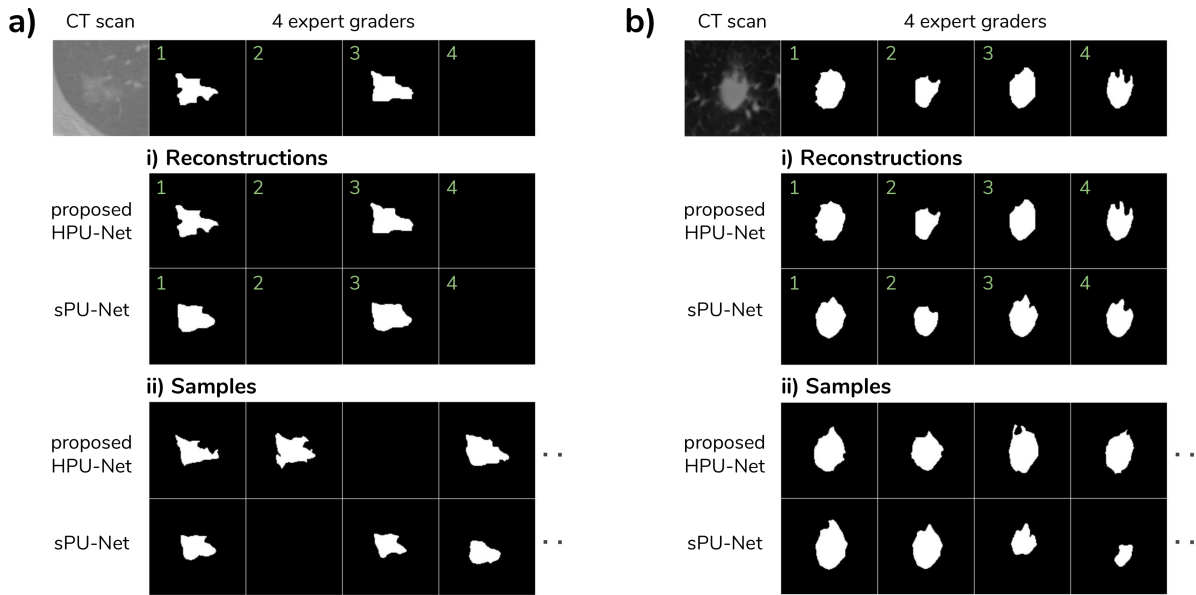


Figure 8.2 | Comparing Reconstructions & Samples on LIDC. Two example CT scans with the 4 available expert gradings. (i) Reconstructions of the 4 graders and (ii) Sampled segmentations. Note that the gradings can be empty, as foreground annotations correspond to supposed abnormal cases only. More cases in Fig. C.3 and C.4.

The *HPU-Net*'s capacity to faithfully learn segmentation distributions with high reconstruction and sample fidelity is also qualitatively evident. Fig. 8.2 compares samples from both models given a pair of CT scans of prospective lung abnormalities. The hierarchical model exhibits enhanced local segmentation structure. Its samples reflect the difficulty to pin-down the boundary of normal vs. abnormal tissue from the image alone (Fig. 8.2a) and also whether or not the salient structure is abnormal. The *sPU-Net*'s samples on the other hand appear much coarser and 'blobby' (Fig. 8.2b), see the panels marked ii) in Fig. 8.2a). The same holds true for both models' posterior reconstructions (panels marked i) in Fig. 8.2a), where a much improved reproduction of fine structure is at display.

In order to explore how the model leverages the hierarchical latent space decomposition, we can use the predicted means μ_i^{prior} for some scales instead of sampling. Fig. 8.3a shows samples for the given CT scans resulting from the process of sampling from the full hierarchy, i.e. from 4 scales in this case. Fig. 8.3b,c show the resulting samples when sampling from the most global or most local scale only. The hierarchical latent space appears to induce the anticipated bias: the global scales determine the coarse structure, which in this case includes the decision on whether or not the structure at hand is abnormal, while the more local scales fill in appropriate local annotations.

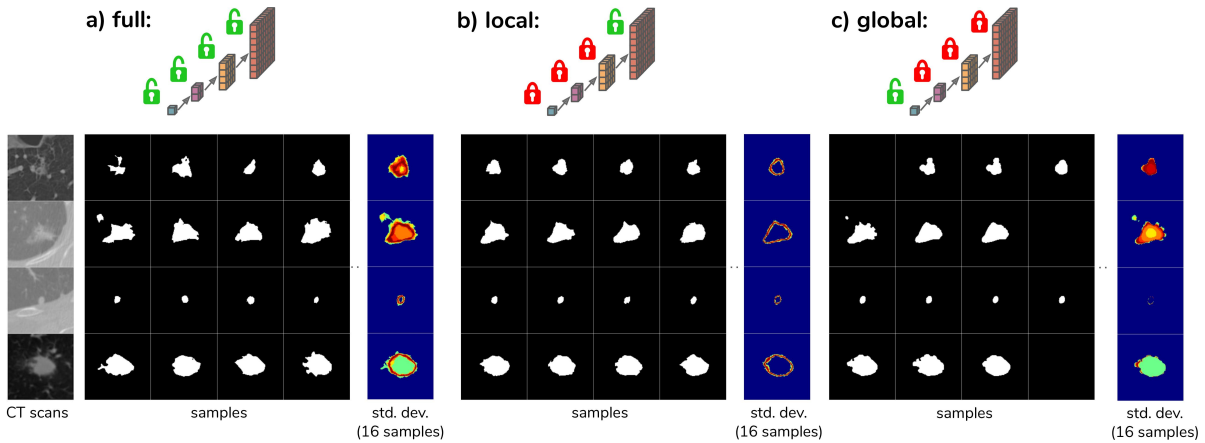


Figure 8.3 | HPU-Net Samples using different Latent Scales on LIDC. Samples and standard deviations across 16 samples given the CT scans on the left. Sampling from (a) the full hierarchy, (b) from only the most local latent scale and (c) from only the most global scale while fixing the respectively remaining scales to their predicted means μ_i^{prior} . Observe in the standard deviations how the local latents alter fine details, mostly at the boundaries, while the global latents can flick the presence of coarser abnormality segmentations on and off. The illustrations above the samples use red padlocks to indicate which scales are fixed.

8.5.2 SNEMI3D: Generative Instance Segmentation of Neurites

As a second dataset we use the SNEMI3D challenge dataset as described in Sec. 8.3. During training we randomly map the instance ids of the cells to one of 15 labels. The HPU-Net, in this case using four latent scales, displays both a strong reconstruction fidelity, $\text{IoU}_{\text{rec}} = 0.60$, as well as a very low Rand error = 0.06. Although we want to caution against a direct comparison between results obtained on our smaller test set (in 2D) against those from the official test set (in 3D), it is interesting to put an eye on the official leader-board, where the best dedicated algorithms reach a Rand Error of ~ 0.025 (e.g. [Lee et al., 2017]) and the human baseline achieved a value of 0.059⁵. For the sPU-Net, employing low dimensional latent spaces ($\mathcal{O} \sim 10$) as before (on other datasets) did not produce satisfactory results. Even when matching the number of global latents of a 4-scale HPU-Net ($\sum_{i=0}^3 2^{2i} = 85$), the sPU-Net struggles with reconstructing instance segmentations of neurites and likewise scores badly in terms of the Rand Error, see Table 8.1c).

From Fig. 8.4 it is evident that the HPU-Net is able to sample coherent instance segmentations of these amorphous structures with largely varying size and shape, resulting in faithful instance segmentations when clustered across samples. In contrast, the sPU-Net has a hard time accommodating for the independently varying instances and also

⁵<http://brainiac2.mit.edu/SNEMI3D/leaders-board>

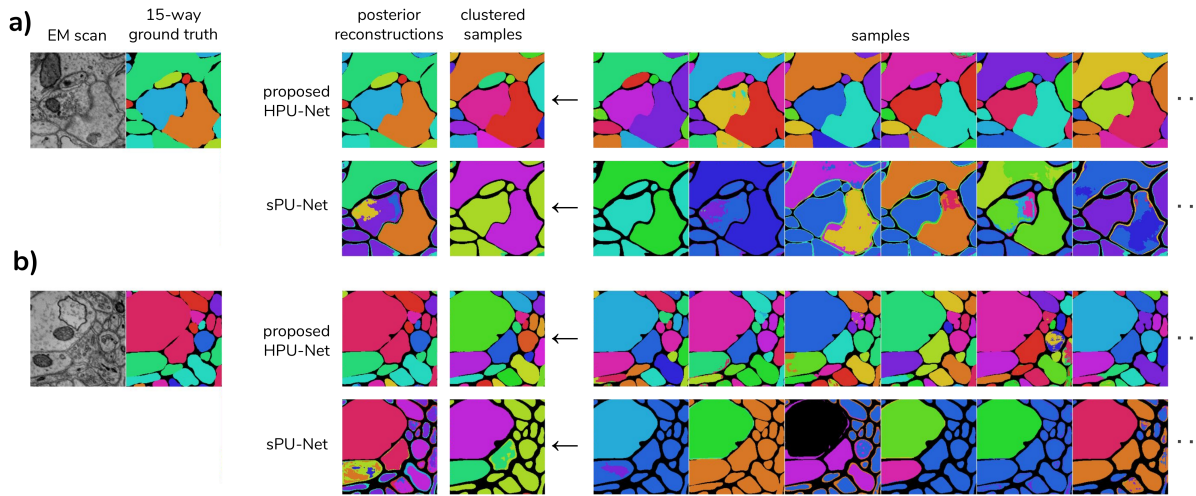


Figure 8.4 | Instance Segmentation of Neurons. From left to right: EM images from SNEMI3D, the ground-truth mapped to 15 random instance ids, the corresponding posterior reconstructions, predicted instance segmentation after clustering as well as 6 samples. Color denotes instance id (one of 15) and background is shown in black. For more examples see Fig. C.5 and C.6 in the appendix. The first row of Fig. a) and b) respectively corresponds to the HPU-Net and the second row to the sPU-Net (using 85 latents).

fails to coherently segment individual instances which is apparent in its samples, the clustering thereof and its reconstructions.

8.5.3 Extrapolation Task on SNEMI3D

In order to further explore the expressiveness of the proposed generative model, we train it to generate extrapolated segmentations given masked images. The masked parts are maximally ambiguous and sensible ways of extrapolating need to be inferred from the unmasked regions. Samples and reconstructions are shown in Fig. 8.5. To be able to visualize the extrapolations across samples we feed in both the image and the ground-truth segmentation of the unmasked region to the prior, so that it can fix the found instance ids (which is not required for this to work). We observe that the model’s generative structure can produce convincing extrapolations, note how the model preserves scale and appearance of unmasked instances, e.g. large cells are more likely to cover larger areas in the masked region and slim cells remain slim and elongated, see third row of samples.

8.5.4 Cityscapes Cars: Generative Instance Segmentation of Cars

In order to test our model’s ability to coherently flip independent regions on natural images, we evaluate it on the task of segmenting car instances on Cityscapes. We train

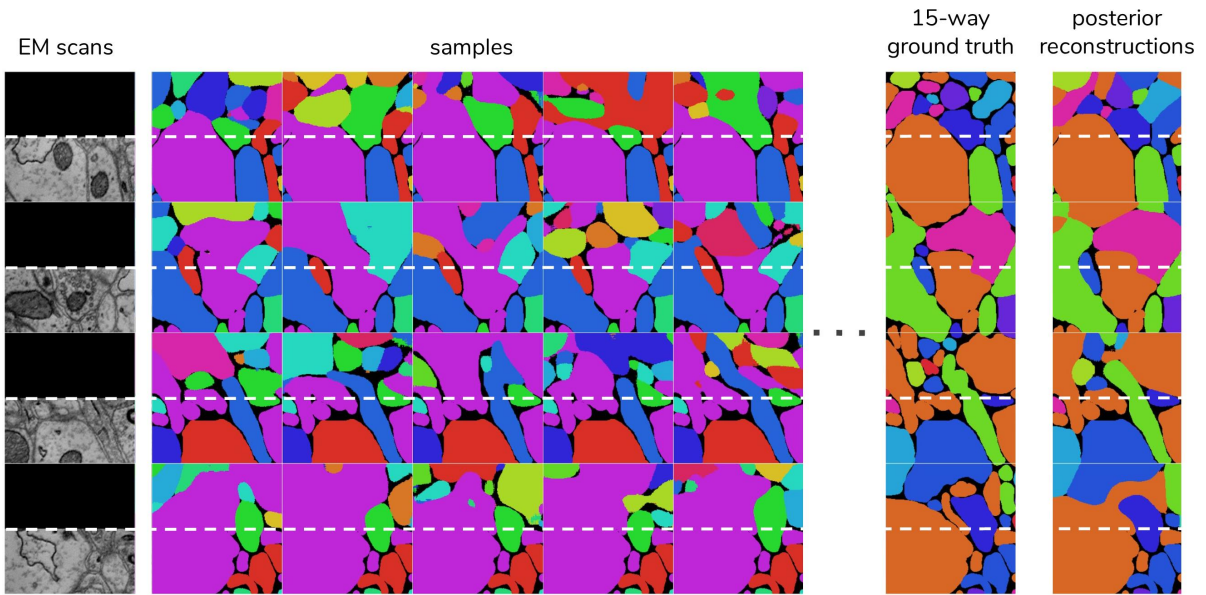


Figure 8.5 | Generative Extrapolation on masked EM Images. Examples show HPU-Net samples and reconstructions. Areas above the dashed line in each row correspond to the masked part. Colors denote instance ids (one of 15) with black for background segmentation.

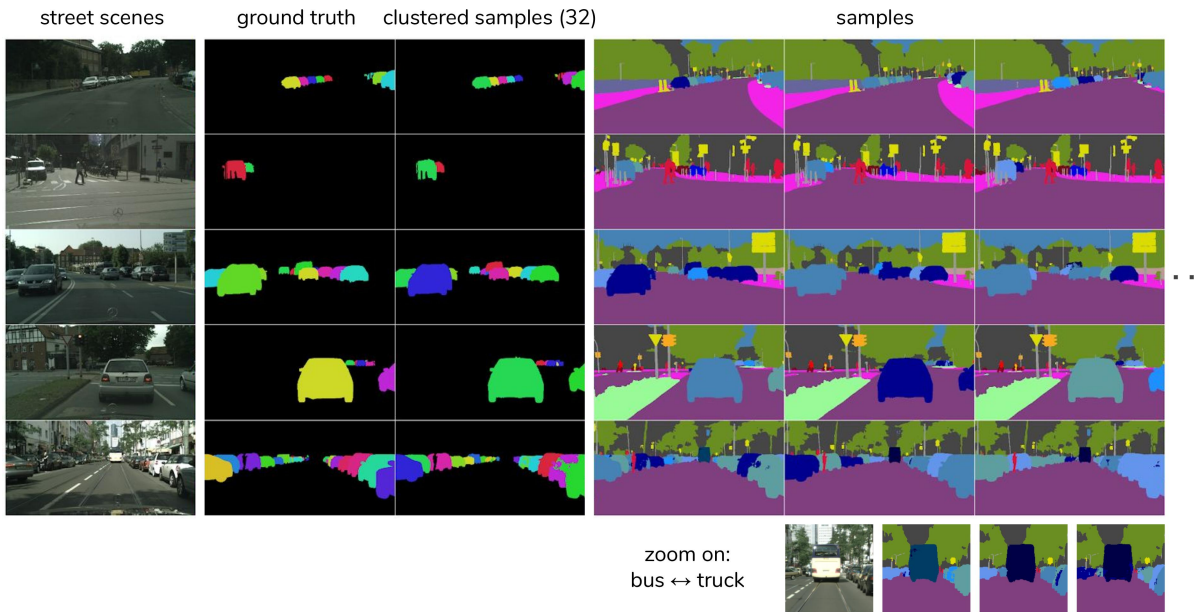


Figure 8.6 | Generative instance segmentation of Cars. Cityscapes test set examples on 512×1024 resolution using the HPU-Net. From left to right: Input street scenes, ground truth car instances, clustered model samples as well as individual samples. Last row that shows crops zoomed in on samples from above. More examples can be found in Fig. C.7 and C.8.

our model to segment all 19 Cityscapes classes while introducing additional alternative car classes that are randomly flipped during training. We run on half-resolution, i.e. 512×1024 (more details in Sec. 8.3 and C.2). At test time we cluster 32 samples per image (see Algorithm 1). Results on the official validation set (our held-out test set) are reported in Table 8.1d). Employing 4 latent scales (with the highest latent resolution at 16×32), we reach an $\text{IoU}_{\text{rec}} = 0.62$, a Rand error = 0.13 and a $\text{AP}_{50} = 46.8$, without ensembling or any other test-time augmentations. While these results are not competitive with top-performing bounding box regressors such as Mask R-CNN [He et al., 2017] ($\text{AP}_{50} = 68.3$ on the Cityscapes test set), which are tailored towards instance segmentation of boxy objects, we observe an arguably solid out-of-the-box performance of the HPU-Net. Direct comparison may further suffer from the post-hoc computation of object-level confidence scores which we are required to carry out for the AP-metric and which bounding-box regressors on the other hand can optimize for during training.

Fig. 8.6 shows predicted and ground truth instance segmentations for five scenes. This task is difficult as aside from varying factors such as appearance and illumination, cars nestled along the road can be heavily occluded and individual cars can cover anything between tiny to large regions of the image. Nonetheless our proposed model can sample individual instance segmentations with good coherence, resulting in strong instance segmentations. Interestingly the model also picks up on ambiguity that is naturally present in the data, e.g. the samples in the first row show coherent flips between parts of the *road* and *sidewalk* and the last row shows coherent flips between *bus* and *truck* annotation for the bus at the end of the road (for which we provide zoomed crops of size 200×200 in Fig. 8.6). Fig. 8.7 shows samples and the standard deviations when sampling from only the most local versus only the most global latent scales. It is apparent that the local latents affect small and distant cars while the global latents control more global factors such as cars close to the observer. This shows that also on this large scale natural image data, the model has learned to separate scales.

8.5.5 Ablation Study

In order to show the effect of some of the main choices we made for the model and the loss formulation, we perform an ablation study on the LIDC lung abnormalities segmentation task. All models are trained with the same training setup and hyper parameters as used in the LIDC experiments (described in Sec. C.2), if not stated differently in the following. First we evaluate the importance of the latent hierarchy. We train 10 random initializations for a model with a global latent scale in the ‘U-Net’s bottom’ that otherwise employs the same model topology as the HPU-Net that we

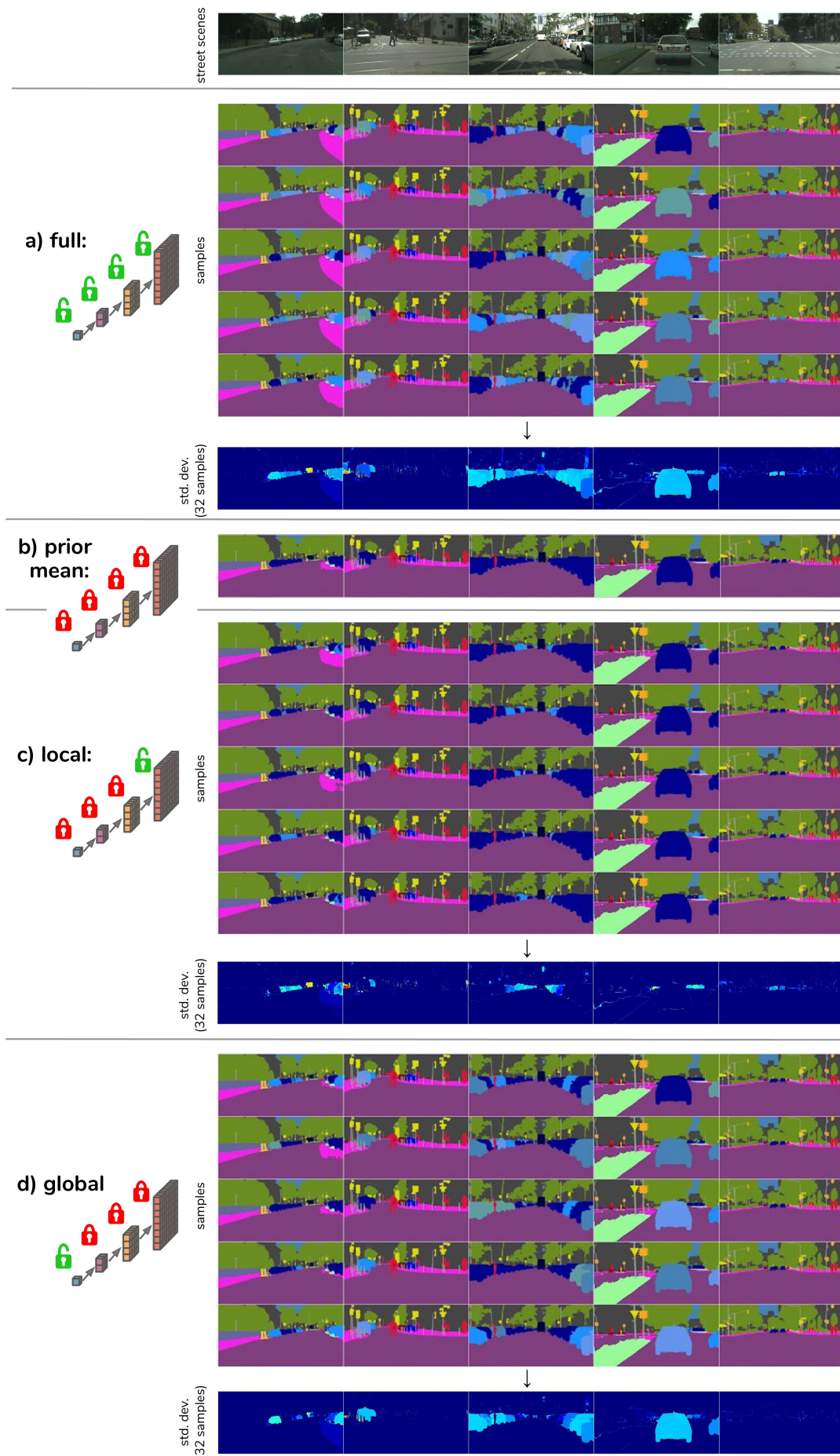


Figure 8.7 | HPU-Net Samples using different Latent Scales on Stochastic Cityscapes. Test set samples for a model trained with 5 distinct latent car ids on resolution 512×1024 . (a) Samples and standard deviation (std. dev.) across 32 samples when sampling from the full hierarchy. (b) Predictions from the prior mean. (c) Samples and std. dev. when sampling only from the most local scale. (d) Samples and std. dev. when sampling only from the most global scale. Note how the global and local scales affect the instance mask generation almost complementarily.

employ on LIDC. For this model we use 85 global latents, i.e. the same number of total latents that the 4-scale hierarchical model employs. In order to arrive at a comparable reconstruction IoU, we found it necessary to raise the reconstruction target κ above the value of 0.05 (employed for the other models) to a value of $\kappa = 0.15$. As reported in Table 8.2, this model performs significantly worse than the HPU-Net in terms of both GED² and the Hungarian-matched IoU, while also suffering from a loss in reconstruction fidelity. As a second model configuration we consider a model with the same topology as the employed HPU-Net, however employing only its most local scale of latents (a spatial grid of size 8×8). The idea is to assess to what degree the latents lower in the hierarchy help coordinate the sampling from the last, most finely resolved grid of latents. The results in Table 8.2 show another significant decrease in the model’s ability to match the ground truth distribution, suggesting that the hierarchy indeed is an important model choice enabling the strong performance in terms of GED² and the Hungarian-matched IoU. Lastly we quantify the effect of employing a top-k loss for the hierarchical model. The last row in Table 8.2 shows the positive effect that the top-k loss formulation has on the reconstruction IoU (IoU_{rec}), while allowing to keep the same level of distribution match (there is a slight increase in Hungarian-matched IoU when ablating the top-k loss, it is however insignificant across 10 random initializations).

Table 8.2 | Ablation study on for the HPU-Net on LIDC. All results are reported on our test set and the given means and standard deviations are taken across 10 random initializations of the same respective model setup and 1000 bootstraps with replacement each. The values reported for κ are normalized per pixel and for comparison the LIDC results reported in Table 8.1 are shown in the first row of this table.

model + loss formulation	IoU_{rec}	GED ²	Hungarian-matched IoU
4-scale hierarchy + GECO ($\kappa = 0.05$) + top-k (k=0.02)	0.97 ± 0.00	0.27 ± 0.01	0.53 ± 0.01
local latents + GECO ($\kappa = 0.05$) + top-k (k=0.02)	0.97 ± 0.00	0.34 ± 0.01	0.45 ± 0.01
global latents + GECO ($\kappa = 0.15$) + top-k (k=0.02)	0.94 ± 0.02	0.40 ± 0.02	0.37 ± 0.02
4-scale hierarchy + GECO ($\kappa = 0.05$)	0.94 ± 0.00	0.27 ± 0.01	0.54 ± 0.01

8.6 Discussion

Targeted at the segmentation of ambiguous medical scans we have previously introduced the Probabilistic U-Net (sPU-Net) which learns an image-global distribution that allows to sample consistent segmentation hypotheses, see Chap. 7. As we show here, this model

however suffers from poor sample and reconstruction fidelity and breaks down altogether in more complex scenarios such as instance segmentation.

With the **Hierarchical Probabilistic U-Net (HPU-Net)**, we propose a model which shows clear quantitative and qualitative evidence for its advantages over the prior art. The proposed model uses a much more flexible generative model and further profits from advances such as improved training procedures for **VAEs** and efficient hard-negative mining. We ablate those choices in experiments conducted on **LIDC** for which we report and discuss results in **Subsec. 8.5.5**). The hierarchical latent space formulation enables to model ambiguities at all scales and affords the learning of complex output interdependencies such as e.g. coherent regions of pixels as found in the task of instance segmentation.

In addition to presenting high-quality results on the segmentation of ambiguous lung CT scans, we achieve strong out of the box performance in instance segmentation of both neurobiological images as well as natural images of street scenes, showing the flexibility and amenability of the proposed model to such tasks. While state-of-the-art deterministic bounding-box regressors [[He et al., 2017](#), [Lin et al., 2017](#)] still perform significantly better on car instance segmentation, they are predominantly based on a pixel-wise refinement of bounding-boxes and are not designed for overlapping or intertwined instances as found in neurobiological instances. Our generative approach could be a way to directly perform dense object-level segmentation, which has recently attracted attention [[Chen et al., 2019](#), [Kirillov et al., 2018](#), [Kulikov and Lempitsky, 2019](#), [Kulikov et al., 2018](#), [Xiong et al., 2019](#)].

The **HPU-Net**'s samples are indicative of model uncertainty for ambiguous cases that it has seen during training, which is expected to benefit prospective down-stream tasks. As such the expressed model uncertainty is valid within the data distribution only and, like many others, the model is not aware if and when it fails out-of-distribution [[Nalisnick et al., 2018](#)]. Aside from allowing to capture multiple scales of variations simultaneously, the latent hierarchy further imposes an inductive bias that mirrors the structure of many medical imaging problems, in which global information can affect top-down decision making, i.e. local annotations in our case. We show this trait in our lung CT scan experiments, where the model learns to separate variations at different scales. Here our model automatically opts to take the decision as to whether the given structure may be abnormal at its most global scale, while reserving more local decisions for local latents, see [Fig. 8.3](#). A similar decomposition is apparent on natural images ([Fig. 8.7](#)). In terms of KL cost, it is more expensive to model global aspects locally, which in combination with the hierarchical model formulation itself, is the mechanism that puts into effect the

separation of scales. Disentangled representations are regarded highly desirable across the board and the proposed model may thus also be interesting for other down-stream applications or image-to-image translation tasks.

In the medical domain the [HPU-Net](#) could be applied in interactive clinical scenarios where a clinician could either pick from a set of likely segmentation hypotheses or may interact with its flexible latent space to quickly obtain the desired results. Additionally, the sampled set of segmentations could be used to propagate uncertainty through the down stream tasks, much like it was done for segmentation ensembles in [\[De Fauw et al., 2018\]](#). The model’s ability to faithfully extrapolate conditioned on prior observations could further be employed in spatio-temporal predictions, such as e.g. predicting tumor therapy response.

Lastly and developed concurrently to the [HPU-Net](#), there is work that similarly extends the [sPU-Net](#) by employing a hierarchical latent structure. In contrast to our work, their segmentation network S is only based on the sampled hierarchical latents, so takes on the form $S = S(\mathbf{z})$ instead of $S = S(X, \mathbf{z})$ as in the [HPU-Net](#). This also implies that their model does not use the high-resolution skip-connections (above scale L) that link the U-Net encoder to its decoder, potentially setting it up for decreased fine-grained details. This work, see [\[Baumgartner et al., 2019\]](#), was published shortly after ours [\[Kohl et al., 2019\]](#).

Chapter 9

Discussion

The contributions brought forth in this thesis are concerned with the algorithmic interpretation of image evidence that may be ambiguous with respect to the target measure of interest.

One way of conveying an interpretation of an image is to assign a label to each pixel or voxel within it, a task referred to as semantic segmentation. A pixel-level understanding of medical images plays a crucial role in many clinical diagnoses and treatments, since e.g. biopsy planning [Epstein et al., 2012], radiation therapy [Borofsky et al., 2017, Nikolov et al., 2018] and tumor surveillance [Kickingereeder et al., 2019] hinge on the precise localization of tissues.

Because medical imaging modalities such as MRI or CT only indirectly measure the molecular identity of tissues, they often only provide ambiguous evidence for target measures of interest [Hameed and Humphrey, 2010, Kitzing et al., 2015, Nagel et al., 2013, Sakala et al., 2017]. For example a lesion may be clearly visible on a scan but the information on whether it is cancerous or not may have been deleted in the imaging process. In such scenarios a group of clinical experts typically produces a set of diverse but plausible segmentations [Armato et al., 2011] that can show significant variation [Borofsky et al., 2017, Bratan et al., 2014].

Having available an estimate of the ensuing distribution holds the potential to improve decision making and care. In real-world clinical practice however, it is common for only a single reader to be involved in the interpretation of medical images, which given their past experience and momentary opinion yields a single subjective outcome. The subjective nature of the current process of detecting and grading lesions is reflected and well documented in present clinical guidelines, such as for example (in the case of prostate imaging) in the ‘Prostate Imaging - Reporting and Data System’ (PIRADS) [Weinreb et al., 2016]. This current clinical practice bars a more principled handling of the inherent

uncertainties of the annotation process and with it the possibility to propagate these uncertainties into down-stream decision making.

A hot question is thus whether learning based systems are a viable approach to improve the interpretation of medical images to that effect. For one, such approaches hold the potential to improve overall performance and reduce subjectivity that arises due to different levels of past experience, since machine learning algorithms can be trained on thousand- or million-fold the cases that a typical clinician is exposed to in a typical career. For another, learning based systems open up the possibility to quantify uncertainties associated with each image analysis that is performed during the process of diagnosis and treatment. Furthermore, stacking up parts of the system responsible for different sub tasks could allow to propagate the individual uncertainties all the way to the final decision that needs to be taken, provided they interface one another in an appropriate way [De Fauw et al., 2018].

In a first step we set aside the modelling of ambiguities and instead explore the utility of a simple but quantitative approach to lesion grading, considering the example of prostate MRI interpretation (see Chap. 5). We show that both a simple cross-validated threshold on a lesion diffusion measurement (ADC) [Posse et al., 1993] and a tree-based machine learning model trained on image-derived lesion features can significantly improve upon the performance of an individual experienced radiologist in subjectively grading lesions during clinical routine. This observation not only quantitatively affirms the discrimination power of ADC, a measure usually qualitatively accounted for in clinical grading, but also ascertains that quantitative and learning based systems can match and surpass the performance of clinical experts, given appropriate training data, which echos the findings of many recent studies [De Fauw et al., 2018, Esteva et al., 2017].

The discipline of medical imaging shoulders significant responsibility but with it also bears the potential to have tangible impact with already comparatively small improvements. For example sparing patients from biopsies that can be fraught with risks, can ultimately preserve a patient's quality of life [Bickelhaupt et al., 2018, Bonekamp et al., 2018]. One way to improve the performance and the robustness of machine learning models that are largely developed on natural image benchmarks, is to carefully account for peculiarities in medical image analyses such as the comparatively large measurement noise, the often considerable image ambiguity [Armato et al., 2011, Kitzing et al., 2015] and the associated noisy image annotations [Borofsky et al., 2017, Bratan et al., 2014].

In Chap. 6 we follow this paradigm and explore the possibility to improve model performance in the face of label noise. We empirically find the standard loss formulation for multi-class segmentation tasks, the Cross Entropy-loss, to result in sub-optimal

performance when training on a prostate MRI dataset, a scenario known to suffer from high ambiguity and consequently large label noise [Hameed and Humphrey, 2010, Kitzing et al., 2015, Nagel et al., 2013, Sakala et al., 2017]. We find that swapping the Cross Entropy-loss for a model of the loss that is learned in an adversarial training scheme [Goodfellow et al., 2014] results in significantly improved test performance and additionally observe growing relative gains when artificially reducing the number of training examples. While the mechanism behind the improvements may require further confirmation, this study emphasizes the utility, if not necessity, of tailoring machine learning methods to the peculiarities of medical image applications and cautions against simple drop-in solutions of deep learning models.

As elaborated upon above, knowing the empirical distribution over plausible image interpretations rather than predicting a single segmentation, could enable more appropriate and informed clinical steps when the presented image evidence is inconclusive. If for example the empirical distribution indicated the possibility of ambiguous interpretations, further clinical steps to resolve the ambiguity, such as biopsies, could be mandated. A range of prior works exists that has sought to capture the appropriate output diversity of deep neural networks given an image. These models either produce only a pixelwise uncertainty corresponding to the distribution’s pixelwise marginals, employ an ensemble of networks or only indirectly condition their output distribution on the given image [Kendall and Gal, 2017, Lakshminarayanan et al., 2017, Rupprecht et al., 2017, Zhu et al., 2017b].

In Chap. 7, we introduce the *Probabilistic U-Net* (PU-Net), a segmentation model that is combined with a conditional variational autoencoder [Jimenez Rezende et al., 2014, Kingma and Welling, 2013]. Using a dataset of ambiguous lung CT scans that provides four expert segmentations for each image [Armato et al., 2011], we show that our model captures the appropriate output distribution significantly better than aforementioned prior work. An additional advantage over producing pixelwise uncertainties lies in the fact that samples from our model are consistent and may thus readily be used in subsequent steps that seek to extract information from predicted segmentation maps. Additionally clinicians could pick an appropriate segmentation sample from a set of samples in order to quickly produce a strong segmentation or alternatively interact with the model’s latent space to swiftly make desired adjustments.

However, because the Probabilistic U-Net’s latent space is image global, it can exhibit difficulty accommodating for multiple independent factors of variation within a single image, for example in the case when multiple lesions are present, and the sampled segmentations can lack in finer detail (i.e. suffer from blobby contours). Additionally,

image ambiguities may be present across different scales and locations, thus affecting different output scales in distinct manners.

In [Chap. 8](#), reflecting the need for a more flexible model, we thus introduce the *Hierarchical Probabilistic U-Net* (HPU-Net). The HPU-Net combines a hierarchical variational autoencoder with a U-Net [[Ronneberger et al., 2015](#)] and gets away with the separate prior network that is present in the Probabilistic U-Net. We show that this model formulation enables sampling and reconstruction of segmentations with high fidelity, i.e. with finely resolved detail, while providing the flexibility to learn complex structured distributions across scales. We demonstrate these abilities on the task of segmenting ambiguous lung CT scans [[Armato et al., 2011](#)] on which we outperform all considered baselines including the Probabilistic U-Net. We additionally consider the task of instance segmentation of neurobiological images [[Kasthuri et al., 2015](#)] and instance segmentation of natural images (Cityscapes [[Cordts et al., 2016](#)]), on which we demonstrate good out of the box performance, highlighting the largely improved flexibility of the model to capture distributions over independently varying image locations.

9.1 Outlook

In this thesis we have sought to improve the handling of ambiguous image evidence in semantic segmentation tasks. We have shown the possibility to outperform an experienced radiologist in the grading of prostate lesions in the case when lesion segmentations are given by an oracle (i.e. an expert annotator) using comparatively simple machine learning techniques and demonstrated an avenue to improve the segmentation performance of a deep segmentation model under noisy labels using a clinical prostate MRI dataset. With the aim of appropriately capturing the plausible distribution over segmentations that is admissible for a given image, we stepped beyond the prediction of deterministic segmentation maps and instead proposed two generative models that allow sampling consistent segmentation hypotheses from a predicted conditional distribution.

These contributions can be viewed as steps in a broader quest towards more quantitative and objective clinical procedures. As elaborated upon in [Chap. 3](#), the diagnosis and eventually the treatment of cancer typically involves a whole range of steps, each chosen so as to bring information to the table, that the previous step might not have been able to produce. Looking ahead and given the advantages of learning based systems, it is clear that the interpretation of evidence and the decision making required in each step could in the future be aided by or perhaps even be carried out by such systems. The arguably hardest part of getting there is endowing algorithms with the soft qualities that every

clinician masters with ease: A radiologist (mostly) recognizes when they are confronted with a case they have never seen before (*out-of-distribution robustness*), a radiologist can (mostly) explain their interpretation and decision (*interpretability*), a radiologist is inherently robust to domain shifts such as images from a new scanner type as they understand anatomy at a more abstract level (*robustness to domain shifts*), a radiologist knows about alternative measures and can actively seek for more evidence (akin to *active learning*) and a radiologist can integrate all the evidence they are presented with into a coherent mental model of the patient.

All of these are desiderata that we should require from models that might be integrated into clinical workflows with increasing autonomy. In the following we briefly discuss the extent to which the models in this thesis may already partially fulfill the criteria and highlight potential future research:

In order to increase trust in algorithmic predictions, but also to debug models and potentially distill scientific knowledge from them, ways to interpret the decision-making of deep neural networks are highly sought after [Kelly et al., 2019]. Perhaps the main obstacle in that quest is simultaneously one of the deep network’s biggest advantages: their highly complex and hierarchical representations which can express functions that live in spaces far too large for a human to parse. In light of this complexity, there are only few techniques that allow for human digestible explanations of deep nets [Lipton, 2016]. One way of producing a level of interpretability, is to enforce meaningful intermediate representations, such as producing semantic segmentations that feed into a classification network, see [De Fauw et al., 2018], or by otherwise constraining the network to depend on low-dimensional representations that can be decoded into something that is meaningful to a human ¹ [Iten et al., 2018].

The *Probabilistic U-Net* and the *Hierarchical Probabilistic U-Net*, open up avenues for both. Instead of producing a per-pixel uncertainty, they predict distributions (and with it the distributions’ spreads) over low-dimensional latent spaces, that carry semantic and, in this looser sense, interpretable meaning. Additionally, because each of the models’ samples are coherent and plausible image interpretations, the samples can readily be fed into down-stream models, similar in spirit to [De Fauw et al., 2018], thus providing a direct handle on how segmentation uncertainty affects down-stream tasks. While it is left as future research to put such a combined system to a test and explore the practical utility of interpreting its latents, it is clear that this formulation allows for straight-forward and principled ways of propagating the uncertainty of the image interpretations.

¹Note that this entails a relaxed notion of interpretability that does not require an understanding of the causal chain that produced those interpretable representations in the first instance.

Both of the proposed generative models are able to capture diverse predictions, admissible given the presented image, and can thus constrain or broaden their interpretation depending on the ambiguity of the image evidence. Nonetheless, the models are not able to indicate their own failure, i.e. flag out-of-distribution cases, that do not fall onto the data manifold they were trained on. This ability is at the focus of recent research and is addressed under many different cloaks such as from the perspective of outlier detection [Nalisnick et al., 2018], adversarial robustness [Arnab et al., 2018, Smith and Gal, 2018], robustness to domain shifts [Ganin et al., 2016] and uncertainty calibration [Lakshminarayanan et al., 2017]. None of these angles unfortunately have produced a fully satisfying answer to the problem (yet) for when the models operate on very high dimensional spaces such as real world medical images. Consequently, the problem still requires further research.

Humans are able to learn good representations from few, but highly representative data points and can actively query for such, drawing on their ability to internalize what they do not know. As mentioned above, this is also a highly desirable feature for machine learning algorithms since it could reduce the number of data points needed to train them [Gal et al., 2017b] and potentially allow for a hypothesis-driven reasoning at test time. In machine learning, calibrated output uncertainties can serve as a proxy for the internal model of ignorance, which we have shown the Probabilistic U-Net to exhibit (see Subsec. 7.7.1). Probing the usefulness of our models in the context of active learning is however left for future research.

Furthermore, looking at the chain of measures typically involved in a clinical diagnosis or by extension the cognitive integration of many environmental cues when driving a car, it is clear that image evidence may be complemented with many other sources of information. Among them e.g. blood tests in clinical scenarios or lidar in autonomous vehicles, all of which are chosen somewhat orthogonally so as to help in disambiguating the prediction tasks. While the proposed models in this thesis can simultaneously condition on different imaging modalities (showing the same view), they are not designed to integrate other (clinical) inputs or temporal inputs (e.g. scans taken across different points in time). The ability to do so promises to improve diagnostic and segmentation performance and may be a crucial step towards more holistic and reliable models. One promising area of research in this direction are conditional generative models that combine the representations of different observations such as conditional neural processes [Garnelo et al., 2018] or generative query networks [Eslami et al., 2018]. First steps towards conditioning a Probabilistic U-Net on several successively taken image observations have

been taken in [Petersen et al., 2019] with the aim to model tumor growth along the temporal axis.

9.2 A Golden Future?

It has only been very few years since the availability of large, annotated data sets coincided with the availability of large-scale computation power as well as the conception of machine learning techniques that allow to profit from the two. Despite remarkable successes, deep learning systems are thus far still narrow in scope in that they are made up of models that are separately trained on and tailored towards individual problems and singular data sources.

Leveraging streams of different data modalities on the other hand, similar to how humans perceive and probe their environment seems to be one natural and promising extension. Not many works however report a successful integration of several different modalities or domains as well as different time horizons. Similarly, current forays into automatic systems for real-world decision making, constitute static input-output mappings. Bringing to mind the dynamic, hypothesis-driven reasoning and active acquisition of relevant data points that is characteristic of human scientists and radiologists, it is clear that current automatic systems largely lack in such capabilities. With the current and arguably strong momentum in machine learning research and ever more data to learn from, there is little doubt, that new holistic systems have the potential of creating ever more reliable, well-calibrated and personalized systems that, in the clinical context, will aid both diagnosis and treatment.

Below we dare a glimpse at a potential mid-term future, in which information about the workings of our bodies is abundant in almost real time and its interpretation is reliable to within well understood bounds of uncertainty.

Fin just turned 65. He was a man of excellent health, something that was also apparent from his decidedly youthful physique. His good condition filled him with a sense of pride, as he had to make a number of painful changes to the irresponsible habits he had picked up in the first two or three decades of his life. The self-administered 180-turn had almost come over night, after he had participated in an early-adopter tech trial program in the late 2020s. The (then) novel whole-body-state simulator that he had signed up to test, had projected a very grim future for his virtual alter ego: Bad health, rapidly degrading looks, a short life. All presented in minute detail and backed up by medical and biological data of a dazzling number of participants.

The number of times his meter readings for blood decomposition, heart rate and brain oxygenation were out of his prescribed ranges ever since, was very low. He knew that for a fact, because his health portal kept track of every single measured point in time. It was a self-learning system that had continually improved as more and more people opted in globally. Yes, it had been hacked a few times, but the encrypted data handling and storage coupled with the strong anonymization guarantees, rendered the captured data useless to the intruders.

Aside from the assurance that the portal was learning from more and more real cases and thus steadily increased its utility to him, it allowed Fin to compare to his de-identified peers. And boy, he was doing well. All his current numbers looked splendid. His projected age surpassed the 110 mark with high confidence. There was even a chance, small but non-negligible, that he may live beyond 130 years. With this prognosis he ranked in the top percentile.

In that sense he had barely reached the mid point of his life and yet he was not surprised when, just a few weeks ago, he learned of a prostatic cell alteration. In fact, he was expecting it: Given his diagnosed genetic predisposition and a family history of prostate cancer, the portal had raised a 95%-confidence warning for occurrence within the next 2 years. Being enrolled in the omni-screening program, Fin regularly visited one of the fully automated check-up pods that were dispersed across the city. The pods performed measurements that his portable meters weren't equipped for, such as a full-body fast MRI. As he stepped out the pod that morning a few weeks ago, he had already received a summary diagnosis from the portal: an integration of all the evidence from MRI, PSA antigen readings and epigenetic blood markers into his personalized model, indicated that an early onset prostatic cell alteration was in progress.

No need to get nervous, as he had already been informed about the best course of action, even before the detected onset. A simulation-based personalized response assessment had determined, that he stood a 74% chance of stopping the cell growth before it even forms a veritable lesion by taking tailored DNA transcription blockers. The blocker protein would be engineered so as to impede transcription of the mutated prostate genome, otherwise leaving the transcription process untouched and thus virtually excluding side-effects. Recent advances in protein design further allowed to spec the drug with a tertiary structure that survives the aggressive micro-environment of his digestive tract -an important requirement for simple and non-invasive oral intake.

Aside from taking the bespoke drug, Fin would not have to abide by any restrictions whatsoever. Continued participation in the omni-screening program would remain part of his routine. His

pod-visits however would now hold the added excitement of localizing the cell-alteration and tracking its change over time, which within well quantifiable limits was possible despite its markedly subtle and ambiguous appearance. He lived with the firm knowledge that even if this treatment would unexpectedly not hit the nail on the head, his portfolio of options was by far and large not exhausted yet.

And really there was not much need to worry: A few weeks into rolling out the portal response and his PSA levels had already begun falling.

List of Own Publications

David Bonekamp*, Simon Kohl*, Manuel Wiesenfarth, Patrick Schelb, Jan Philipp Radtke, Michael Götz, Philipp Kickingereder, Kaneschka Yaqubi, Bertram Hitthaler, Nils Gählert, Tristand Anselm Kuder, Fenja Deister, Martin Freitag, Markus Hohenfellner, Boris A Hadaschik, Heinz-Peter Schlemmer, and Klaus H. Maier-Hein. Radiomic machine learning for characterization of prostate lesions with mri: comparison to adc values. *Radiology*, 289(1):128–137, 2018.

Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017a.

Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for prostate cancer detection. *Machine Learning for Health Workshop, Advances in Neural Information Processing Systems*, 2017b.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.

Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *Med-NeurIPS Workshop, Advances in Neural Information Processing Systems*, 2019.

* shared first authorship with equal contributions.

‘Adversarial Networks for Prostate Cancer Detection’ is a revised workshop version of ‘Adversarial networks for the detection of aggressive prostate cancer’.

Appendix A

Finding Discriminative MRI Features

With kind permission by the Radiological Society of North America (RSNA), Appendix A reproduces parts of the following publication:

David Bonekamp, Simon Kohl, Manuel Wiesenfarth, Patrick Schelb, Jan Philipp Radtke, Michael Götz, Philipp Kickingereeder, Kaneschka Yaqubi, Bertram Hitthaler, Nils Gählert, Tristan Anselm Kuder, Fenja Deister, Martin Freitag, Markus Hohenfellner, Boris A Hadaschik, Heinz-Peter Schlemmer and Klaus H Maier-Hein. “Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC values.” *Radiology* 289, no. 1 (2018): 128-137,

cited in the following as [Bonekamp et al., 2018].

A.1 Systematic and Targeted MR Imaging/TRUS-Fusion Biopsies

All men underwent transperineal grid-directed biopsy performed under general anesthesia with rigid software registration using BiopSee (MEDCOM, Darmstadt, Germany). Fusion-biopsy of MR imaging-suspicious lesions was performed first (interquartile range (IQR) 3–5 cores, median 4 per lesion) followed by systematic saturation biopsy (20–26 cores, median 23 cores), as previously described [Hadaschik et al., 2011, Radtke et al., 2016]. This biopsy approach combining targeted biopsies and transperineal systematic saturation biopsies has been validated against and confirmed concordance to radical prostatectomy specimen [Radtke et al., 2016]. A median of 29 biopsies (IQR 24–33) were taken per patient with the number of biopsies adjusted to prostate volume. Histopathological results of targeted regions and whole gland assessment served as standard of reference.

A.2 Cohort Inclusion Criteria and Demographics

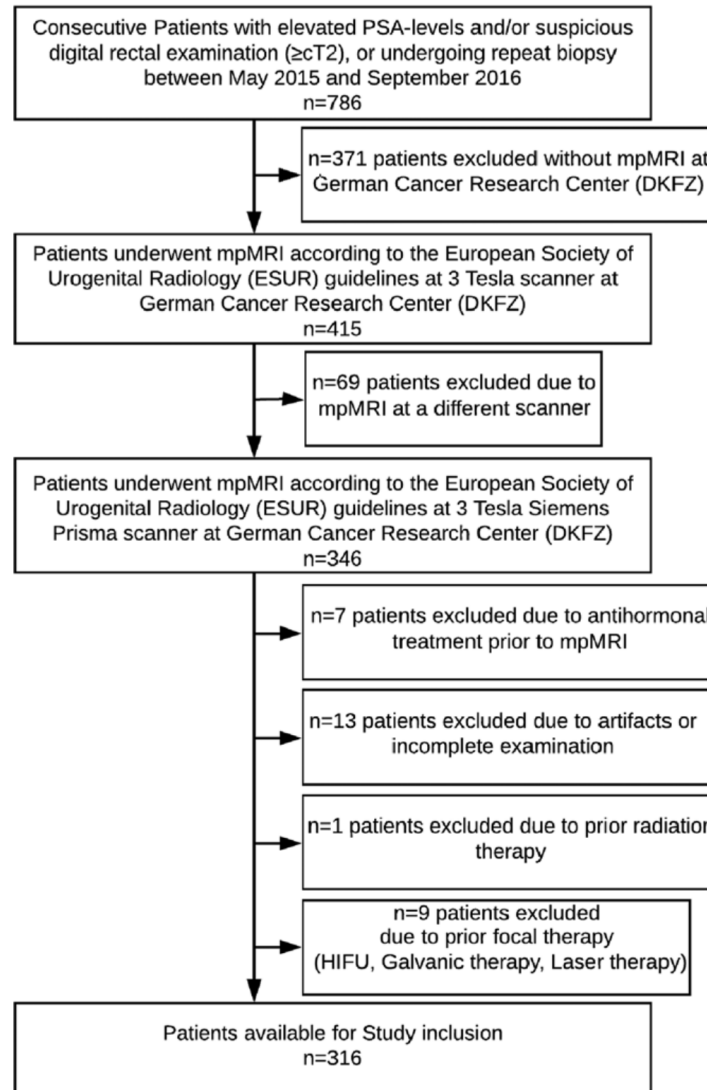


Figure A.1 | Patient Inclusion and Exclusion Flow. Diagram for inclusion of patients into the study, as published in [Bonekamp et al., 2018].

Variable	Training Cohort (<i>n</i> = 183)	Test Cohort (<i>n</i> = 133)
Median age (y) [*]	64.5 (59–71)	63 (58–71)
Median PSA (ng/mL) [*]	6.6 (4.9–9.5)	7.5 (5.4–11)
Median PSA density [*]	0.16 (0.10–0.26)	0.16 (0.11–0.23)
No. of patients without MRI-detected lesions	26	12
No. of patients with MRI-detected lesions [†]	157 (100)	121 (100)
1 lesion	86 (55)	47 (39)
2 lesions	58 (37)	53 (44)
3 lesions	11 (7)	18 (15)
4 lesions	2 (1)	3 (2)
No. of patients with specified maximum Gleason score [†]		
No prostate cancer	76 (42)	50 (38)
6 (3+3)	35 (19)	34 (25)
7a (3+4)	49 (27)	31 (23)
7b (4+3)	8 (4)	7 (5)
8 (4+4)	4 (2)	8 (6)
9a (4+5)	7 (4)	2 (2)
9b (5+4)	4 (2)	1 (1)
No. of patients with specified MRI index lesion [†]		
No lesion	26 (14)	12 (9)
PI-RADS 2	11 (6)	1 (1)
PI-RADS 3	42 (23)	30 (23)
PI-RADS 4	60 (33)	54 (40)
PI-RADS 5	44 (24)	36 (27)
No. of MRI-detected lesions negative for sPC [†]	163 (67)	159 (73)
No. of MRI-detected lesions positive for sPC [†]	80 (33)	60 (27)
Peripheral zone	54 (22)	37 (17)
Transition zone	26 (11)	23 (10)
No. of lesions with specified MRI assessment [†]		
Total	243 (100)	219 (100)
PI-RADS 2	21 (9)	4 (2)
PI-RADS 3	80 (33)	82 (37)
PI-RADS 4	91 (37)	88 (40)
PI-RADS 5	51 (21)	45 (21)
No. of MRI-detected lesions with specified zone distribution [†]		
Peripheral zone	144 (59)	131 (60)
Transition zone	75 (31)	79 (36)
Anterior fibromuscular stroma	17 (7)	9 (4)
Central zone	7 (3)	0 (0)

Note.—PSA = prostate-specific antigen, PI-RADS = Prostate Imaging Reporting and Data System, sPC = clinically significant prostate cancer.

^{*} Data in parentheses are the interquartile range.

[†] Data in parentheses are percentages.

Figure A.2 | Demographic and Clinical Characteristics of Included Patients as published in [Bonekamp et al., 2018].

A.3 Detailed Results

Cohort and Method	Sensitivity (%) [*]	Specificity (%) [*]	No. of FP Lesions [†]	No. of FN Lesions [†]	Reduction Misclassification [‡]	<i>P</i> Value [§]
Global Model						
Training						
Radiologist	79 (63/80)	52 (84/163)	79/163 (reference)	17/80 (reference)		
mADC	79 (63/80)	67 (110/163)	53/163 (26)	17/80 (0)	26/243 (10.7)	.008 #
RML	79 (63/80)	63 (103/163)	60/163 (19)	17/80 (0)	19/243 (7.8)	.035 #
Test						
Radiologist	88 (53/60)	50 (79/159)	80/159 (reference)	7/60 (reference)		
mADC	90 (54/60)	62 (99/159)	60/159 (20)	6/60 (1)	21/219 (9.6)	.048
RML	97 (58/60)	58 (93/159)	66/159 (14)	2/60 (5)	20/219 (9.1)	.176
Combined Zone-specific Models						
Training						
Radiologist	79 (63/80)	52 (84/163)	79/163 (reference)	17/80 (reference)		
mADC	79 (63/80)	67 (110/163)	53/163 (26)	17/80 (0)	26/243 (10.7)	.010 #
RML	79 (63/80)	58 (94/163)	69/163 (10)	17/80 (0)	10/243 (4.1)	.289 [#]
Test						
Radiologist	88 (53/60)	50 (79/159)	80/159 (reference)	7/60 (reference)		
mADC	92 (55/60)	62 (98/159)	61/159 (19)	5/60 (2)	21/219 (9.6)	.038
RML	92 (55/60)	53 (84/159)	75/159 (5)	5/60 (2)	7/219 (3.2)	.596

Note.—FP = false-positive, FN = false-negative, mADC = mean apparent diffusion coefficient, RML = radiomic machine learning.

^{*} Data in parentheses are raw data.

[†] Data in parentheses show the reduction in the number of lesions compared to the reference value (radiologist method).

[‡] Reduction in FP plus FN lesions compared to the reference, divided by the number of FP plus FN lesions. The number in parentheses is the ratio of FP plus FN reduction, divided by all lesions and expressed as a percentage.

[§] McNemar test for differences in specificity to radiologist performance.

^{||} Statistical significance.

[#] *P* values derived after bootstrap analysis have to be regarded with caution.

Figure A.3 | Performance on a per Lesion Basis. Diagnostic performance of the radiologist interpretation (PIRADS), mean ADC (mADC), and the RF ensemble (RML) Learning on the train and the test cohort, as published in [Bonekamp et al., 2018].

Cohort and Method	Sensitivity (%) [*]	Specificity (%) [*]	No. Patients with FP Lesion(s) [†]	No. Patients with FN Lesion(s) [†]	Reduction Misclassification [‡]	P Value [§]
Global Model						
Training						
Radiologist	86 (54/63)	57 (68/120)	52/120 (reference)	9/63 (reference)		
mADC	86 (54/63)	67 (80/120)	40/120 (12)	9/63 (0)	12/183 (6.6)	.180
RML	83 (52/63)	62 (74/120)	46/120 (6)	11/63 (-2)	-4/183 (-2.2)	.429
Test						
Radiologist	89 (40/45)	43 (38/88)	50/88 (reference)	5/45 (reference)		
mADC	93 (42/45)	51 (45/88)	43/88 (7)	3/45 (2)	9/133 (6.8)	.496
RML	96 (43/45)	51 (45/88)	43/88 (7)	2/45 (3)	10/133 (7.5)	.496
Combined Zone-specific Models						
Training						
Radiologist	86 (54/63)	57 (68/120)	52/120 (reference)	9/63 (reference)		
mADC	86 (54/63)	68 (81/120)	39/120 (13)	9/63 (0)	13/183 (7.1)	.134
RML	84 (53/63)	54 (65/120)	55/120 (-3)	10/63 (-1)	-4/183 (-2.2)	.749
Test						
Radiologist	89 (40/45)	43 (38/88)	50/88 (reference)	5/45 (reference)		
mADC	93 (42/45)	51 (45/88)	43/88 (7)	3/45 (2)	9/133 (6.8)	.530
RML	93 (42/45)	38 (33/88)	55/88 (-5)	3/45 (2)	-3/133 (-2.3)	.530

Note.—FP = false-positive, FN = false-negative, mADC = mean apparent diffusion coefficient, RML = radiomic machine learning.

^{*} Data in parentheses are raw data.

[†] Data in parentheses show the reduction in the number of patients compared to the reference value (radiologist method).

[‡] Reduction in FP plus FN lesions compared to the reference, divided by the number of FP plus FN lesions. The number in parentheses is the ratio of FP plus FN reduction, divided by all lesions and expressed as a percentage.

[§] McNemar test for differences in specificity to radiologist performance.

^{||} P values derived after bootstrap analysis have to be regarded with caution.

Figure A.4 | Performance on a per Patient Basis. Diagnostic performance of the radiologist interpretation (PIRADS), mean ADC (mADC), and the RF ensemble (RML) Learning on the train and the test cohort, as published in [Bonekamp et al., 2018].

Appendix B

The Probabilistic U-Net

B.1 Metrics

In the [LIDC-IDRI](#) dataset, given that we have $m = 4$ ground truth samples and n samples from the models, we employ the following statistic:

$$\hat{D}_{\text{GED}}^2(P_{\text{gt}}, P_{\text{out}}) = \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m d(S_i, Y_j) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d(S_i, S'_j) - \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m d(Y_i, Y'_j). \quad (\text{B.1})$$

Here $d(x, y) = 1 - \text{IoU}(x, y)$, where x and y are the predicted and ground truth masks of the lesion. In the case that both are empty masks, we define its distance to be 0, so that the metric rewards the agreement on lesion absence.

On the Cityscapes task, given that we have defined the settings, we have full knowledge about the ground truth distribution, which is a mixture of $M = 32$ Dirac delta distributions. Hence, we do not need to sample from it, but use it directly in the estimator:

$$\hat{D}_{\text{GED}}^2(P_{\text{gt}}, P_{\text{out}}) = \frac{2}{n} \sum_{i=1}^n \sum_{j=1}^M d(S_i, Y_j) \omega_j - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d(S_i, S'_j) - \sum_{i=1}^M \sum_{j=1}^M d(Y_i, Y'_j) \omega_i \omega_j, \quad (\text{B.2})$$

where ω_j is the weight for the j -th mixture, which is a delta distribution containing all the density in Y_j . Here the distance d depends on the average IoU of the 10 switchable classes only. Predicting one of such classes that is not present in the ground truth leads to a 0 score, which will be one of the terms over which we average. The computed average does not account for classes that are not present in both prediction and ground truth.

B.2 How models fit the ground truth distribution

In this section we analyse the frequency in which each mode of the Cityscape task is targeted by each model, and how much that varies from the ground truth distribution. We report the mode-wise and pixel-wise marginal occurrence frequencies of the sampled segmentation variants. In the mode-wise case, each sample is matched to its closest ground truth mode (using 1-IoU as the distance function). Then, the frequency of each

Table B.1 | Numerical (mean) results of the Probabilistic U-Net on LIDC. The full distributions are shown in Fig. 7.4a.

# Samples	1	4	8	16
\hat{D}_{GED}^2	0.811	0.388	0.321	0.287

mode is computed by counting the number of samples that most closely match that mode. In the pixel-wise case, the marginal frequencies $p(\text{predicted class}|\text{ground-truth class})$ are obtained by counting all pixels across all images and corresponding samples that show a valid pixel hypothesis given the ground-truth, normalized by the number of respective uni-modal ground-truth pixels. In Fig. B.1 we present the results for U-Net Ensemble and Dropout U-Net, in Fig. B.2 we show the results for M-Heads and Image2Image VAE, finally in Fig. B.3 we present the results for our approach.

B.3 Sampling LIDC masks using different models

Fig. B.4-B.8 show samples of our proposed model as well as all the baselines given the same input images. For reference the expert segmentations are shown in the four rows just below the images. Table B.1 shows the numerical results from Fig. 7.4a.

B.4 Sampling Cityscapes segmentations using our model

Fig. B.9 shows samples of our proposed model on the Cityscapes dataset, and Table B.2 shows the numerical results from Fig. 7.4b, so that new approaches can be compared to those.

Table B.2 | Numerical (mean) results of the Probabilistic U-Net on Stochastic Cityscapes. The full distributions are shown in Fig. 7.4b.

# Samples	1	4	8	16
\hat{D}_{GED}^2	0.874	0.337	0.248	0.206

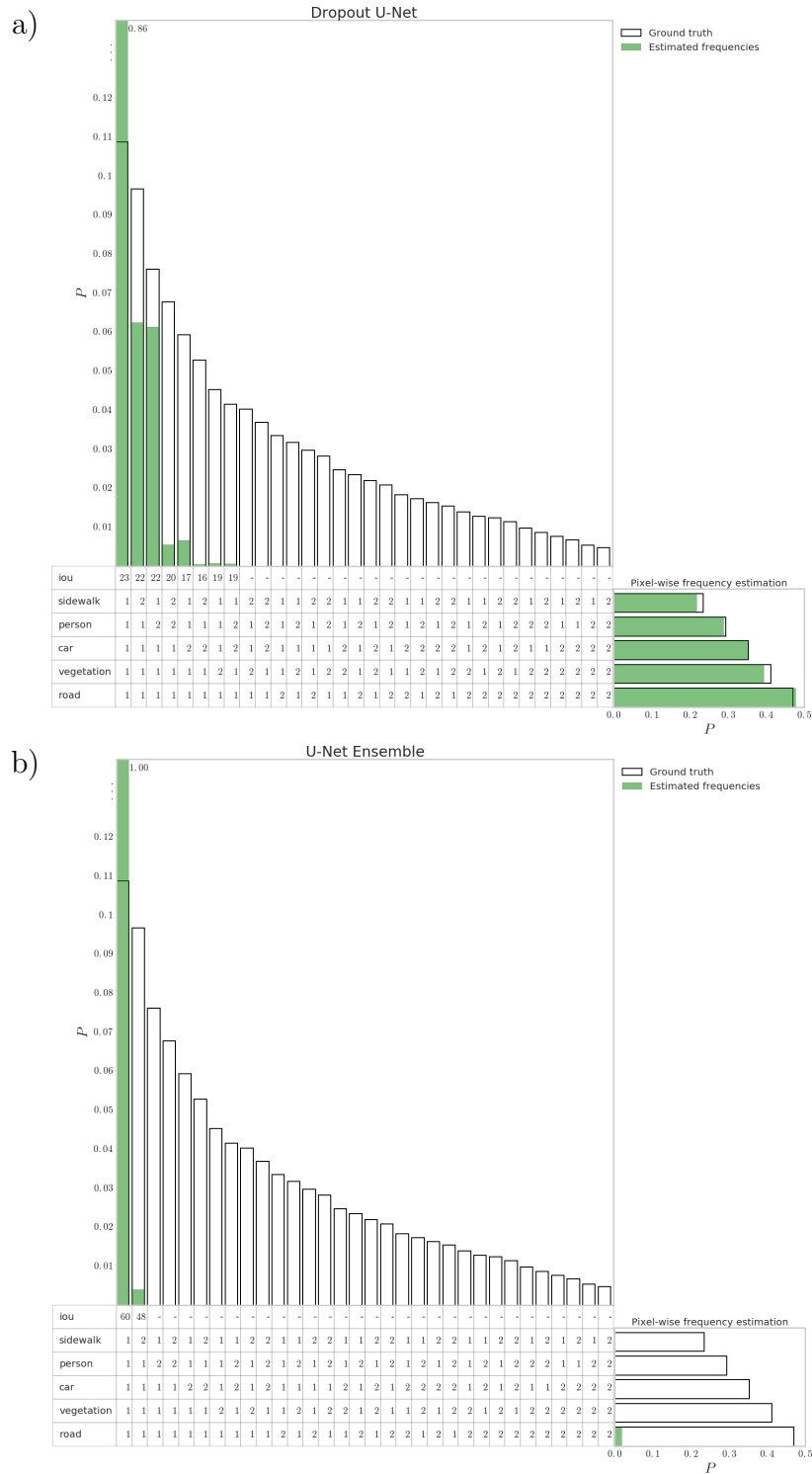


Figure B.1 | Probability calibration of the Dropout U-Net and U-Net Ensemble. The vertical histogram shows the mode-wise occurrence frequencies of samples in comparison to the ground-truth probability of the modes, and the horizontal histogram reports the pixel-wise marginal frequencies, i.e. the sampled pixel-fractions for each new stochastic class (e.g. sidewalk 2) with respect to the corresponding existing one (sidewalk).

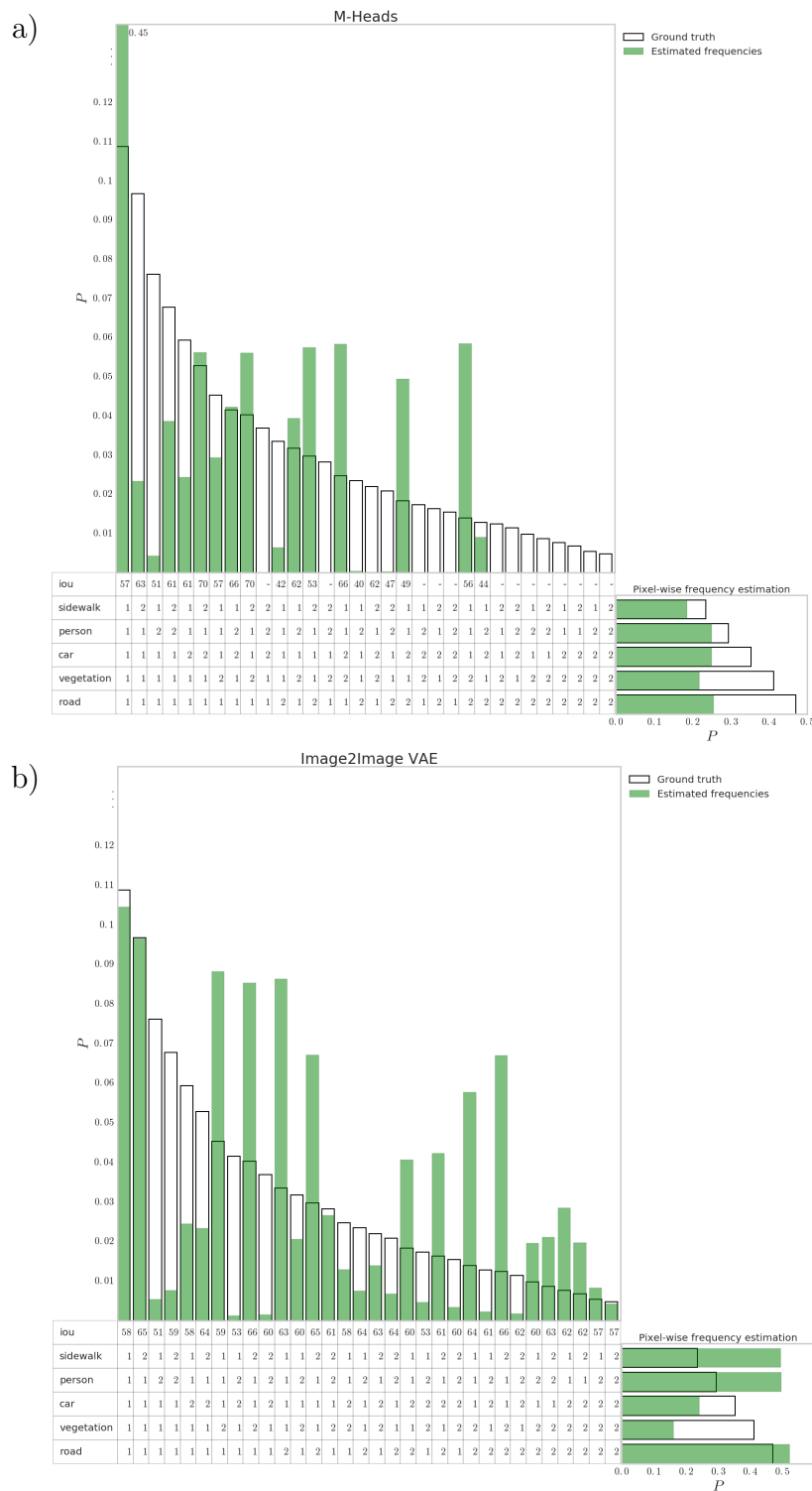


Figure B.2 | Probability calibration of M-Heads and Image2Image VAE. The vertical histogram shows the mode-wise occurrence frequencies of samples in comparison to the ground-truth probability of the modes, and the horizontal histogram reports the pixel-wise marginal frequencies, i.e. the sampled pixel-fractions for each new stochastic class (e.g. sidewalk 2) with respect to the corresponding existing one (sidewalk)

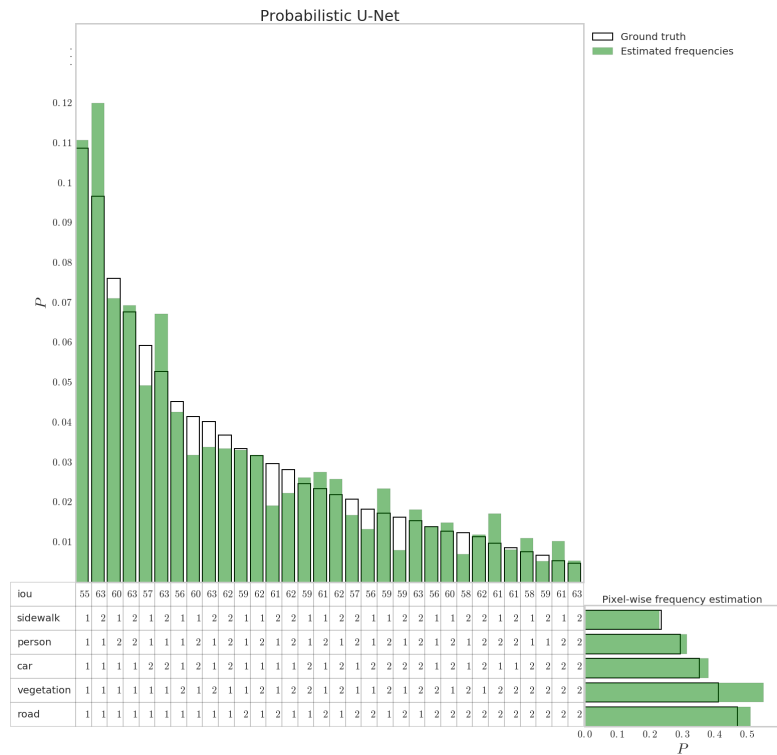


Figure B.3 | Probability calibration of the Probabilistic U-Net. The vertical histogram shows the mode-wise occurrence frequencies of samples in comparison to the ground-truth probability of the modes, and the horizontal histogram reports the pixel-wise marginal frequencies, i.e. the sampled pixel-fractions for each new stochastic class (e.g. sidewalk 2) with respect to the corresponding existing one (sidewalk).

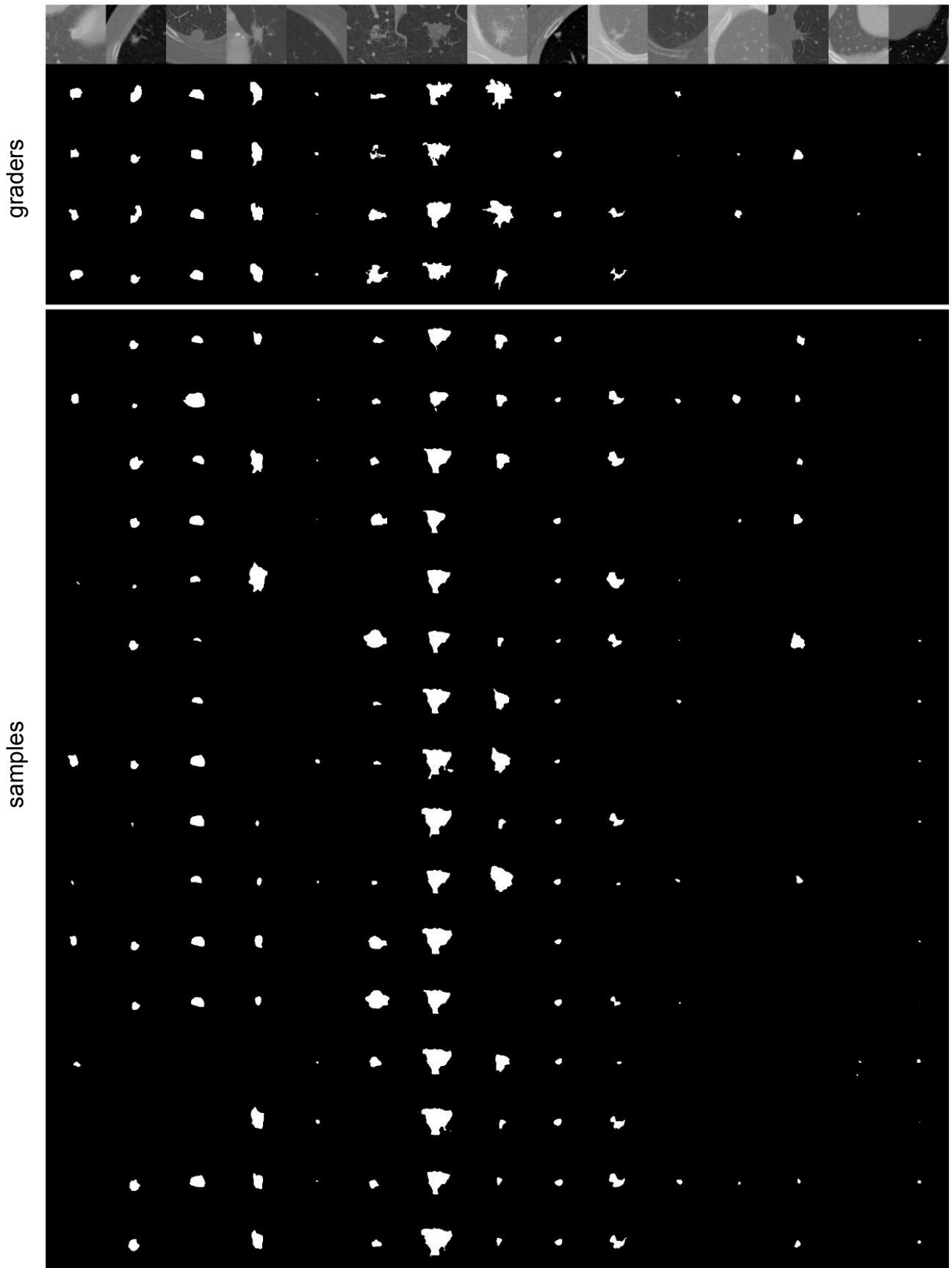


Figure B.4 | LIDC samples from the Probabilistic U-Net. The upper panel shows LIDC test set images from 15 different subjects alongside the respective ground-truth masks by the 4 graders. The panel below gives the corresponding 16 random samples from the network.

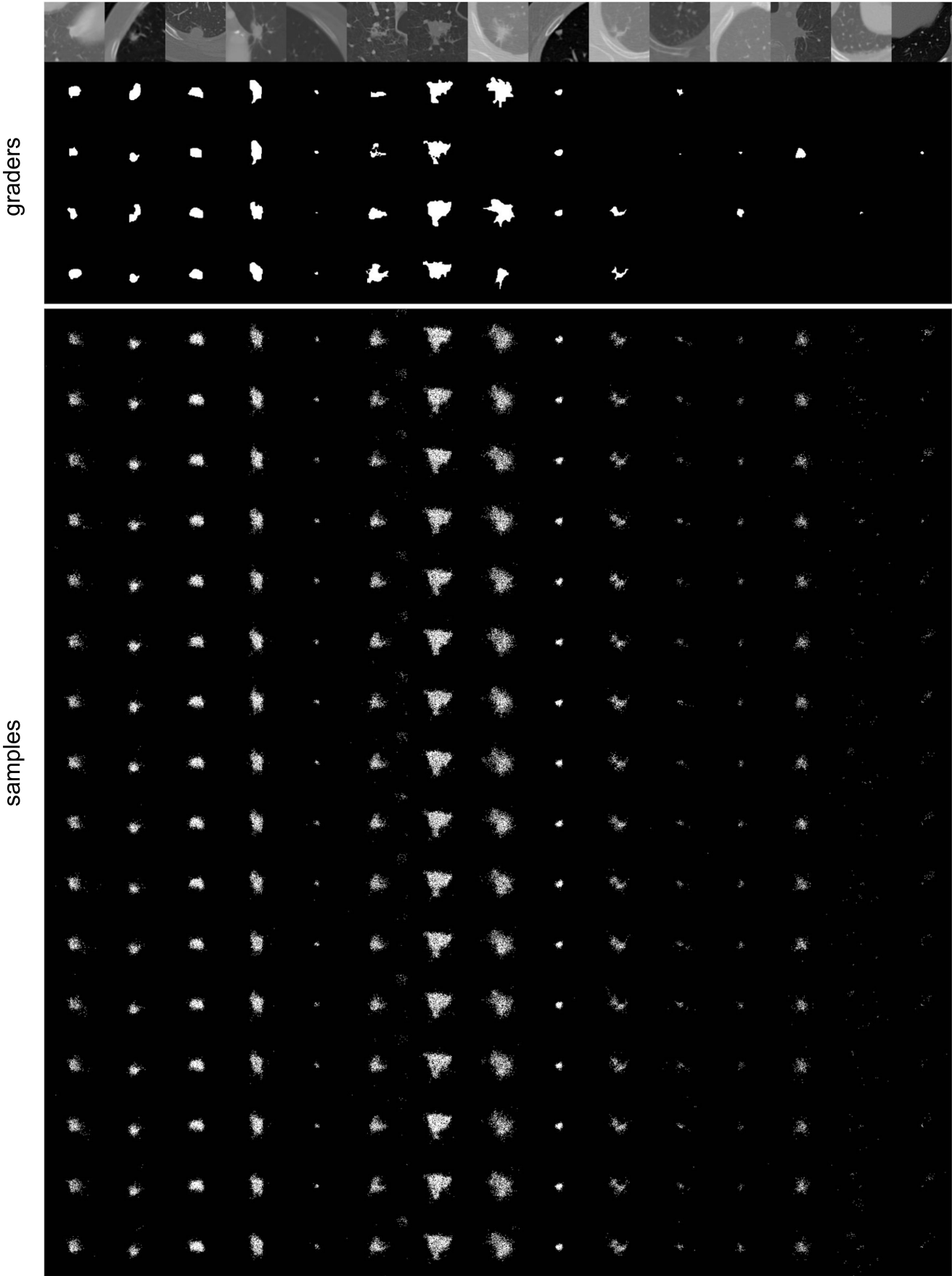


Figure B.5 | LIDC samples from the Dropout U-Net. Same layout as Fig. B.4.

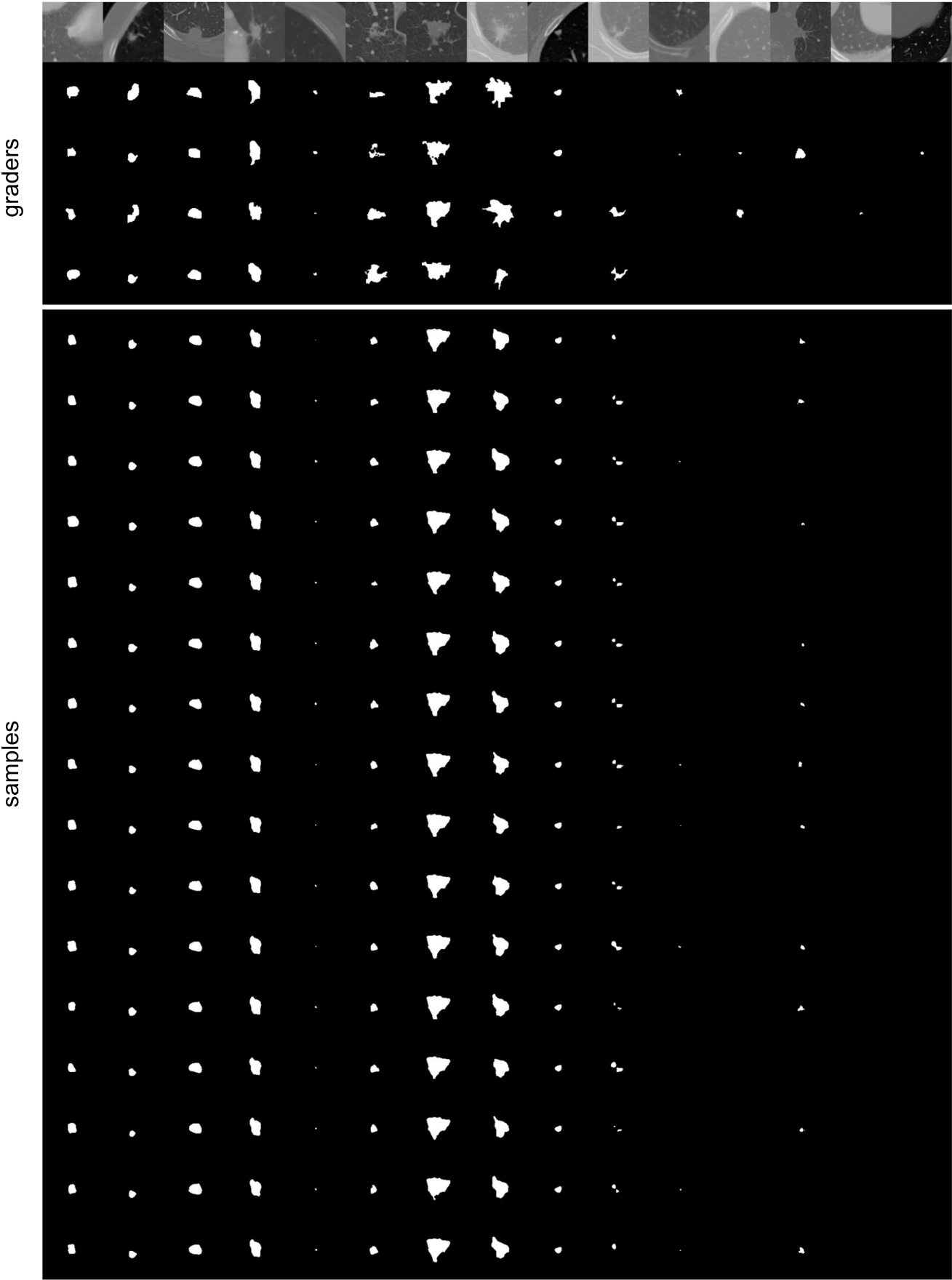


Figure B.6 | LIDC samples from the U-Net Ensemble. Same layout as Fig. B.4.

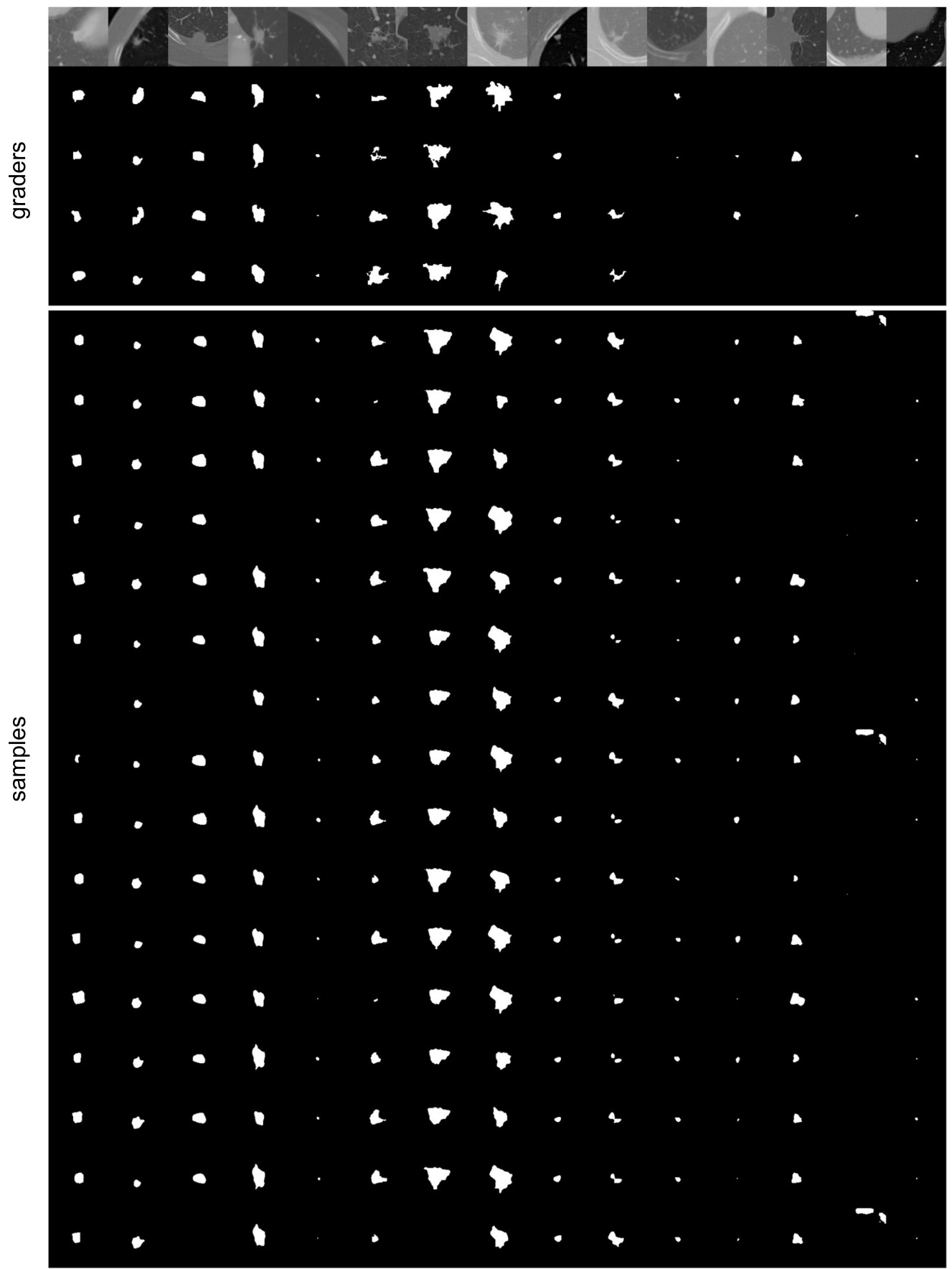


Figure B.7 | LIDC samples from the M-Heads model (using a network with 16 heads). Same layout as Fig. B.4.

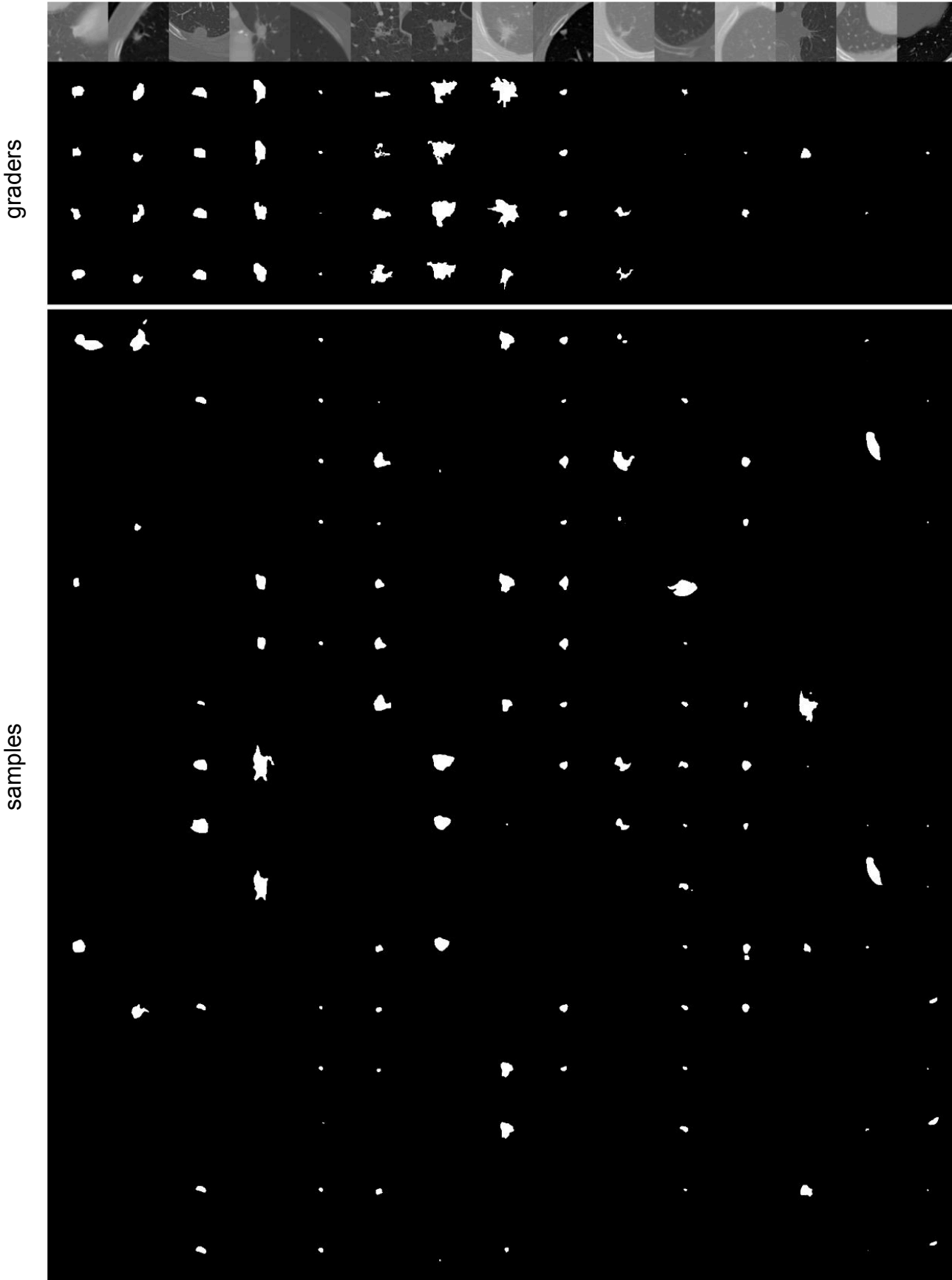


Figure B.8 | LIDC samples from the Image2Image VAE. Same layout as Fig. B.4.

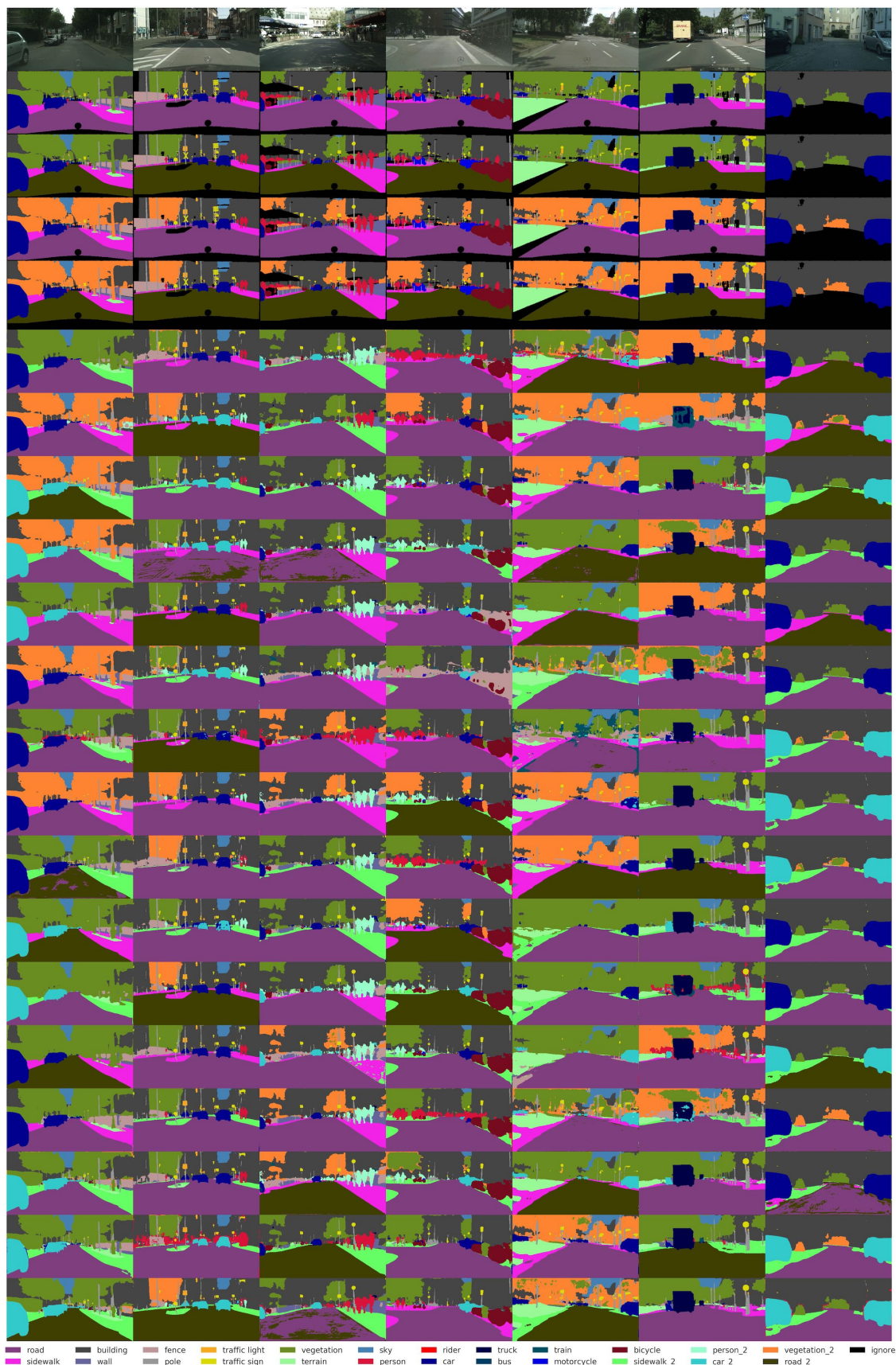


Figure B.9 | Stochastic Cityscapes samples from the Probabilistic U-Net. The first row shows Cityscapes images, the following 4 rows show 4 out of the 32 ground truth modes with black pixels denoting pixels that are masked during evaluation. The remaining 16 rows show random samples of the network.

B.5 Training Details

In this section we describe the architecture settings and training procedure for both experiments.

B.5.1 Lung Abnormalities Segmentation

We only use those lesions that were specified as a polygon (outline) in the XML files of the LIDC dataset, disregarding the ones that only have center of shape. That is, according to the LIDC-IDRI paper [Armato et al., 2011] we use the ones that are larger than 3mm, and filtering out the others, that are clinically less relevant [Armato et al., 2011]. We also filter out each Dicom file whose absolute value of SliceLocation differs from the absolute value of ImagePositionPatient[-1]. Finally we assume that two masks from different graders correspond to the same lesion if their tightest bounding boxes overlap.

During training image-grader pairs are drawn randomly. We apply augmentations to the image tiles (180×180 pixels size): random elastic deformation, rotation, shearing, scaling and a randomly translated crop that results in a tile size of 128×128 pixels. The U-Net architecture we use is similar to [Ronneberger et al., 2015] with the exception that we down- and up-sample feature maps by using bilinear interpolations. The cores of all models are identical and feature 4 down- and up-sampling operations, at each scale the blocks comprise three convolutional layers with 3×3 -kernels, each followed by a ReLU-activation. In our model, both the prior and the posterior (as well as the posterior in Image2Image VAE) nets have the same architecture as the U-Net’s encoder path, i.e. they are made up to the same number of blocks and type of operations. Their last feature maps are global average pooled and fed into a 1×1 convolution that predicts the Gaussian distributions parameterized by mean and standard deviation. The architecture last layers, corresponding to $f_{\text{comb.}}$, comprise the appropriate number of 1×1 -kernels and are activated with a softmax. The base number of channels is 32 and is doubled or respectively halved at each down- or up-sampling transition. All individual models share this core component and for ease of comparability we let all models undergo the same training schedule: the training proceeds over 240k iterations with an initial learning rate of $1e^{-4}$ that is lowered to $1e^{-6}$ in 5 steps. All weights of all models are initialized with orthogonal initialization having the gain (multiplicative factor) set to 1, and the bias terms are initialized by sampling from a truncated normal with $\sigma = 0.001$. We use a batch-size of 32, weight-decay with weight $1e^{-5}$ and optimize using the Adam optimizer with default settings [Kingma and Ba, 2014]. A KL weight of $\beta = 10$ with a latent space of 3 dimensions gave best validation results for the baseline Image2Image VAE, and $\beta = 1$ and a 6D latent space performed well for the Probabilistic U-Net, although the performances were alike across the hyperparameters tried on the validation set.

B.5.2 Stochastic Cityscapes Street Scene Segmentation

We down-sample the Cityscapes images and label maps to a size of 256×512 . Similarly to above, we apply random elastic deformation, rotation, shearing, scaling, random translation and additionally impose random color augmentations on the images during training. The U-Net cores in this task are identical to the ones above, but process an additional feature scale (implying one additional up- and one additional down-sampling

operation). The training procedure is also equivalent to the previous experiment, also using 240k iterations, except that here we employ a batch-size of 16, and the initial learning rate of $1e^{-4}$ is lowered to $1e^{-5}$ in 3 steps. The Cityscapes dataset includes ignore label masks for each image with which we mask the loss during training, and the metric during evaluation. A KL weight of $\beta = 1$ and 3D latents gave best validation results for the Image2Image VAE and a $\beta = 1$ and 6D latents performed best for the Probabilistic U-Net (although 3-5D performed similarly).

Appendix C

The Hierarchical Probabilistic U-Net

C.1 Instance Segmentation Post-Processing

For the instance segmentation experiments we post-process the clustered samples to remove tiny regions that sometimes appear at segmentation borders. For each cluster (instance) found via [Algorithm 1](#), we check whether it survives an erosion operation with an $n \times n$ -structure element. If the given erosion eliminates the cluster, we replace each pixel within the cluster in question by its majority neighbour label. The majority neighbour label is determined from a $m \times m$ -box centered at the given pixel. The cluster label that is to be replaced as well as background labels are ignored while finding the majority label. If this results in 0 valid neighbour labels, we keep the current pixel’s label in SNEMI3D and use the background label in Cityscapes. In both SNEMI3D and Cityscapes, we chose $n = 5$ and $m = 11$. Painting in the majority label is carried out on the fly.

Training The HPU-Net is trained using the GECO-objective ([Eq. 8.12](#)) and a stochastic top-k reconstruction loss. As described in [Subsec. 8.2.3](#), the k th percentile employed for the top-k objective is fixed across tasks to 2% of each batch’s pixels. The GECO-objective aims at matching a reconstruction target value κ . For each experiment we chose κ sufficiently low so as to correspond to a strong reconstruction performance while resulting in a training schedule that is not dominated by the reconstruction term over the entire course of the training (e.g. if κ is chosen too high, the Lagrange multiplier λ , and thus the learning pressure it exerts, mounts and remains on the reconstruction term rather than moving over on the KL terms). The desired behavior of the reconstruction objective \mathcal{L}_{rec} and the Lagrange multiplier λ can be observed in [Fig. C.1](#) and [Fig. C.2](#), where λ rises until \mathcal{L}_{rec} matches κ , after which λ drops and the pressure on the KL-terms increases.

In contrast to the regular cross-entropy employed in semantic segmentation, the reconstruction error here is not averaged but summed over individual pixels (before being averaged across batch instances). This is because the likelihood is assumed to factorize over the pixels of an image and so their log-likelihood is summed over. For comparability we however report \mathcal{L}_{rec} and κ per pixel (e.g. in [Fig. C.1](#), [Fig. C.2](#) and in [Table 8.2](#)).

The precise training setups for each of the tasks and models are reported below. Note that the training objectives for all models encompass an additional weight-decay term that is weighted by a factor of $1e^{-5}$.

C.2 Training Details by Dataset

C.2.1 LIDC-IDRI Lung CT scans

During training on LIDC, image-grader pairs are drawn randomly. Similar to [Subsec. B.5.1](#), we apply random augmentations¹ to the image and label tiles (180×180 pixels size) including random elastic deformation, rotation, mirroring, shearing, scaling and a randomly translated crop that results in a tile size of 128×128 pixels. Any padding added to the images and labels during the augmentation process is masked from the loss during training.

In order to evaluate the Probabilistic U-Net on additional metrics than those employed in [\[Kohl et al., 2018\]](#) and in order to bootstrap from 10 model instantiations, we retrain a re-implementation of the model with the exact same hyperparameters and setup as described in [Subsec. B.5.1](#), i.e. we employ a 5-scale model, with three 3×3 -convolutions per encoder and decoder-scale, a separate prior and posterior net that mirror the used U-Net’s encoder as well as 6 global latents and three final 1×1 convolutions. Moreover we employ an identical ELBO-formulation ($\beta = 1$), train with identical batch-size of 32, number of iterations ($240k$) and learning rate schedule $0.5e^{-5} \rightarrow 1e^{-6}$.

On LIDC, the [HPU-Net](#) uses 8 latent scales resulting in a global 1×1 -‘U-Net bottom’ and 3 res-blocks per encoder and decoder scale. The base number of channels is 24 and until the fourth down-sampling the number of channels is doubled after each down-sampling operation, resulting in a maximum width of 192 channels. The U-Net’s decoder mirrors this setup. We train the [HPU-Net](#) with an initial learning rate of $1e^{-4}$ that is lowered to $0.5e^{-5}$ in 4 steps over the course of $240k$ iterations. The employed batch-size is 32. The [HPU-Net](#) is trained with the GECO-objective using $\kappa = 0.05$.

[Fig. C.1](#) shows how the top-k reconstruction term \mathcal{L}_{rec} , the Lagrange multiplier λ , as well as the individual KL-terms (and their sum) progress in the course of training for the 10 random model initializations reported in [Table 8.1](#). As mentioned above, GECO structures the dynamics such that λ puts pressure on \mathcal{L}_{rec} until it reaches its target value κ . After that the training objective holds the reconstruction term at κ while optimizing for lower overall [Kullback-Leibler divergence \(KL\)](#). The KL is a measure for how much more information the posterior distribution carries compared to the prior, a quantity that we aim to minimize. Note that the KL-sum is very similar for all models, but the way the KL splits across the hierarchy can differ. The models that end up using the global latents profit from a slightly lower overall KL indicating that this decomposition is more efficient, e.g. it is more efficient not to repeat global information in the local latents when it is already provided by global latents etc.

C.2.2 SNEMI3D neocortex EM slices

During training on [SNEMI3D](#) we randomly sample a latent (class) id for each cell in each image. We limit the number of instance ids to 15 and just like on [LIDC](#) we apply random augmentations including random elastic deformation, rotation, mirroring, shearing, scaling and a randomly translated crop. Any padding added to the images and labels during the augmentation process is masked from the loss during training.

¹We use the code available at <https://github.com/deepmind/multidim-image-augmentation/>.

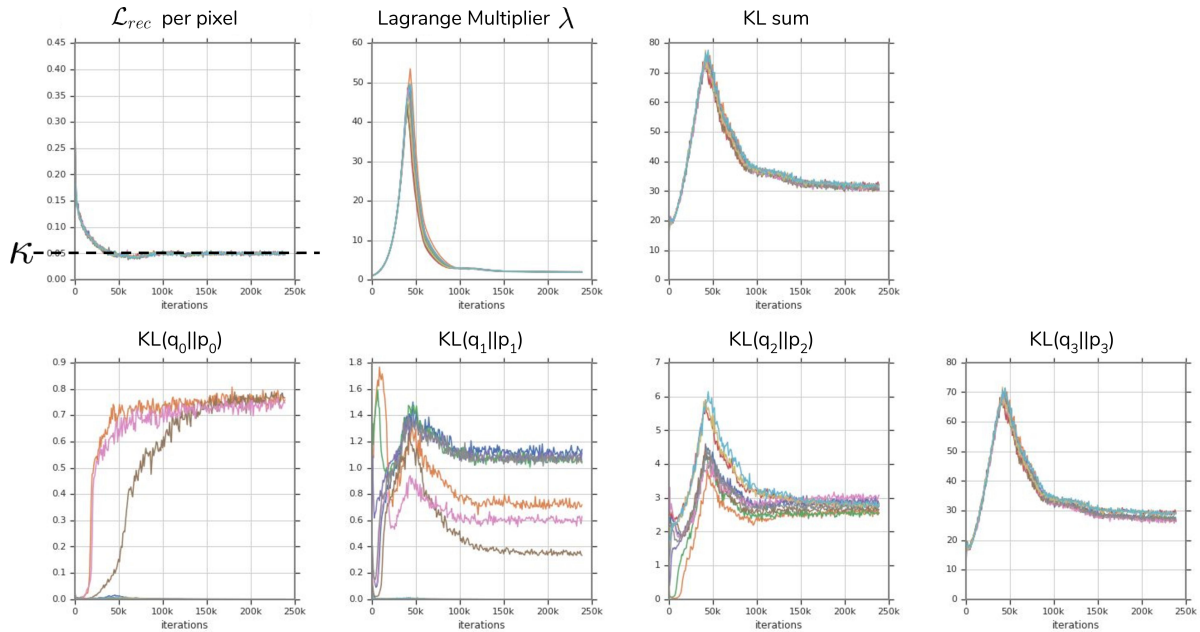


Figure C.1 | Losses during training on LIDC. Components of the learning objective in the course of the LIDC training for 10 random initializations.

For the standard Probabilistic U-Net we employ a 9 scale architecture and a base number of 24 channels, that until the 4th down-sampling, is doubled after each down-sampling operation, resulting in a maximum width of 192 channels. The sPU-Net again uses three 3×3 -convolutions per encoder and decoder scale, while the HPU-Net employs three res-blocks. The HPU-Net also employs 32 base channels, a total of 9 scales interleaved with four (scalar) latent scales, resulting in a total of 85 latents. This is also the number of global latents that we used for the sPU-Net, since employing low numbers of latents, such as ~ 10 as used on the datasets in Chap. 7 never converged (even working with 85 global latents does not make for a very stable training). Both models are trained for 450k iterations with a batch-size of 24, and an initial learning rate of $1e^{-4}$ that is lowered to $1e^{-7}$ in 5 steps. The HPU-Net is trained with the GECO-objective using $\kappa = 1.20$.

Fig. C.2 again shows how the top-k reconstruction term \mathcal{L}_{rec} , the Lagrange multiplier λ , as well as the individual KL-terms (and their sum) progress in the course of training for the 10 random model initializations reported in Table 8.1. Again the KL sums to a similarly low value across models with different decompositions across the four scales.

C.2.3 Cityscapes Car Instances

We resample the Cityscapes images and labels to half-resolution, i.e. 512×1024 . During training we randomly sample a (latent) instance id for each car in the image, where we limit the total number of car ids to 5. We apply random deformations including random color augmentations, elastic deformation, rotation, mirroring, shearing, scaling and a randomly translated crop. Any padding added to the images and labels during the augmentation process is masked from the loss during training alongside any such

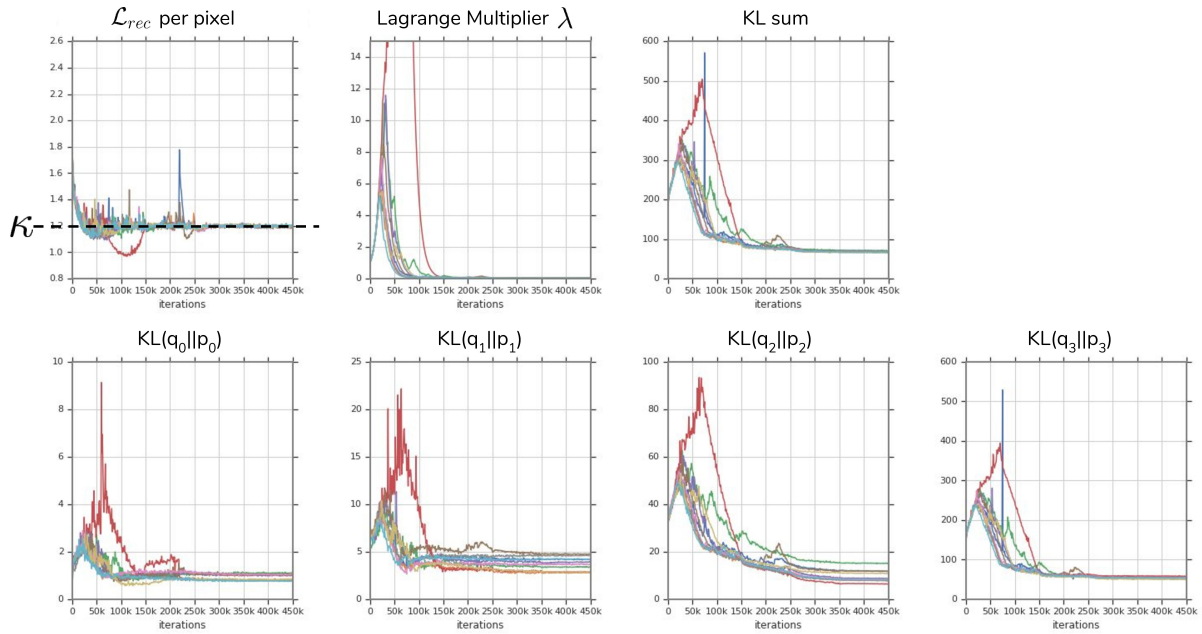


Figure C.2 | Losses during training on SNEMI3D. Components of the learning objective in the course of the SNEMI3D training for 10 random initializations.

pixels that are marked as part of the ‘ignore’-class in the dataset (pixels that can’t be attributed to one of the provided 19 classes).

We train a [HPU-Net](#) with 9 scales, resulting in a 2×4 -‘U-Net bottom’ and 4 latent scales. Using another scale (so 5 latent scales and a number of 10 overall scales) did not significantly change the results and due to the image aspect ratio of 1:2, does not result in a fully global latent scale either. The employed model uses two res-blocks for each encoder and decoder scale and we train the model with a batch-size of 128 for 100k iterations using TPU accelerators and spatial batch partitioning. We use an initial learning rate of $2e^{-4}$ that is halved after 70k iterations. The base number of channels is 32 and until the fourth down-sampling the number of channels are doubled after each down-sampling operation, resulting in a maximum width of 256 channels. The [HPU-Net](#) is trained with the GECO-objective using $\kappa = 0.77$.

C.3 GED^2 on LIDC subset B

On ‘Subset B’ the [sPU-Net](#) achieves a $GED^2 = 0.52 \pm 0.09$ while the [HPU-Net](#) achieves as $GED^2 = 0.38 \pm 0.02$. Both values result from the set of 10 models used for the LIDC results in [Table 8.1](#) (again using 1000 bootstraps with replacement).

Table C.1 | Baseline Test Set Results on LIDC. The mean and standard deviations for the baselines are calculated from one random initialization and 1000 bootstraps with replacement. Both GED² and the Hungarian-matched IoU are calculated using 16 samples.

LIDC	GED ²	Hung.-m. IoU
U-Net Ensemble	0.49 ± 0.01	0.50 ± 0.01
Dropout U-Net	0.47 ± 0.00	0.24 ± 0.00
M-Heads	0.33 ± 0.00	0.47 ± 0.00
Image2Image VAE	0.41 ± 0.01	0.44 ± 0.00

C.4 GED² and Hungarian-matched IoU for Baselines on LIDC

C.5 LIDC, SNEMI3D and Cityscapes: Extra Examples

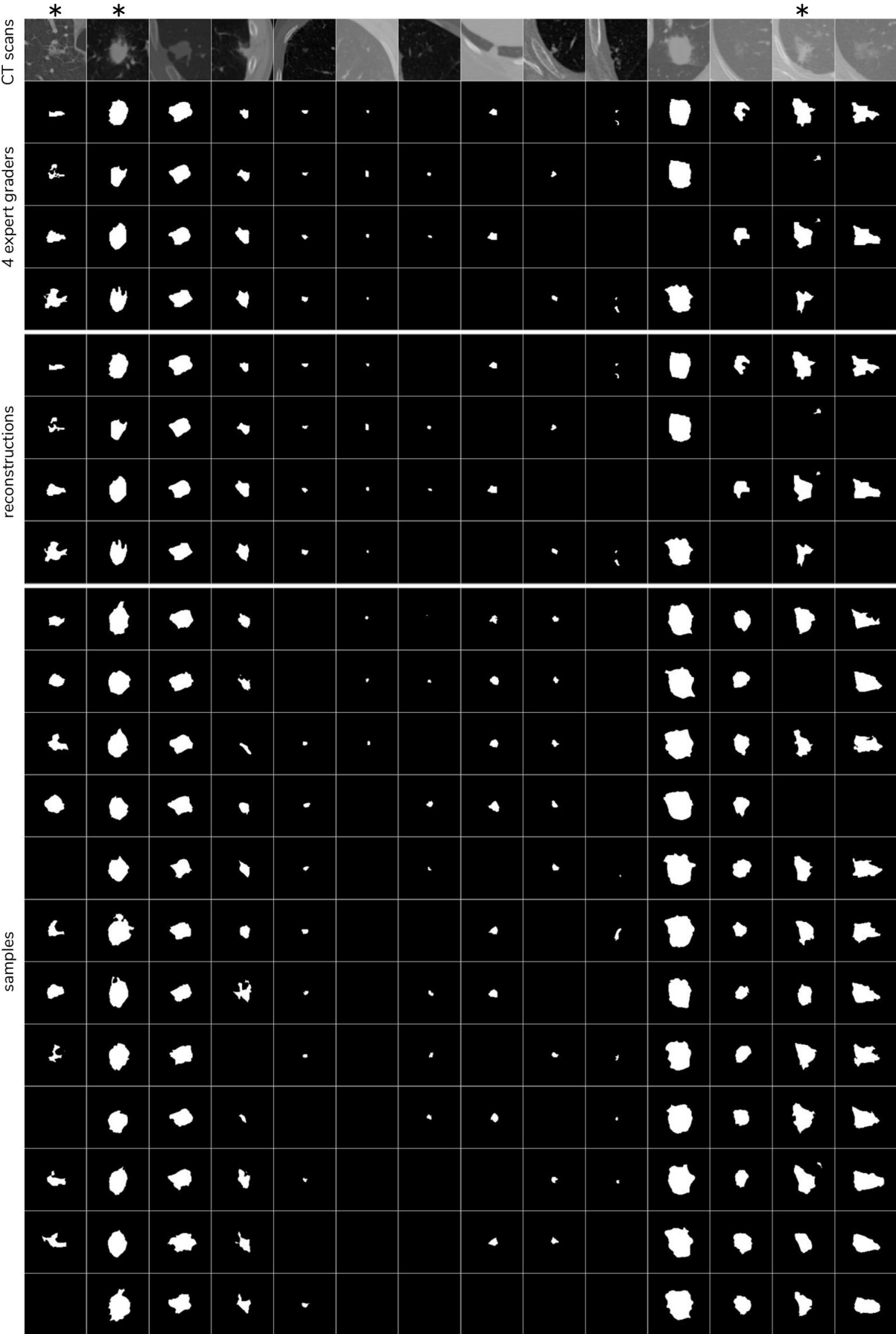


Figure C.3 | HPU-Net examples on LIDC. Qualitative results on test set. An asterisk (*) denotes cases that we also use in Fig. 8.3.

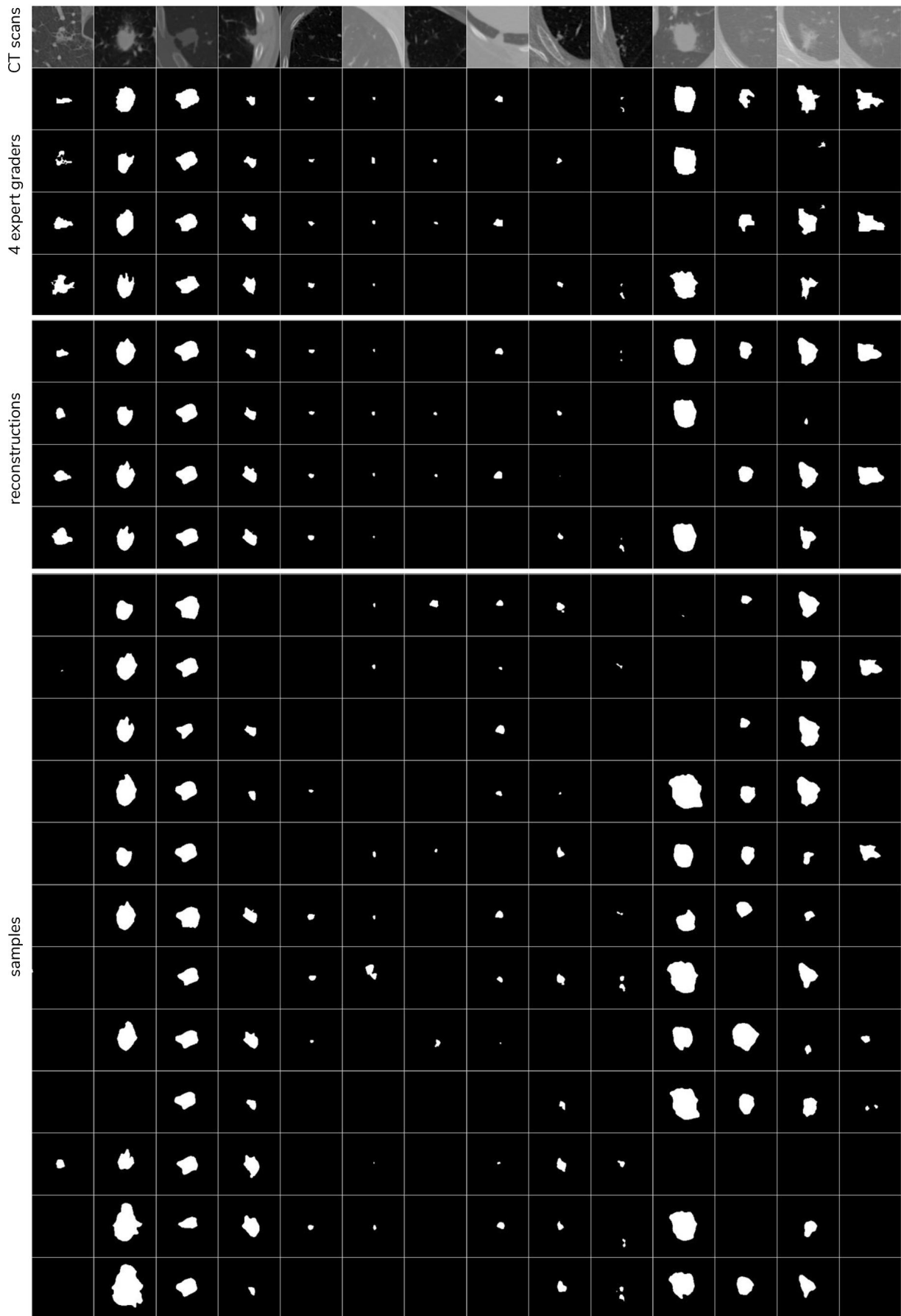


Figure C.4 | sPU-Net examples on LIDC. Qualitative results on our test set.

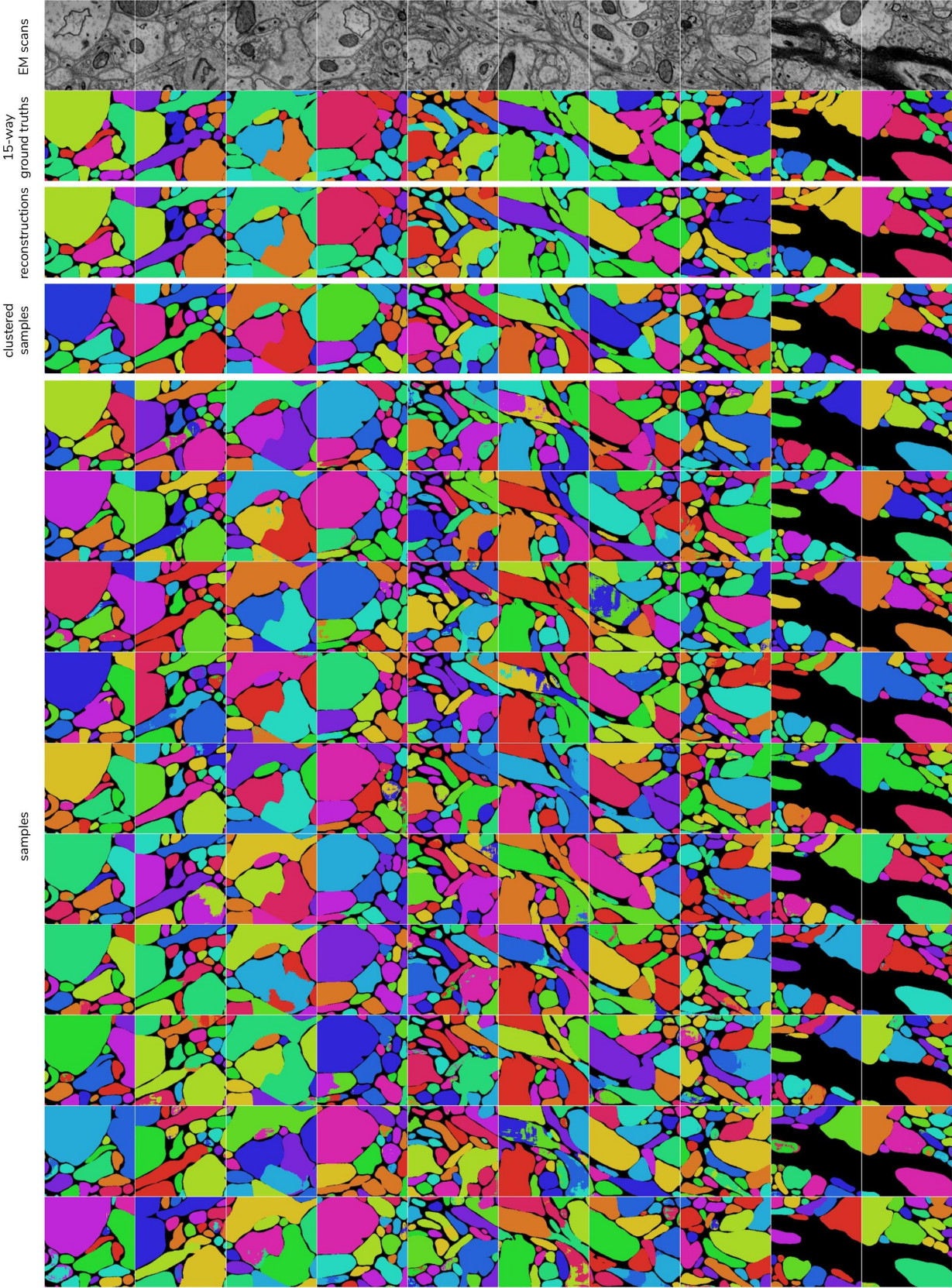


Figure C.5 | HPU-Net examples on SNEMI3D. Qualitative results on our test set.

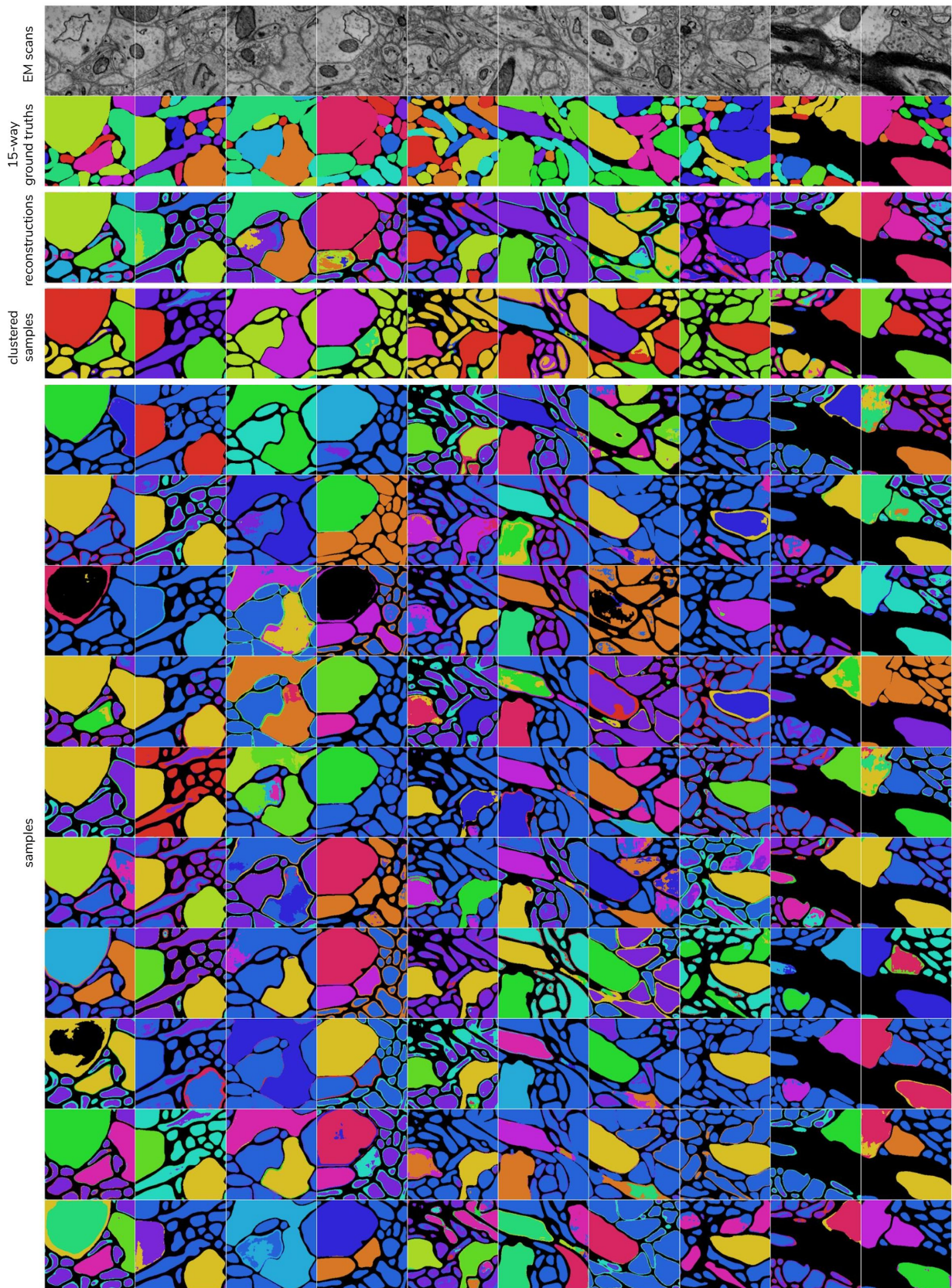


Figure C.6 | sPU-Net examples on SNEMI3D. Qualitative results on our test set.

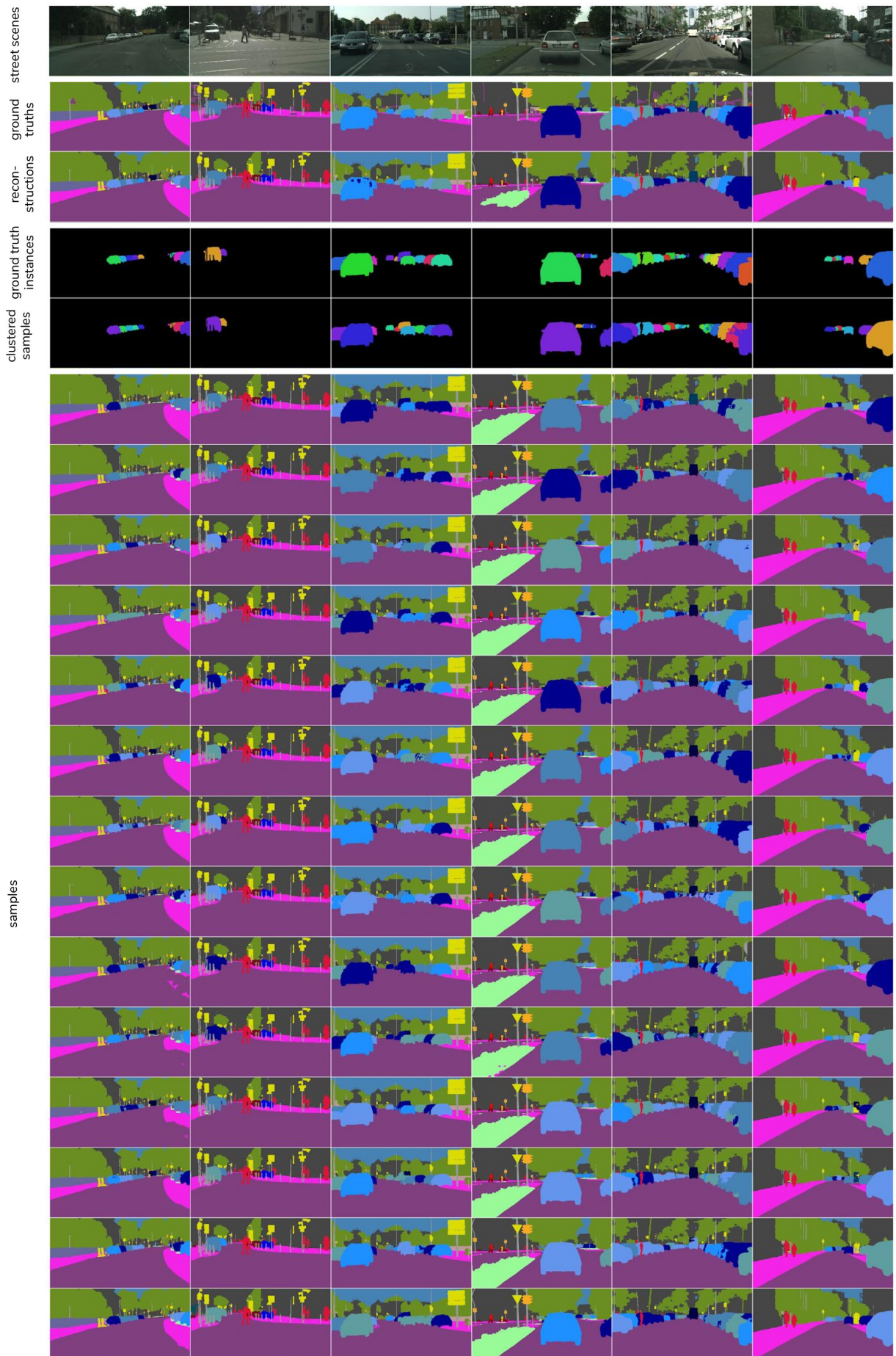


Figure C.7 | HPU-Net examples for Instance Segmentation on Cityscapes. Qualitative results on our test set for a model trained with 5 distinct latent car ids on resolution 512×1024 . The 5 car ids take on different shades of blue. Samples show good consistency across individual car instances resulting in high-quality instance segmentations, see the 4th row from the top. Note how the model flips other natural ambiguous regions aside from cars e.g. *street* \leftrightarrow *sidewalk* in the first scene and *truck* \leftrightarrow *bus* in the second last.

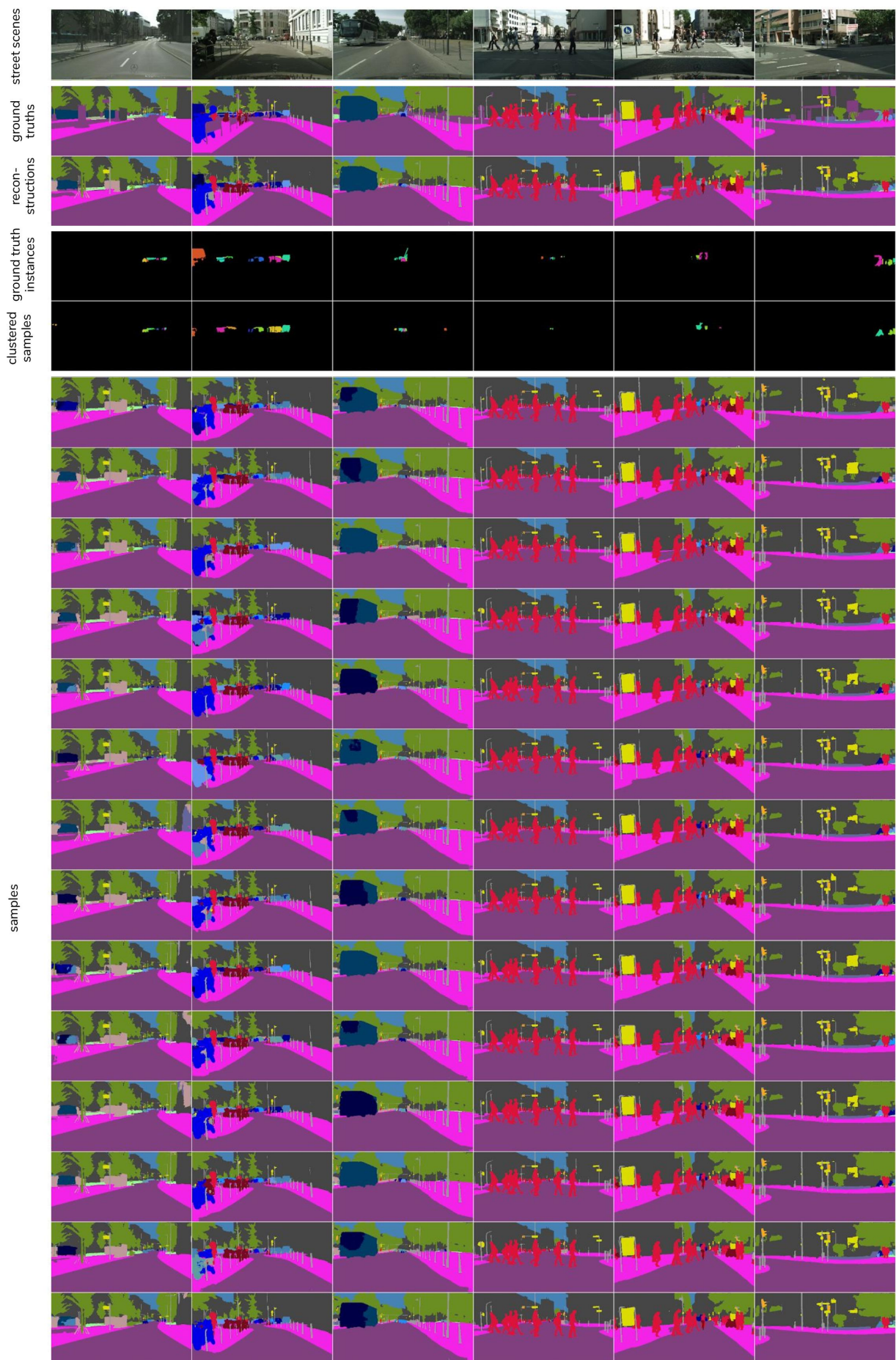


Figure C.8 | HPU-Net examples for Instance Segmentation on Cityscapes (difficult cases). Qualitative results on our test set for a model trained with 5 distinct latent car ids on resolution 512×1024 . The 5 car ids take on different shades of blue. Here we show the top difficult cases in the test set in terms of the Rand error, which shows the difficulty of segmenting individual cars when they are very distant in the scene or heavily occluded.

References

- Hugo JW Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5:4006, 2014.
- Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G Hindley, Alex Freeman, et al. Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study. *The Lancet*, 389 (10071):815–822, 2017.
- American Cancer Society. Key Statistics for Prostate Cancer, 2016a. URL <https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html>. Accessed: 2019-07-05.
- American Cancer Society. Tests for Prostate Cancer, 2016b. URL <https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/how-diagnosed.html>. Accessed: 2019-07-14.
- American Cancer Society. Cancer Facts and Figures 2019, 2019. URL <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf>. Accessed: 2019-07-05.
- James Anderson. An Introduction to Routine and Special Staining, 2019. URL <https://www.leicabiosystems.com/pathologyleaders/an-introduction-to-routine-and-special-staining/>. Accessed: 2019-07-15.
- Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 9:142, 2015.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- III Armato, G. Samuel, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, and Laurence P. Clarke. Data from lidc-idri. the cancer imaging archive. <http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX>, 2015.

- Samuel G Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.
- Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018.
- Jelle O Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J Fütterer. Esur prostate mr guidelines 2012. *European radiology*, 22(4):746–757, 2012.
- Jan Barghaan. A Closer Look at Microscopy, 2015. URL <https://thepathologist.com/inside-the-lab/a-closer-look-at-microscopy>. Accessed: 2019-07-15.
- Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse m-best solutions in markov random fields. In *European Conference on Computer Vision*, pages 1–16. Springer, 2012.
- Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlemaier, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. *arXiv preprint arXiv:1906.04045*, 2019.
- Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.
- Sebastian Bickelhaupt, Paul Ferdinand Jaeger, Frederik Bernd Laun, Wolfgang Lederer, Heidi Daniel, Tristan Anselm Kuder, Lorenz Wuesthof, Daniel Paech, David Bonekamp, Alexander Radbruch, et al. Radiomics based on adapted diffusion kurtosis imaging helps to clarify most mammographic findings suspicious for cancer. *Radiology*, 287(3):761–770, 2018.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.
- David Bonekamp, Simon Kohl, Manuel Wiesenfarth, Patrick Schelb, Jan Philipp Radtke, Michael Götz, Philipp Kickingereder, Kaneschka Yaqubi, Bertram Hitthaler, Nils Gählert, et al. Radiomic machine learning for characterization of prostate lesions with mri: comparison to adc values. *Radiology*, 289(1):128–137, 2018.
- Eran Borenstein and Shimon Ullman. Combined top-down/bottom-up segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 30(12):2109–2125, 2008.
- Samuel Borofsky, Arvin K George, Sonia Gaur, Marcelino Bernardo, Matthew D Greer, Francesca V Mertan, Myles Taffel, Vanesa Moreno, Maria J Merino, Bradford J Wood, et al. What are we missing? false-negative cancers at multiparametric mr imaging of the prostate. *Radiology*, 286(1):186–195, 2017.

- Diane Bouchacourt, Pawan K Mudigonda, and Sebastian Nowozin. Disco nets: Dissimilarity coefficients networks. In *Advances in Neural Information Processing Systems*, pages 352–360, 2016.
- Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 105–112. IEEE, 2001.
- Flavie Bratan, Christelle Melodelima, Rémi Souchon, Au Hoang Dinh, Florence Mège-Lechevallier, Sébastien Crouzet, Marc Colombel, Albert Gelet, and Olivier Rouvière. How accurate is multiparametric mr imaging in evaluation of prostate cancer volume? *Radiology*, 275(1):144–154, 2014.
- Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- H Ballentine Carter, Alan W Partin, Patrick C Walsh, Bruce J Trock, Robert W Veltri, William G Nelson, Donald S Coffey, Eric A Singer, and Jonathan I Epstein. Gleason score 6 adenocarcinoma: should it be labeled as cancer? *Journal of Clinical Oncology*, 30(35):4294, 2012.
- Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11 (Jul):2079–2107, 2010.
- Aritrick Chatterjee, Geoffrey Watson, Esther Myint, Paul Sved, Mark McEntee, and Roger Bourne. Changes in epithelium, stroma, and lumen space correlate more strongly with gleason pattern and are stronger predictors of prostate adc changes than cellularity metrics. *Radiology*, 277(3):751–762, 2015.
- Chao Chen, Vladimir Kolmogorov, Yan Zhu, Dimitris Metaxas, and Christoph Lampert. Computing the m most probable modes of a graphical model. In *Artificial Intelligence and Statistics*, pages 161–169, 2013.
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017a.
- Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017b.
- Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

- Ni Chen and Qiao Zhou. The evolving gleason grading system. *Chinese Journal of Cancer Research*, 28(1):58, 2016.
- Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. *arXiv preprint arXiv:1903.12174*, 2019.
- Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.
- Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- Matthew R Cooperberg and Peter R Carroll. Trends in management for patients with localized prostate cancer, 1990-2013. *Jama*, 314(1):80–82, 2015.
- Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.
- Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- Stuart Currie, Nigel Hoggard, Ian J Craven, Marios Hadjivassiliou, and Iain D Wilkinson. Understanding mri: basic mr physics for physicians. *Postgraduate medical journal*, 89(1050):209–223, 2013.
- Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.
- Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.
- Jeffrey De Fauw, Sander Dieleman, and Karen Simonyan. Hierarchical autoregressive image models with auxiliary decoders. *arXiv preprint arXiv:1903.04933*, 2019.
- Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
- Olivio F Donati, Yousef Mazaheri, Asim Afaq, Hebert A Vargas, Junting Zheng, Chaya S Moskowitz, Hedvig Hricak, and Oguz Akin. Prostate cancer aggressiveness: assessment with whole-lesion histogram analysis of the apparent diffusion coefficient. *Radiology*, 271(1):143–152, 2013.
- Olivio F Donati, Asim Afaq, Hebert Alberto Vargas, Yousef Mazaheri, Junting Zheng, Chaya S Moskowitz, Hedvig Hricak, and Oguz Akin. Prostate mri: evaluating tumor volume and apparent diffusion coefficient as surrogate biomarkers for predicting tumor gleason score. *Clinical Cancer Research*, 20(14):3705–3711, 2014.
- Jonathan I Epstein, Zhaoyong Feng, Bruce J Trock, and Phillip M Pierorazio. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified gleason grading system and factoring in tertiary grades. *European urology*, 61(5):1019–1024, 2012.
- Jonathan I Epstein, Michael J Zelefsky, Daniel D Sjoberg, Joel B Nelson, Lars Egevad, Cristina Magi-Galluzzi, Andrew J Vickers, Anil V Parwani, Victor E Reuter, Samson W Fine, et al. A contemporary prostate cancer grading system: a validated alternative to the gleason score. *European urology*, 69(3):428–435, 2016.
- Seyed Mohammadali Eslami. Generative probabilistic models for object segmentation. 2014.
- SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.
- Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.
- Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.
- Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019.

- Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.
- Navid Farahani, Anil V Parwani, and Liron Pantanowitz. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathol Lab Med Int*, 7: 23–33, 2015.
- Duc Fehr, Harini Veeraraghavan, Andreas Wibmer, Tatsuo Gondo, Kazuhiro Matsumoto, Herbert Alberto Vargas, Evis Sala, Hedvig Hricak, and Joseph O Deasy. Automatic classification of prostate cancer gleason scores from multiparametric magnetic resonance images. *Proceedings of the National Academy of Sciences*, 112(46):E6265–E6273, 2015.
- Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- Francois Fleuret and Donald Geman. Coarse-to-fine face detection. *International Journal of computer vision*, 41(1-2):85–107, 2001.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
- Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.
- Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *2009 IEEE 12th international conference on computer vision*, pages 670–677. IEEE, 2009.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.
- Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems*, pages 3581–3590, 2017a.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017b.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

- Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and SM Eslami. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018.
- Dariu M Gavrilă and Vasanth Philomin. Real-time object detection for” smart” vehicles. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 87–93. IEEE, 1999.
- Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2015.
- Corrado Gini. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208:73–79, 1936.
- Shoshana B Ginsburg, Ahmad Algohary, Shivani Pahwa, Vikas Gulani, Lee Ponsky, Hannu J Aronen, Peter J Boström, Maret Böhm, Anne-Maree Haynes, Phillip Brenner, et al. Radiomic features for prostate cancer detection on mri differ between the transition and peripheral zones: preliminary findings from a multi-institutional study. *Journal of Magnetic Resonance Imaging*, 46(1):184–193, 2017.
- Donald F Gleason. Classification of Prostatic Carcinomas. *Cancer Chemotherapy Reports*, 50(3):125–128, 1966.
- Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016.
- Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.
- Matthew D Greer, Anna M Brown, Joanna H Shih, Ronald M Summers, Jamie Marko, Yan Mee Law, Sandeep Sankineni, Arvin K George, Maria J Merino, Peter A Pinto, et al. Accuracy and agreement of piradsv2 for prostate cancer mpMRI: a multireader study. *Journal of Magnetic Resonance Imaging*, 45(2):579–585, 2017.
- Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. In *Advances In Neural Information Processing Systems*, pages 3549–3557, 2016.
- Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton. Who said what: Modeling individual labelers improves classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

- Danna Gurari, Kun He, Bo Xiong, Jianming Zhang, Mehrnoosh Sameki, Suyog Dutt Jain, Stan Sclaroff, Margrit Betke, and Kristen Grauman. Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s). *International Journal of Computer Vision*, 126(7):714–730, 2018.
- Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2012.
- Boris A Hadaschik, Timur H Kuru, Corina Tulea, Philip Rieker, Ionel V Popeneciu, Tobias Simpfendorfer, Johannes Huber, Pawel Zogal, Dogu Teber, Sascha Pahernik, et al. A novel stereotactic prostate biopsy system integrating pre-interventional magnetic resonance imaging and live ultrasound fusion. *The Journal of urology*, 186(6):2214–2220, 2011.
- Omar Hameed and Peter A Humphrey. Pseudoneoplastic mimics of prostate and bladder carcinomas. *Archives of pathology & laboratory medicine*, 134(3):427–443, 2010.
- Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.
- NL Hansen, BC Koo, AY Warren, C Kastner, and T Barrett. Sub-differentiating equivocal pi-rads-3 lesions in multiparametric magnetic resonance imaging of the prostate to improve cancer detection. *European journal of radiology*, 95:307–313, 2017.
- Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.
- NE Hawass. Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. *The British journal of radiology*, 70(832):360–366, 1997.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- Xuming He, Richard S Zemel, and Debajyoti Ray. Learning and incorporating top-down cues in image segmentation. In *European conference on computer vision*, pages 338–351. Springer, 2006.
- Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
- MA Hill, Peter O’Neill, and William G McKenna. Comments on potential health effects of mri-induced dna lesions: quality is more important to consider than quantity. *European Heart Journal–Cardiovascular Imaging*, 17(11):1230–1238, 2016.
- Geoffrey Hinton and Drew Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*. Citeseer, 1993.
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates for optical flow with multi-hypotheses networks. *arXiv preprint arXiv:1802.07095*, 2018.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Fabian Isensee, Jens Petersen, Simon AA Kohl, Paul F Jäger, and Klaus H Maier-Hein. nnu-net: Breaking the spell on successful medical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
- Raban Iten, Tony Metger, Henrik Wilming, Lídia Del Rio, and Renato Renner. Discovering physical concepts with neural networks. *arXiv preprint arXiv:1807.10300*, 2018.
- Paul F Jaeger, Simon AA Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H Maier-Hein. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. *arXiv preprint arXiv:1811.08661*, 2018.
- Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.
- Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*., 2014.

- Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.
- Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015.
- Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195, 2019.
- Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, pages 5580–5590, 2017.
- Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.
- Farzad Khalvati, Alexander Wong, and Masoom A Haider. Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. *BMC medical imaging*, 15(1):27, 2015.
- Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.
- Philipp Kickingereder, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, et al. Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet Oncology*, 20(5):728–740, 2019.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd international conference on Learning Representations (ICLR)*., 2013.
- Diederik P Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *Neural Information Processing Systems (NIPS)*., 2014.
- Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow.(nips), 2016. URL <http://arxiv.org/abs/1606.04934>, 2016.
- Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

- Alexander Kirillov, Bogdan Savchynskyy, Dmitriy Schlesinger, Dmitry Vetrov, and Carsten Rother. Inferring m-best diverse labelings in a single one. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1814–1822, 2015a.
- Alexander Kirillov, Dmytro Shlezinger, Dmitry P Vetrov, Carsten Rother, and Bogdan Savchynskyy. M-best-diverse labelings for submodular energies and beyond. In *Advances in Neural Information Processing Systems*, pages 613–621, 2015b.
- Alexander Kirillov, Alexander Shekhovtsov, Carsten Rother, and Bogdan Savchynskyy. Joint m-best-diverse labelings as a parametric submodular minimization. In *Advances in Neural Information Processing Systems*, pages 334–342, 2016.
- Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *arXiv preprint arXiv:1801.00868*, 2018.
- Yu Xuan Kitzing, Adilson Prando, Celi Varol, Gregory S Karczmar, Fiona Maclean, and Aytakin Oto. Benign conditions that mimic prostate carcinoma: Mr imaging features with histopathologic correlation. *Radiographics*, 36(1):162–175, 2015.
- Lev Borisovich Klebanov, Viktor Beneš, and Ivan Saxl. *N-distances and their applications*. Charles University in Prague, the Karolinum Press, 2005.
- Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017a.
- Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for prostate cancer detection. *Machine Learning for Health Workshop, Advances in Neural Information Processing Systems*, 2017b.
- Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.
- Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019.
- Thijs Kooi, Geert Litjens, Bram Van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35: 303–312, 2017.
- Sven Kosub. A note on the triangle inequality for the jaccard distance. *arXiv preprint arXiv:1612.02696*, 2016.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.
- Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. *arXiv preprint arXiv:1904.05257*, 2019.
- Victor Kulikov, Victor Yurchenko, and Victor Lempitsky. Instance segmentation by deep coloring. *arXiv preprint arXiv:1807.10007*, 2018.
- M Pawan Kumar, PHS Ton, and Andrew Zisserman. Obj cut. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 18–25. IEEE, 2005.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6405–6416, 2017.
- J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- KC Latchamsetty, Jr LS Borden, CR Porter, M Lacrampe, M Vaughan, E Lin, N Conti, JL Wright, and JM Corman. Experience improves staging accuracy of endorectal magnetic resonance imaging in prostate cancer: what is the learning curve? *The Canadian journal of urology*, 14(1):3429–3434, 2007.
- Paul C Lauterbur et al. Image formation by induced local interactions: examples employing nuclear magnetic resonance. 1973.
- Denis Le Bihan and E Breton. Imagerie de diffusion in-vivo par résonance magnétique nucléaire. *Comptes-Rendus de l'Académie des Sciences*, 93(5):27–34, 1985.
- Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.
- Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

- Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pages 2119–2127, 2016.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- Alan H Lipkus. A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, 26(1-3):263–265, 1999.
- Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.
- Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.
- Stacy Loeb, Yasin Folkvaljon, David Robinson, Ingela Franck Lissbrant, Lars Egevad, and Pär Stattin. Evaluation of the 2015 gleason grade groups in a nationwide population-based cohort. *European urology*, 69(6):1135–1141, 2016.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- Pauline Luc, Camille Couprie, and Soumith Chintala *et al.* Semantic Segmentation using Adversarial Networks. *arXiv preprint arXiv:1611.08408*, 2016.
- Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.
- Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. *arXiv preprint arXiv:1902.02102*, 2019.
- Mark Hammer. MRI Physics: Diffusion-Weighted Imaging, 2013. URL <http://xrayphysics.com/dwi.html#adc>. Accessed: 2019-07-13.
- Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

- Donald W McRobbie, Elizabeth A Moore, Martin J Graves, and Martin R Prince. *MRI from Picture to Proton*. Cambridge university press, 2017.
- J Melia, R Moseley, RY Ball, DFR Griffiths, K Grigor, P Harnden, M Jarmulowicz, LJ McWilliam, R Montironi, M Waller, et al. A uk-based investigation of inter-and intra-observer reproducibility of gleason grading of prostatic biopsies. *Histopathology*, 48(6):644–654, 2006.
- Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. *arXiv preprint arXiv:1812.01608*, 2018.
- Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- Drew Moghanaki, Baris Turkbey, Neha Vapiwala, Behfar Ehdaie, Steven J Frank, Patrick W McLaughlin, and Mukesh Harisinghani. Advances in prostate cancer mri and pet/ct for staging and radiotherapy treatment planning. In *Seminars in radiation oncology*, volume 27, page 21. NIH Public Access, 2017.
- Paul C Moldovan, Thomas Van den Broeck, Richard Sylvester, Lorenzo Marconi, Joaquim Bellmunt, Roderick CN van den Bergh, Michel Bolla, Erik Briers, Marcus G Cumberbatch, Nicola Fossati, et al. What is the negative predictive value of multiparametric magnetic resonance imaging in excluding prostate cancer at biopsy? a systematic review and meta-analysis from the european association of urology prostate cancer guidelines panel. *European urology*, 72(2):250–266, 2017.
- Berrend G Muller, Joanna H Shih, Sandeep Sankineni, Jamie Marko, Soroush Rais-Bahrami, Arvin Koruthu George, Jean JMCH de la Rosette, Maria J Merino, Bradford J Wood, Peter Pinto, et al. Prostate cancer: interobserver agreement and accuracy with the revised prostate imaging reporting and data system at multiparametric mr imaging. *Radiology*, 277(3):741–750, 2015.
- James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.
- Klaas NA Nagel, Martijn G Schouten, Thomas Hambroek, Geert JS Litjens, Caroline MA Hoeks, Bennie ten Haken, Jelle O Barentsz, and Jurgen J Fütterer. Differentiation of prostatitis and prostate cancer by using diffusion-weighted mr imaging and mr-guided biopsy at 3 t. *Radiology*, 267(1):164–172, 2013.
- Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don’t know? *arXiv preprint arXiv:1810.09136*, 2018.
- National Cancer Institute. Morphology & Grade, 2019. URL <https://training.seer.cancer.gov/prostate/abstract-code-stage/morphology.html>. Accessed: 2019-07-09.
- Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cían Hughes, Harry Askham, Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

- Gabriel Nketiah, Mattijs Elschot, Eugene Kim, Jose R Teruel, Tom W Scheenen, Tone F Bathen, and Kirsten M Selnæs. T2-weighted mri-derived textural features reflect prostate cancer aggressiveness: preliminary results. *European radiology*, 27(7):3050–3059, 2017.
- Marco Nolden, Sascha Zelzer, Alexander Seitel, Diana Wald, Michael Müller, Alfred M Franz, Daniel Maleike, Markus Fangerau, Matthias Baumhauer, Lena Maier-Hein, et al. The medical imaging interaction toolkit: challenges and advances. *International journal of computer assisted radiology and surgery*, 8(4):607–620, 2013.
- Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv preprint arXiv:1705.01936*, 2017.
- Juan Nunez-Iglesias, Ryan Kennedy, Stephen M Plaza, Anirban Chakraborty, and William T Katz. Graph-based active learning of agglomeration (gala): a python library to segment 2d and 3d neuroimages. *Frontiers in neuroinformatics*, 8:34, 2014.
- Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.
- Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *arXiv preprint arXiv:1903.07291*, 2019.
- Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. Machine learning methods for quantitative radiomic biomarkers. *Scientific reports*, 5:13087, 2015.
- Michelle Peckham, Adele Knibbs, and Steve Paxton. The Histology Guide, 2013. URL <https://www.histology.leeds.ac.uk/what-is-histology/index.php>. Accessed: 2019-07-15.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Jens Petersen, Paul F Jäger, Fabian Isensee, Simon AA Kohl, Ulf Neuberger, Wolfgang Wick, Jürgen Debus, Sabine Heiland, Martin Bendszus, Philipp Kickingereder, et al. Deep probabilistic modeling of glioma growth. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 806–814. Springer, 2019.
- Thibaut Pierre, Francois Cornud, Loïc Colléter, Frédéric Beuvon, Frantz Foissac, Nicolas B Delongchamps, and Paul Legmann. Diffusion-weighted imaging of the prostate: should we use quantitative metrics to better characterize focal lesions originating in the peripheral zone? *European radiology*, 28(5):2236–2245, 2018.
- Robert A Pooley. Fundamental physics of mr imaging. *Radiographics*, 25(4):1087–1099, 2005.
- Stefan Posse, Charles A Cuenod, and D Le Bihan. Human brain: proton diffusion mr spectroscopy. *Radiology*, 188(3):719–725, 1993.

- Prostate Cancer UK. How Prostate Cancer is diagnosed, 2018. URL https://prostatecanceruk.org/media/2494030/how-prostate-cancer-is-diagnosed_ifm.pdf. Accessed: 2019-07-05.
- P Puech, A Sufana Iancu, B Renard, A Villers, and L Lemaitre. Detecting prostate cancer with mri—why and how. *Diagnostic and interventional imaging*, 93(4):268–278, 2012.
- Johann Radon. On the determination of functions from their integral values along certain manifolds. *IEEE transactions on medical imaging*, 5(4):170–176, 1986.
- Jan P Radtke, Constantin Schwab, Maya B Wolf, Martin T Freitag, Celine D Alt, Claudia Kesch, Ionel V Popeneciu, Clemens Huetttenbrink, Claudia Gasch, Tilman Klein, et al. Multiparametric magnetic resonance imaging (mri) and mri–transrectal ultrasound fusion biopsy for index tumor detection: correlation with radical prostatectomy specimen. *European urology*, 70(5):846–853, 2016.
- Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290, 2019.
- William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- C Radhakrishna Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982.
- Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019.
- Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548*, 2018.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- Danilo Jimenez Rezende and Fabio Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.
- Stacey Ronaghan. The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark , 2018. URL <https://medium.com/@srnghn/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>. Accessed: 2019-07-19.
- 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*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. URL <http://arxiv.org/abs/1505.04597>.

- Bruce Rosen and Lawrence Wald. Mr image encoding, 2006. MIT OpenCourseWare <http://ocw.ateneo.net/courses/health-sciences-and-technology/hst-584j-magnetic-resonance-analytic-biochemical-and-imaging-techniques-spring-2006/readings/imageencoding.pdf>.
- Andrew B Rosenkrantz, James S Babb, Samir S Taneja, and Justin M Ream. Proposed adjustments to pi-rads version 2 decision rules: impact on prostate cancer detection. *Radiology*, 283(1):119–129, 2016a.
- Andrew B Rosenkrantz, Luke A Ginocchio, Daniel Cornfeld, Adam T Froemming, Rajan T Gupta, Baris Turkbey, Antonio C Westphalen, James S Babb, and Daniel J Margolis. Interobserver reproducibility of the pi-rads version 2 lexicon: a multicenter study of six experienced prostate radiologists. *Radiology*, 280(3):793–804, 2016b.
- Andrew B Rosenkrantz, Abimbola Ayoola, David Hoffman, Anunita Khasgiwala, Vinay Prabhu, Paul Smereka, Molly Somberg, and Samir S Taneja. The learning curve in prostate mri interpretation: self-directed learning versus continual reader feedback. *American Journal of Roentgenology*, 208(3):W92–W100, 2017.
- Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International journal of computer assisted radiology and surgery*, 13(6):925–933, 2018.
- Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.
- Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *International Conference on Computer Vision (ICCV)*, 2017.
- Michelle D Sakala, Raymond B Dyer, and Rafel Tappouni. The” erased charcoal” sign. *Abdominal Radiology*, 42(3):981, 2017.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.
- Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.
- Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.
- Wolfgang Schlegel, Christian P Karger, and Oliver Jäkel. *Medizinische Physik: Grundlagen–Bildgebung–Therapie–Technik*. Springer-Verlag, 2018.

- Thomas Serre, Lior Wolf, and Tomaso Poggio. Object recognition with features inspired by visual cortex. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF BRAIN AND COGNITIVE SCIENCES, 2006.
- Aditya Siddhant and Zachary C Lipton. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*, 2018.
- Siemens Healthineers. MAGNETOM Prisma, 2019. URL <https://www.siemens-healthineers.com/magnetic-resonance-imaging/3t-mri-scanner/magnetom-prisma>. Accessed: 2019-07-05.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Deborah S Smith and William J Catalona. Interexaminer variability of digital rectal examination in detecting prostate cancer. *Urology*, 45(1):70–74, 1995.
- Lewis Smith and Yarín Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015a.
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015b.
- Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.
- Geoffrey A Sonn, Richard E Fan, Pejman Ghanouni, Nancy N Wang, James D Brooks, Andreas M Loening, Bruce L Daniel, Katherine J To’o, Alan E Thong, and John T Leppert. Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *European urology focus*, 2017.
- Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34, 1948.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Edward O Stejskal and John E Tanner. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The journal of chemical physics*, 42(1):288–292, 1965.

- Akira Suga, Keita Fukuda, Tetsuya Takiguchi, and Yasuo Ariki. Object recognition and segmentation using sift and graph cuts. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.
- Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.
- Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.
- Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.
- Kinh Tieu and Paul Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2):17–36, 2004.
- Shimon Ullman, Michel Vidal-Naquet, and Erez Sali. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682, 2002.
- Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems*, pages 5596–5605, 2017.
- Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.
- Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.
- HA Vargas, AM Hötker, DA Goldman, CS Moskowitz, T Gondo, Kazuhiro Matsumoto, B Ehdaie, S Woo, SW Fine, VE Reuter, et al. Updated prostate imaging reporting and data system (pirads v2) recommendations for the detection of clinically significant prostate cancer using multiparametric mri: critical evaluation using whole-mount pathology as standard of reference. *European radiology*, 26(6):1606–1612, 2016.

- Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 839–847, 2017.
- A Vignati, S Mazzetti, V Giannini, F Russo, E Bollito, Francesco Porpiglia, M Stasi, and Daniele Regge. Texture features on t2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness. *Physics in Medicine & Biology*, 60(7):2685, 2015.
- Paul Viola, Michael Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1:511–518, 2001.
- Alex Waibel. Consonant recognition by modular construction of large phonemic time-delay neural networks. In *Advances in neural information processing systems*, pages 215–223, 1989.
- Jing Wang, Chen-Jiang Wu, Mei-Ling Bao, Jing Zhang, Xiao-Ning Wang, and Yu-Dong Zhang. Machine learning-based analysis of mr radiomics can help to improve the diagnostic performance of pi-rads v2 in clinically relevant prostate cancer. *European radiology*, 27(10):4082–4090, 2017.
- Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903, 2004.
- Jeffrey C Weinreb, Jelle O Barentsz, Peter L Choyke, Francois Cornud, Masoom A Haider, Katarzyna J Macura, Daniel Margolis, Mitchell D Schnall, Faina Shtern, Clare M Tempany, et al. Pi-rads prostate imaging–reporting and data system: 2015, version 2. *European urology*, 69(1):16–40, 2016.
- Andreas Wibmer, Hedvig Hricak, Tatsuo Gondo, Kazuhiro Matsumoto, Harini Veeraghavan, Duc Fehr, Junting Zheng, Debra Goldman, Chaya Moskowitz, Samson W Fine, et al. Haralick texture analysis of prostate mri: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores. *European radiology*, 25(10):2840–2850, 2015.
- Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6): 80–83, 1945.
- Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.
- John Winn and Nebojsa Jojic. Locus: Learning object classes with unsupervised segmentation. In *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, volume 1, pages 756–763. IEEE, 2005.
- Sungmin Woo, Chong Hyun Suh, Sang Youn Kim, Jeong Yeon Cho, and Seung Hyup Kim. Diagnostic performance of prostate imaging reporting and data system version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. *European urology*, 72(2):177–188, 2017.

- Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.
- Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016a.
- Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016b.
- Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. *arXiv preprint arXiv:1901.03784*, 2019.
- Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, pages 8778–8788, 2018.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017a.
- Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017b.

List of figures

1.1	Ambiguity in Lung CT Scans.	2
3.1	Anatomy of the Prostate.	17
3.2	Current Clinical Prostate Cancer Diagnosis.	18
3.3	Exemplary PIRADS Assessments on DWI images.	21
3.4	Prostate MRI Ambiguities.	23
3.5	Gleason Pattern.	24
4.1	Classification of Image Analysis at different Granularities.	31
4.2	Radiomic Feature Extraction.	32
4.3	Deep CNN Architectures for Semantic Segmentation.	35
5.1	Radiomics Workflow.	50
5.2	Radiologist-Performed Example Segmentations.	51
5.3	ROC Curves for Prostate Lesion Classificaion.	55
5.4	Test Cohort Example Case 1.	57
5.5	Test Cohort Example Case 2.	58
6.1	Training under Ambiguous Images.	64
6.2	Schematic of Adversarial Training for Semantic Segmentation.	66
6.3	Prostate MRI Example Cases.	71
6.4	Performance Comparison in the small Dataset Limit.	73
7.1	The Probabilistic U-Net.	77
7.2	Baseline architectures.	82
7.3	Qualitative results of the Probabilistic U-Net.	84
7.4	Squared energy distance results.	85
7.5	Visualization of the latent space for the lung abnormalities segmentation.	88
7.6	Visualization of the latent space for the Cityscapes task.	89
7.7	Calibration of Mode Frequencies of the Probabilistic U-Net.	90

7.8	Ablation analysis.	91
7.9	Histograms of lesion presences.	93
8.1	The Hierarchical Probabilistic U-Net.	101
8.2	Qualitative Comparison of Reconstructions & Samples on LIDC	111
8.3	HPU-Net Samples using different Latent Scales on LIDC.	112
8.4	Instance Segmentation of Neurons.	113
8.5	Generative Extrapolation on masked EM Images.	114
8.6	Generative instance segmentation of Cars.	114
8.7	HPU-Net Samples using different Latent Scales on Stochastic Cityscapes.	116
A.1	Patient Inclusion and Exclusion Flow.	134
A.2	Demographic and Clinical Characteristics of Included Patients.	135
A.3	Patient Inclusion and Exclusion Flow.	136
A.4	Patient Inclusion and Exclusion Flow.	137
B.1	Probability calibration of the Dropout U-Net and U-Net Ensemble.	141
B.2	Probability calibration of M-Heads and Image2Image VAE.	142
B.3	Probability calibration of the Probabilistic U-Net.	143
B.4	LIDC samples from the Probabilistic U-Net.	144
B.5	LIDC samples from the Dropout U-Net.	145
B.6	LIDC samples from the U-Net Ensemble.	146
B.7	LIDC samples from the M-Headsmodel.	147
B.8	LIDC samples from the Image2Image VAE.	148
B.9	Stochastic Cityscapes samples from the Probabilistic U-Net.	149
C.1	Losses during training on LIDC.	155
C.2	Losses during training on SNEMI3D.	156
C.3	HPU-Net examples on LIDC.	158
C.4	sPU-Net examples on LIDC.	159
C.5	HPU-Net examples on SNEMI3D.	160
C.6	sPU-Net examples on SNEMI3D.	161
C.7	HPU-Net examples for Instance Segmentation on Cityscapes.	162
C.8	HPU-Net examples for Instance Segmentation on Cityscapes (difficult cases).	163

List of tables

3.1	Gleason Score Categories.	25
5.1	Top 10 Most Important Features.	59
6.1	Quantitative Results for Prostate Tumor Segmentation.	72
7.1	Predicting ambiguity of LIDC images.	92
8.1	Test set results: HPU-Net vs. sPU-Net	110
8.2	Ablation study for the HPU-Net on LIDC	117
B.1	LIDC results of the Probabilistic U-Net.	140
B.2	Stochastic Cityscapes results of the Probabilistic U-Net.	140
C.1	Baseline Test Set Results	157

Acronyms

ADC Apparent Diffusion Coefficient. 11, 17, 19–23, 27, 46, 47, 49–52, 57–61, 68, 74, 122, 136, 137

AUC Area Under the Curve. 54, 55, 58

B1500 Diffusion-weighted image with $b = 1500 \text{ s mm}^{-2}$. 49–52, 57, 58, 68

CE Cross Entropy. xiv, 34, 38, 63, 64, 70, 72–74, 122, 123

cGAN conditional Generative Adversarial Network. 44

CNN Convolutional Neural Network. 2, 3, 29–35, 39, 41–43, 61

CT Computed Tomography. 1, 2, 11–13, 29, 41, 98

cVAE conditional Variational Auto-Encoder. 40, 43, 75, 76, 79

DRE Digital Rectal Examination. 16–18, 26, 27, 48, 57, 58

DSC Sørensen–Dice Coefficient. 69, 70, 72

DWI Diffusion-weighted Imaging. 10, 11, 17, 20–23, 25, 46, 48–50

ELBO Evidence Lower Bound. 42, 80, 103

EM Electron Microscopy. 99, 110, 113

FCN fully convolutional network. 35, 36, 61

FN False Negative. 55–57, 70, 81

FP False Positive. 55–57, 70, 81

GAN Generative Adversarial Network. 43, 44, 63, 65, 67, 74, 78, 95, 99

- GED** Generalized Energy Distance. 81, 85, 108
- GS** Gleason Score. 18, 19, 24–27, 47, 68, 69
- HPU-Net** Hierarchical Probabilistic U-Net. 97, 99, 103–105, 109–119, 153–156, 188, 189
- IDRI** Image Database Resource Initiative. 2, 83, 85, 86, 88, 139, 150
- IoU** Intersection over Union. 70, 81, 92, 108, 109
- KL** Kullback-Leibler divergence. 42, 80, 86, 102, 150, 151, 153, 154
- LIDC** Lung Image Database Consortium. xv, 2, 83–86, 88, 92, 99, 105, 108, 109, 118, 139, 140, 144–148, 150, 154, 188, 189
- mADC** mean Apparent Diffusion Coefficient. 47, 52, 54–61
- MC** Monte Carlo. 40–42
- ML** Machine Learning. 32, 49, 60, 63
- mpMRI** Multi-parametric MRI. 17, 19, 20, 25–27, 45, 46, 56, 61, 74
- MRI** Magnetic Resonance Imaging. 1, 3, 4, 7–10, 12, 15, 20, 22, 23, 29, 41, 45, 47–49, 54, 55, 57, 60, 61, 73
- NAN** Not-A-Number. 109
- PCa** Prostate Cancer. 16, 17, 19, 56, 61
- PIRADS** The Prostate Imaging Reporting and Data System. 18–23, 26, 45–49, 51, 54, 57, 58, 60, 61, 121, 136, 137
- PSA** Prostate-specific Antigen. 16–18, 27, 48, 51, 57, 58
- PZ** Peripheral Zone. 20, 22, 23, 26, 49, 54, 57, 58, 60, 61, 68
- RF** Random Forest. 50, 52–56, 58–60, 136, 137
- RML** Radiomic Machine Learning. 54–58, 60, 61
- ROC** Receiver Operator Characteristic. 50, 54, 55, 60

RoI Region of Interest. 33

SGD Stochastic Gradient Descent. 64, 65

SNEMI3D 3D Segmentation of Neurites in EM images. 99, 106, 109, 110, 112, 113, 154

SotA State-of-the-Art. 30, 34, 36, 37

sPU-Net standard Probabilistic U-Net. 99, 100, 102, 106, 108–113, 117, 119, 155, 156, 189

T1w T1-weighted. 8, 17, 20, 22, 23

T2w T2-weighted. 8, 17, 20, 22, 23, 46–52, 57–60, 68, 74

TP True Positive. 70, 81

TRUS-biopsy Trans-rectal Ultra-Sound Guided Biopsy. 18–20, 26, 27, 46, 48, 61, 68

TZ Transitional Zone. 20, 22, 23, 26, 49, 54, 57–60, 68

VAE Variational Auto-Encoder. 42–44, 78, 95, 99, 118

VOI Volume of Interest. 49, 52