Machine Learning for MRI-guided Prostate Cancer Diagnosis and Interventions

by

Alireza Mehrtash

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Machine Learning for MRI-guided Prostate Cancer Diagnosis and Interventions

submitted by **Alireza Mehrtash** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Electrical and Computer Engineering.**

Examining Committee:

Purang Abolmaesumi, Department of Electrical and Computer Engineering
Supervisor

Leonid Sigal, Department of Computer Science
University Examiner

Zhen Jane Wang, Department of Electrical and Computer Engineering
University Examiner

James Duncan, Yale University
External Examiner

Additional Supervisory Committee Members:

Robert Rohling, Department of Electrical and Computer Engineering
Supervisory Committee Member

William Wells, Brigham and Women's Hospital, Harvard Medical School
Supervisory Committee Member

Tina Kapur, Brigham and Women's Hospital, Harvard Medical School
Supervisory Committee Member

Abstract

Prostate cancer is the second most prevalent cancer in men worldwide. Magnetic Resonance Imaging (MRI) is widely used for prostate cancer diagnosis and guiding biopsy procedures due to its ability in providing superior contrast between cancer and adjacent soft tissue. Appropriate clinical management of prostate cancer critically depends on meticulous detection and characterization of the disease and precise biopsy procedures if necessary.

The goal of this thesis is to develop computational methods to aid radiologists in diagnosing prostate cancer in MRI and planning necessary interventions. To this end, we have developed novel methods for assessing probability of clinically significant prostate cancer in MRI, localizing biopsy needles in MRI, and providing segmentation of structures such as the prostate gland. The proposed methods in this thesis are based on supervised machine learning techniques, in particular deep convolutional neural networks (CNNs). We have also developed methodology that is necessary in order for such deep networks to eventually be useful in clinical decision-making workflows; this spans the areas of domain adaptation, confidence calibration, and uncertainty estimation for CNNs. We used domain adaptation to transfer the knowledge of lesion segmentation learned from MRI images obtained using one set of acquisition parameters to another. We also studied predictive uncertainty in the context of medical image segmentation to provide model confidence (i.e expectation of success) at inference time. We further proposed parameter ensembling by perturbation for calibration of neural networks.

Lay Summary

Prostate cancer is the first most diagnosed cancer in North American men and the second most common cancer in men worldwide. Early detection of prostate cancer increases the chances of long-term survival. Magnetic Resonance Imaging (MRI) can aid doctors in better screenings of prostate cancer. However, prostate cancer screening with MRI is not 100% accurate and often leads to missing high-risk patients and unnecessary aggressive treatment for low-risk patients. The purpose of this thesis is to develop reliable computational methods to aid physicians for better diagnosis and treatment of prostate cancer patients.

Preface

This thesis is primarily based on five manuscripts resulting from the collaboration among multiple researchers. The manuscripts have been modified accordingly to present a consistent thesis.

A study described in Chapter 2 has been published in:

- Alireza Mehrtash, Alirea Sedghi, Mohsen Ghafoorian, Mehdi Taghipour, Clare M. Tempany, William M. Wells III, Tina Kapur, Parvin Mousavi, Purang Abolmaesumib, Andrey Fedorov. Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. *Medical Imaging 2017: Computer-Aided Diagnosis. International Society for Optics and Photonics*, 10134: 101342A, 2017.

The contribution of the author was in developing, implementing, and evaluating the method. Drs. Sedghi and Ghafourian contributed in developing and implementing the proposed method. Drs. Tempany, and Taghipour provided clinical insight. Drs. Fedorov, Abolmaesumi, Mousavi, and Kapur helped with their valuable suggestions in improving the methodology.

A version of Chapter 3 has been published in:

- Alireza Mehrtash, Mohsen Ghafoorian, Guillaume Pernelle, Alireza Ziaei, Friso G. Heslinga, Kemal Tuncali, Andriy Fedorov, Ron Kikinis, Clare M Tempany, William M Wells, Purang Abolmaesumi, Tina Kapur. Automatic needle segmentation and localization in MRI With 3-D convolutional neural networks: application to MRI-targeted prostate biopsy. *IEEE Transactions on Medical Imaging*, 38(4):1026-1036, 2018.

The contribution of the author was in developing, implementing, and evaluating the method. Drs. Ghafoorian and Pernelle provided valuable scientific inputs to improve the proposed method. Dr. Ziaei and F.G. Heslinga created the needle segmentation ground truth. Dr. Tuncali performed the biopsy procedures, with technical support from Dr. Fedorov. Dr. Tempany provided clinical insight for MR-guided prostate biopsy procedure. Profs. Kapur,

Wells, Abolmaesumi, and Kikinis helped with their valuable suggestions in improving the methodology.

A version of Chapter 4 has been published in:

- Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, Andriy Fedorov, Purang Abolmaesumi, Bram Platel, William M Wells. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer 2017.

Dr. Ghafoorian and the author equally contributed in developing, implementing, and evaluating the method. Prof. de Leeuw helped with data preparation and annotation. Drs. Wells, Platel, Abolmaesuimi, Fedorov, van Ginneken, Tempany, Guttmann, and Pesteie helped with their valuable suggestions in improving the methodology.

A version of Chapter 6 has been published in:

- Alireza Mehrtash, William M. Wells III, Clare M. Tempany, Purang Abolmaesumi, Tina Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 2020.

The contribution of the author was in developing, implementing, and evaluating the method. Profs. Kapur, Abolmaesumi, Tempany, and Wells helped with their valuable suggestions in improving the methodology. All co-authors contributed to the editing of the manuscript.
A version study presented in Chapter 7 will be published in proceedings of 2020 Conference on Neural Information Processing Systems (NeurIPS):

- Alireza Mehrtash, Purang Abolmaesumi, Polina Goland, Tina Kapur, Demian Wassermann, William M. Wells III. PEP: Parameter Ensembling by Perturbation. *NeurIPS 2020.*

The contribution of the author was in developing, implementing, and evaluating the method. Prof. Wells contributed to the mathematical derivation of the local analysis. Drs. Abolmaesumi, Wassermann, Goland, and Kapur helped with their valuable contributions and suggestions in improving the methodology.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2D	Two Dimensional
3D	Three Dimensional
ADC	Apparent Diffusion Coefficient
AUC	Area Under the Curve
BN	Batch Normalization
CADe	Computer-aided detection
CADx	Computer-aided diagnosis
CE	Cross-entropy
CNN	Convolutional Neural Network
CT	Computed Tomography
CZ	Central Zone
DCE	Dynamic Contrast-enhanced
DWI	Diffusion Weighted Imaging
EM	Expectation-maximization
FCN	Fully Convolutional Neural Network
FLAIR	Fluid-attenuated Inversion Recovery
GGG	Gleason Grade Group
GPU	Graphical Processing Unit
HD	Hausdorff Distance
MCD	Monte Carlo Dropout
mpMRI	Multi-parametric Magnetic Resonance Imaging
MRI	Magnetic Resonance Imaging

NLL	Negative Log Likelihood
NN	Neural Network
PCa	Prostate Cancer
PEP	Parameter Ensembling by Perturbation
PI-RADS	Prostate Imaging Reporting and Data System
PK	Pharmacokinetics
PZ	Peripheral Zone
ReLu	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TRUS	Transrectal Ultrasound
TS	Temperature Scaling
TZ	Transition Zone
WMH	White Matter Hyperintensities

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Purang Abolmaesumi. I am grateful for his invaluable insight, guidance, and continuous support during my PhD studies. I particularly appreciate him for giving me the freedom to find and investigate different ideas.

I feel very fortunate to have Dr. Tina Kapur as my co-supervisor, guide, and friend. She is not just an outstanding mentor, she is a wonderful human being. One of a kind! Thank you, Tina! Thank you for believing in me and for always being there for me. Thank you for your endless support to make this dream come true.

Many thanks to my wonderful co-supervisor Prof. William (Sandy) M. Wells. I was so lucky to have the chance to work with you. I enjoyed all the machine learning chats we had over the past years and I learned a lot from you.

I am thankful for my committee members, Prof. Robert Rohling, Prof. Leonid Sigal, and Prof. Zhen Jane Wang for reading my thesis and providing me with constructive comments and insightful feedback.

I thankfully acknowledge the grants that supported my research during the course of my studies from the National Center for Image Guided Therapy (NIH P41EB015898), the Natural Science and Engineering Research Council of Canada (NSERC), and the Canadian Institutes of Health Research (CIHR).

I am so grateful to my mentors at SPL over the past years. I would like to thank Prof. Ron Kikinis for his full support since I joined SPL. I would like to thank Prof. Clare M. Tempany for always being supportive of my research. Moreover, I would like to thank her for keeping me up-to-date with all the clinical insights that I needed for my research in prostate cancer. I would like to thank Dr. Andrey Fedorov from whom I learned a lot about working with medical imaging data. I would like to thank all the other amazing SPL members that I worked with them over the past years. Thank you Steve, Jay, Junichi, Brunilda, and Danielle. I would also like to thank the amazing researchers that I met at SPL and worked on several ideas together. Thank you Alireza Sedghi, Joeky, Mehdi, Friso, and Prashin.

I feel lucky to met Mohsen Ghafoorian at SPL in November 2016. Despite

his short stay in Boston, we managed to do lots of fruitful research together. Mohsen is such a talent and always comes up with the brightest ideas. Beyond this, he is an amazing friend. Thank you Mohsen!

My sincere thanks go to all the RCL members at UBC. I enjoyed working on a medical imaging challenge with Mehran, Amir, and Jorden. I had an amazing time and learned a lot from you guys. I am also very happy that I met Mehran at RCL. I miss working with him and all the brainstorming that we had during our coffee breaks.

Special thanks to Saeideh and Rasool for taking us from the airport the night we arrived in Vancouver and taking the best care of us for our whole stay in Vancouver. Because of you, we felt at home.

I offer my gratitude to my amazing parents-in-law for their full support and outstanding care throughout my studies. Thank you, Fereshteh and Ali! I am very fortunate to have you in my life. Special thanks to my brother-in-law Amir for all his encouragements.

I am so grateful to my wonderful sisters. Thank you Tahmineh for your unconditional support, selfless care, and devotion which helped me to fulfill my dream. Thank you Manjijeh for always believing in me and encouraging me. I want to express my gratitude to my brother-in-law Homayoun, whom has always been a source of inspiration for me. I would also like to thank Kiana and Roxana who have always been the best nieces one could have ever wished for.

And finally, I would have not been able to finish my PhD without the daily love that I received from my better half Roya and my adorable daughters Diana and Amitis. Thank you Roya for your unwavering love, support, encouragement, and patience in every step of this journey.

To the loving memory of Fatemeh Khorgami and Javad Mehrtash

To Roya, Diana, and Amitis

Chapter 1

Introduction

1.1 Clinical Background

The prostate is a walnut-shaped gland that is part of the male reproductive system (Figure 1.1). It is located in the pelvis at the base of the urinary bladder and surrounds the urethra. The prostate produces the seminal fluid that combines with sperm from the testes. The alkaline nature of the prostatic fluid helps in reducing the acidity of the vaginal environment which could extend the lifespan of sperm. The prostate is composed of both globular and fibromuscular tissues are enclosed in a surface termed the prostatic capsule or prostatic fascia [171]. Along the urethra, from superior to inferior, the prostate is composed of three primary regions, the base (below the bladder), the midgland, and the apex (inferior part in the vicinity of the urogenital diaphragm). Histologically, the prostate is divided into four primary zones: anterior fibromuscular stroma (AS), the transition zone (TZ), the central zone (CZ), and the peripheral zone (PZ). AS contains no glandular tissue. CZ and TZ surround the ejaculatory ducts and the proximal urethra, respectively. TZ, CZ, and PZ contain about 5%, 20%, and $70 - 80\%$ of the glandular tissue, respectively [183].

Prostate cancer is the second most frequently diagnosed cancer in men and the fifth leading cause of cancer mortality worldwide [144]. In the United States, it is the most frequently diagnosed, noncutaneous male malignancy and the second leading cause of cancer-related mortality among men in the United States [165]. Statistics of prostate cancer frequency, morbidity, and mortality can be examined in many different ways. It is a very common cancer, as it is a “tumor of aging," but it has a very low disease-specific

Parts of Sections 1.1, 1.2, and 1.3 are adapted from Wenya Linda Bi, Ahmed Hosny, Matthew B. Schabath, Maryellen L. Giger, Nicolai J. Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F. Dunn, Raymond H. Mak, Rulla M. Tamimi, Clare M. Tempany, Charles Swanton, Udo Hoffmann, Lawrence H. Schwartz, Robert J. Gillies, Raymond Y. Huang, Hugo J. W. L. Aerts. Artificial intelligence in cancer imaging: clinical challenges and applications. CA: A Cancer Journal for Clinicians. Wiley Periodicals, Inc. on behalf of American Cancer Society 2019.

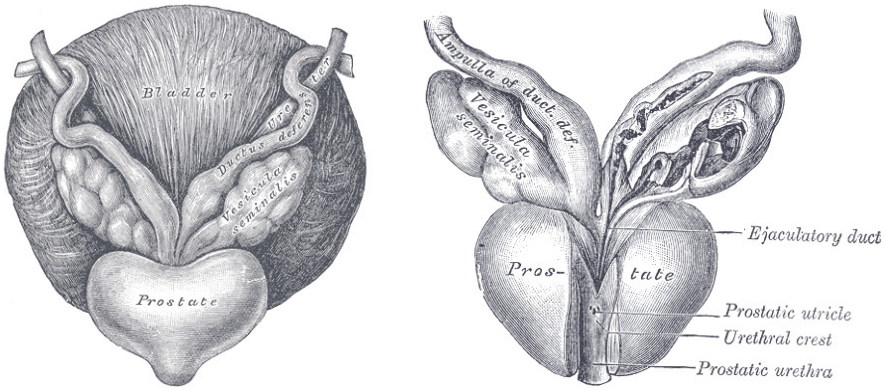


Figure 1.1: Prostate anatomy [53].

mortality, all of which reinforce its characterization as a complex public health concern that impacts a large population. Although prostate cancer is a serious disease, most men diagnosed with prostate cancer do not die of it [56]. The key clinical problems in prostate cancer diagnosis today include 1) overdiagnosis and overtreatment resulting from an inability to predict the aggressiveness and risk of a given cancer; and 2) inadequate targeted biopsy sampling, leading to misdiagnosis and to disease progression in men with seemingly low-risk prostate cancer. In a meta-analysis [111], the reported rate of overdiagnosis of nonclinically significant prostate cancer was as high as 67%, leading to unnecessary treatment and associated morbidity. Because of this range of clinical behavior, it is necessary to differentiate men who have clinically significant tumors (those with a biopsy Gleason score 7 or higher and/or tumor volume > 0.5 ml) [195] as candidates for therapy from those who have clinically insignificant tumors and can safely undergo active surveillance. It has been noted that potential survival benefits from aggressively treating early-stage prostate cancer are undermined by harm from the unnecessary treatment of indolent disease.

The current screening procedures for prostate cancer include a digital rectal examination (DRE) and the prostate-specific antigen (PSA) blood test (which has recently been downgraded because of high false positive rates). DRE can only detect late-stage prostate cancer in the PZ. Hence it lacks the required sensitivity for early-stage cancer and cancer in other zones. PSA elevated levels often indicate the presence of prostate cancer. On the other hand, patients with prostatitis or benign prostatic hyperplasia can also have higher than normal PSA levels. Hence, while PSA screening

is a sensitive diagnostic test, it lacks the required specificity. An abnormal screening indicates the possibility of prostate cancer, and random systematic (sextant) biopsies of the entire organ are performed on the patient under the guidance of transrectal ultrasound (TRUS). These biopsies randomly sample a very small part of the gland and the results sometimes miss the most aggressive tumor within the gland [36, 58, 155].

1.2 Magnetic Resonance Imaging for Prostate Cancer

Multi-parametric Magnetic Resonance Imaging (mpMRI) provides the required soft tissue contrast for detection and localization of suspicious clinically significant prostate lesions and gives information about tissue anatomy, function, and characteristics (Figure 1.2). Importantly, it has superior capabilities to detect the "clinically significant disease." Recent years have seen a growth in the volume of mpMRI examination of prostate cancer due to its ability to detect these lesions and allow targeted biopsy sampling. A large population study from the UK suggested that use of mpMRI as a triage before primary biopsy can reduce the number of unnecessary biopsies by a quarter and decrease overdiagnosis of clinically insignificant disease [4]. This was further validated in the and on smaller data sets than would be optimal. In the multinational PRECISION study of 500 patients [81], men randomized to mpMRI prior to biopsy experienced a significant increase in the detection of clinically significant disease over the current standard of care, which employs a 10-12 core transrectal ultrasound-guided biopsy (38% vs 26%).

MRI has demonstrated value in not just detecting and characterizing clinically significant prostate cancer, but also in guiding biopsy needles to the suspicious targets [4]. MRI has been incorporated into the biopsy procedure in two different ways. The first is MR/ultrasound fusion biopsy in which targets for biopsy are identified in diagnostic mpMRI, displayed in the TRUS image (using MR/Ultrasound fusion) during a routine sextant biopsy, and additionally sampled. The second is in-bore biopsy that is performed inside the bore of an MR scanner; the diagnostic mpMRI with marked targets is overlaid on a rapid acquisition intra-operative MR image (using MR/MR fusion) and targeted sampling is performed by the internationalist. A study of over 1000 men undergoing biopsy for suspected prostate cancer [164] showed targeted MR/ultrasound fusion biopsy, compared with standard extended-sextant TRUS biopsy, was associated with increased detection of high-risk prostate cancer and decreased detection of low-risk prostate cancer. Smaller

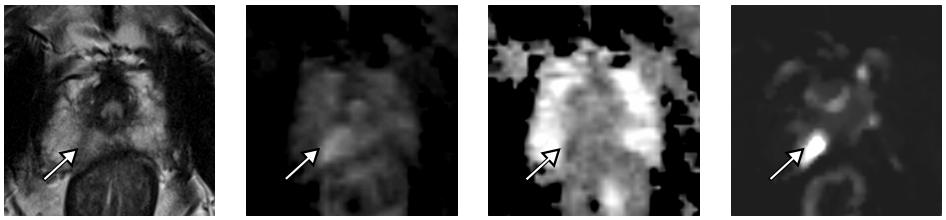


Figure 1.2: Multiparametric MRI of a patient with clinically significant prostate cancer. Arrows mark the lesion location. (a) Axial T2-weighted MR. (b) computed high–b value (1400 sec/mm^2) diffusion-weighted MR. (c) ADC map (d) K^{Trans} parametric map from dynamic contrast enhanced T1-weighted MRI.

studies for in-bore biopsies have reported requiring significantly fewer cores and revealed a significantly higher percentage of cancer involvement per biopsy core [136, 139, 187, 196].

1.3 Machine Learning in Prostate Cancer Imaging

The growing trend towards mpMRI has introduced a demand for experienced radiologists to interpret the exploding volumes of oncological prostate MRIs. Furthermore, reading challenging cases and reducing the high rate of interobserver disagreements on findings is a remaining challenge for prostate MRI. In 2015, the European Society of Urogenital Radiology, American College of Radiology, and AdmeTech foundation published the second version of Prostate Imaging Reporting and Data system (PI-RADS). These provide guidelines for radiologists in reading and interpreting the prostate mpMRI, which aim to increase the consistency of interpretation and communication of mpMRI findings. Over the past ten years, machine learning models have been developed as Computer-aided detection (CADe) and Computer-aided diagnosis (CADx) systems to detect, localize, and characterize prostate tumors [109]. In conjunction with PI-RADS, accurate CAD systems can increase the inter-rater reliability and improve the diagnostic accuracy of mpMRI reading and interpretation [48]. In preliminary analyses, it has been shown that the addition of a CADx system can improve the performance of radiologists in prostate cancer interpretation.

The clinical motivation of this thesis is to aid radiologists in the detection and classification of prostate cancer. While clinically prostate cancer has

swung between extremes of under- and over- diagnosis and treatment, the underlying computer vision questions have remained unchanged; are there distinct patterns in MRI images of patients suspected of prostate cancer that can be automatically detected to help detect and biopsy the cancer? will a pattern recognition method developed for one set of MRI images work when the acquisition protocol or scanner is upgraded? Is this addressable with available technology or are methodological improvements needed? Based on these questions we developed objectives for this thesis which are described in the next section.

Computational methods mostly based on supervised machine learning have been successfully applied to imaging modalities such as MRI and ultrasound to detect suspicious lesions and differentiate clinically significant cancers from the rest. Recent application of deep learning in prostate cancer screening and aggressive cancer diagnosis has produced promising results. Preliminary work in mpMRI CADx systems focused primarily on classic supervised machine learning methodologies, including combinations of feature extractors and shallow classifiers. In this category of machine leanring systems, feature engineering plays a central role in the overall performance of the CAD system. Combinations of CADe and CADx systems have been reported that use intensity, anatomical, Pharmacokinetics (PK), texture, and blobness features [106]. PK metrics can be extracted from a time signal analysis of intravenous contrast passing through a given volume of tissue. They include parameters such as wash-in and wash-out. Texture features are also signal based and depend heavily on imaging techniques. Others used intensity features calculated from mpMRI sequences, including T2-weighted, ADC, high b-value DWI, and a T2 estimation map by proton density image[106], or only using features extracted from PK analysis and DTI parameter maps [125]. Similar image-based features were included into CAD systems [19, 49, 129, 161] and many of these systems use support vector machines (SVMs) for classification [21, 94, 125, 188].

In the past few years, advancements in Deep Learning (DL) have dominated the field of computer-assisted PCa detection using mpMRI or ultrasound information, as individual modalities [192]. Most of the research has utilized Convolutional Neural Networks (CNN) and various optimization strategies for achieving state-of-the-art performance in PCa detection. Several groups have taken advantage of the multi-sequence nature of the mpMRI data by stacking each modality as input channels similar to RGB images [22, 110], and integrating their information early in the training. Kiraly *et al.* [87] used Fully Convolutional Networks (FCN) for localization and classification of prostate lesions and achieved Area Under the Curve (AUC) of 0.83 by training on 202

patients. Schelb *et al.* [154] proposed a U-Net architecture on bi-parametric prostate MRI (T2-Weighted and ADC), and achieved performance similar to that of Prostate Imaging Reporting and and Data System (PI-RADS), the clinical standard of mpMRI scoring [183]. In another recent study, Sedghi et al. [159] demonstrated the potential of integration of multimodal information from MRI and temporal ultrasound to improve prostate cancer detection. By using Fully Convolutional Neural Networks (FCNs) as the architecture of choice, they created cancer probability maps in the entire imaging planes immediately

The results of the ongoing research in the use of machine learning for the detection and characterization of prostate cancer are promising and demonstrate ongoing improvement. The recent body of research in prostate cancer image analysis shows a transition from feature engineering and classic machine learning methods towards deep learning and the use of large training sets. Unlike lung and breast cancers, clinical routines in prostate cancer have not yet adopted regulated CAD systems. However, the recently achieved results of deep learning techniques on mid-size datasets such as the PROSTATEx benchmark are promising. As it is now evident there has been a rapid growth in prostate MR exam volumes worldwide and increasing demand for accurate interpretations. Accurate CAD systems will improve the diagnostic accuracy of prostate MRI readings which will result in better care for individual patients, as fewer patients with benign and indolent tumors (false positives) will need to undergo invasive biopsy and/or radical prostatectomy procedures which can lower their quality of life. On the other hand, early detection of prostate cancer improves the prognosis of patients with clinically significant prostate cancer. Computer-assisted detection and diagnosis systems of prostate cancer help clinicians by potentially reducing the chances of either missing or overdiagnosing suspicious targets on diagnostic MRIs, although this merits additional validation in trials before routine clinical incorporation.

1.4 Objectives

The main objective of this thesis is to develop reliable machine learning models and algorithms that can improve MRI-guided prostate cancer diagnosis and interventions. We start by building solutions and applications to facilitate prostate cancer diagnosis and interventions. We investigate the problem of distinguishing the normal gland from cancer using mpMRI images of the prostate. We study methods for localizing biopsy needle tip and trajectory in MRI scans obtained for MRI-guided prostate biopsy.

We then address common challenges regarding data in real clinical setups. We propose approaches to improve our diagnosis system by recognizing that uncertainty in biopsy location is an issue. This led us to model the error and use probabilistic inference to accommodate this error. We develop transfer learning methodologies to address the domain shift problem.

Furthermore, we study the problem of prostate segmentation in the context of weakly-supervised learning and uncertainty estimation. Prostate segmentation in MRI is an important preprocessing step for several tasks such as automated fusion of imaging data for targeted biopsy (guided by MRI alone or by MRI-ultrasound fusion), quantification of PSA density in assessing treatment response, and dose planning in radiotherapy. We propose a methodology for weakly-supervised segmentation with partially annotated ground truth. We further investigate methods to improve the calibration of deep segmenters through ensembling.

Finally, we propose a novel methodology for ensembling based on parameter perturbation.

1.5 Contributions

This thesis is an attempt to develop techniques that are essential for MRI-guided prostate cancer diagnosis and interventions. In the course of achieving this objective the following contributions were made:

- Developing a novel deep neural network for diagnosing clinically significant prostate cancer in mpMRI. The method uses diffusion and dynamic contrast images together with information about the location of the suspicious target to predict the probability of clinically significant cancer.

- Developing a novel probabilistic framework to model the uncertainty regarding the location of the biopsy samples. Also, developing a novel Gaussian weighted loss function as a form of data augmentation (label imputation) to train FCNs with sparse biopsy locations. The framework provides posterior probabilities of latent true biopsy locations given the image, model, observed biopsy location, and the biopsy outcome.

- Developing novel method for fast automatic needle tip and trajectory localization and visualization in MRI for prostate biopsies. The proposed method has a performance comparable with human inter-observer concordance, reinforcing the clinical acceptability of it.

- Developing a novel transfer learning technique for domain adaptation of networks trained with one set of MRI acquisition parameters. This is an essential step for deployment of machine learning models in practice where imaging parameters are changing. The proposed method is capable of tuning the deep network to the new domain. Here, we perform experiments on brain MRI images to assess the contributions. Since there are no prior assumptions regarding the specific problem of white matter hyperintensities (WMH) segmentation, we anticipate that the proposed method can be generalized to other medical imaging problems including prostate cancer diagnosis with MRI. However, confirmation of this requires multi-domain prostate MRI datasets and further experimentation.

- Proposing a novel method for weakly-supervised semantic segmentation with point and scribble supervision in FCNs. We also propose partial Dice loss, a variant of Dice loss function for deep weakly-supervised segmentation with sparse pixel-level annotations. Here, in addition to prostate segmentation, we evaluate the proposed method with heart and kidney segmentation problems.

- Developing a novel technique based on ensembling for confidence calibration and predictive uncertainty estimation for deep medical image segmentation. Also, proposing a novel entropy-based metric to predict the segmentation quality of foreground structures, which can be further used to detect out-of-distribution test inputs. We evaluate our contributions across three medical image segmentation applications of the prostate, the heart, and the brain.

- Proposing a new technique for confidence calibration uncertainty estimation of neural networks without the need for network modification or several rounds of training. We proposed parameter ensembling by perturbation (PEP) which prepares an ensemble of parameter values as perturbations of the optimal parameter set from training by a Gaussian with a single variance parameter.

1.6 Thesis Outline

The rest of this thesis is divided into six chapters as outlined below:

CHAPTER 2: PROSTATE CANCER DIAGNOSIS IN MRI

In this chapter, we introduce novel deep learning techniques for clinically significant prostate cancer detection and diagnosis in mpMRI. We train, validate, and test deep CNNs on patients suspected of having prostate cancer. We propose two different styles of CNN architectures for cancer diagnosis: patch-based method and fully convolutional neural network (FCN). For the patch-based method, we use 3D convolutional neural networks and fully connected layers. For this model, in addition to image features, we also feed location features of the suspicious to the network. Our results suggest that for the proposed architecture, the combination of diffusion weighted MRI (DWI) and parametric maps from dynamic contrast-enhanced (DCE) MRI serve as the best imaging features for diagnosing prostate cancer. The second proposed architecture, FCNs, makes it feasible to do prediction on whole gland in a single inference. Partial cross-entropy loss is used to train FCNs on sparse ground truth locations. Furthermore, we studied methods to address sparsity of training data and also location uncertainty of ground truth deep cancer classifiers. We observe that Gaussian weighted loss improves the area under the receiver operating characteristic curve and the proposed biopsy location adjustment substantially improves the sensitivity of the models.

CHAPTER 3: BIOPSY NEEDLE LOCALIZATION IN MRI

Image-guidance improves tissue sampling during biopsy by allowing the physician to visualize the tip and trajectory of the biopsy needle relative to the target in MRI, CT, ultrasound, or other relevant imagery. A system for fast automatic needle tip and trajectory localization and visualization in MRI was developed and tested in the context of an active clinical research program in prostate biopsy at Brigham and Women's hospital. Needle tip and trajectory were annotated on 583 T2-weighted intra-procedural MRI scans acquired after needle insertion for 71 patients who underwent transperenial MRI-targeted biopsy procedure at our institution. The images were divided into two independent training-validation and test sets at the patient level. A deep 3-dimensional fully convolutional neural network model was developed, trained and deployed on these samples. The accuracy of the proposed method, as tested on previously unseen data, was 2.80 mm average in needle tip detection, and 0.98° in needle trajectory angle. An observer study was designed in which independent annotations by a second observer, blinded to the original observer, were compared to the output of the proposed method. The resultant error was comparable to the measured inter-observer concordance, reinforcing the clinical acceptability of the proposed method.

CHAPTER 4: TRANSFER LEARNING FOR DOMAIN ADAPTATION IN MRI

It is well known that variations in MRI acquisition protocols result in different appearances of normal and diseased tissue in the images. Convolutional neural networks (CNNs), which have shown to be successful in many medical image analysis tasks, are typically sensitive to the variations in imaging protocols. Therefore, in many cases, networks trained on data acquired with one MRI protocol, do not perform satisfactorily on data acquired with different protocols. This limits the use of models trained with large annotated legacy datasets on a new dataset with a different domain which is often a recurring situation in clinical settings. In this study, we investigated the following central questions regarding domain adaptation in medical image analysis: Given a fitted legacy model, 1) How much data from the new domain is required for a decent adaptation of the original network?; and, 2) What portion of the pre-trained model parameters should be retrained given a certain number of the new domain training samples? To address these questions, we conducted extensive experiments in white matter hyperintensity segmentation task. We trained a CNN on legacy MR images for a specific task and evaluated the performance of the domain-adapted network on the same task with images from a different domain. We then compared the performance of the model to the surrogate scenarios where either the same trained network is used or a new network is trained from scratch on the new dataset. The domain-adapted network tuned only by two training examples achieved a performance substantially outperforming a similar network trained on the same set of examples from scratch.

CHAPTER 5: WEAKLY-SUPERVISED MEDICAL IMAGE SEGMENTATION

Fully Convolutional neural networks (FCNs) including U-Nets, have achieved state-of-the-art results in semantic segmentation for numerous medical imaging applications. Training deep models for segmentation requires high-quality pixel-level ground truth annotations, which is time-consuming and expensive. Partial annotations such as points or scribbles can be used as less expensive alternatives. In this chapter, we study weakly-supervised FCN-based segmentation methods that can be trained with only a single annotated point or only a single annotated scribble per slice of a medical image volume. We propose the use of a partial Dice loss function in our methods because it

encourages higher Dice values for collections of pixels where ground truth is known. Furthermore, we systematically compare partial Dice loss with partial cross-entropy loss in terms of segmentation quality and demonstrate statistically significant performance improvement. We evaluate the proposed methods through extensive experiments in five segmentation tasks across three medical image domains - images of the prostate, the kidney, and the heart. Among these applications, our methods with a single point or a single scribble supervision achieve $51\%-95\%$ and $86\%-97\%$ of the performance of the fully supervised training, respectively.

CHAPTER 6: UNCERTAINTY ESTIMATION IN SEGMENTATION

Fully convolutional neural networks (FCNs), and in particular U-Nets, have achieved state-of-the-art results in semantic segmentation for numerous medical imaging applications. Moreover, batch normalization and Dice loss have been used successfully to stabilize and accelerate training. However, these networks are poorly calibrated i.e. they tend to produce overconfident predictions for both correct and erroneous classifications, making them unreliable and hard to interpret. In this chapter, we study predictive uncertainty estimation in FCNs for medical image segmentation. We make the following contributions: 1) We systematically compare cross-entropy loss with Dice loss in terms of segmentation quality and uncertainty estimation of FCNs; 2) We propose model ensembling for confidence calibration of the FCNs trained with batch normalization and Dice loss; 3) We assess the ability of calibrated FCNs to predict the segmentation quality of structures and detect out-of-distribution test examples. We conduct extensive experiments across three medical image segmentation applications of the prostate, the heart, and the brain to evaluate our contributions. The results of this study offer considerable insight into the predictive uncertainty estimation and out-of-distribution detection in medical image segmentation and provide practical recipes for confidence calibration. Moreover, we consistently demonstrate that model ensembling improves confidence calibration.

CHAPTER 7: PEP: PARAMETER ENSEMBLING BY PERTURBATION

Ensembling is recognized as an effective approach for increasing the predictive performance and calibration of deep networks. We introduce a new approach, Parameter Ensembling by Perturbation (PEP), that constructs an ensemble of parameter values as random perturbations of the optimal parameter set from training by a Gaussian with a single variance parameter.

The variance is chosen to maximize the log-likelihood of the ensemble average ($\mathbb{L}$) on the validation data set. Empirically, and perhaps surprisingly, $\mathbb{L}$ has a well-defined maximum as the variance grows from zero (which corresponds to the baseline model). Conveniently, calibration level of predictions also tends to grow favorably until the peak of $\mathbb{L}$ is reached. In most experiments, PEP provides a small improvement in performance, and, in some cases, a substantial improvement in empirical calibration. We show that this "PEP effect" (the gain in log-likelihood) is related to the mean curvature of the likelihood function and the empirical Fisher information. Experiments on ImageNet pre-trained networks including ResNet, DenseNet, and Inception showed improved calibration and likelihood. We further observed a mild improvement in classification accuracy on these networks. Experiments on classification benchmarks such as MNIST and CIFAR-10 showed improved calibration and likelihood, as well as the relationship between the PEP effect and overfitting; this demonstrates that PEP can be used to probe the level of overfitting that occurred during training. In general, no special training procedure or network architecture is needed, and in the case of pre-trained networks, no additional training is needed.

CHAPTER 8: CONCLUSION AND FUTURE WORKS

This chapter includes a short summary followed by a discussion of the methods for prostate cancer diagnosis and interventions. It also includes suggestions for future works.

Chapter 2

Prostate Cancer Diagnosis in MRI

2.1 Introduction and Background

Prostate cancer is the most frequently diagnosed noncutaneous male malignancy and the second leading cause of cancer-related mortality among men in the United States [165]. Magnetic resonance imaging (MRI) is widely used for prostate cancer detection, localization, diagnosis, and guidance for biopsy procedures due to its ability in providing superior contrast between cancer and adjacent soft tissue [183]. Convolutional neural networks (CNNs) have been successfully used for prostate cancer detection, localization, and characterization in medical images [192]. Machine learning models are often trained with pathology results from biopsy procedures [7]. Training deep CNNs with sparse labels can be addressed with both patch-based [22, 110, 156, 190] and fully convolutional neural network (FCN) models [25, 71, 87, 88, 154]. In patch-based training, samples with the biopsy points at their center are created and the network will act as a binary classifier. Hence, the model output will be two neurons corresponding to the binary output. The input of FCNs is the whole image slice containing the prostate and the output is a probability map image with the same dimensions as the input image. FCNs can be used to create a cancer probability map for the whole prostate. Furthermore, FCN architectures allow efficient training and learning of contextual features and provide a computationally efficient method to estimate cancer probability for the whole volume. Biopsy ground truth location can become noisy due to the registration and sampling errors during biopsy procedures. Further noise can be introduced as a result of inter-modality image registration or in

Section 2.3 of this chapter is adapted from Alireza Mehrtash, Alirea Sedghi, Mohsen Ghafoorian, Mehdi Taghipour, Clare M. Tempany, William M. Wells III, Tina Kapur, Parvin Mousavi, Purang Abolmaesumib, Andrey Fedorov. Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. Medical Imaging 2017: Computer-Aided Diagnosis. International Society for Optics and Photonics, 10134: 101342A, 2017.

the course of annotating biopsy locations on multiparametric MRI (mpMRI). Noisy ground truth can have adverse effects on both training and inference of computer-assisted diagnosis systems.

In this chapter, we study the problem of clinically significant prostate cancer diagnosis with CNNs. We develop both patch-based and FCN models for diagnosis. We further propose a probabilistic framework to include biopsy location uncertainty into the inference. In summary, we make the following contributions:

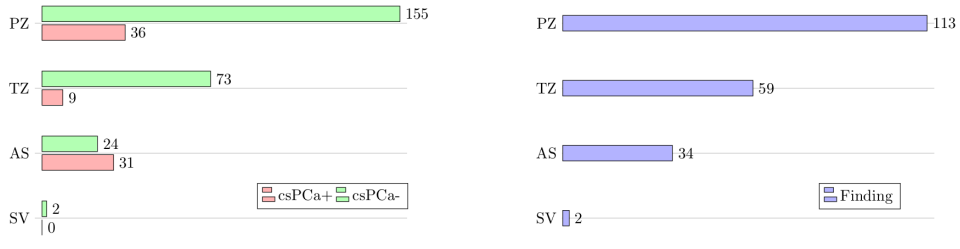
- We present a 3D CNN tailored for the task of diagnosis of clinically significant prostate cancer of suspicious findings in mpMRI. The proposed network benefits from the explicit addition of location-aware features (zonal information of the finding).

- We propose an FCN for end-to-end diagnosis and segmentation of clinically significant cancer tissues. To do so we present a Gaussian weighted loss as a label imputation mechanism for training FCNs with sparse biopsy data. We compare the proposed loss function with partial cross-entropy (CE) [27, 159, 177] where biopsy locations are used for loss calculation in optimization. We observe that FCNs trained with the proposed loss function perform achieves better classification results compared to those trained partial CE loss.

- We propose a probabilistic framework for modeling ground truth location uncertainty. By using priors on observed biopsy locations we calculate the probability of the true latent biopsy locations. Using the posterior biopsy location probability distributions, we adjust the biopsy location for nearby positive findings (lesions). We observe that compared to baselines, updated biopsy location improves sensitivity significantly through detecting lesions where the biopsy location was displaced.

- We train and validate our proposed methods on PROSTATEx dataset [7, 106].

The rest of this chapter is organized as follows: in Section 2.2, we describe the PROSTATEx dataset that was used for this study. Section 2.3 presents the proposed 3D patch-based CNN model for cancer diagnosis. Section 2.4 covers the proposed FCN model and the probabilistic framework for clinically significant cancer segmentation and diagnosis. Section 3.5 presents a discussion and our conclusions from this chapter.



(a) Training data

(b) Test data

Figure 2.1: Distribution of training and test datasets of the PROSTATEx challenge. **(a)** Training samples: the distribution of lesion findings shows that the training dataset is not balanced in terms of both zonal distribution and the clinical significance of the finding. **(b)** Test samples are not balanced in terms of zones.

2.2 Data

The training dataset consisted of 204 patients with 330 suspicious lesion findings, and the test dataset consisted of 140 patients with 208 findings. For each of the findings, assignment to one out of a possible four prostate anatomic regions was available. These anatomic regions are: the peripheral zone (PZ), which comprises $70-80\%$ of the glandular tissue and accounts for $\approx 70\%$ of prostate cancers; the transition zone (TZ), which comprises 5% of the glandular tissue and accounts for $\approx 25\%$ of prostate cancers; the central zone (CZ), which comprises 20% of the glandular tissue and accounts for $\approx 5\%$ of prostate cancers; and the non-glandular anterior fibromuscular stroma (AS) [109]. The training and test samples in the PROSTATEx challenge were from PZ, TZ, AS, and seminal vesicles (SV) as illustrated in Figure 2.1.

2.3 Patch-based Cancer Classifier

2.3.1 Preprocessing

After minor data cleaning that consisted of excluding patients with incomplete series and also SV findings, we selected 201 subjects with 321 findings for training and validation purposes. In order to augment and balance the training dataset, we used flipping and translation of the original data. As a result of data augmentation, we generated 5-fold cross-validation datasets with 10,000 training and 2,000 validation samples for each fold. For training-

validation splitting, we used stratified sampling based on pathology outcome and the prostate zone to make the subgroups homogeneous. Image intensities were normalized to be within the range of [0,1]. 3D patches of size $40 \times 40 \times 40$ mm for T2, $32 \times 32 \times 12$ for DWI and DCE-MRI images, centered at finding locations served as training image patches.

2.3.2 Network Architecture

Our CNN architecture, illustrated in Figure 2.2, included three input streams: ADC maps and maximum b-value from DWI, and K^{trans} from DCE-MRI. Similar to the work of Ghafoorian et al. [45], we added explicit zone information to the first dense layer. The DCE-MRI and DWI streams with input sizes of $(32 \times 32 \times 12)$ had 9 convolutional layers combining of $(3 \times 3 \times 1)$ and $(3 \times 3 \times 3)$ filter sizes. Max-pooling layers of size $2 \times 2 \times 1$ were applied in selected middle layers. At the end of each stream, the output of the last convolutional layer was connected to a dense layer. The neurons of this layer were concatenated with the zonal information of the finding and applied to another set of three fully connected layers. Leaky rectified linear unit [114] function, which allows a small, non-zero gradient when the unit is not active, was used as the non-linearity element.

2.3.3 Training

For training the network, we used the stochastic gradient descent algorithm with the Adam update rule [86], a mini-batch size of 64, and a binary cross-entropy loss function. We initialized the CNN weights randomly from a Gaussian distribution using the He method [59]. We also batch-normalized [70] the intermediate responses of all layers to accelerate the convergence. To prevent overfitting, in addition to the batch-normalization, we used drop-out with 0.25 probability as well as L_2 regularization with $\lambda_2 = 0.005$ penalty on neuron weights. We used an early stopping policy by monitoring validation performance and picked the best model with the highest accuracy on the validation set. Cross-validation was used to find the best combination of input channels and the number of filters for convolutional layers.

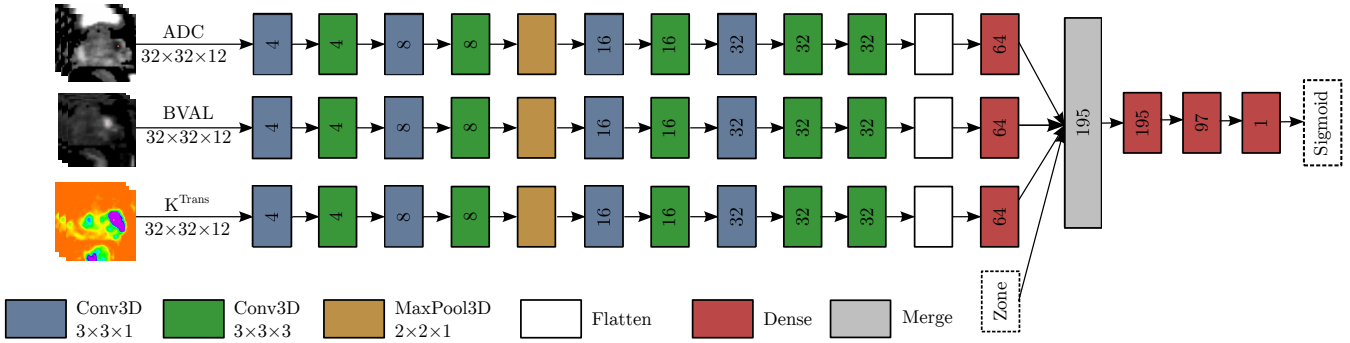


Figure 2.2: Architecture of the proposed 3D CNN for the PROSTATEx Challenge for detection of clinically significant cancer. The network uses a combination of ADC map, maximum B-Value (BVAL) from DWI, and K^{trans} from DCE-MRI with zone information.

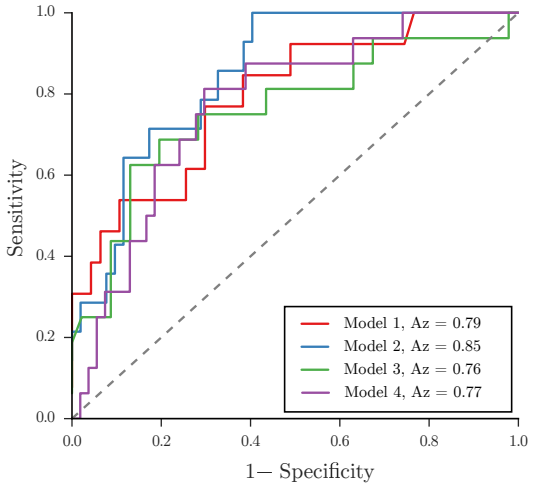
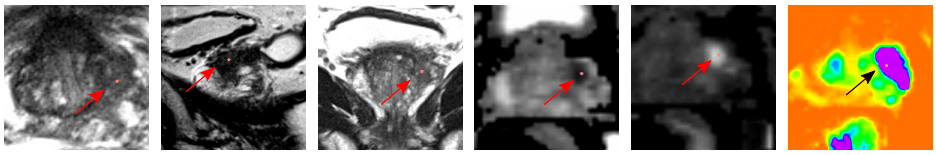


Figure 2.3: Comparison of classifiers trained with architecture in Figure 2.2 on different folds of cross-validation.

2.3.4 Results

Our training-validation results indicate that the combination of ADC, maximum b-Value, and K^{trans} modalities in combination with zonal information of the lesion leads to the best performance characterized by the area under the curve (A_z) of the receiver operating characteristic (ROC) curve. Figure 2.3 shows the results of training on different folds of our cross-validation. For test data prediction we combined the prediction of the best 4 out of the 5 models by averaging the outputs of the models. Figure 2.4 shows an example of a true positive finding in the validation set.

This network was evaluated by the organizers of the PROSTATEx challenge on a held-out test set containing 206 findings from 140 patients and achieved and area under the curve (AUC) of receiver operating characteristic curve (ROC) of 0.80. This is within the range of our validation results, indicating that the proposed model generalized well on the test data. Our results are also comparable with the A_z values of 0.79 and 0.83 achieved by an experienced human reader for PI-RADS v1 and PI-RADS v2, respectively [80]. The proposed method ranked 6^{th} out of 72 entries in the challenge [7].



(a) (b) (c) (d) (e) (f)

Figure 2.4: An example of a PZ true positive in the validation set. Only (d-f) modalities with zone information (zone=PZ) were used by the network to predict the clinical significance of the finding.

2.4 FCN Classifier and Uncertainty in Biopsy Location

Here, we consider MRI-guided cancer diagnosis with sparse biopsy ground truth as a weakly-supervised binary classification problem. The input images $I_i \in \mathbb{R}^n$ are n-dimensional. Sparse labels $z_i \in \{0, 1\}$ are reported biopsy results where 0 corresponds to benign or clinically insignificant cancer (Gleason score $\leq$ 3+3) and 1 corresponds to clinically significant cancer (Gleason score $\geq$ 3+4). Each biopsy label comes with a reported (observed) biopsy coordinate $x_i^o \in \mathbb{R}^3$ that can be noisy. Figure 2.5 visually illustrates the problem and the proposed method.

2.4.1 Gaussian Weighted Loss

Deep neural networks are often optimized using maximum likelihood estimation:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_i \ln(p(z_i|I_i, x_i^o, \theta)) \tag{2.1}$$

In this problem the labeled data is sparse. Partial CE loss [27, 177] can be used to train an FCN with partially labeled data. We can consider labeling the adjacent pixels having the same label as the reported biopsy points. Label imputation can be done around the biopsy point, by considering a conditional probability between pseudo-label sample pairs $(x_i, \hat{y}_i)$ and biopsy location (x_i^o, y_i) such that the pseudo-labels have the same class label as the observed label. We assume conditional probability $p(x_i^o|x_i)$ has a Gaussian $p(x_i^o|x_i) \sim \mathcal{N}(x - x^o, \Sigma)$ distribution. Using this, we can rewrite Equation 2.1 as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_i \underset{p(x_i|x_i^o, I_i, \theta, y_i)}{\mathbb{E}} [\ln(p(z_i|I_i, x_i, \theta))] \tag{2.2}$$

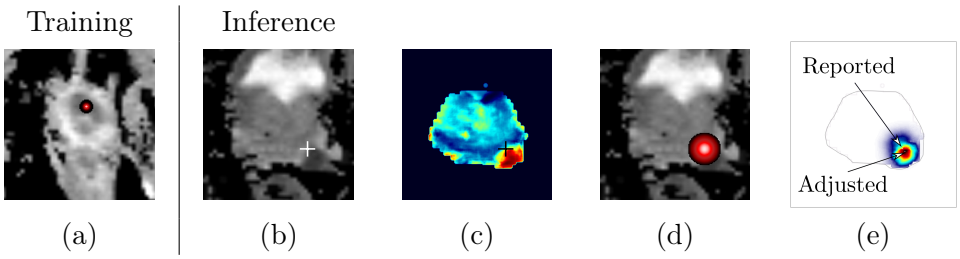


Figure 2.5: Method overview. For the training time (a) we propose a weighted cross-entropy (CE) loss to handle sparse biopsy ground truth. For the inference time (b-d), we propose a probabilistic framework which models noise in observed locations and makes adjustments. FCN architecture is used for the seamless rendering of the cancer probability maps (c). (a) shows a sample train image (ADC) together with a Gaussian loss weight centered on biopsy location. The weight is applied to the pixel-level samples of the CE loss. The trained network with optimized parameters, $\hat{\theta}$, is then used for inference. (b) shows a sample test image, I_i, with reported (observed) biopsy position, x_i^o, marked on the left peripheral zone. (c) shows the network prediction for probability of cancer at each pixel $p(z_i = 1|I_i, x_i, \hat{\theta})$. $z_i = 1$ denotes cancer outcome for biopsy. (d) shows the input image overlaid with a Gaussian denoting $p(x_i^p|x_i)$, the probability of latent true biopsy given observed biopsy location. (e) shows the probability distribution for latent true biopsy location $p(x_i|x_i^o, I_i, \hat{\theta}, z_i = 1)$. Using (e), the presumably misplaced reported location of the biopsy can be adjusted. The proposed network uses multi-modal inputs and here for simplicity we only show ADC inputs.

that can be further expanded into:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_i \sum_{x_i} p(x_i^i|x_i^o, I_i, \theta, z_i) \cdot \ln(p(z_i|I_i, x_i, \theta)) \tag{2.3}$$

$p(x_i^i|x_i^o, I_i, \theta^n, z_i)$ can be considered as a weight that will be applied during the training to the pixel samples. The weighted loss function in FCNs can be interpreted as an alternative for shift augmentation in patch-based training.

2.4.2 Location Uncertainty-aware Inference

The observed location of biopsy points x_i^o can be noisy. By modeling the noise in biopsy locations as Gaussian, we can formulate the latent true biopsy

coordinate $x_i \in \mathbb{R}^3$ as:

$$p(x_i^o|x_i) \sim \mathcal{N}(x_i - x_o, \Sigma) \tag{2.4}$$

Given the image I_i, classifier estimates the probability of cancer at the location x_i^o, as $p(z_i|I_i, x_i, \hat{\theta}^n)$. We can form a conditional probability to represent the most probable latent location of biopsy:

$$p(x_i|x_i^o, I_i, \hat{\theta}, z_i) = \frac{p(z_i|I_i, x_i, \hat{\theta}) \cdot p(x_i^o|x_i)}{\sum_{x_i} p(z_i|I_i, x_i, \hat{\theta}) \cdot p(x_i^o|x_i)} \tag{2.5}$$

For each pixel in image the probability of latent location given the cancer outcome is positive $p(x_i|x_i^o, I_i, \hat{\theta}, z_i = 1)$ or negative $p(x_i|x_i^o, I_i, \hat{\theta}, z_i = 0)$ can be calculated. To improve the sensitivity of the model, and reducing the chance of missing cancer, the reported biopsy can be adjusted to the most probable latent biopsy location x^* given the outcome is positive:

$$x^* = \underset{x}{\operatorname{argmax}}\, p(x_i|x_i^o, I_i, \hat{\theta}, z_i = 1). \tag{2.6}$$

2.4.3 Experimental Setup

Data and Preprocessing

We used 203 patients with 325 suspicious lesion from the training set of the PROSTATEx dataset [106]. Stratified 6-fold cross validation was used for the training and validation of the proposed methods. Registration with mutual information maximization [191] was used to adjust possible misalignments between mpMRI sequences. We followed the same registration procedure that was done by Kiraly et al. [87]. Registered images were then resampled to the resolution of $0.5 \times 0.5 \times 3$ mm. All axial slices were then cropped at the center to create images of size 224×224 pixels as the input size of the FCN. Image intensities were normalized to be within the range of [0,1].

Model & Training

We used an architecture similar to U-Net [147] but with three input channels and fewer kernel filters at each layer. The inputs of the model are ADC and high b-value images from DWU and K^{trans} from DCE-MRI. The input and output of the model have sizes of $224 \times 224 \times 3$, and $224 \times 224 \times 2$, respectively. The network has the same number of layers as the original U-Net. The sizes of kernels for the encoder section of the network are 16, 16, 32, 32, 32, 32, 64, 64, 128, and 128. The parameters of the convolutional layers

were initialized randomly from a Gaussian distribution [59]. For optimization, stochastic gradient descent with the Nesterov Adam update rule [34] was used. A mini-batch of 16 examples was used during the training. The initial learning rate was set to 0.0005 and it was reduced by a factor of 0.5 if the average validation loss did not improve by 0.001 in 5 epochs. We used 50 epochs for the training of the models with an early stopping policy. For each training run, the model checkpoint was saved at the epoch where the validation loss was lowest. For each of the validation folds, the model was trained 5 times with partial CE and 5 times with Gaussian weighted CE, each with random weight initialization and random shuffling of the training data. We used ensembling by averaging network predictions to boost performance and calibration of the models [96].

Experiments

We compare the classification performance of models trained with partial CE loss with those trained with Gaussian-weighted CE. We generate 2D Gaussian weighted with σ values of 0.5, 1, and 2. For all trained models, we calculate $p(x_i|x_i^o, I_i, \hat{\theta}, z_i = 1)$ with σ values of 5, 9, and 15 and find the adjusted biopsy location x^*. We then compare the baseline predictions at the reported biopsy locations with probabilities at the adjusted biopsy location. For statistical tests and calculating 95% confidence intervals (CI), we use bootstrapping ($n = 1000$).

2.4.4 Results

Table 2.1 compares the classification performances of the models trained with partial CE with those trained with Gaussian CE loss with $\sigma = 2$. The area under the receiver operating characteristic curve (ROC) (A_z) was improved from 0.74 to 0.78 by including adjacent labels in loss calculation. Gaussian CE with σ values of 0.5 and 1 achieved A_z (95% CI) of 0.78 (0.72−0.84) and 0.77 (0.71−0.83), respectively. Both A_zs significantly better than the baseline with partial CE. Table 2.1 also compares the performance of original (observed) biopsy points with adjusted locations with different σ values. As expected, biopsy location adjustment acts in favor of finding lesions and increases sensitivity. The increase in sensitivity is at the cost of a notable reduction in the specificity of the models.

Figure 2.6 provides some examples of the proposed probability framework location uncertainty estimation and biopsy location adjustments.

2.5 Discussion and Conclusion

In this chapter, we studied training deep cancer classifiers using sparse biopsy annotations. Moreover, we modeled biopsy location uncertainty and proposed a method for improving the sensitivity of the models by biopsy location adjustment. The FCN was trained with DWI and DCE-MRI data as input and biopsy ground truth as targets for clinically significant prostate cancer detection. We proposed a Gaussian weighted loss function as a form of data augmentation (label imputation) to train FCNs with sparse biopsy locations. Furthermore, we proposed a probabilistic framework to model the uncertainty regarding the location of the biopsy samples. The framework provides posterior probabilities of latent true biopsy locations given the image, model, observed biopsy location, and the biopsy outcome. The models were assessed on 325 biopsy locations. The results of our experiments show that the proposed weighted loss provides better classification compared to the baselines that only used given sparse labels in their loss function. We showed improved sensitivity and area under ROC curves by adjusting biopsy location adjustments at the expense of more false positives.

The main limitation of our work is the relatively small number of cases that were used for the development and validation of our system. Only about one fourth of the biopsies were clinically significant cancers. Not only cancerous samples were limited, but also they were heterogeneous in terms of severity of cancer and the lesion sizes. Despite this, the achieved results are promising and should be validated by larger patient populations and preferably with independent test sets.

Future work will explore the use of the posterior probabilities on latent true biopsy coordinates for improving training procedures. Through an expectation-maximization (EM) framework, Equation 2.5 can be used as an $E-$step to re-estimate probability distribution on biopsy locations given the prior knowledge and classifier's output. Maximum likelihood estimation (Equation 2.1) can be updated to include the current knowledge of the distribution where samples come from (M$-$step).

Table 2.1: Classification quality of models for diagnosing clinically significant prostate cancer in MRI evaluated on reported biopsy locations (n=325). Models trained with partial cross-entropy loss are compared with those trained with Gaussian cross-entropy loss. The results of inference time biopsy location adjustments are also provided for multiple Gaussian kernel sizes.

Biopsy Location	TP	TN	FN	FP	Sens (%)	Spec (%)	F-1	A_z (95%CI)
	Partial Cross-entropy Loss							
Original	37	210	40	38	48.05	84.68	0.49	0.74 (0.67-0.80)
Adjusted ($\sigma = 5$)	55	161	55	22	71.43	64.92	0.50	0.76 (0.70−0.82)
Adjusted ($\sigma = 9$)	68	116	9	132	88.31	46.77	0.49	0.77 (0.70−0.83)†
Adjusted ($\sigma = 15$)	72	79	5	168	93.50	31.85	0.45	0.78 (0.71−0.84) †
	Gaussian Cross-entropy Loss ($\sigma = 2$)							
Original	41	212	36	36	53.25	85.48	0.53	0.78 (0.72−0.83)
Adjusted ($\sigma = 5$)	47	198	30	50	61.04	79.84	0.54	0.79 (0.74−0.85) †
Adjusted ($\sigma = 9$)	57	174	20	74	74.03	70.16	0.55	0.77 (0.72−0.83)
Adjusted ($\sigma = 15$)	65	127	12	121	78.00	51.21	0.49	0.79 (0.73−0.84) †

TP = true positives; TN = true negatives; FN = false negatives; FP = false positives; Sens = sensitivity; Spec = specificity;

† Difference are statistically significant (p-value<0.01).

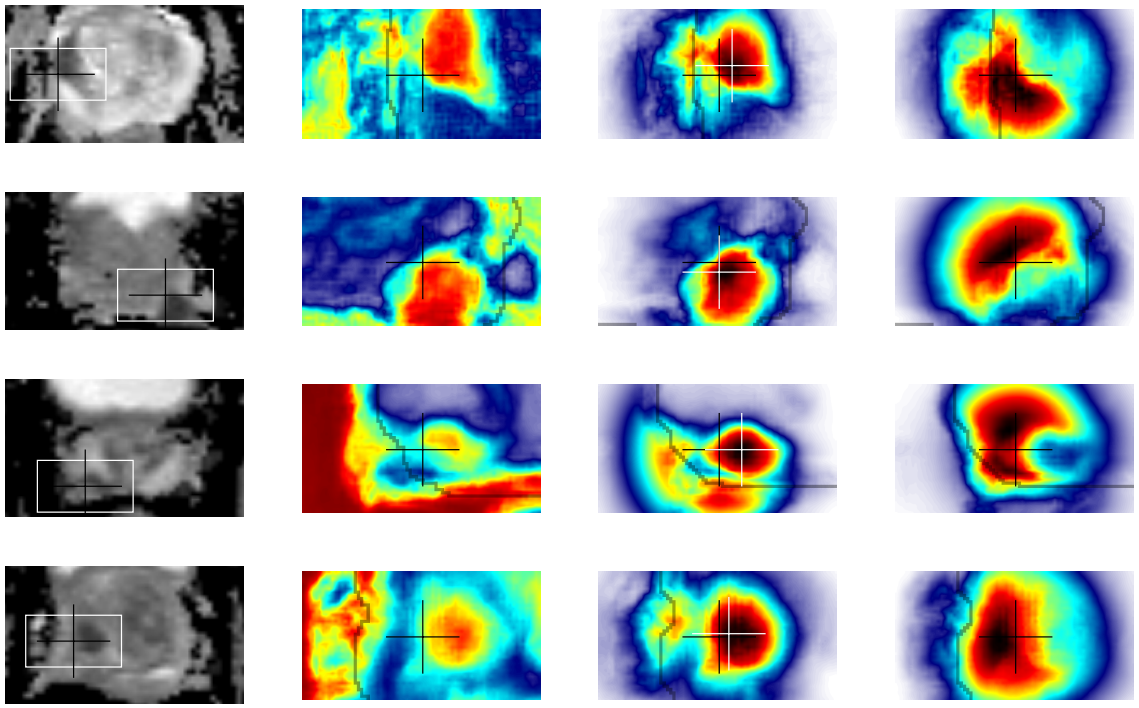


Figure 2.6: Examples of biopsy location adjustments. The first column shows the input ADC images with given biopsy locations (black crossharis). The b-value and K^{Trans} images were used for inference but not shown here. For all the of the four examples, the most probable latent location for cancer location was found (white crosshairs) using Equation 2.6. Last two columns show calculated true latent biopsy location probabilities given the results is clinically significant (third column), or insignificant (fourth column). The top two rows show false positive predictions that turned into true positives by the proposed adjustment. The bottom two rows show true negatives that turned into false positives by adjustment.

Chapter 3

Biopsy Needle Localization in MRI

3.1 Introduction and Background

When screening indicates the possibility of prostate cancer in an individual, the standard of care includes non-targeted systematic (sextant) biopsies of the entire organ under the guidance of transrectal ultrasound (TRUS). These biopsies randomly sample a very small part of the gland and the results sometimes miss the most aggressive tumor within the gland [36, 58, 155]. MRI has demonstrated value in not just detecting and localizing the cancer, but also in guiding biopsy needles to the suspicious targets [4]. In particular, biopsies with intra-operative MRI guidance require significantly fewer cores than with the standard TRUS approach and reveal a significantly higher percent of cancer involvement per biopsy core [136, 139, 187, 196].

Accurately placing needles in suspicious target tissue is critical for the success of a biopsy procedure. An intra-operative MRI allows the physician to check the position and trajectory of the needle relative to the suspicious target in a three-dimensional (3D) stack of cross-sectional images and to make needed adjustments. Physicians achieve targeting accuracy in the range of 3–6 mm for MRI-guided prostate biopsy, which is adequate for the task since clinically significant prostate cancer lesions are typically larger than 0.5 mL in volume or 9.8 mm in diameter (assuming spherical lesions) [90, 170, 182]. Automatic localization of the needle tip and trajectory can aid the physician by providing rapid 3D visualization that reduces their cognitive load and the duration of the procedure. In addition, for the realization of robot-guided percutaneous needle placement procedures, accurate and automatic needle

This chapter is adapted from Alireza Mehrtash, Mohsen Ghafoorian, Guillaume Pernelle, Alireza Ziaei, Friso G. Heslinga, Kemal Tuncali, Andriy Fedorov, Ron Kikinis, Clare M Tempany, William M Wells, Purang Abolmaesumi, Tina Kapur. Automatic needle segmentation and localization in MRI With 3-D convolutional neural networks: application to MRI-targeted prostate biopsy. IEEE Transactions on Medical Imaging, 38(4):1026-1036, 2018.

localization is a necessary part of the feedback loop [145].

While MRI is the imaging modality of choice for identifying suspicious biopsy targets because of its ability to provide superior soft tissue contrast, it poses two types of challenges in the needle localization task. The first challenge is that parts of a needle may appear substantially different from others in an MRI scan while also being difficult to distinguish from surrounding tissue [141]. This variability in grayscale appearance of needles confounds automatic segmentation algorithms and is addressed in this study.

Today, aside from the proposed work, there are no automatic solutions for the segmentation of needles from MRI images [32, 167]. Even manual segmentation from MRI is tedious and error-prone, and to the best of our knowledge, not attempted in clinical or research programs. The second challenge, while not addressed in this study, is worth noting; an MRI does not directly show the geometric location of a needle. Instead, the needle is detected through a loss of signal due to the susceptibility artifact that occurs at the interfaces of materials with substantially different magnetic resonance properties, and is commonly referred to as the needle artifact. Studies report a displacement between the actual needle tip and the needle tip artifact [167]. For brevity the term needle is used instead of needle artifact in this study. Needle trajectory is defined as the set of points connecting the center of the artifact across a stack of axial cross-sections. The needle tip is the center of needle artifact at the most distal plane.

Several approaches have been suggested in the literature for segmentation and localization of needle-like i.e. elongated tubular objects in medical images. Segmentation of tortuous and branched structures, such as blood vessels [105, 193], white matter tracts [57, 133] or nerves [174] are the targets of many reported methods. Other methods target straight or bent catheters [65, 116, 137]. Based on the clinical application, the proposed techniques have been applied to different image modalities including ultrasound [3, 13, 65], computed tomography [52, 128], and MRI [116, 137] for the purpose of localization after insertion or real-time guidance during insertion. Many attempts have been made to incorporate hand crafted and kernel-based methods to segment and localize the objects which can be considered as line detection algorithms. The reported methods are based on 3D Hough transforms [13, 65], on model based and raycasting-based search [116, 137], orthogonal 2-dimensional projections [3], generalized radon transforms [132], and random sample consensus (RANSAC) [184].

Deep convolutional neural networks (CNNs) use the power of representation learning for complex pattern recognition tasks [51]. Deep model representations are learned through multiple levels of abstraction in a su-

pervised training scheme, as opposed to hand-crafting of features. CNNs have been extensively used in medical image analysis and have outperformed conventional methods for many tasks [107]. For instance, CNNs have been shown to achieve outstanding performance for segmentation [45], localization [29], cancer diagnosis [120], quality assessment [2], and vessel segmentation [189].

In this work, we propose a CNN-based system for automatic segmentation and localization of biopsy needles in MRI images. The proposed system uses CNNs to extract hierarchical representations from MRI to segment needles for the purpose of tip and trajectory localization. An asymmetric 3D fully convolutional neural network with in-plane pooling and up-sampling layers was designed to handle the anisotropic nature of the needle MR images. The proposed asymmetry in the network design is computationally efficient and allows the whole volumetric MR images to be used for the process of training. A large dataset of MRI acquired in transperineal prostate biopsy procedures was used for developing the system; 583 volumetric T2-weighted MRI from 71 biopsy procedures (on 71 distinct patients) were used to design, optimize, train and test the deep learning models.

The performance of CNNs and other supervised machine learning methods is measured against that of experienced humans, which is known to be variable for medical image analysis tasks; observer studies are used to establish ranges for human performance, against which automated CNNs can be rated. An observer study was conducted to compare the quality of the predictions against a second observer.

To promote further research and facilitate reproduction of the results, the resultant trained deep learning model is publicly available via DeepInfer[119], an open-source deployment platform for trained models. To the best of our knowledge, we are the first and only group to attempt fully automatic segmentation and localization of needles in MRI. Since there are no prior assumptions regarding the prostate images, the proposed method can be generalized and adopted in other clinical procedures for needle segmentation and localization in MRI.

The rest of this chapter is organized as follows: in Section 3.2 we describe the methods for this study including the clinical workflow of in-gantry MRI-targeted prostate biopsy and details of the proposed deep learning system. Section 3.3 and 3.4 cover the experimental setup and results, respectively, of applying the proposed system to the MRI-targeted biopsy procedure. Section 3.5 presents a discussion and our conclusions from this study.

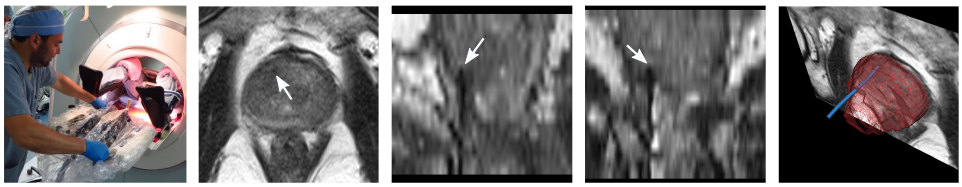


Figure 3.1: Transperineal in-gantry MRI-targeted prostate biopsy procedure: (a) The patient is placed in the supine position in the MRI gantry, and his legs are elevated to allow for transperineal access. The skin of the perineum is prepared and draped in a sterile manner, and the needle guidance template is positioned. (b), (c) and (d): Axial, sagittal and coronal views of intraprocedural T2-W MRI with needle tip marked by white arrow. (e) 3D rendering of the needle (blue), segmented by our method, and visualized relative to the prostate gland (purple), and an MRI cross-section that is orthogonal to the plane containing the needle tip.

3.2 Methods

3.2.1 MRI-Targeted Biopsy Clinical Workflow

The general workflow of an in-gantry transperineal MRI-targeted prostate biopsy involves imaging in two stages: a) the preoperative stage during which multiparametric MRI consisting of T1, T2, diffusion weighted, and dynamic contrast enhanced images are acquired and the cancer suspicious targets are marked, and b) the intraoperative stage during which the patient is immobilized on the table top inside the gantry of the MRI scanner and tissue samples are acquired transperineally with a biopsy needle under intraoperative MRI guidance. At the beginning of the intra-operative stage, anesthesia is administered to the patient and a grid template affixed to his perineum to facilitate targeted sampling. Intra-operative MR images are acquired as needed to optimize the skin entry point and depth for each needle insertion. One or more biopsy samples are taken for each target, depending on the sample quality. Samples are sent for histological analysis, and the institutional post-operative care protocol is followed for the patient. At our site, almost 600 such procedures have been performed under intravenous conscious sedation; one to five biopsy samples are obtained using an off-the-shelf 18-gauge side-cutting MR-compatible core biopsy needle and the patient discharged, on an average, two hours later [38, 136, 179].

3.2.2 Data

The data used in this study consists of 583 intraprocedural MRI scans obtained from 71 patients who underwent transperineal MRI-guided biopsy between December 2010 and September 2015. This retrospective study was HIPAA compliant and institutional review board approval (IRB) and informed consent was obtained. The patients in this cohort had prostate MRI lesions suspicious for new cancer, recurrent cancer after prior therapy, or lesions suspicious for higher grade cancer than their initial diagnosis. Each of the intraprocedural MRI scans is an axial fast spin echo (FSE) T2-weighted volume of size range $256–320 \times 204–320 \times 18–30$ voxels, with voxel spacing in the range of 0.53–0.94 mm in-plane and slice thickness of 3.6–4.8 mm. The acquisition parameters for the FSE sequence were set as follows: repetition time (TR) is 3000 ms, echo time (TE) is 106 ms, and flip angel (FA) is 120 degrees [136]. The imaging time is about one minute and is performed after needle insertion to visualize it relative to the target. These scans were acquired on either a conventional wide-bore, 3T MR scanner (Verio, Siemens Healthcare, Erlangen, Germany) or a ceiling-mounted version of it (IMRIS/Siemens Verio; IMRIS, Minnetonka, Minn).

3.2.3 Data Annotation

A custom needle annotation software tool was used by an expert human rater to interactively mark the needle trajectory and tip on each of the 583 MRI [137]. These annotations are also referred to as ground truth. In these images, a needle can be identified by the dark susceptibility artifact around its shaft, as seen in Figure 3.1(c) and (d). The annotation tool allowed the human rater to place several control points ranging from the tip of the needle to its base. Those control points were then used to fit a Bézier curve which represents a trajectory of the needle artifact. Thus the manual needle trajectory relies only on the observer input (and not on the underlying gray scale values). Ground truth needle segmentation label maps were then generated by creating a 4 mm diameter cylinder around the Bézier curve to cover the hypointense artifact that surrounds the needle shafts, as seen in Figure 3.1(e).

It should be noted that even for experienced human observers, there can be ambiguity in picking the axial plane containing the needle tip due to the large slice thickness and partial volume effects. In addition, there are cases, as shown in Figure 3.2, where the needle susceptibility artifact consists of two hypointense regions separated by a hyperintense one (instead of a single hypointese region). The human observer followed the needle carefully across

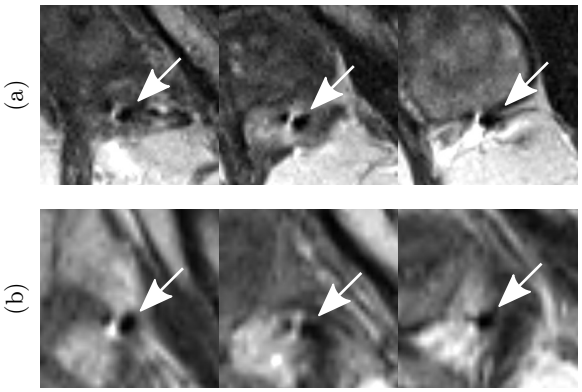


Figure 3.2: (a) and (b): Examples of needle induced susceptibility artifacts in MRI where instead of a single hypointense (dark) region, there are two hypointense regions separated by a hyperintense (bright) region. In such cases, the human expert followed the needle carefully across several slices to ensure the integrity of the annotation. The arrow marks the needle identified by the expert.

Table 3.1: Number of patients and needle MRIs for training/validation and test sets.

set	training & validation		test	
	# patients	# images	# patients	# images
size	50	410	21	173

several slices to ensure the integrity of the annotation.

The annotated images were split at the patient level into 70% training/cross-validation for algorithm development and 30% for final testing (Table 3.1).

3.2.4 Data Preprocessing

Prior to training the CNN models, the data was preprocessed in four steps: resampling, cropping, padding, and intensity normalization, as follows.

Resampling: First, the data was resampled to a common resolution of $0.88 \times 0.88 \times 3.6$ mm. The MR images and ground truth segmentation maps were resampled with linear and nearest neighbor interpolation methods respectively. SimpleITK implementation of the interpolation methods were used for image resampling [113].

Cropping: Second, to constrain the search area for the needle tip, each MR image was cropped to a cube of size $165.4 \times 165.4 \times 165.6$ mm ($188 \times 188 \times 46$ voxels) around the center of the prostate gland. The size of the box was chosen to be large enough to easily accommodate the size of the largest expected diseased gland, and small enough to fit in the GPU memory for efficient processing. Even though a very coarsely selected bounding box that contains the prostate gland is sufficient for this step, we used a separate deep network, a customized variant of 2D U-Net architecture, that was readily available to us to perform the segmentation automatically [119][1].

Padding: Third, the borders of the cropped volume were padded by 50 mm (14 pixels) in the z direction. The zero padding in the z direction is required to accommodate the reduction in spatial dimension of the output of 3D convolutional filters. As a result of convolution operation, for in-plane directions (x and y) the output of the final layer of the network (green box in Figure 3.3) will be 88 pixels smaller than the input image (blue box in Figure 3.3).

Intensity Normalization: Fourth, to reduce the heterogeneity of the grayscale distribution in the data, intensities were truncated and re-scaled to the range between 0.1% and 99% quantiles of the intensity histogram and then normalized to the range of $[0, 1]$.

[1] `http://www.deepinfer.org/models/prostate-segmenter/`

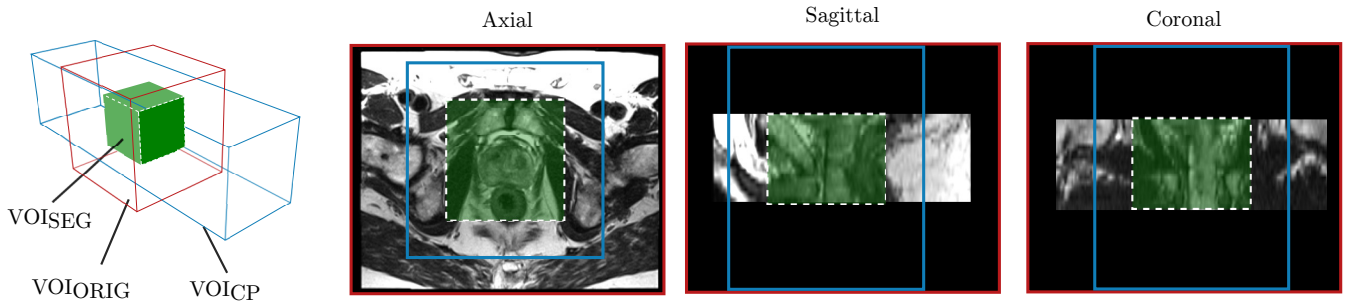


Figure 3.3: Original, cropped and padded, and segmentation volumes of interests (VOIs) (a) The original grayscale volume (VOI_{ORIG}, red box) is cropped in x and y directions and padded in the z direction to a volume of size $164.5 \times 164.5 \times 165.6$ mm ($188 \times 188 \times 46$ voxels) centered on the prostate gland (VOI_{CP}, blue box). VOI_{CP} is used as the network input. The network output segmentation map is of size $88 \times 88 \times 64.8$ mm ($100 \times 100 \times 18$ voxels) (VOI_{SEG}, green box). The adjusted voxel spacing for the volumes is $0.88 \times 0.88 \times 3.6$ mm. (b), (c), (d) show axial, sagittal and coronal views respectively of a patient case overlayed with the boundaries of the volumes VOI_{ORIG}, VOI_{CP} and VOI_{SEG}.

3.2.5 Convolutional Neural Networks

In this chapter a binary classification model based on CNNs is proposed for needle segmentation and localization in prostate MR images. The deep network architecture is composed of sequential convolutional layers $l \in [1, L]$. At each convolutional layer l, the input feature map (image) is convolved by a set of K kernels $\boldsymbol{W}_l = \{W^1, ..., W^K\}$ and biases $\boldsymbol{b}_l = \{b^1, ..., b^K\}$ to generate a new feature map. A non-linear activation function f is then applied to this feature map to generate the output $\boldsymbol{Y}_l$ which is the input for the next layer. The nth feature map of the output of the l^{th} layer can be expressed by:

$$Y_l^n = f(\sum_{k=1}^{K} \boldsymbol{W}_l^{n,k} * \boldsymbol{Y}_{l-1}^k + b_l^n), \tag{3.1}$$

The concatenation of the feature maps at each layer provides a combination of patterns to the network, which become increasingly complex for deeper layers. Training of the CNN is usually done through several iterations of stochastic gradient descent (SGD), in which several samples of training data (a batch) is processed by the network. At each iteration, based on the calculated loss the network parameters (kernel weights and biases) are optimized by SGD in order to decrease the loss.

Medical image segmentation can be formulated as a pixel-level classification problem which can be solved by convolutional neural networks. Leveraging the volumetric nature of the data through the inter-slice dependence of 2D slices is a key factor in 3D biomedical image classification and segmentation problems. Representation learning for segmentation in 3D has been done in different ways: directly by the use of 3D convolutional filters, multi-view CNNs with 2D images, and recurrent architectures [107]. 3D convolutional filters can be used in 3D architectures known as fully convolutional neural networks (FCNs) or through patch-based sliding-window methods[27].

The use of FCNs for image segmentation allows for end-to-end learning, with each pixel of the input image being mapped by the FCN to the output segmentation map. This class of neural networks has shown great success for the task of semantic segmentation [112]. During training, the FCN aims to learn representations based on local information. Although patch-based methods have shown promising results in segmentation tasks [45], FCNs have the advantage of reduction in the computational overhead of sliding-window-based computation. The efficiency of FCNs in prediction time makes them better suited for procedures such as intera-operative imaging where time is an important factor. One drawback of 3D FCNs is the memory constraint of the

graphical processing units (GPUs) to hold the large parameters during the optimization process which limits the input size, number of model parameters and number of mini-batches in stochastic gradient descent iterations. In addition to CNNs, recurrent neural networks (RNNs) have also been successfully used for segmentation tasks by feeding prior information from adjacent locations such as nearby slices or nearby patches into the classifier [6].

The CNN proposed in this work is a 3D FCN. FCNs for segmentation usually consist of an encoder (contracting) path and a decoder (expanding) path [8, 147]. The encoder path consists of repeated convolutional layers followed by activation functions with max-pooling layers on selected feature maps. The encoder path decreases the resolution of the feature maps by computing the maximum of small patches of units of the feature maps. However, good resolution is critical for accurate segmentation, therefore in the decoder path, up-sampling is performed to restore the initial resolution, but the feature maps are concatenated to keep the computation and memory requirements tractable. As a result of multiple convolutional layers and max-pooling operations the feature maps are reduced and the intermediate layers of an FCN become successively smaller. Therefore, following the convolutions, an FCN uses inverse convolutions (or backward convolutions) to up-sample the intermediate layers until the input resolution is matched [35, 112]. FCNs with skip-connections are able to combine high level abstract features with low level high resolution features which has been shown to be successful in segmentation tasks [27].

3.2.6 Network Architecture

We present a fully automatic approach for needle localization by segmentation in prostate MRI based on a 14-layers deep anisotropic 3D FCN with skip-connections (Figure 3.4). The network architecture is inspired by the 3D U-Net model [27]. We improved the network architecture to efficiently handle the anisotropic nature of MRI volumes and for the specific problem of needle segmentation in MRI. Due to the time constraints in intraoperative imaging, MRIs taken during the interventional procedure often have thick slices but high resolution in the axial plane which leads to anisotropic voxels. Pooling and up-sampling were only applied to the in-plane axes (x and y) to handle the anisotropic nature of the needle MRI. The proposed asymmetry in the network design is computationally efficient and allows the whole volumetric MRI to be used for training.

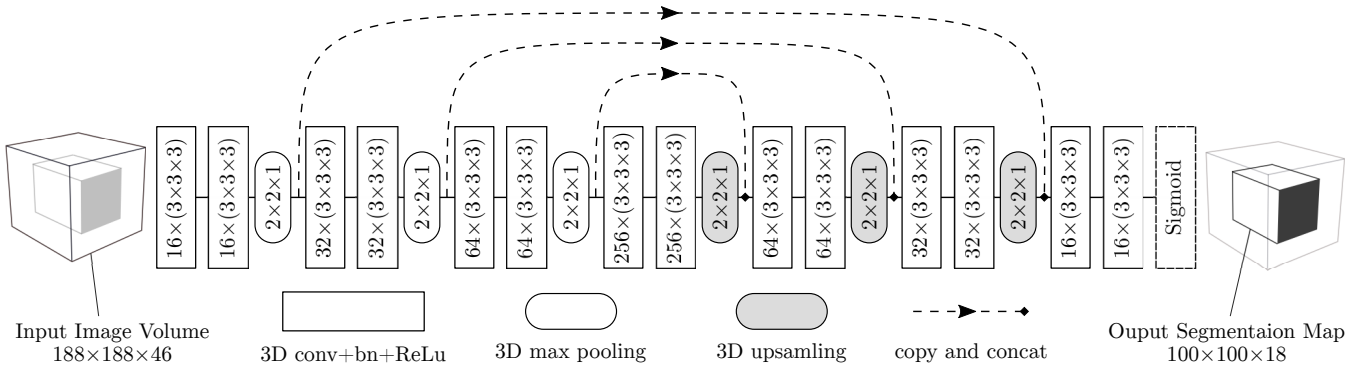


Figure 3.4: Schematic overview of the anisotropic 3D fully convolutional neural network for needle segmentation and localization in MRI. Network architecture consisting of 14 convolutional layers, 3 max-pooling and 3 up-sampling layers. Convolutional layers were applied without padding while max-pooling layers halved the size of their inputs only in in-plane directions. The parameters including the kernel sizes and number of kernels are explained in each corresponding box. Shortcut connections insures combination of low-level and high-level features. The input to the network is the 3D volume image with the prostate gland at the center ($188 \times 188 \times 46$) and the output segmentation map has the size of $100 \times 100 \times 18$.

As illustrated in Figure 3.4, the proposed network consists of 14 convolution layers. Each convolution layer has a kernel size of $(3 \times 3 \times 3)$ with stride of size 1 in all three dimensions. Since the input of the proposed network is a T2-weighted MRI, the number of channels for the first layer is equal to one. After each convolutional layer, a rectified linear unit (ReLu) $f(x) = max(0, x)$ is used as the nonlinear activation function except for the last layer [55] where a sigmoid function $S(x) = e^x(e^x + 1)^{-1}$ is used to map the output to a class probability between 0 and 1. There are 3 max-pooling and 3 up-sampling layers of size $(2 \times 2 \times 1)$ in the encoder and decoder paths respectively. The network has a total of 3,231,233 trainable parameters. The input to the network is the 3D volume image with the prostate gland at the center $(188 \times 188 \times 46)$ and the output segmentation map size is $100 \times 100 \times 18$ which corresponds to a receptive field of size $88 \times 88 \times 65$ mm.

3.2.7 Training

During training of the proposed network, we aimed to minimize a loss function that measures the quality of the segmentation on the training examples. This loss $\mathcal{L}_t$ over N training volumes can be defined as:

$$\mathcal{L}_t = -\frac{1}{N}\sum_{n=1}^{N}\left(\frac{2|X_n \cap Y_n|}{|X_n|+|Y_n|+s}\right), \tag{3.2}$$

where X_n is the output segmentation map, Y_n is the ground truth obtained from expert manual segmentation for the n^{th} training volume, and s (set to 5), is the smoothing coefficient which prevents the denominator from being zero. This loss function has demonstrated utility in image segmentation problems where there is a heavy imbalance between the classes, as in our case where most of the data is considered background [124].

We used a SGD algorithm with the Adam update rule [86] which was implemented in the Keras framework [26]. During the training we used a mini-batch of 4 image volumes. The initial learning rate was set to 0.001. Learning rate was reduced by a factor of 0.8 if the average of validation Dice score did not improve by 10^{-5} in five epochs. The parameters of the convolutional layers were initialized randomly from a Gaussian distribution using the He method [59]. To prevent overfitting, in addition to the batch-normalization [70], we used drop-out with 0.1 probability as well as L_2 regularization with $\lambda_2 = 10^{-5}$ penalty on convolutional layers except the last one. Training was performed on 410 MRI scans from 50 patients using five-fold cross validation with splitting at the patient level. Each training sample was a 3D patch

(also referred to as input volume or VOI_{CP}) of size $188 \times 188 \times 46$ voxel. Data augmentation was performed by flipping the 3D volumes horizontally (left to right), which doubled the amount of training examples [44]. Cross-validation was used to optimize and tune the hyperparameters including CNN architecture, training scheme, and finding the best epoch (model checkpoint) for the test-time deployment. For each cross-validation fold, we used 100 as the maximum number of epochs for training and an early stopping policy by monitoring validation performance. This resulted in five trained models, one from each of the cross-validation folds, that are aggregated later with the ensembling method (described in Section 3.3.3) for test-time prediction.

3.3 Experimental Setup

3.3.1 Observer Study

We designed an observer study in which a second observer, blinded to the annotations by the first observer (the ground truth), segmented the needle trajectory on the test set (n = 173 images) using the same annotations tools as the first observer. We compared the performance of both the proposed automatic system and the second observer with the first observer (ground truth).

3.3.2 Evaluation Metrics

We evaluated the accuracy of the system by measuring how well it localizes the tip of the needle, and how well it segments the entire trajectory of the needle. In addition to measuring the tip and angular deviation errors which are commonly used to quantify targeting accuracy of percutaneous needle insertion procedures [13], we report the number of axial planes contained in the tip error because of the high anisotropy of the data set. We used the Hausdorff distance to measure the quality of the segmentation of the entire length of needle (beyond the tip error) [116, 137].

- Tip deviation error ΔP: The ground truth needle tip position was determined as the center of the needle artifact in the most distal plane of the needle segmentation image $P(x, y, z)$. Tip deviation ΔP is quantified as the 3D Euclidean distance between the prediction $\hat{P}$ and manually specified ground truth P in millimeters.

- Tip axial plane detection error ΔA: The tip plane detection error is the absolute value of the distance between the ground truth axial plane

index A containing the needle tip and the predicted axial plane index $\hat{A}$ in voxels.

- Hausdorff Distance HD: Trajectory accuracy was calculated by measuring the directed Hausdorff distance between two N-D sets of predicted $\hat{X}$ and ground truth X needles defined with

$$d_{\mathrm{H}}(X, \hat{X}) = \max\{ \sup_{x \in X} \inf_{\hat{x} \in \hat{X}} d(x, \hat{x}), \sup_{\hat{x} \in \hat{X}} \inf_{x \in X} d(x, \hat{x}) \}, \tag{3.3}$$

  where sup represents the supremum and inf the infimum, and x and $\hat{x}$ are points from X and $\hat{X}$ respectively.

- Angular deviation error $\Delta\theta$: The true needle direction θ was defined as the angle between the needle shaft and the axial plane. The angular deviation between the ground truth needle direction θ and the predicted needle direction $\hat{\theta}$ quantifies the accuracy of needle direction prediction ($\Delta\theta = |\theta - \hat{\theta}|$).

3.3.3 Test-time Augmentation

Test-time augmentation seeks to improve classification by analyzing multiple augmentations or variants of the same image and averaging out the results. Recently, it has been used to improve pulmonary nodule classification from CT [168], detection of lacunes from MRI [44], and prostate cancer diagnosis from MRI [74]. We performed test-time augmentation by flipping 3D volumes horizontally which doubles the test data.
We conducted experiments to quantify the impact of training-time and test-time augmentation and performed analysis to measure the statistical significance of the results. Paired comparison of needle tip localization errors for unequal cases was performed using Wilcoxon signed-rank test (two-tailed). For reporting statistical analysis results, statistical significance was set at 0.05.

3.3.4 Ensembling

As reported in Section 3.2.7, cross-validation resulted in five trained models. Combined with test-time augmentation, this results in 10 segmentation maps for each test case, i.e. for each of the five trained models, there is one prediction for the test image, and one for its flipped variant. To obtain the final binary prediction, we used an iterative ensembling or voting mechanism to aggregate the results of 10 predictions at the voxel level (Algorithm 1).

Algorithm 1 Overview of the ensembling algorithm

Require: S, the sum image of n predictions

Constant: $\nu = 100$, min number of voxels in a needle
Initialize: $\tau = \lceil \frac{n}{2} \rceil$, min vote threshold
Initialize: B=0, output segmentation image, same shape as S
repeat
  for voxel $s[i] \in S$ and $b[i] \in B$ **do**
    if $s[i] \geq \tau$ **then**
      $b[i] \leftarrow 1$ {voxel is needle}
    else
      $b[i] \leftarrow 0$ {voxel is not needle}
    end if
  end for
  $\tau \leftarrow \tau - 1$ {relax threshold if no needle found}
  $c \leftarrow \sum_{b[j] \in B} b[j]$ {c is total number of needle voxels}
until $c \geq \nu$ or $\tau = 0$

The input to the algorithm is S, the sum image of all predictions. In an iterative procedure, the binary segmentation map B, is generated by thresholding S using τ. τ is initialized at the value of majority votes $\tau = \lceil \frac{n}{2} \rceil$, where n is the number of predictions. ν is a constant that represents the minimum size of a needle which was measured at 100 voxels over the 410 needles in the training set. An iterative procedure reduces τ until a needle is found.

Finally, to obtain the needle tip and trajectory the binary segmentation map is converted to 3D points in space by getting the center of the bounding box of the needle in each axial slice. The most distal point in the z-axis is considered the needle tip.

3.3.5 Implementation and Deployment

The proposed algorithm for needle localization and segmentation was implemented in Keras V2.0 [26] with Tensorflow back-end [1] and trained on an Nvidia GTX Titan X GPU with 12 GB of memory, hosted on a machine running Ubuntu 14.04 operating system on a 3.4 GHz Intel(R) Core(TM) i7-6800K CPU with 64 GB of memory. Training of large volumetric 3D networks was enabled and accelerated by the efficient cudNN[2] implementation of deep neural network layers. The trained models were deployed in the

[2] `https://developer.nvidia.com/cudnn`

Table 3.2: Needle tip localization error (mm) for test cases for proposed CNN method and the second observer*.

	$\overline{\Delta P}$	$\sigma_{\Delta P}$	$RMS(\Delta P)$	$M(\Delta P)$
CNN	2.80	3.59	4.54	0.88
Second observer	2.89	4.05	4.98	0.88

* $\overline{\Delta P}$, $\sigma_{\Delta P}$, $RMS(\Delta P)$ and $M(\Delta P)$ are the mean, standard deviation, root mean square, and median of the needle tip deviation, respectively. First row indicates the error of the CNN against ground truth, and second row indicates error of the second observer against ground truth.

open-source DeepInfer toolkit and are publicly available for download and use from the DeepInfer model repository[3].

3.4 Results

We tested the proposed method on a previously unseen test set of 173 MRI volumes from 21 patients. Figure 3.5 visually illustrates the localization of a single needle and the corresponding measured quality metrics in an example that is representative of the results of the proposed system. In the rest of this section, we present quantitative results of the performance of the proposed system against the ground truth, and also that of a second observer against the ground truth.

3.4.1 Tip Localization

The average tip localization errors for the proposed automatic system and the second observer relative to the ground truth are presented in Table 3.2. Corresponding box plots are presented in Figure 3.6. The median needle tip deviation for both the CNN and the second observer was 0.88 mm (1 pixel in the transaxial plane). Perfect matching of the predicted needle tip (0 mm deviation) was achieved for 32 (18% of test images) and 46 needles (27% of test images) for the CNN and second observer, respectively.

[3] `http://www.deepinfer.org/models/prostate-needle-finder/`

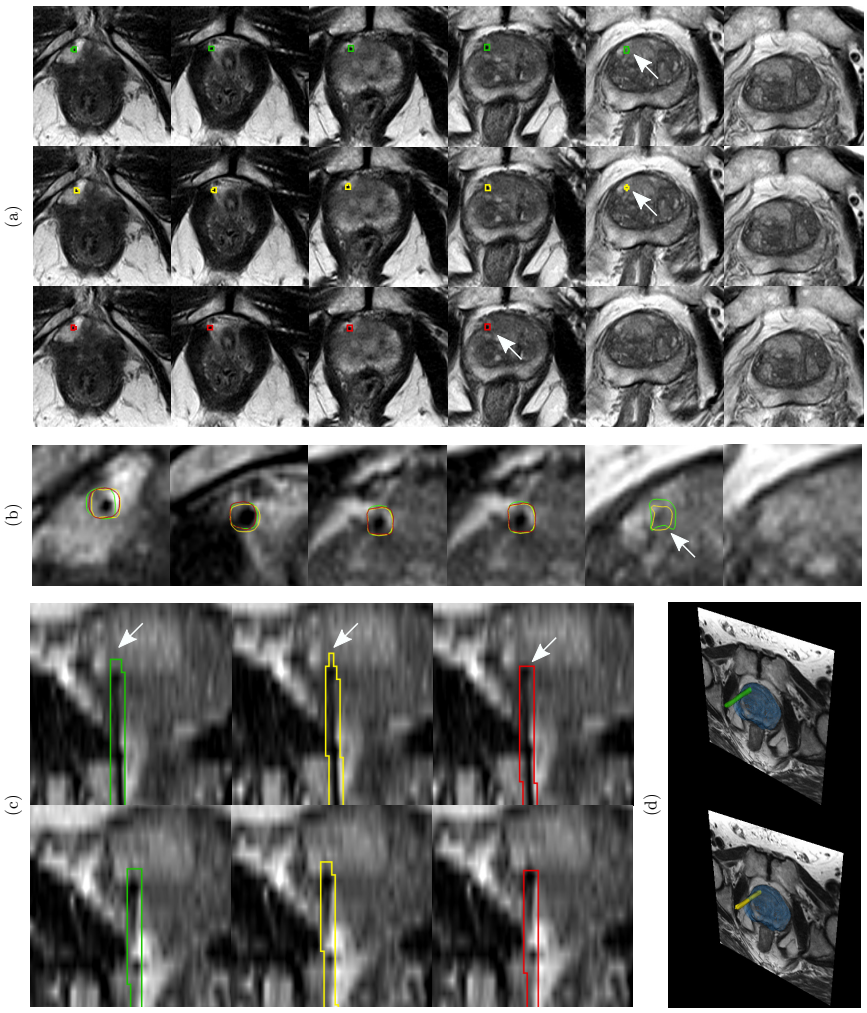


Figure 3.5: An example test case. Green, yellow, and red contours show the needle segmentation boundaries of the ground truth, the proposed system, and the second observer respectively. The arrows mark the needle tips. (a) First row shows ground truth. Second row shows predictions of the proposed system. Third row second observer annotations. (b) Zoomed view of slices in (a). (c) Coronal views. (d) 3D rendering of the needle relative to the prostate gland (blue), ground truth and CNN predictions. For the proposed CNN, the measured needle tip localization error (ΔP), tip axial plane detection error (ΔA), Hausdorff distance (HD), and angular deviation error ($\Delta\theta$) are 1.76 mm, 0 voxels, 1.24 mm, and 0.30° respectively.

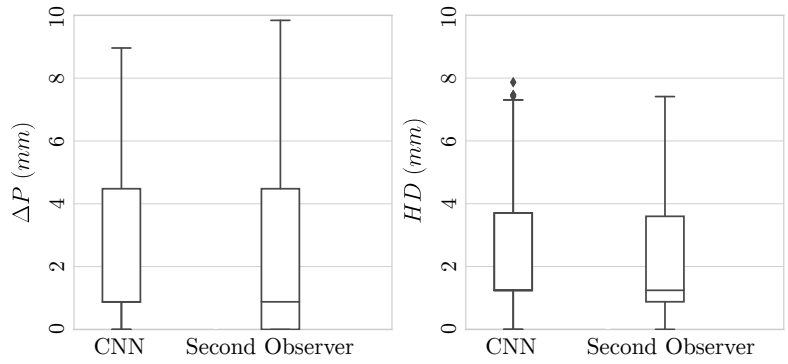


Figure 3.6: Box plots of the needle tip deviation error and Hausdorff distance (HD) in millimeters for the test cases. Distances of automatic (CNN) and second observer are shown which are comparable. The median tip localization error and the median HD distance for both CNN and second observer are 0.88 mm (1 pixel in transaxial plane) and 1.24 mm, respectively.

3.4.2 Tip Axial Plane Detection

The bar chart in Figure 3.7 summarizes the accuracy results for needle tip axial plane detection. For 113 images (65%) the algorithm detected the correct axial slice ($\Delta A = 0$) containing the needle tip which is comparable to the agreement between the two observers (108 cases (62%)). The algorithm missed the needle tip by one slice ($\Delta A = 1$) on 44 images (25%), by two slices ($\Delta A = 2$) for 9 images (5%) and by three or more slices ($\Delta A \geq 3$) for 7 images (4%). The bar chart shows that the performance of the CNN and the second observer are in the same range.

3.4.3 Trajectory Localization

Table 3.3 presents the results of needle trajectory localization error in terms of directional Hausdorff distance (HD) for the test cases for both the CNN and the second observer. Trajectory localization errors are summarized as the mean, standard deviation, root mean square, and median of the error. Corresponding box plots are presented in Figure 3.6.

3.4.4 Needle Direction

Table 3.4 presents the needle direction error in terms of angular deviation ($\Delta\theta$) for the test cases for both the CNN and the second observer. Needle

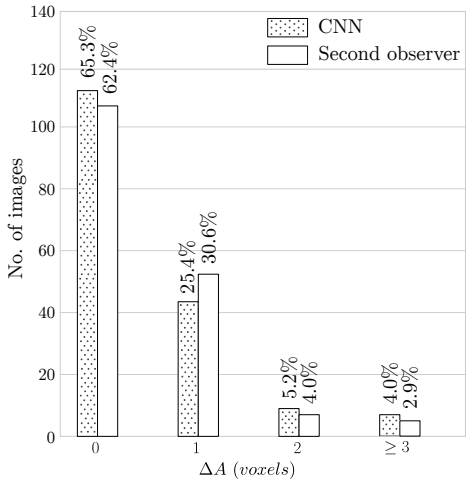


Figure 3.7: Bar charts of needle tip axial plane localization error (ΔA). Needle tip axial plane distance error of the automatic (CNN) method and second observer are shown. The results of the automatic CNN method are comparable with the second-observer.

angular deviation errors are summarized as the mean, standard deviation, root mean square, and median of the error.

3.4.5 Data Augmentation

Table 3.5 summarizes the impact of training-time and test-time augmentation on system performance as measured by the mean and standard deviation of the needle tip localization error in millimeters. The bottom right cell indicates best performance when both training-time and test-time augmentation are used. While both training and test-time augmentation did demonstrate smaller averages of needle tip deviation errors and fewer number of failures, we did not find improvements to be statistically significant.

3.4.6 Execution Time

The execution time of the proposed system was measured for inference on the test set of 173 volumes in the same environment that was described in Section 3.3.5. The average localization time using the proposed system was 29 seconds. This includes preprocessing, running five models on the original MRI volume and the flipped version, ensembling, and resampling back to

Table 3.3: Trajectory localization error averaged over test cases for proposed CNN method and the second observer* (units are in millimeters).

	$\overline{HD}$	σ_{HD}	$RMS(HD)$	$M(HD)$
CNN	3.00	3.15	4.35	1.24
Second observer	2.29	2.82	3.63	1.24

* $\overline{HD}$, σ_{HD}, $RMS(HD)$, and $M(HD)$ are the mean, standard deviation, root mean square, and median of the needle trajectory localization Hausdorff distance, respectively.

Table 3.4: Needle direction error quantified as the deviation angle averaged over test cases for proposed CNN method and the second observer* (units are in degrees).

	$\overline{\Delta\theta}$	$\sigma_{\Delta\theta}$	$RMS(\Delta\theta)$	$M(\Delta\theta)$
CNN	0.98	1.1	1.47	0.68
Second observer	0.97	1.04	1.43	0.75

* $\overline{\Delta\theta}$, $\sigma_{\Delta\theta}$, $RMS(\Delta\theta)$, and $M(\Delta\theta)$ are the mean, standard deviation, root mean square, and median of the needle deviation angle, respectively.

the original spatial resolution of the input image. In comparison, the second observer annotated a needle in 52 seconds on average.

3.5 Discussion

Automatic localization of needle tip and visualization of needle trajectories relative to the target can aid interventionalists in percutaneous needle placement procedures. Furthermore, accurate needle tip and trajectory localization is necessary for robot-guided needle placement. To the best of our knowledge, this is the first report of a fully automatic system for biopsy needle segmentation and localization in MRI with deep convolutional neural networks. A fairly large dataset of 583 MRI volumes from 71 patients suspected of prostate cancer was used to design, optimize, and test the proposed system. The system achieves human expert level performance for MRI-targeted prostate biopsy procedures. The results on an unseen test set show a mean accuracy error of 2.8 mm in detection of needle tip, 96% detection of axial tip plane within 2 slices, mean Hausdorff distance of 3 mm in needle trajectory, and a

Table 3.5: Impact of training-time and test-time augmentation on performance*.

	$\overline{\Delta P} \pm \sigma_{\Delta P}$	
	Train-Time Augmentation	
	Without	With
Without Test-Time Augmentation	4.92 ± 13.22	3.07 ± 3.70†
With Test-Time Augmentation.	3.93 ± 8.83	2.80 ± 3.59

* This table quantifies the system performance as measured by the mean and standard deviation of the needle tip localization error in millimeters.
† This model failed to segment one needle out of 173 in the test set. All other models did not miss any needles.

mean 0.98° error in needle trajectory angle, all of which lie within the range of agreement between human experts as shown by an observer study.

Our results support the findings of other studies in using 3D fully convolutional neural networks including 3D U-Net and its variants, for biomedical image segmentation to achieve promising results [107]. Additionally, the deployed trained model segments and localizes a needle in a 3D MRI volume in 29 seconds which makes it viable for adoption in the clinical workflow of MRI-targeted prostate biopsy procedures. The results of experiments to quantify the effect of data augmentation demonstrated smaller averages of needle tip deviation errors. However, unlike Ghafoorian et al. [44], we did not find the improvements to be statistically significant. Further analysis on larger test sets is required to statistically assess the effect of data augmentation for the needle segmentation problem. By preserving the ratio between in-plane resolution and slice thickness with anisotropic max-pooling and down sampling, we were able to train and deploy our model with whole 3D MRI volumes as inputs to the networks.

CNNs tend to be sensitive to the variations in MRI acquisition protocols. Variations in parameters during the acquisition of the MRI volumes result in different appearances of tissue and needle artifact [46]. Although we used a fairly large dataset of 583 MRI volumes in our experiments, and these MRI were acquired on two different MRI scanners, they were all obtained in a single institution using substantially similar MRI protocols. Therefore it is a reasonable conclusion that the performance of the trained models will degrade when applied to data acquired using substantially different MRI parameters. Domain adaptation and transfer learning techniques can be used to address this issue [46]. Moreover, due to the large slice thickness of 3.6

mm and partial volume effect, in many cases there is ambiguity in identifying the correct axial plane containing the needle tip. In this study we used the first observer as the gold standard and compared the second observer and the proposed method with it. Ideally, we would have had multiple observers and used majority voting for needle segmentation and tip localization.

3.6 Conclusion

We presented a new method for biopsy needle localization in MRI. A deep 3D fully convolutional neural network model was developed, trained and deployed using 583 T2-weighted MRI scans for 71 patients. The accuracy of the proposed method, as tested on previously unseen data, was 2.80 mm average in needle tip detection, and 0.98° in needle trajectory angle. We further designed an observer study in which independent annotations by a second observer, blinded to the original observer, were compared to the output of the proposed method. We showed that 3D convolutional neural networks, designed with some attention to domain knowledge, can effectively segment and localize needles from in-gantry MRI-targeted prostate biopsy images. The results of this study suggest that our proposed system can be used to detect and localize biopsy needles in MRI within the range of clinical acceptance and human-expert performance.

Chapter 4

Transfer Learning for Domain Adaptation in MRI

4.1 Introduction and Background

Deep neural networks have been extensively used in medical image analysis and have outperformed the conventional methods for specific tasks such as segmentation, classification, and detection [107]. For instance on brain MR analysis, convolutional neural networks (CNN) have been shown to achieve outstanding performance for various tasks including white matter hyperintensities (WMH) segmentation [45], tumor segmentation [77], microbleed detection [33], and lacune detection [44]. Although many studies report excellent results on specific domains and image acquisition protocols, the generalizability of these models on test data with different distributions is often not investigated and evaluated. Therefore, to ensure the usability of the trained models in real-world practice, which involves imaging data from various scanners and protocols, domain adaptation remains a valuable field of study. This becomes even more important when dealing with Magnetic Resonance Imaging (MRI), which demonstrates high variations in soft tissue appearances and contrasts among different protocols and settings.

Mathematically, a domain D can be expressed by a feature space χ and a marginal probability distribution $P(X)$, where $X = \{x_1, ..., x_n\} \in \chi$ [135]. A supervised learning task on a specific domain $D = \{\chi, P(X)\}$, consists of a pair of a label space Y and an objective predictive function $f(.)$ (denoted by $T = \{Y, f(.)\}$). The objective function $f(.)$ can be learned from the training data, which consists of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$. After the training process, the learned model denoted by $\tilde{f}(.)$ is used to predict the

This chapter is adapted from Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, Andriy Fedorov, Purang Abolmaesumi, Bram Platel, William M Wells. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer 2017.

label for a new instance x. Given a source domain D_S with a learning task T_S and a target domain D_T with learning task T_T, transfer learning is defined as the process of improving the learning of the target predictive function $f_T(.)$ in D_T using the information in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$ [135]. We denote $\tilde{f}_{ST}(.)$ as the predictive model initially trained on the source domain D_S, and domain-adapted to the target domain D_T.

In the medical image analysis literature, transfer classifiers such as adaptive SVM and transfer AdaBoost, are shown to outperform the common supervised learning approaches in segmenting brain MRI, trained only on a small set of target domain images [186]. In another study, a machine learning-based sample weighting strategy was shown to be capable of handling multi-center chronic obstructive pulmonary disease images [24]. Recently, also several studies have investigated transfer learning methodologies on deep neural networks applied to medical image analysis tasks. Some studies used networks pre-trained on natural images to extract features and followed by another classifier, such as a Support Vector Machine (SVM) or a random forest [37]. Other studies [163, 176] performed layer fine-tuning on the pre-trained networks for adapting the learned features to the target domain.

Considering the hierarchical feature learning fashion in CNN, we expect the first few layers to learn features for general simple visual building blocks, such as edges, corners, and simple blob-like structures, while the deeper layers learn more complicated abstract task-dependent features. In general, the ability to learn domain-dependent high-level representations is an advantage enabling CNNs to achieve great recognition capabilities. However, it is not obvious how these qualities are preserved during the transfer learning process for domain adaptation. For example, it would be practically important to determine how much data on the target domain is required for domain adaptation with sufficient accuracy for a given task, or how many layers from a model fitted on the source domain can be effectively transferred to the target domain. Or more interestingly, given a number of available samples on the target domain, what layer types and how many of those can we afford to fine-tune. Moreover, there is a common scenario in which a large set of annotated legacy data is available, often collected in a time-consuming and costly process. Upgrades in the scanners, acquisition protocols, etc., as we will show, might make the direct application of models trained on the legacy data unsuccessful. To what extent these legacy data can contribute to a better analysis of new datasets, or vice versa, is another question worth investigating.

In this chapter, we aim towards answering the questions discussed above. At the time of running experiments of this study, we did not have access

Table 4.1: Number of patients for the domain adaptation experiments.

	Source Domain			Target Domain		
Set	Train	Validation	Test	Train	Validation	Test
Size	200	30	50	100	26	33

to multi-domain prostate cancer datasets. Hence, we chose to perform experiments on brain WMH segmentation problem where we used transfer learning methodology for domain adaptation of models trained on legacy MRI data. Since there are no prior assumptions regarding the specific problem of WMH segmentation, we expect that the proposed method can be generalized to other medical imaging domain adaptation problems including prostate cancer diagnosis with MRI. However, confirmation of such expectations requires multi-domain datasets and further experimentation.

4.2 Materials and Method

4.2.1 Dataset

Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUN DMC) [185] is a longitudinal study of patients diagnosed with small vessel disease. The baseline scans acquired in 2006 consisted of fluid-attenuated inversion recovery (FLAIR) images with voxel size of 1.0×1.2×5.0 mm and an inter-slice gap of 1.0 mm, scanned with a 1.5 T Siemens scanner. However, the follow-up scans in 2011 were acquired differently with a voxel size of 1.0×1.2×3.0 mm, including a slice gap of 0.5 mm. The follow-up scans demonstrate a higher contrast as the partial volume effect is less of an issue due to thinner slices. For each subject, we also used 3D T1 magnetization-prepared rapid gradient-echo (MPRAGE) with voxel size of 1.0×1.0×1.0 mm which is the same among the two datasets. We should note that even though the two scanning protocols are only different on the FLAIR scans, it is generally accepted that the FLAIR is by far the most contributing modality for WMH segmentation. Reference WMH annotations on both datasets were provided semi-automatically, by manually editing segmentations provided by a WMH segmentation method [43] wherever needed.

The T1 images were linearly registered to FLAIR scans, followed by brain extraction and bias-filed correction operations. We then normalized the image intensities to be within the range of [0, 1].

In this study, we used 280 patient acquisitions with WMH annotations

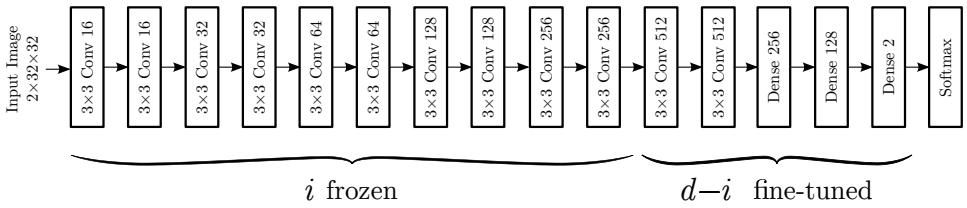


Figure 4.1: Architecture of the convolutional neural network used in our experiments. The shallowest i layers are frozen and the rest $d - i$ layers are fine-tuned. d is the depth of the network which was 15 in our experiments.

from the baseline as the source domain, and 159 scans from all the patients that were rescanned in the follow-up as the target domain. Table 4.1 shows the data split into the training, validation, and test sets. It should be noted that the same patient-level partitioning which was used on the baseline, was respected on the follow-up dataset to prevent potential label leakages.

4.2.2 Sampling

We sampled 32×32 patches to capture local neighborhoods around WMH and normal voxels from both FLAIR and T1 images. We assigned each patch with the label of the corresponding central voxel. To be more precise, we randomly selected 25% of all voxels within the WMH masks, and randomly selected the same number of negative samples from the normal appearing voxels inside the brain mask. We augmented the dataset by flipping the patches along the y axis. This procedure resulted in training and validation datasets of size ∼1.2m and ∼150k on the baseline, and ∼1.75m and ∼200k on the followup.

4.2.3 Network Architecture and Training

We stacked the FLAIR and T1 patches as the input channels and used a 15-layer architecture consisting of 12 convolutional layers of 3×3 filters and 3 dense layers of 256, 128, and 2 neurons, and a final softmax layer. We avoided using pooling layers as they would result in a shift-invariance property that is not desirable in segmentation tasks, where the spatial information of the features is important to be preserved. The network architecture is illustrated in Figure 4.1.

To tune the weights in the network, we used the Adam update rule [86] with a mini-batch size of 128 and a binary cross-entropy loss function. We

used the Rectified Linear Unit (ReLU) activation function as the non-linearity and the He method [59] that randomly initializes the weights drawn from a $\mathcal{N}(0, \sqrt{\frac{2}{m}})$ distribution, where m is the number of inputs to a neuron. Activations of all layers were batch-normalized to speed up the convergence [70]. A decaying learning rate was used with a starting value of 0.0001 for the optimization process. To avoid over-fitting, we regularized our networks with a drop-out rate of 0.3 as well as the L_2 weight decay with λ_2=0.0001. We trained our networks for a maximum of 100 epochs with an early stopping policy. For each experiment, we picked the model with the highest area under the curve on the validation set.

We trained our networks with a patch-based approach. At segmentation time, however, we converted the dense layers to their equivalent convolutional counterparts to form a fully convolutional network (FCN). FCNs are much more efficient as they avoid the repetitive computations on neighboring patches by feeding the whole image into the network. We prefer the conceptual distinction between dense and convolutional layers at the training time, to keep the generality of experiments for classification problems as well (e.g., testing the benefits of fine-tuning the convolutional layers in addition to the dense layers). Patch-based training allows class-specific data augmentation to handle domains with hugely imbalanced class ratios (e.g., WMH segmentation domain).

4.2.4 Domain Adaptation

To build the model $\tilde{f}_{ST}(.)$, we transferred the learned weights from $\tilde{f}_S$, then we froze the shallowest i layers and fine-tuned the remaining $d-i$ deeper layers with the training data from D_T, where d is the depth of the trained CNN. This is illustrated in Figure 4.1. We used the same optimization update-rule, loss function, and regularization techniques as described in Section 4.2.3.

4.2.5 Experiments

On the WMH segmentation domain, we investigated and compared three different scenarios: 1) Training a model on the source domain and directly applying it on the target domain; 2) Training networks on the target domain data from scratch; and 3) Transferring model learned on the source domain onto the target domain with fine-tuning. To identify the target domain dataset sizes where transfer learning is most useful, the second and third scenarios were explored with different training set sizes of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 25, 50 and 100 cases. We extensively expanded the third scenario

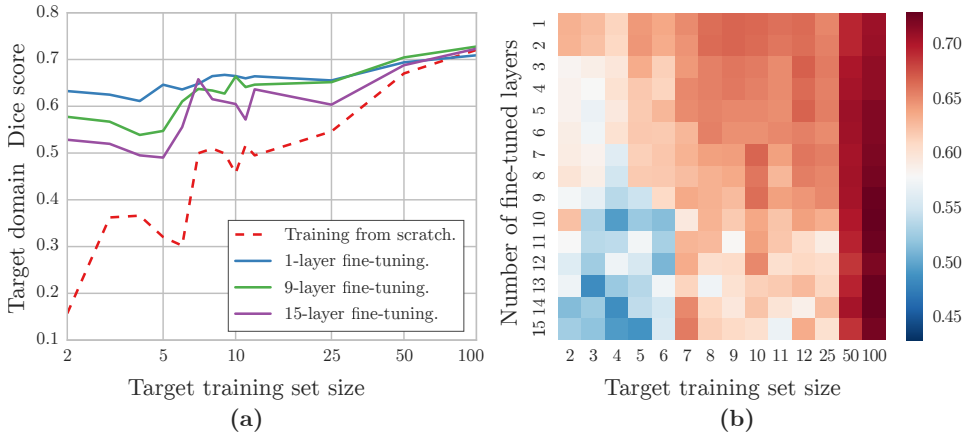


Figure 4.2: **(a)** The comparison of Dice scores on the target domain with and without transfer learning. A logarithmic scale is used on the x axis. **(b)** Given a deep CNN with d=15 layers, transfer learning was performed by freezing the i initial layers and fine-tuning the last $d - i$ layers. The Dice scores on the test set are illustrated with the color-coded heatmap. On the map, the number of fine-tuned layers are shown horizontally, whereas the target domain training set size is shown vertically.

investigating the best freezing/tuning cut-off for each of the mentioned target domain training set sizes. We used the same network architecture and training procedure among the different experiments. The reported metric for the segmentation quality assessment is the Dice score.

4.3 Results

The model trained on the set of images from the source domain ($\tilde{f}_S$), achieved a Dice score of 0.76. The same model, without fine-tuning, failed on the target domain with a Dice score of 0.005. Figure 4.2(a) demonstrates and compares the Dice scores obtained with three domain-adapted models to a network trained from scratch on different target training set sizes. Figure 4.2(b) illustrates the target domain test set Dice scores as a function of target domain training set size and the number of abstract layers that were fine-tuned. Figure 4.3 presents and compares qualitative results of WMH segmentation of several different models of a single sample slice.

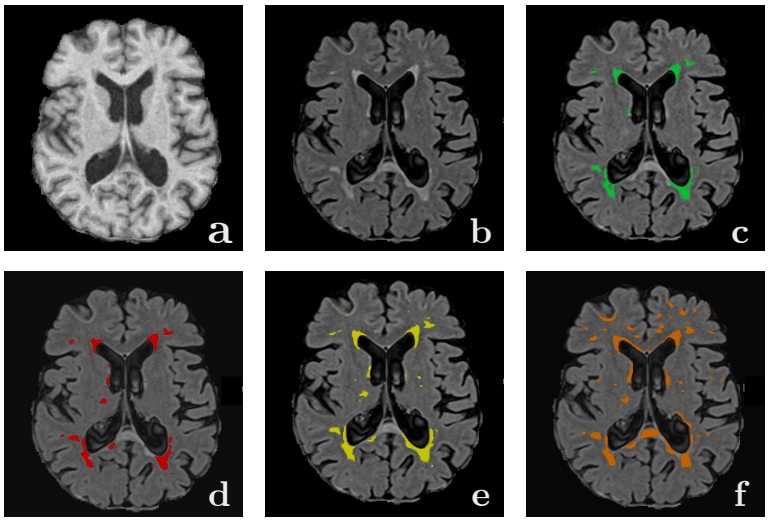


Figure 4.3: Examples of the brain WMH MRI segmentations. **(a)** Axial T1-weighted image. **(b)** FLAIR image. **(c-f)** FLAIR images with WMH segmented labels: **(c)** reference (green) WMH. **(d)** WMH (red) from a domain adapted model ($\tilde{f}_{ST}(.)$) fine-tuned on five target training samples. **(e)** WMH (yellow) from model trained from scratch ($\tilde{f}_{T}(.)$) on 100 target training samples. **(f)** WMH (orange) from model trained from scratch ($\tilde{f}_{T}(.)$) on 5 target training samples.

4.4 Discussion and Conclusion

We observed that while $\tilde{f}_S$ demonstrated a decent performance on D_S, it totally failed on D_T. Although the same set of learned representations is expected to be useful for both as the two tasks are similar, the failure comes to no surprise as the distribution of the responses to these features is different. Observing the comparisons presented by Figure 4.2(a), it turns out that given only a small set of training examples on D_T, the domain adapted model substantially outperforms the model trained from scratch with the same size of training data. For instance, given only two training images, $\tilde{f}_{ST}$ achieved a Dice score of 0.63 on a test set of 33 target domain test images, while $\tilde{f}_T$ resulted in a dice of 0.15. As Figure 4.2(b) suggests, with only a few D_T training cases available, the best results can be achieved by fine-tuning only the last dense layers, otherwise enormous number of parameters compared to the training sample size would result in over-fitting. As soon as more training data becomes available, it makes more sense to fine-tune

the shallower representations (e.g., the last convolutional layers). It is also interesting to note that tuning the first few convolutional layers is rarely useful considering their domain-independent characteristics. Even though we did not experiment with training-time fully convolutional networks such as U-Net [147], arguments can be made that the same conclusions would apply to such architectures.

Chapter 5

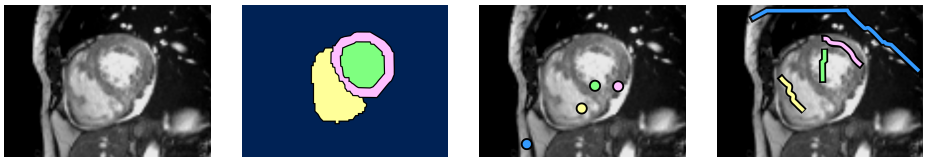
Weakly-supervised Medical Image Segmentation

5.1 Introduction and Background

Fully convolutional neural networks (FCNs), and in particular U-Nets [27, 147] have been increasingly used for semantic segmentation of both normal organs and lesions and achieved top ranking results in medical imaging challenges [92, 126]. The use of FCNs for image segmentation allows efficient training and learning of contextual features. FCNs are commonly trained using masks for which the ground truth is available for all of the pixels of the whole input image. Creating accurate pixel-level labels for medical images is time-consuming, expensive, and requires a high level of expertise. Annotation cost can be substantially reduced by using weaker supervision, i.e. by not needing full pixel-level labels.

Different weakly-supervised methods have been proposed for learning semantic segmentation with various forms of weak annotations. Unlabeled data can be at the level of images or pixels. In medical imaging research, much of the focus has been on studying image-level unlabeled data, i.e when training data consist of images with full annotations and images without any labels [9, 12, 23, 42, 130, 197, 198]. Other forms of weak supervision have also been studied, including response evaluation criteria in solid tumors (RECIST) [18], bounding boxes [143], and image-level tags [134]. Less studied are the cases with pixel-level unlabeled data, i.e. annotations in forms of partial labels [84, 199], scribbles [20, 73], or points [148]. Examples of full and partial annotations for cardiac MRI segmentation are shown in Figure 5.1.

A growing body of literature deals with the problem of training segmentation models with a mixture of labeled and unlabeled medical imaging data. Most methods propose hybrid loss functions with both supervised and semi-supervised components [9, 20, 23]. Different forms of pseudo-labels have been used as weak annotation for learning from unlabeled data. These methods



(a) (b) (c) (d)

Figure 5.1: Sample cardiac MRI image (a) with different forms of annotations (b-d); Yellow, purple, green, and blue colors correspond to the right ventricle, endocardium, left ventricle, and background, respectively. Fully supervised training of FCNs for semantic segmentation requires annotation of all pixels (b). The goal of this study is to develop weakly-supervised segmentation methods for training FCNs with a single point (c) and scribble (d). In this study, points refer to single-pixel marks for each class on each image slice. Scribbles have a width of one pixel. In this example, the sizes of points and scribbles are exaggerated for better visualization.

include initial coarse segmentation [18, 73], self-learning [9], and uncertainty-aware label propagation [130, 197]. Other methods use prior knowledge about anatomical structures [198, 199], adversarial training [130], conditional random field (CRF) post-processing [9, 20, 73, 143], and attention-based learning [23, 130].

Bai et al. [9] and Baur et al. [12] were among the first to leverage unlabeled data for improving FCN performance for medical image segmentation tasks. Baur et al. [12] proposed a semi-supervised method for multiple sclerosis lesion segmentation by adding an auxiliary manifold embedding loss to the supervised Dice loss. Using an iterative self-learning method Bai et al. [9] showed improvements in cardiac segmentation quality by including unlabeled MRI samples. Cai et al. [18] proposed a weakly supervised slice-propagated segmentation method for lymph node segmentation with RECIST annotations. Can et al. [20] presented a scribble-supervised learning framework which includes a region growing step for creating uncertainty maps for labels and an Expectation-Maximization approach for learning the network parameters. In another scribble-supervised segmentation work, Ji et al. [73] used partial cross-entropy (CE) [177] and dense CRF loss for the task of brain tumor segmentation. Kervadec et al. [84] proposed a novel loss function that includes constraints on sizes of structures and showed promising results for cardiac segmentation in MRI. Sedai et al. [158] and [197] successfully used uncertainty-aware pseudo-labels for semi-supervised segmentation of retinal layer and atrium, respectively. Zhen et al. proposed a semi-supervised adversarial

learning model with atlas priors for liver segmentation in CT scans [198]. Zhu et al. [199] proposed a prior aware loss function by regularizing the organ size distributions of the model output for an abdominal segmentation problem in CT scans.

In this chapter, we study the problem of weakly-supervised semantic segmentation with point and scribble supervision in FCNs. Specifically, we explore how far we can go with a single annotated point or a single annotated scribble per slice of a volumetric data set. We also propose partial Dice loss, a variant of Dice loss [124] for deep weakly-supervised segmentation with sparse pixel-level annotations. We furthermore compare partial Dice loss with partial CE [27, 159, 177] in terms of segmentation quality. Finally, we assess point and scribble-supervised segmentation on five different semantic segmentation tasks from medical images of the heart, the prostate, and the kidney. In a majority of these experiments, partial Dice loss provides statistically significant performance improvement over partial CE. The use of single point supervision results in 51%−95% of the performance of fully supervised training and the use of single scribble supervision achieves 86%−97% of the performance of fully supervised training.

5.2 Method

Semantic segmentation can be formulated as a pixel-level classification problem. In this setup, the pixels in the training image and label pairs can be considered as N i.i.d data points $\mathcal{D} = \{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^{N}$, where $\boldsymbol{x} \in \mathbb{R}^M$ is the M-dimensional input and each pixel y_i in $\boldsymbol{y} \in \mathbb{R}^M$, can be one and only one of the k possible classes, $k \in \{1, ..., K\}$. In weakly-supervised learning with partial annotations, the ground truth for labels is only available for a subset of pixels in $\boldsymbol{y}$. Here, we use FCNs for image segmentation, which allows for end-to-end learning, with each pixel of the input image being mapped by the FCN to the output segmentation map and it is more straightforward to implement segment-level loss functions such as Dice loss in this architecture. Neural networks can be formulated as parametric conditional probability models, $p(y_j = k|x_j, \theta)$, where the parameter set θ is chosen to minimize a loss function, and $p(\hat{y}_i = k|x_i, \theta)$ is the probability of pixel i belonging to class k. Subsequently, $p(y_i = k|x_i, \theta^*)$ is used for inference, where θ^* is the optimized parameter set.

The Dice coefficient was originally developed as a measure of similarity between two sets; it is twice the size of the intersection divided by the sum of the sizes of the two sets. The soft Dice loss function [124] is a generalized

measure where the probabilistic output of a segmenter is compared to the training data, set memberships are augmented with label probability, and a smoothing factor is added to the denominator to make the loss function differentiable.

With the Dice loss, the parameter set θ is chosen to minimize the negative of weighted Dice of different structures. Here, we propose the partial Dice loss, in which the parameter set is chosen to minimize the negative of Dice of different structures over only the pixels where the ground truth is known:

$$\mathcal{L}_{PDL} = -2\sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \left[m_i \cdot p(\hat{y}_i = k|x_i, \theta) \cdot (y_i = k)\right]}{\sum_{i=1}^{N} m_i \cdot \left[p(\hat{y}_i = k|x_i, \theta) + (y_i = k)\right] + \epsilon}, \tag{5.1}$$

where m_i is the the mask that is applied to each pixel. m_i is 1 for pixels where ground truth (partial annotation) is available and 0 for unlabeled pixels. $(y_i = k)$ is the binary indicator which denotes if the class label k is the correct class of the ith pixel, N is the number of pixels that are used in each mini-batch, and ϵ is the smoothing factor.

A similar masked Dice loss function was previously proposed as part of a semi-supervised training approach in ASDNet [130]. There, masking was used to filter out unconfident pseudo-label pixels which were generated through label propagation for images without any ground truth. Here, masking is used to limit the Dice loss calculation ($\mathcal{L}_{PDL}$ above) to only the labeled pixels in a weakly-supervised framework where labels (or ground truth annotations) are available only for small subsets of pixels via points and scribbles.

Similar to partial Dice loss, partial CE loss, $\mathcal{L}_{PCL}$, can be defined to calculate loss only over the sparse ground truth. Partial CE has been successfully used in segmentation of biomedical [27] and natural images [177]. With weighted partial CE loss, the parameter set θ is chosen to maximize the average log likelihood over the training pixels where ground truth is known:

$$\mathcal{L}_{PCL} = -\sum_{i=1}^{N} m_i \cdot \ln\left(p(\hat{y}_i = k|x_i, \theta)\right) \cdot (y_i = k)\,, \tag{5.2}$$

where m_i is the the mask that is applied to each pixel, $p(\hat{y}_i = k|x_i, \theta)$ is the probability of pixel belonging to class k, $(y_i = k)$ is the binary indicator which denotes if the class label k is the correct class of ith pixel, and N is the number of pixels that are used in each mini-batch.

5.3 Applications and Data

We performed experiments on five different semantic segmentation tasks from medical images of the heart, the kidney, and the prostate. These five segmentations include three structures in the heart - the left ventricle, the right ventricle, and the endocardium, along with the kidney, and the prostate gland, as described next. For heart segmentation, data from the MICCAI 2017 ACDC challenge for automated cardiac diagnosis were used [194]. This is a four-class segmentation task; cine MR images (CMRI) of patients are to be segmented into the left ventricle, the endocardium, the right ventricle, and the background. This dataset consists of end-diastole and end-systole images of 100 patients. We used only the end-diastole images in our study. For kidney segmentation, data from the MICCAI 2019 KiTS challenge for kidney tumor segmentation were used [61]. The training dataset consists of 210 arterial phase abdominal CT of kidney cancer patients. In this study, we only considered the problem of kidney segmentation and not tumors. Hence, we considered healthy and cancerous kidney tissue as the same class. For prostate segmentation, the public PROSTATEx dataset [106] together with 40 prostate gland full annotations from Meyer et al. [123] were used. This is a two-class segmentation task; Axial T2-weighted images of men suspected of having prostate cancer are to be segmented into the prostate gland and background. For all three segmentation tasks, the patients were split into training (40%), validation (10%), and testing (50%) sets. Prostate and cardiac images were resampled to common in-plane resolutions of 0.5×0.5 mm and 2×2 mm, respectively. Kidney images were resampled to the resolution of $1 \times 1 \times 1$ mm. All axial slices were then cropped at the center to create images of size 224×224 pixels as the input size of the FCN. Image grayscales were normalized to be within the range of [0,1].

5.4 Experimental Setup

5.4.1 Partial Annotation Generation

For all the training and validation data, partial annotations were generated automatically in form of single points and single scribbles per slice per class. Points were generated by randomly sampling a single pixel from each of the foreground classes and the background class for all of the 2D slices of each 3D volume. A single scribble was generated at each 2D slice by first randomly sampling for the start and endpoints of the scribble. Then the A* path search algorithm was used to construct a path between these points [16]. Scribbles

were generated for foreground and background classes only on 2D slices where there was a foreground class. Slices with no foreground were left unlabeled. Figure 5.1 shows examples of automatically generated points and scribbles and their corresponding full mask.

5.4.2 Training

For all experiments, we used a baseline FCN model similar to the two-dimensional U-Net architecture [147] but with fewer kernel filters at each layer. The input and output of the FCN have a size of 224×224 pixels. The network has the same number of layers as the original U-Net but with fewer kernels. The number of kernels for the encoder section of the network were 8, 8, 16, 16, 32, 32, 64, 64, 128, and 128. The parameters of the convolutional layers were initialized randomly from a Gaussian distribution [59]. For optimization, stochastic gradient descent with the Adam update rule [86] was used. During the training, we used a mini-batch of 16 examples. The initial learning rate was set to 0.005 and it was reduced by a factor of $0.5 - 0.8$ if the average validation Dice score did not improve by 0.001 in 10 epochs. We used 1000 epochs for training the models with an early stopping policy. For weakly-supervised training experiments, partial annotations were used for both training and validation labels. For each run, the model checkpoint was saved at the epoch where the validation loss was lowest. For each of the three segmentation problems, and for each type of partial ground truth, the model was trained 10 times with partial CE and 10 times with Dice loss, each with random weight initialization and random shuffling of the training data. Ensembling [121] was used to combine the test predictions. The same ensembling procedure was used for fully supervised training. All the deep models were implemented and optimized using the Keras framework [26].

5.4.3 Partial Loss Functions

CE loss and Dice loss are the two most commonly used loss functions in training FCNs for semantic segmentation. Dice loss [124] is robust to class imbalance and directly optimizes the model for semantic segmentation performance. CE indirectly improves segmentation through pixel-level classification and models trained with CE loss generally produce better-calibrated class probabilities [121]. Here, we compare the segmentation quality of models trained with partial annotations with partial CE loss with those trained with partial Dice Loss. We also compare weakly-supervised training with fully-supervised segmentation with Dice loss. We assess the segmentation

quality of the model with the Dice coefficient and 95^{th} percentile Hausdorff distance (H95).

5.5 Results

Partial annotations were generated as points and scribbles for all five segmentation tasks. For held-out test images, Dice and HD95 were calculated. Bootstrapping (n=1000) was performed and 95% confidence intervals (CIs) were calculated. P-values of less than 0.05 were regarded as statistically significant. Table 5.1 provides the proportions of partial annotations to full labels; it also compares the averages of Dice coefficients of foreground segments for single and ensemble models trained with partial Dice loss, partial CE loss, and models trained with full masks.

For cardiac segmentation with point supervision, Dice coefficients of endocardium and left ventricle were significantly better for models trained with partial Dice loss. Models trained with partial CE showed significantly better performance in the right ventricle. Point and scribble supervision achieved ranges of 71%−97% and 78%−97% of the performance of fully supervised annotation.

For prostate segmentation with point or scribble supervision, no statistically significant differences were found between models trained with either Dice loss or CE. Point and scribble supervision achieved 74% and 90% of the performance of fully supervised annotation.

For kidney segmentation, partial Dice loss was significantly better for both point and scribbles. Point and scribble supervision achieved ranges of 42%−51% and 84%−87% of the performance of fully supervised annotation.

Table 5.2 compares the averages of HD95 of foreground segments for single and ensemble models trained with partial Dice loss, partial CE loss, and models trained with full masks. For the endocardium, the prostate gland, and the kidney, models trained with partial Dice loss and point supervision showed significantly better segmentation in terms of HD95. For the right ventricle, models trained with partial Dice loss and scribble showed significantly better results. The differences between the two partial loss functions were not statistically significant for the rest of the segments. Figure 5.2 visually compares the ensemble models trained with point supervision and full masks through some representative examples over the five segmentation tasks.

Table 5.1: Segmentation quality of models in terms of the Dice coefficient (95% CI) of foreground structures: Weakly-supervised training with partial annotations (points and scribbles) is compared with fully supervised training. Models trained with partial CE loss (PCL) [27] are compared with those that were trained with the proposed partial Dice loss (PDL). Fractions of partial annotations to full labels are given (abbreviated to fr.). Boldface indicates statistically significant differences between model pairs (p-value<0.05).

	R. Ventricle	Endocardium	L. Ventricle	Prostate	Kidney
Point Supervision					
fr.	0.19%	0.27%	0.20%	0.02%	0.13%
PCL	**0.94 (0.87−0.96)**	0.64 (0.45−0.79)	0.69 (0.40−0.87)	0.71 (0.56−0.85)	0.38 (0.22−0.64)
PDL	0.92 (0.87−0.96)	**0.70 (0.49−0.82)**	**0.74 (0.44−0.87)**	0.71 (0.52−0.84)	**0.46 (0.31−0.63)**
Scribble Supervision					
fr.	2.93%	4.91%	2.56%	0.75%	2.07%
PCL	0.91 (0.89−0.95)	0.80 (0.70−0.89)	0.91 (0.80−0.95)	0.81 (0.69−0.88)	0.76 (0.60−0.91)
PDL	**0.94 (0.92−0.96)**	**0.83 (0.71−0.89)**	0.91 (0.73−0.96)	0.83 (0.72−0.90)	**0.78 (0.66−0.88)**
Full Supervision					
	0.97 (0.94−0.98)	0.90 (0.74−0.95)	0.95 (0.90−0.97)	0.96 (0.92−0.97)	0.90 (0.77−0.96)

Table 5.2: Segmentation quality of models in terms of 95^{th} Hausdorff distance (95% CI) of foreground structures. Models trained with partial cross-entropy (PCL) [27] are compared with those that were trained with the proposed partial Dice loss (PDL). Boldface indicates statistically significant differences between model pairs (p-value<0.05).

	R. Ventricle	Endocardium	L. Ventricle	Prostate	Kidney
Point Supervision					
PCL	6.8 (2.0−14.2)	15.4 (8.0−29.7)	29.3 (15.1−70.0)	17.4 (11.4−22.7)	35.7 (11.7−57.3)
PDL	5.9 (2.0−13.4)	**8.9 (4.0−15.2)**	25.0 (10.2−54.7)	**14.1 (8.0−25.5)**	**31.1 (11.0−53.8)**
Scribble Supervision					
PCL	4.9 (2.8−11.5)	4.1 (2.8−10.8)	7.4 (2.8−15.2)	11.9 (6.1−31.6)	29.1 (3.3−52.8)
PDL	**3.7 (2.0−10.3)**	3.6 (2.0−12.0)	8.3 (2.0−22.4)	13.1 (6.1−20.6)	27.3 (4.1−49.1)
Full Supervision					
	2.3 (2.0−10.2)	2.2 (2.0−2.8)	3.7 (2.0−14.3)	2.3 (1.6−3.4)	20.7 (1.4−47.2)

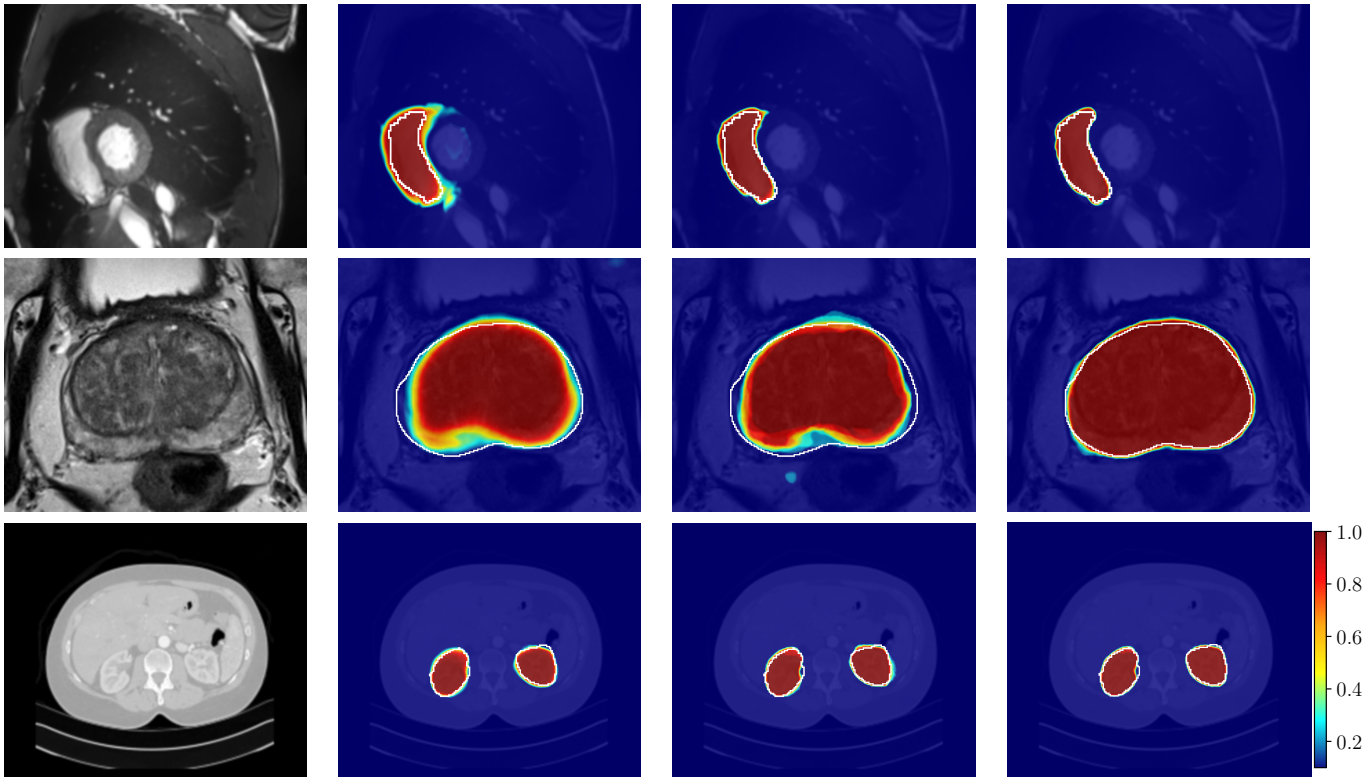


Figure 5.2: Examples of segmentation from scribble-supervised training of models with partial cross-entropy loss (CE), partial Dice loss (DSC), and models trained with full masks. The rows from top to bottom show the results for segmentation of the right ventricle, the prostate gland, and the kidney, respectively.

5.6 Discussion and Conclusion

Through extensive experiments, we have assessed weakly-supervised segmentation with partial annotations for medical image segmentation with FCNs. We proposed partial Dice loss for weakly-supervised segmentation with partial annotations, specifically points and scribbles. Moreover, we compared partial Dice loss with partial CE loss and compared them with fully supervised segmentation. We have performed these assessments using five segmentation tasks across three medical image domains tasks to ensure the generalizability of the findings. The results show that in a majority of the experiments, partial Dice provides statistically significant improvement in segmentation quality over partial CE in terms of Dice coefficient and 95^{th} percentile Hausdorff distance.

Further work needs to be carried out to include self-learning methodologies in the proposed weakly-supervised FCN framework. Such self-learning methods can be combined by uncertainty estimation methods [121] to produce confidence-aware pseudo-labels that can be used to further boost performance.

We conclude that partial annotations including points and scribbles are a promising direction for weakly-supervised segmentation using FCNs.

Chapter 6

Uncertainty Estimation for Image Segmentation

6.1 Introduction and Background

Fully convolutional neural networks (FCNs), and in particular the U-Net [147], have become a de facto standard for semantic segmentation in general and in medical image segmentation tasks in particular. The U-Net architecture has been used for segmentation of both normal organs and lesions and achieved top ranking results in several international segmentation challenges [76, 92, 126]. Despite numerous applications of U-Nets, very few works have studied the capability of these networks in capturing predictive uncertainty.

Predictive uncertainty or prediction confidence is described as the ability of a decision-making system to provide an expectation of success (i.e. correct classification) or failure for the test examples at inference time. Using a frequentist interpretation of uncertainty, predictions (i.e. class probabilities) of a *well-calibrated* model should match the probability of success of those inferences in the long run [54]. For instance, if a well-calibrated brain tumor segmentation model classifies 100 pixels each with the probability of 0.7 as cancer, we expect 70 of those pixels to be correctly classified as cancer. However, a poorly calibrated model with similar classification probabilities is expected to result in many more or less correctly classified pixels. Miscalibration frequently occurs in many modern neural networks (NNs) that are trained with advanced optimization methods [54]. Poorly-calibrated NNs are often highly confident in misclassification [5]. In some applications, for example, medical image analysis, or automated driving, overconfidence can be dangerous.

The soft Dice loss function [124], also known as Dice loss, is a generalized measure where the probabilistic output of a segmenter is compared to the

This chapter is adapted from Alireza Mehrtash, William M. Wells III, Clare M. Tempany, Purang Abolmaesumi, Tina Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2020.

training data, set memberships are augmented with label probability, and a smoothing factor is added to the denominator to make the loss function differentiable. With the Dice loss, the model parameter set is chosen to minimize the negative of weighted Dice of different structures. Dice loss is robust to class imbalance and has been successfully applied in many segmentation problems [173]. Furthermore, Batch Normalization (BN) effectively stabilizes convergence and also improves performance of networks for natural image classification tasks [70]. BN and Dice loss have made FCN optimization seamless. The addition of BN to the U-Net has improved optimization and segmentation quality [27]. However, it has been reported that both BN and Dice loss have adverse effects on calibration quality [15, 54, 153]. Consequently, FCNs trained with BN and Dice loss do not produce well-calibrated probabilities leading to poor uncertainty estimation. In contrast to Dice loss, cross-entropy loss provides better calibrated predictions and uncertainty estimates, as it is a strictly proper scoring rule [50]. Yet, the use of cross-entropy as the loss function for training FCNs can be challenging in situations where there is a high class imbalance, e.g., where most of an image is considered background [173]. Hence, it is of great significance and interest to study methods for confidence calibration of FCNs trained with BN and Dice loss.

Another important aspect of uncertainty estimation is the ability of a predictive model to distinguish *in-distribution* test examples (i.e. those similar to the training data) from *out-of-distribution* test examples (i.e. those that do not fit the distribution of the training data) [62]. The ability of the models to detect out-of-distribution inputs is specifically important for medical imaging applications as deep networks are sensitive to *domain shift*, which is a recurring situation in medical imaging [46]. For instance, networks trained on one MRI protocol often do not perform satisfactorily on images obtained with slightly different parameters or out-of-distribution test images. Hence, in the face of an out-of-distribution sample, an ideal model knows and announces *"I do not know"* and seeks human intervention – if possible – instead of a silent failure. Figure 6.1 shows an example of out-of-distribution detection from a U-Net model that was trained with BN and Dice loss for prostate gland segmentation before and after confidence calibration.

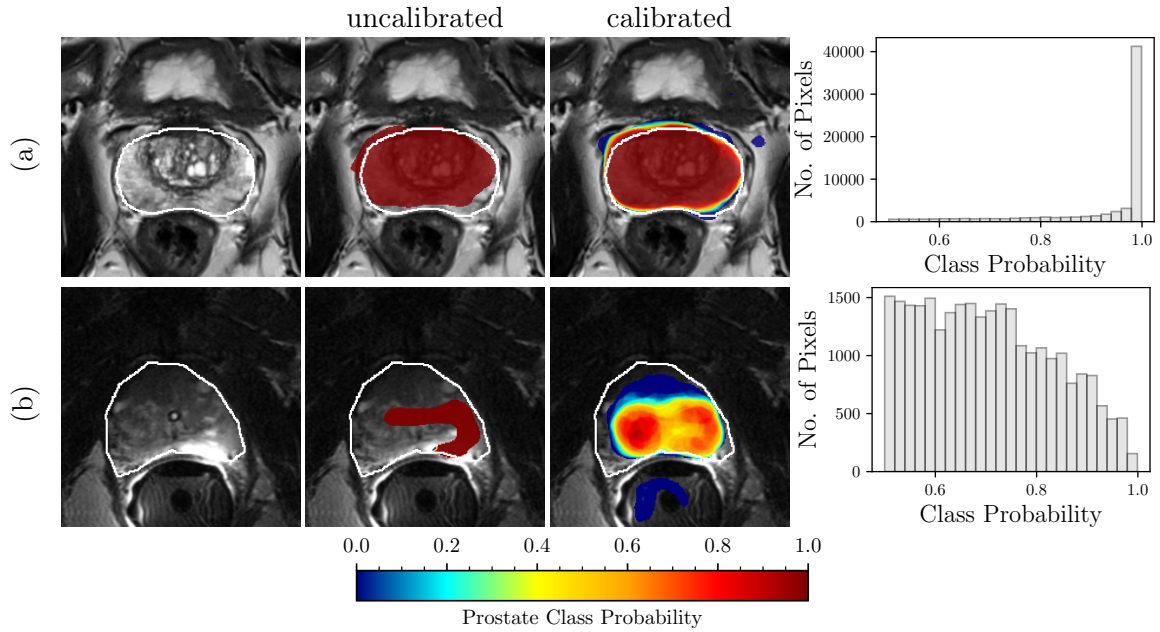


Figure 6.1: Calibration and out-of-distribution detection. Models for prostate gland segmentation were trained with T2-weighted MR images acquired using phased-array coils. The results of inference are shown for two test examples imaged with: (a) phased-array coil (in-distribution example), and (b) endorectal coil (out-of-distribution example). The first column shows T2-weighted MRI images with the prostate gland boundary drawn by an expert (white line). The second column shows the MRI overlaid with uncalibrated segmentation predictions of an FCN trained with Dice loss. The third column shows the calibrated segmentation predictions of an ensemble of FCNs trained with Dice loss. The fourth column shows the histogram of the calibrated class probabilities over the predicted prostate segment of the whole volume. Note that the bottom row has a much wider distribution compared to the top row, indicating that this is an out of distribution example. In the middle column, prediction prostate class probabilities ≤ 0.001 has been masked out.

6.2 Related Works

There has been a recent growing interest in uncertainty estimation and confidence measurement with deep NNs. Although most studies on uncertainty estimation have been done through Bayesian modeling of the NN, there has been some recent interest in using non-Bayesian approaches such as ensembling methods. Here, we first briefly review Bayesian and non-Bayesian methods and then review the recent literature for uncertainty estimation for semantic segmentation applications.

In the Bayesian approach, the deterministic parameters of the NN are replaced by prior probability distributions. Using Bayesian inference, given the data samples, a posterior probability distribution over the parameters is calculated. At inference time, instead of single scalar probability, the Bayesian NN gives probability distributions over the output label probabilities [115], which models NN predictive uncertainty. Gal and Ghahramani [41] proposed to use dropout [169] as a Bayesian approximation. They proposed *Monte Carlo dropout* (MC dropout) in which dropout layers are applied before every weight together with non-linearities. The probabilistic Gaussian process is approximated at inference time by running the model several times with active dropout layers. Implementing MC dropout is straightforward and has been applied in several application domains including medical imaging [102]. In a similar Bayesian approach, Teye et al. [180] showed that training NNs with BN [70] can be used to approximate inference of Bayesian NNs. For networks with BN and without dropout, *Monte Carlo Batch Normalization* (MCBN) can be considered an alternative to MC dropout. In another Bayesian work, Heo et al. [64] proposed a method that allows the attention model to leverage uncertainty. By learning the *Uncertainty-aware Attention* (UA) with variational inference, they improved both model calibration and performance in attention models. Seo et al. [160] proposed a variance-weighted loss function that enables learning single-shot calibration scores. In combination with stochastic depth and dropout, their method can improve confidence calibration and classification accuracy. Recently, Liao et al. [104] proposed a method for modeling such uncertainty in intra-observer variability of 2D echocardiography using the proposed cumulative density function probability method.

Non-Bayesian approaches have been proposed for probability calibration and uncertainty estimation. Guo et al. [54] studied the problem of confidence calibration in deep NNs. Through experiments, they analyzed different parameters such as depth, width, weight decay, and BN and their effect on calibration. They also used temperature scaling to easily calibrate trained

models. Ensembling has been used as an effective tool to improve classification performance of deep NNs in several applications including medical image segmentation [78, 118]. Following the success of ensembling methods [31] in improving baseline performance, Lakshminarayanan proposed *Deep Ensembles* in which model averaging was used to estimate predictive uncertainty [96]. By training collections of models with random initialization of parameters and adversarial training, they provided a simple approach to assess uncertainty. This observation motivated some of the experimental design in our work. Unlike MC dropout, using Deep Ensembles does not require network architecture modification. In [96] authors showed that Deep Ensembles outperforms MC dropout on two image classification problems. On the downside, Deep Ensembles requires retraining a model from scratch, which is computationally expensive for large datasets and complex models.

Predictive uncertainty estimation has been studied specifically for the problem of semantic segmentation with deep NNs. Bayesian SegNet [82] was among the first that addressed uncertainty estimation in FCNs by using MC dropout. They applied MC dropout by adding dropout layers after the pooling and upsampling blocks of the three innermost layers of the encoder and decoder sections of the SegNet architecture. Using similar approaches for uncertainty estimation, Kwon et al. [95] and Sedai et al. [157] used Bayesian NNs for uncertainty quantification in segmentation of ischemic stroke lesions and visualization of retinal layers, respectively. Sander et al. [153] applied MC dropout to capture instance segmentation uncertainty in ambiguous regions and compared different loss functions in terms of the resultant miscalibration. Kohl et al. [89] proposed a *Probabilistic U-Net* that combined an FCN with a conditional variance autoencoder to provide multiple segmentation hypotheses for ambiguous images. In similar work, Hu et al. [66] studied uncertainty quantification in the presence of multiple annotations as a result of inter-observer disagreement. They used a probabilistic U-Net to quantify uncertainty in the segmentation of lung abnormalities. Baumgartner et al. [11] presented a probabilistic hierarchical model where separate latent variable are used for different resolutions and variational autoencoder is used for inference. Rottmann and Schubert [149] proposed a prediction quality rating method for segmentation of nested multi-resolution street scene images by measuring both pixel-wise and segment-wise measures of uncertainty as predictive metrics for segmentation quality. Recently, Karimi et al. [79] used ensembling for uncertainty estimation of difficult to segment regions and used this information to improve clinical target volume estimation in prostate ultrasound images. In another recent work, Jungo and Reyes [75] studied uncertainty estimation for brain tumor and skin lesion segmentation tasks.

In conjunction with uncertainty estimation and confidence calibration, several works have studied out-of-distribution detection [30, 62, 100, 103, 162]. In a non-Bayesian approach, Hendrycks and Gimpel [62] used softmax prediction probability baseline to effectively predict misclassificaiton and out-of-distribution in test examples. Liang et al. [103] used temperature scaling and input perturbations to enhance the baseline method of Hendrycks and Gimpel [62]. In the context of a generative NN scheme, Lee et al.[100] used a loss function that encourages confidence calibration and this resulted in improvements in out-of-distribution detection. Similarly, DeVries and Taylor [30] proposed a hybrid with a confidence term to improve out-of-distribution detection. Shalev et al. [162] used multiple semantic dense representations of the target labels to detect misclassified and adversarial examples.

6.3 Contributions

In this chapter, we study predictive uncertainty estimation for semantic segmentation with FCNs and propose ensembling for confidence calibration and reliable predictive uncertainty estimation of segmented structures. In summary, we make the following contributions:

- We analyze the choice of loss function for semantic segmentation in FCNs. We compare the two most commonly used loss functions in training FCNs for semantic segmentation: cross-entropy loss and Dice loss. We train models with these loss functions and compare the resulting segmentation quality and predictive uncertainty estimation. We observe that FCNs trained with Dice loss perform significantly better segmentation compared to those trained with cross-entropy but at the cost of poor calibration.

- We propose model ensembling [96] for confidence calibration of FCNs trained with Dice loss and batch normalization. By training collections of FCNs with random initialization of parameters and random shuffling of training data, we create an ensemble that improves both segmentation quality and uncertainty estimation. We also compare ensembling with MC dropout [41, 82]. We empirically quantify the effect of the number of models on calibration and segmentation quality.

- We propose to use average entropy over the predicted segmented object as a metric to predict segmentation quality of foreground structures, which can be further used to detect out-of-distribution test inputs.

Table 6.1: Number of patients for training, validation, and test sets used in this study.

Application	Brain		Heart	Prostate	
Dataset	CBICA	TCIA	ACDC	PROSTATEx	PROMISE12†
# Training	66	–	40	16	–
# Validation	22	–	10	4	–
# Test	–	102	50	20	35

† Used only for out-of-distribution detection experiments.

Our results demonstrate that object segmentation quality correlates inversely with the average entropy over the segmented object and can be used effectively for detecting out-of-distribution inputs.

- We demonstrate our method for uncertainty estimation and confidence calibration on three different segmentation tasks from MRI images of the brain, the heart, and the prostate. Where appropriate, we report the statistical significance of our findings.

6.4 Applications & Data

Table 6.1 shows the number of patient images in each dataset and how we split these into training, validation, and test sets. In the following subsections, we briefly describe each segmentation task, data characteristics, and preprocessing.

6.4.1 Brain Tumor Segmentation Task

For brain tumor segmentation, data from the MICCAI 2017 BraTS challenge [10, 122] was used. This is a four-class segmentation task; multiparametric MRI of brain tumor patients are to be segmented into enhancing tumor, non-enhancing tumor, edema, and background. The training dataset consists of 190 multiparametric MRI (T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and FLAIR sequences) from brain tumor patients. The dataset is further subdivided into two sets: CBICA and TCIA. The images in CBICA set were acquired at the Center for Biomedical Image Computing and Analytics (CBICA) at the University of Pennsylvania [10]. The images in the TCIA set were acquired across multiple institutions and hosted by the National Cancer

Institute, The Cancer Imaging Archive (TCIA). The CBICA subset was used for training and validation and the TCIA subset was reserved as the test set.

6.4.2 Ventricular Segmentation Task

For heart ventricle segmentation, data from the MICCAI 2017 ACDC challenge for automated cardiac diagnosis was used [194]. This is a four-class segmentation task; cine MR images (CMRI) of patients are to be segmented into the left ventricle, the myocardium, the right ventricle, and the background. This dataset consists of end-diastole (ED) and end-systole (ES) images of 100 patients. We used only the ED images in our study.

6.4.3 Prostate Segmentation Task

For prostate segmentation, the public datasets, PROSTATEx [106] and PROMISE12 [108] were used. This is a two-class segmentation task; Axial T2-weighted images of men suspected of having prostate cancer are to be segmented into the prostate gland and the background. For PROSTATEx dataset, 40 images with annotations from Meyer et al. [123] were used. All these images were acquired at the same institution. PROSTATEx dataset was used for both training and testing purposes, and PROMISE12 dataset was set aside for test only. PROMISE12 dataset is a heterogeneous multi-institutional dataset acquired using different MR scanners and acquisition parameters. We used the 50 training images for which ground truth is available.

6.4.4 Data Pre-processing

Prostate and cardiac images were resampled to the common in-plane resolution of 0.5×0.5 mm and 2×2 mm, respectively. Brain images were resampled to the resolution of $1 \times 1 \times 2$ mm. All axial slices were then cropped at the center to create images of size 224×224 pixels as the input size of the FCN. Image intensities were normalized to be within the range of [0,1].

6.5 Methods

6.5.1 Model

Semantic segmentation can be formulated as a pixel-level classification problem, which can be solved by convolutional neural networks [107]. The pixels in the training image and label pairs can be considered as N i.i.d data points $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, where $\boldsymbol{x} \in \mathbb{R}^M$ is the M-dimensional input and y can

be one and only one of the k possible classes $k \in \{1, ..., K\}$. The use of FCNs for image segmentation allows for end-to-end learning, with each pixel of the input image being mapped by the FCN to the output segmentation map. Compared to FCNs, patch-based NNs are much slower at inference time as they require sliding window mechanisms for predicting each pixel [112]. Moreover, it is more straightforward to implement segment-level loss functions such as Dice loss in FCN architectures. FCNs for segmentation usually consist of an encoder (contracting) path and a decoder (expanding) path [112, 147]. FCNs with skip-connections are able to combine high level abstract features with low-level high-resolution features, which has been shown to be successful in segmentation tasks [27, 147]. NNs can be formulated as parametric conditional probability models, $p(y_j|x_j, \theta)$, and the parameter set θ is chosen to minimize a loss function. Both cross-entropy (CE) and negative of Dice Similarity Coefficient (DSC), known as Dice loss, have been used as loss functions for training FCNs. Class weights are used for optimization convergence and dealing with the class imbalance issue. With CE loss, parameter set is chosen to maximize the average log-likelihood over training data:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\omega_k \ln\left(p(\hat{y}_i = k|x_i, \theta)\right) \cdot (y_i = k)\,, \tag{6.1}$$

where $p(\hat{y}_i = k|x_i, \theta)$ is the probability of pixel i belonging to class k, $(y_i = k)$ is the binary indicator which denotes if the class label k is the correct class of ith pixel, ω_k is the weight for class k, and N is the number of pixels that are used in each mini-batch.

With the Dice loss, the parameter set is chosen to minimize the negative of weighted Dice of different structures:

$$\mathcal{L}_{DSC} = -2\sum_{k=1}^{K}\frac{\omega_k \sum_{i=1}^{N}\left[p(\hat{y}_i = k|x_i, \theta) \cdot (y_i = k)\right]}{\sum_{i=1}^{N}\left[p(\hat{y}_i = k|x_i, \theta) + (y_i = k)\right] + \epsilon}, \tag{6.2}$$

where $p(\hat{y}_i = k|x_i, \theta)$ is the probability of pixel belonging to class k, $(y_i = k)$ is the binary indicator which denotes if the class label k is the correct class of ith pixel, ω_k is the weight for class k, N is the number of pixels that are used in each mini-batch, and ϵ is the smoothing factor to make the loss function differentiable. Subsequently, $p(y_i|x_i, \theta^*)$ is used for inference, where θ^* is the optimized parameter set.

6.5.2 Calibration Metrics

The output of an FCN for each input pixel is a class prediction $\hat{y_j}$ and its associated class probability $p(y_j|x_j, \theta)$. The class probability can be considered the model confidence or probability of correctness and can be used as a measure for predictive uncertainty at the pixel level. Strictly proper scoring rules are used to assess the calibration quality of predictive models [50]. In general, scoring rules assess the quality of uncertainty estimation in models by awarding well-calibrated probabilistic forecasts. Negative log-likelihood (NLL), and Brier score [17], are both strictly proper scoring rules that have been previously used in several studies for evaluating predictive uncertainty [41, 54, 96]. In a segmentation problem, for a collection of N pixels, NLL is calculated as:

$$NLL = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\ln\left(p(\hat{y}_i = y_k|x_i, \theta)\right)\cdot(\hat{y}_i = y_k) \tag{6.3}$$

Brier score (Br) measures the accuracy of probabilistic predictions:

$$Br = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\left[p(\hat{y}_i = y_k|x_i, \theta) - (\hat{y}_i = y_k)\right]^2 \tag{6.4}$$

In addition to NLL and Brier score, we directly assess the predictive power of a model by analyzing test examples confidence values versus their measured expected accuracy values. To do so, we use reliability diagrams as visual representations of model calibration and Expected Calibration Error (ECE) as summary statistics for calibration [54, 127]. Reliability diagrams plot expected accuracy as a function of class probability (confidence). The reliability diagram of a perfectly calibrated model is the identity function. For expected accuracy measurement, the samples are binned into N groups and the accuracy and confidence for each group are computed. Assuming D_m to be indices of samples whose confidence predictions are in the range of $\left(\frac{m-1}{M}, \frac{m}{M}\right]$, the expected accuracy of the D_m is $Acc(D_m) = \frac{1}{|D_m|}\sum_{i\in D_m}\mathbf{1}(\hat{y}_i = y_i)$. The average confidence on bin D_m is calculated as $\overline{P}(D_m) = \frac{1}{|D_m|}\sum_{i\in D_m} p(\hat{y}_i = y_i|x_i, \theta)$. ECE is calculated by summing up the weighted average of the differences between accuracy and the average confidence over the bins:

$$\text{ECE} = \sum_{m=1}^{M}\frac{|D_m|}{N}\left|ACC(D_m) - \overline{P}(D_m)\right|, \tag{6.5}$$

where N is the total number of samples. In other words, ECE is the average of gaps on the reliability diagram.

6.5.3 Confidence Calibration with Ensembling

We propose to empirically determine whether ensembling [31] results in confidence calibration of otherwise poorly calibrated FCNs trained with Dice loss. To this end, similar to the Deep Ensembles method [96], we train M FCNs with random initialization of the network parameters and random shuffling of the training dataset in mini-batch stochastic gradient descent. However, unlike the Deep Ensemble method, we do not use any form of adversarial training. We train each of the M models in the ensemble from scratch and then compute the probability of the ensemble p_E as the average of the baseline probabilities as follows:

$$p_E(y_j = k|x_j) = \frac{1}{M}\sum_{m=1}^{M} p(y_j = k|x_j, \theta_m^*), \tag{6.6}$$

where $p(y_i = k|x_i, \theta_m^*)$ are the individual probabilities.

6.5.4 Segment-level Predictive Uncertainty Estimation

For segmentation applications, besides the pixel-level confidence metric, it is desirable to have a confidence metric that captures model uncertainty at the segment-level. Such a metric would be very useful in clinical applications for decision making. For a well-calibrated system, we anticipate that a segment-level confidence metric can predict the segmentation quality in the absence of ground truth. The metric can be used to detect out-of-distribution samples and hard or ambiguous cases. Such metrics have been previously proposed for street scene segmentation [149]. Given the pixel-level class predictions $\hat{y}_i$ and their associated ground truth class y_i for a predicted segment $\hat{\mathcal{S}}_k = \{s \in (x_i, \hat{y}_i)|\hat{y}_i = k\}$, we propose to use the average of pixel-wise entropy values over the predicted foreground [4] segment $\hat{\mathcal{S}}_k$ as a scalar metric for volume-level confidence of that segment as:

[4]Following the convention in the semantic segmentation literature, we are using foreground and background labels regardless of the fact that the problem is binary or k-class segmentation [112].

$$\begin{aligned}\overline{\mathcal{H}(\hat{\mathcal{S}}_k)} = -\frac{1}{\left|\hat{\mathcal{S}}_k\right|} \sum_{i \in \hat{\mathcal{S}}_k} [p(\hat{y}_i \\ = k|x_i, \theta) \cdot \ln\left(p(\hat{y}_i = k|x_i, \theta)\right) + (1 - p(\hat{y}_i \\ = k|x_i, \theta)) \cdot \ln\left(1 - p(\hat{y}_i = k|x_i, \theta)\right)].\end{aligned} \tag{6.7}$$

In calculating the average entropy of $\hat{\mathcal{S}}_k$, we assumed binary classification: the probability of belonging to class k, $p(\hat{y}_i = k|x_i, \theta)$ and the probability of belonging to other classes $1 - p(\hat{y}_i = k|x_i, \theta)$.

6.6 Experiments

6.6.1 Training Baselines

For all of the experiments, we used a baseline FCN model similar to the two-dimensional U-Net architecture [147] but with fewer kernel filters at each layer. The input and output of the FCN have a size of 224×224 pixels. Except for the brain tumor segmentation that used a three-channel input (T1CE, T2, FLAIR), for the rest of the problems the input was a single channel. The network has the same number of layers as the original U-Net but with fewer kernels. The number of kernels for the encoder section of U-Net were 8, 8, 16, 16, 32, 32, 64, 64, 128, and 128. The parameters of the convolutional layers were initialized randomly from a Gaussian distribution [59]. For each of the three segmentation problems, the model was trained 100 times with CE and 100 times with Dice loss, each with random weight initialization and random shuffling of the training data. For the models that were trained with Dice loss, the softmax activation function of the last layer was substituted with sigmoid function as it improved the convergence substantially. For CE loss, class weights ω_k, were calculated as inverse frequencies of the class labels for the combined pixels in training and validation data. For Dice loss, uniform class weights, ω_k, were used for all the foreground segments, except for the myocardium class in heart segmentation where the class weight was three times the other two foreground classes. For optimization, stochastic gradient descent with the Adam update rule [86] was used. During the training, we used a mini-batch of 16 examples for prostate segmentation and 32 examples for brain tumor and cardiac segmentation tasks. The initial learning rate was set to 0.005 and it was reduced by a factor of $0.5 - 0.8$ if the average validation Dice score did not improve by 0.001 in 10 epochs. We used 1000 epochs for training the models with an early stopping policy. For each run,

the model checkpoint was saved at the epoch where the validation DSC was the highest.

6.6.2 Cross-entropy vs. Dice

CE loss aims to minimize the average negative log-likelihood over the pixels, while Dice loss improves segmentation quality in terms of the Dice coefficient directly. As a result, we expect to observe models trained with CE to achieve a lower NLL and models trained with Dice loss to achieve better Dice coefficients. Here, our main focuses are to observe the segmentation quality of a model that is trained with CE in terms of Dice loss and the calibration quality of a model that was trained with Dice loss. We compare models trained with CE with those trained with Dice on three segmentation tasks.

6.6.3 MC dropout

MC dropout was implemented by modifying the baseline network as it was done in Bayesian SegNet [82]. Dropout layers were added to the three inner-most encoder and decoder layers with a dropout probability of 0.5. At inference time, Monte Carlo sampling was done with 50 samples and the mean of the samples was used as the final prediction.

6.6.4 Confidence Calibration

We used ensembling (Equation 6.6) to calibrate batch normalized FCNs trained with Dice loss. For the three segmentation problems, we made ensemble predictions and compared them with baselines in terms of calibration and segmentation quality. For calibration quality, we compared NLL, Brier score, and ECE%. For segmentation quality, we compared dice and 95^{th} percentile Hausdorff distance. Moreover, for calibration quality assessment we calculated the metrics on two sets of samples from the held-out test datasets: 1) the whole test dataset (all pixels of the test volumes) 2) pixels belonging to dilated bounding boxes around the foreground segments. The foreground segments and the adjacent background around them usually have the highest uncertainty and difficulty. At the same time, background pixels far from foreground segments show less uncertainty but outnumber the foreground pixels. Using bounding boxes removes most of the correct (certain) background predictions from the statistics and will lead to a better highlighting of the differences among models. For all three problems, we constructed bounding boxes of the foreground structures. The boxes are then

dilated by 8 mm in each direction of the in-plane axes and 2 slices (which translates to 4mm to 20mm) in each direction of the out-of-plane axis.

We also measured the effect of ensembles by calculating $p_E(y|x)$ (Equation 6.6) for ensembles with number of models (M) of 1, 2, 5, 10, 25, and 50. To provide better statistics and reduce the effect of chance in reporting the performance, for each ensemble, we sampled the 100 baseline models n times and reported the averages and 0.95 CI of the NLL and Dice. For example, for M=50, instead of reporting the means of NLL and Dice on a single set of 50 models (out of the 100 trained models), we sampled n sets of 50 models and reported the averages and 0.95 CI of the NLL and Dice. For prostate and heart segmentation tasks n was set to 50 and for brain tumor segmentation n was set to 10.

6.6.5 Segment-level Predictive Uncertainty

For each of the segmentation problems, we calculated volume-level confidence for each of the foreground labels and $\overline{\mathcal{H}(\hat{\mathcal{S}})}$ (Equation 6.7) vs. Dice. For prostate segmentation, we are also interested in observing the difference between the two datasets of PROSTATEx test set (which is the same as the source domain) and PROMISE-12 set (which can be considered as a target set).

Finally, in all the experiments, for statistical tests and calculating 95% confidence intervals (CI), we used bootstrapping (n=100). P-values of less than 0.01 were regarded as statistically significant. In all the presented tables, boldfaced text indicates the best results for each instance and shows that the differences are statistically significant.

6.7 Results

Table 6.2 compares the calibration quality and segmentation performance of baselines and ensembles (M=50) trained with CE loss with those that were trained with Dice loss and those that were calibrated with MC dropout. The averages and 95% CI values for NLL, Brier score, and ECE% for the bounding boxes around the segments are provided. Table 6.2 also compares the averages and 95% CI values of Dice coefficients of foreground segments for baselines trained with cross-entropy loss, Dice loss, and baselines calibrated with ensembling (M=50) for the whole volume. Calibration quality results for whole volumes and segmentation quality results in terms of Hausdorff distances are provided in Tables I and II of the Supplementary Material,

respectively. For all tasks across all segments, in terms of segmentation performance, baselines trained with Dice loss outperform those trained with CE loss and ensembles of models trained with Dice loss outperform all the other models. For all three segmentation tasks, calibration quality was significantly better in terms of NLL and ECE% for baseline (single) models trained with CE comparing to those that were trained with Dice loss. However, the direction of change for Brier score was not consistent among models trained with CE vs models trained with Dice loss. For bounding boxes of brain tumor and prostate segmentation, the Brier scores were significantly better for models trained with Dice loss compared to those trained with CE, while in the case of the heart segmentation was the opposite. The ensemble models show significantly better calibration qualities for all metrics across all tasks. In all cases ensembling outperformed baselines and MC dropout models in terms of calibration quality. We observe that ensembling improves the calibration quality of the models trained with Dice loss significantly. MC dropout consistently improves the calibration quality of the models trained with Dice loss. However, for models trained with CE loss, MC dropout only improves the calibration quality of prostate application models and not brain and heart applications.

The graphs in Figure 6.2 show the quantitative improvement in the calibration and segmentation as a function of the number of models in the ensemble, for each of the three segmentation applications of the prostate, the heart, and the brain tumors. As we see, for the prostate, the heart, and the brain tumor segmentation, using even five ensembles (M=5) of baselines trained with Dice loss can reduce the NLL by about 66%, 44%, and 62%, respectively. Qualitative examples for improvement as a function of number of models in ensemble are provided in the Supplementary Material Figures 5 and 6.

Figure 6.3 provides scatter plots of Dice coefficient vs. the proposed segment-level predictive uncertainty metric, $\overline{\mathcal{H}(\hat{\mathcal{S}})}$ (Equation 6.7), for models trained with Dice loss and calibrated with ensembling (M=50). For better visualization, Dice values were logit transformed $\text{logit}(p) = \ln(\frac{p}{1-p})$ as in [131]. In all three segmentation tasks, we observed a strong correlation ($0.77 \leq r \leq 0.92$) between logit of Dice coefficient and average of entropy over the predicted segment. For the prostate segmentation task, a clustering is obvious among the test set from the source domain (PROSTATEx dataset) and those from the target domain (PROMISE12). Investigation of individual cases reveals that most of the poorly segmented cases, which were predicted correctly using $\overline{\mathcal{H}(\hat{\mathcal{S}})}$, can be considered out-of-distribution examples as they

were imaged with endorectal coils.

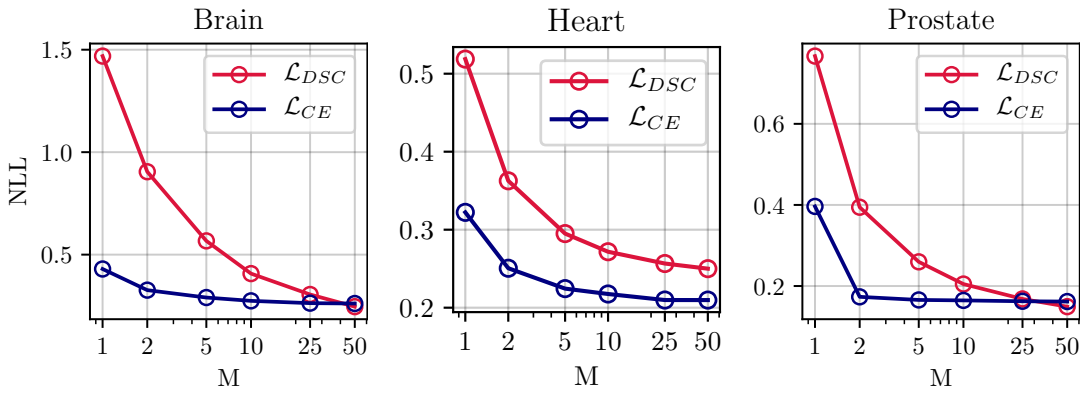


Figure 6.2: Improvements in calibration as a function of the number of models in the ensemble for baselines trained with cross-entropy and Dice loss functions. Calibration quality in terms of NLL improves as number of models M increases for prostate, heart, and brain tumor segmentation. For each task an ensemble of size M=10 trained with Dice loss outperforms the baseline model (M=1) trained with cross-entropy in terms of NLL. Same plot with 0.95 CIs and for both whole volume and bounding box measurements are given in Figure 4 of the Supplementary Material.

Table 6.2: Calibration quality and segmentation performance for baselines trained with cross-entropy ($\mathcal{L}_{CE}$) are compared with those that were trained with Dice loss ($\mathcal{L}_{DSC}$) and those that were calibrated with ensembling (M=50) and MC dropout. Boldfaced font indicates the best results for each application (model) and shows that the differences are statistically significant.

	Calibration Quality †			Segmentation Performance (Average Dice Score (95% CI)) ‡		
Application (Model)	NLL (95% CI)	Brier (95% CI)	ECE% (95% CI)	Segment I*	Segment II*	Segment III*
Brain ($\mathcal{L}_{CE}$)	0.52 (0.16−1.66)	0.23 (0.08−0.62)	8.11 (1.54−26.23)	0.37 (0.00−0.84)	0.47 (0.07−0.82)	0.58 (0.03−0.87)
Brain (MCD⋆ $\mathcal{L}_{CE}$)	0.81 (0.16−2.62)	0.36 (0.08−0.92)	13.41 (0.80−43.26)	0.34 (0.00−0.81)	0.34 (0.03−0.76)	0.54 (0.02−0.86)
Brain (EN $\mathcal{L}_{CE}$)	**0.29 (0.11−0.71)**	0.15 (0.05−0.40)	**3.28 (0.52−10.06)**	0.49 (0.00−0.92)	0.59 (0.11−0.86)	0.68 (0.04−0.91)
Brain ($\mathcal{L}_{DSC}$)	0.62 (0.17−2.70)	0.23 (0.06−0.55)	13.20 (2.60−33.55)	0.45 (0.00−0.89)	0.60 (0.10−0.90)	0.67 (0.07−0.91)
Brain (MCD $\mathcal{L}_{DSC}$)	1.14 (0.28−4.04)	0.18 (0.06−0.49)	8.96 (2.41−23.87)	0.43 (0.00−0.88)	0.58 (0.08−0.89)	0.64 (0.03−0.91)
Brain (EN $\mathcal{L}_{DSC}$)	0.31 (0.16−0.78)	**0.14 (0.08−0.35)**	3.71 (0.94−15.27)	**0.51 (0.00−0.93)**	**0.66 (0.11−0.91)**	**0.74 (0.16−0.92)**
Heart ($\mathcal{L}_{CE}$)	0.36 (0.16−1.18)	0.17 (0.09−0.41)	5.75 (1.42−17.99)	0.77 (0.17−0.91)	0.73 (0.45−0.86)	0.91 (0.63−0.97)
Heart (MCD $\mathcal{L}_{CE}$)	0.36 (0.17−1.10)	0.17 (0.09−0.41)	5.70 (1.39−17.93)	0.78 (0.27−0.90)	0.73 (0.47−0.86)	0.92 (0.64−0.97)
Heart (EN $\mathcal{L}_{CE}$)	**0.23 (0.13−0.58)**	**0.13 (0.07−0.30)**	**2.51 (0.58−10.15)**	0.81 (0.18−0.93)	0.77 (0.56−0.88)	**0.93 (0.79−0.97)**
Heart ($\mathcal{L}_{DSC}$)	0.62 (0.17−2.70)	0.23 (0.06−0.55)	13.20 (2.60−33.55)	0.84 (0.14−0.96)	0.81 (0.49−0.90)	0.92 (0.64−0.97)
Heart (MCD $\mathcal{L}_{DSC}$)	0.41 (0.17−1.51)	0.45 (0.11−0.81)	36.79 (6.17−70.58)	0.84 (0.12−0.96)	0.78 (0.04−0.89)	0.91 (0.61−0.97)
Heart (EN L_{DSC})	0.31 (0.16−0.78)	0.14 (0.08−0.35)	3.71 (0.94−15.27)	**0.87 (0.12−0.96)**	**0.83 (0.59−0.91)**	**0.93 (0.71−0.98)**
Prostate ($\mathcal{L}_{CE}$)	0.40 (0.22−0.79)	0.25 (0.13−0.47)	8.08 (1.60−25.50)	0.83 (0.62−0.91)	–	–
Prostate (MCD $\mathcal{L}_{CE}$)	0.30 (0.14−0.69)	0.16 (0.08−0.30)	5.23 (0.70−12.75)	0.77 (0.49−0.89)		
Prostate (EN $\mathcal{L}_{CE}$)	0.16 (0.13−0.25)	0.09 (0.06−0.16)	4.12 (1.92−7.04)	0.87 (0.68−0.92)	–	–
Prostate ($\mathcal{L}_{DSC}$)	0.74 (0.31−1.60)	0.11 (0.06−0.27)	5.72 (3.20−12.57)	0.88 (0.72−0.93)	–	–
Prostate (MCD $\mathcal{L}_{DSC}$)	0.48 (0.22−1.03)	0.11 (0.07−0.25)	5.23 (2.75−11.60)	0.86 (0.67−0.93)	–	–
Prostate (EN $\mathcal{L}_{DSC}$)	**0.15 (0.07−0.25)**	**0.07 (0.04−0.14)**	**2.02 (0.48−3.89)**	**0.90 (0.76−0.95)**	–	–

† The presented calibration quality metrics are calculated for bounding boxes. For whole volume results see Table I of the Supplementary Material.

‡ Comparison between Hausdorff distance of different models is provided in Table II of the Supplementary Material.

* For brain application segments, I, II, and III correspond to non-enhancing tumor, edema, and enhancing tumor, respectively. For heart application segments, I, II, and III correspond to the right ventricle, the myocardium, and the left ventricle, respectively. For prostate application segment I corresponds to the prostate gland.

⋆ MCD stands for Monte Carlo Dropout.

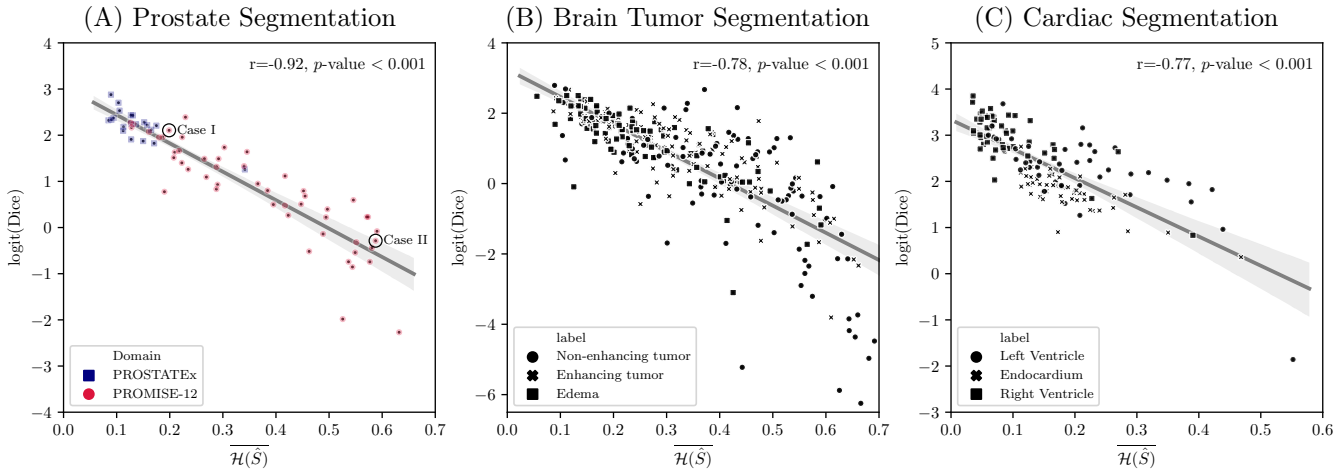


Figure 6.3: Segment-level predictive uncertainty estimation: Top row: Scatter plots and linear regression between Dice coefficient and average of entropy over the predicted segment $\overline{\mathcal{H}(\hat{\mathcal{S}})}$. For each of the regression plots, Pearson's correlation coefficient (r) and 2-tailed p-value for testing non-correlation are provided. Dice coefficients are logit transformed before plotting and regression analysis. For the majority of the cases in all three segmentation tasks, the average entropy correlates well with Dice coefficient, meaning that it can be used as a reliable metric for predicting the segmentation quality of the predictions at test-time. Higher entropy means less confidence in predictions and more inaccurate classifications leading to poorer Dice coefficients. However, in all three tasks there are few cases that can be considered outliers. (A) For prostate segmentation, samples are marked by their domain: PROSTATEx (source domain), and the multi-device multi-institutional PROMISE12 dataset (target domain). As expected, on average, the source domain performs much better than the target domain, meaning that average entropy can be used to flag out-of-distribution samples.

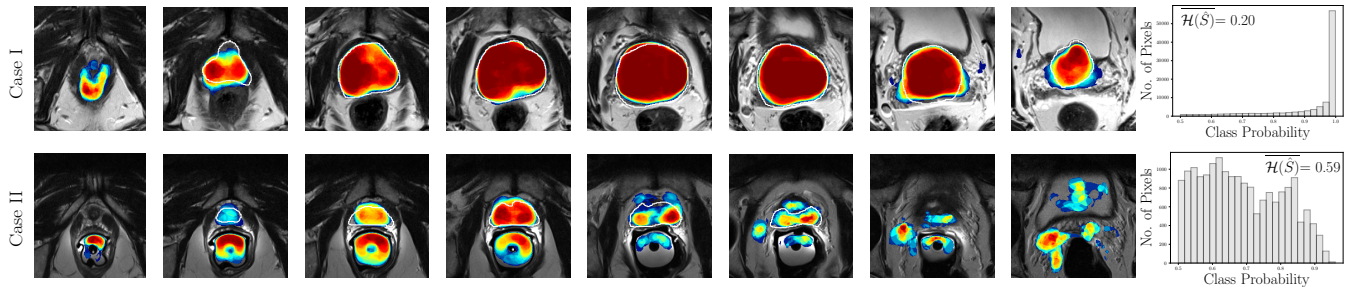


Figure 6.4: The two bottom rows correspond to two of the cases from the PROMISE12 dataset are marked in (A): Case I and Case II; These show the prostate T2-weighted MRI at different locations of the same patient with overlaid calibrated class probabilities (confidences) and histograms depicting distribution of probabilities over the segmented regions. The white boundary overlay on prostate denotes the ground truth. The wider probability distribution in Case II associates with a higher average entropy which correlates with a lower Dice score. Case-I was imaged with phased-array coil (same as the images that was used for training the models), while Case II was imaged with endorectal coil (out-of-distribution case in terms of imaging parameters). The samples in scatter plots in (B) and (C) are marked by their associated foreground segments. The color bar for the class probability values is given in Figure 6.1. Qualitative examples for brain and heart applications and scatter plots for models trained with cross-entropy are given in Figures 7 and 8 of the Supplementary Material, respectively.

6.8 Discussion

Through extensive experiments, we have rigorously assessed uncertainty estimation for medical image segmentation with FCNs. Furthermore, we proposed ensembling for confidence calibration of FCNs trained with Dice loss. We have performed these assessments using three common medical image segmentation tasks to ensure the generalizability of the findings. The results consistently show that for baseline (single) models, cross-entropy loss is better than Dice loss in terms of uncertainty estimation in terms of NLL and ECE%, but falls short in segmentation quality. We then showed that ensembling with $M \geq 5$ notably calibrates the confidence of models trained with Dice loss and CE loss. Importantly, we also observed that in addition to NLL reduction, the segmentation accuracy in terms of the Dice coefficient and Hausdorff distance was also improved through ensembling. We also showed that ensembling outperforms MC dropout in estimating the uncertainty of deep image segmenters. This confirms previous findings in the image classification literature [96]. Consistent with the results of previous studies [92], we observed that segmentation quality improved with ensembling. The results of our experiments for comparing cross-entropy with Dice loss are in line with the achieved results of Sanders et al. [153].

Importantly, we demonstrated the feasibility of constructing metrics that can capture predictive uncertainty of individual segments. We showed that the average entropy of segments can predict the quality of the segmentation in terms of Dice coefficient. Preliminary results suggest that calibrated FCNs have the potential to detect out-of-distribution samples. Specifically, for prostate segmentation, the ensemble correctly predicted the cases where it failed due to differences in imaging parameters (such as different imaging coils). However, it should be noted that this is an early attempt to capture the segment-level quality of segmentation and the results thus need to be interpreted with caution. The proposed metric can be improved by adding prior knowledge about the labels. Furthermore, it should be noted that the proposed metric does not encompass any information on number of samples used in that estimation.

As introduced in the methods section, some loss functions are "proper scoring rules", a desirable quality that promotes well-calibrated probabilistic predictions. The Deep Ensembles method has a proper scoring rule requirement for the baseline loss function [96]. The question arises: *"Is the Dice loss a proper scoring rule?"* Here, we argue that there is a fundamental mismatch in the potential usage of the Dice loss for scoring rules. Scoring rules are functions that compare a probabilistic prediction with an outcome. In the

context of binary segmentation, an outcome corresponds to a binary vector of length n, where n is the number of pixels. The difficulty with using scoring rules here is that the corresponding probabilistic prediction is a distribution on binary vectors. However, the predictions made by deep segmenters are collections of n label probabilities. This is in contrast to distributions on binary vectors, which are more complex and in general characterized by probability mass functions with 2^n parameters, one for each of the 2^n possible outcomes (the number of possible binary segmentations). The essential problem is that deep segmenters do not predict distributions on outcomes (binary vectors). One potential workaround is to say that the network does predict the required distributions, by constructing them as the product of the marginal distributions. This, though, has the problem that the predicted distributions will not be similar to the more general data distributions, so in that sense, they are bound to be poor predictions.

We used segmentation tasks in the brain, the heart and the prostate to assess uncertainty estimation. Although each of these tasks was performed on MRI images, there were subtle differences between them. The brain segmentation task was performed on three-channel input (T1 contrast-enhanced, FLAIR, and T2) while the other two were performed on single-channel input (T2 for prostate and Cine images for the heart). Moreover, the number of training samples, the size of the target segments, and the homogeneity of samples were different in each task. Only publicly available datasets were used in this study to allow others to easily reproduce these experiments and results. The ground truth was created by experts and independent test sets were used for all experiments. For prostate gland segmentation and brain tumor segmentation tasks, we used multi-scanner, multi-institution test sets. For all three tasks, the boundaries of the target segments were commonly identified as areas of high uncertainty estimate. Compared to the prostate and heart applications, we observed lower segmentation quality in the brain tumor application. Segmentation of lesions (in this case brain tumors) is generally a harder problem compared to the segmentation of organs (in this case the heart, and the prostate gland). This is partly because lesions are more heterogeneous. This is partly due to the fact that lesions are more heterogeneous. However, as shown in Figure 6.3 the calibrated models successfully predicted the segmentation quality and total failures (where the model failed to predict any meaningful structure – e.g. Dice score ≤ 0.05.

Our focus was not on achieving state-of-the-art results on the three mentioned segmentation tasks, but on using these to understand and improve the uncertainty prediction capabilities of FCNs. Since we performed several rounds of training with different loss functions, we limited the number of

parameters in the models to speed up each training round; we carried out experiments with 2D CNNs (not 3D), used fewer convolutional filters in our baseline compared to the original U-Net, and performed limited (not exhaustive) hyperparameter tuning to allow reasonable convergence. 2D U-Nets have been used extensively to segment 3D images and we used these to conduct the experiments reported above. 2D vs 3D is one of the many design choices or hyper-parameters of constructing deep networks for semantic segmentation, without a clear-cut answer that 2D U-Nets are always better for 2D images and 3D U-Nets are always better for 3D images. In fact, in some applications, 2D networks have outperformed 3D networks [92]. However, in the case of confidence calibration using deep ensembles, preliminary experiments (that we have included in Appendix F of the Supplementary Material) indicate no difference between using 3D U-Nets or 2D U-Nets. A comprehensive empirical study on this topic would be quite interesting.

In this chapter, we compared calibration qualities of two commonly used loss functions and showed that loss function directly affects calibration quality and segmentation performance. As stated earlier, calibration quality is an important metric that provides information about the quality of the predictions. We think it is important for users of deep networks to be aware of the calibration qualities associated with different loss functions, and to that end, we think that it would be interesting to investigate the calibration and segmentation quality of other commonly used loss functions such as combinations of Dice loss and cross-entropy loss, as well as the recently proposed Lovász-Softmax loss [14] that we think is promising for medical image segmentation.

For the proposed segment-level predictive uncertainty measure (Equation 6.7), we assumed binary classification and entropy of the foreground class was calculated by considering every other class as background. However, there are neighborhood relationships between classes and adjacent pixels that could be further integrated using measures such as multi-class entropy or similar strategies such as the Wasserstein losses [39].

There remains a need to study calibration methods that, unlike ensembling, do not require training from scratch which is time-consuming. In this study, we only investigated uncertainty estimation for MR images. Although parameter changes occur more often in MRI comparing to computed tomography (CT), it would still be very interesting to study uncertainty estimation in CT images. Parameter changes in CT can also be a source of failure in CNNs. For instance, changes in slice thickness or use of contrast can result in failures in predictions and it is highly desirable to predict such failures through model confidence.

We believe that our research will serve as a base for future studies on uncertainty estimation and confidence calibration for medical image segmentation. Further study of the sources of uncertainty in medical image segmentation is needed. Uncertainty has been classified as aleatoric or epistemic in medical applications [69] and Bayesian modeling [83]. Aleatoric refers to types of uncertainties that exist due to noise or the stochastic behavior of a system. In contrast, epistemic uncertainties are rooted in limitation in knowledge about the model or the data. In this study, we consistently observed higher levels of uncertainty at specific locations such as boundaries. For example in the prostate segmentation task, single and multiple raters often have higher inter- and intra-disagreements in the delineation of the base and apex of the prostate rather than at the mid-gland boundaries [108]. Such disagreements can leave their traces on models that are trained using ground truth labels with such discrepancies. It has been shown that with enough training data from multiple raters, deep models are able to surpass human agreements on segmentation tasks [107]. However, few works have been done on the correlation of ground truth quality and model uncertainty that results from rater disagreements [172, 178].

6.9 Conclusion

We conclude that model ensembling can be used successfully for confidence calibration of FCNs trained with Dice Loss. Also, the proposed average entropy metric can be used as an effective predictive metric for estimating the performance of the model at test-time when the ground-truth is unknown.

Chapter 7

PEP: Parameter Ensembling by Perturbation

7.1 Introduction and Background

Deep neural networks have achieved remarkable success on many classification and regression tasks [98]. In the usual usage, the parameters of a conditional probability model are optimized by maximum likelihood on large amounts of training data [51]. Subsequently the model, in combination with the optimal parameters, is used for inference. Unfortunately, this approach ignores uncertainty in the value of the estimated parameters; as a consequence over-fitting may occur and the results of inference may be overly confident. In some domains, for example medical applications, or automated driving, overconfidence can be dangerous [5].

Probabilistic predictions can be characterized by their level of *calibration*, an empirical measure of consistency with outcomes, and work by Guo et al. shows that modern neural networks (NN) are often poorly calibrated, and that a simple one-parameter *temperature scaling* method can improve their calibration level [54]. Explicitly Bayesian approaches such as *Monte Carlo Dropout* (MCD) [41] have been developed that can improve likelihoods or calibration. MCD approximates a Gaussian process at inference time by running the model several times with active dropout layers. Similar to the MCD method [41], Teye et al. [180] showed that training NNs with batch normalization (BN) [70] can be used to approximate inference with Bayesian NNs. Directly related to the problem of uncertainty estimation, several works have studied out-of-distribution detection. Hendrycks and Gimpel [62] used softmax prediction probability baseline to effectively predict misclassification and out-of-distribution in test examples. Liang et al. [103] used temperature scaling and input perturbations to enhance the baseline

This chapter is is adapted from: Alireza Mehrtash, Purang Abolmaesumi, Polina Goland, Tina Kapur, Demian Wassermann, William M. Wells III. PEP: Parameter Ensembling by Perturbation. NeurIPS 2020.

method of Hendrycks and Gimpel [62]. In a recent work, Rohekar et al. [146] proposed a method for confounding training in deep NNs by sharing neural connectivity between generative and discriminative components. They showed that using their BRAINet architecture, which is a hierarchy of deep neural connections, can improve uncertainty estimation. Hendrycks et al. [63] showed using pre-training can improve uncertainty estimation. Thulasidasan et al. [181] showed that mixed up training can improve calibration and predictive uncertainty of models. Corbière et al. [28] proposed *True Class Probability* as an alternative for classic Maximum Class Probability. They showed that learning the proposed criterion can improve model confidence and failure prediction. Raghu et al. [142] proposed a method for direct uncertainty prediction that can be used for medical second opinions. They showed that deep NNs can be trained to predict uncertainty scores of data instances with high human reader disagreement.

Ensemble methods [31] are regarded as a straightforward way to increase the performance of base networks and have been used by the top performers in imaging challenges such as ILSVRC [175]. The approach typically prepares an ensemble of parameter values that are used at inference-time to make multiple predictions, using the same base network. Different methods for ensembling have been proposed for improving model performance, such as M-heads [101] and Snapshot Ensembles [67]. Following the success of ensembling methods in improving baseline performance, Lakshminarayanan et al. proposed *Deep Ensembles* in which model averaging is used to estimate predictive uncertainty [96]. By training collections of models with random initialization of parameters and adversarial training, they provided a simple approach to assess uncertainty.

Deep Ensembles and MCD have both been successfully used in several applications for uncertainty estimation and calibration improvement. However, Deep Ensembles requires retraining a model from scratch for several rounds, which is computationally expensive for large datasets and complex models. Moreover, Deep Ensembles cannot be used to calibrate pre-trained networks for which the training data is not available. MCD requires network architecture to have dropout layers. In many modern networks, BN removes the need for dropout [70]. Hence, there is a need for network modification if the original architecture does not have dropout layers. It is also challenging or not feasible in some cases to use MCD on out-of-the-box pre-trained networks. Parameter (weight) perturbation at training time has been used to good effect in variational Bayesian deep learning [85] and to improve adversarial robustness [72].

In this chapter, we propose Parameter Ensembling by Perturbation (PEP),

a simple ensembling approach that uses random perturbations of the optimal parameters from a single training run. Unlike MCD which needs dropout at training, PEP can be applied to any pre-trained network without restrictions on the use of dropout layers. Unlike Deep Ensembles, PEP needs only one training run. PEP can provide improved log-likelihood and calibration for classification problems, without the need for specialized or additional training, substantially reducing the computational expense of ensembling. We show empirically that the log-likelihood of the ensemble average ($\mathbb{L}$) on hold-out validation and test data grows initially from that of the baseline model to a well-defined peak as the spread of the parameter ensemble increases. We also show that PEP may be used to probe curvature properties of the likelihood landscape. We conduct experiments on deep and large networks that have been trained on ImageNet (ILSVRC2012) [150] to assess the utility of PEP for improvements on calibration and log-likelihoods. The results show that PEP can be used for probability calibration on pre-trained networks such as DenseNet [68], Inception [175], ResNet [60], and VGG [166]. Improvements in the log-likelihood range from small to significant but they are almost always observed in our experiments. To compare PEP with MCD and Deep Ensembles, we run experiments on classification benchmarks such as MNIST and CIFAR-10 which are small enough for us to re-train and add dropout layers. Finally, We perform further experiments to study the relationship between over-fitting and the "PEP effect," (the gain in log-likelihood over the baseline model) where we observe larger PEP effects for models with higher levels of over-fitting.

In this chapter, we limit our experiments to computer vision benchmarks such as MNIST, CIFAR-10, and ImageNet. The proposed PEP method and the theoretical developments apply to deep NNs in general. Here, we do not run any experiments on medical images. However, we expect that the proposed method can be generalized and adopted well for medical imaging applications in general and prostate cancer diagnosis in MRI in particular. As we showed in Chapter 6, deep NNs trained on medical images are often poorly calibrated. PEP provides an affordable method to calibrate such models without the additional cost of training. Importantly, PEP does not require access to training data. This could facilitate the calibration of models trained for medical applications where security and privacy are top priorities.

To the best of our knowledge, this is the first report of using ensemble of perturbed deep nets as an accessible and computationally inexpensive method for calibration and performance improvement. Our method is potentially most useful when the cost of training from scratch is too high in terms of effort or carbon footprint.

7.2 Method

In this section, we describe the PEP model and analyze local properties of the resulting PEP effect (the gain in log-likelihood over the comparison baseline model). In summary PEP is formulated in the Bayes' network (hierarchical model) framework; it constructs ensembles by Gaussian perturbations of the optimal parameters from training. The single variance parameter is chosen to maximize the likelihood of ensemble average predictions on validation data, which, empirically, has a well-defined maximum. PEP can be applied to any pre-trained network; only one standard training run is needed, and no special training or network architecture is needed.

7.2.1 Baseline Model

We begin with a standard discriminative model, e.g., a classifier that predicts a distribution on y_i given an observation x_i,

$$p(y_i; x_i, \theta) \ . \tag{7.1}$$

Training is conventionally accomplished by maximum likelihood,

$$\theta^* \doteq \operatorname*{argmax}_\theta \mathcal{L}(\theta) \qquad \text{where the log-likelihood is:} \qquad \mathcal{L}(\theta) \doteq \sum_i \ln L_i(\theta) \ , \tag{7.2}$$

and $L_i(\theta) \doteq p(y_i; x_i, \theta)$ are the individual likelihoods. Subsequent predictions are made with the model using θ^*.

7.2.2 Hierarchical Model

Empirically, different optimal values of θ are obtained on different data sets; we aim to model this variability with a very simple parametric model – an isotropic normal distribution with mean and scalar variance parameters,

$$p(\theta; \bar{\theta}, \sigma) \doteq N(\theta; \bar{\theta}, \sigma^2 I) \ . \tag{7.3}$$

The product of Eqs. 7.1 and 7.3 specifies a joint distribution on y_i and θ; from this we can obtain model predictions by marginalizing over θ, which leads to

$$p(y_i; x_i, \bar{\theta}, \sigma) = \mathbb{E}_{\theta \sim N(\bar{\theta}, \sigma^2 I)} \left[p(y_i; x_i, \theta) \right] \ . \tag{7.4}$$

We approximate the expectation by a sample average,

$$p(y_i; x_i, \bar{\theta}, \sigma) \approx \frac{1}{m} \sum_j p(y_i; x_i, \theta_j) \qquad \text{where} \qquad \theta_{j=1}^m \underset{\text{IID}}{\leftarrow} N(\bar{\theta}, \sigma^2 I), \tag{7.5}$$

i.e., the predictions are made by averaging over the predictions of an ensemble. The log-likelihood of the ensemble prediction as a function of σ is then

$$\mathbb{L}(\sigma) \doteq \sum_i \ln \frac{1}{m} \sum_j L_i(\theta_j) \qquad \text{where} \qquad \theta_{j=1}^m \underset{\text{IID}}{\leftarrow} N(\bar{\theta}, \sigma^2 I) \tag{7.6}$$

(dependence on $\bar{\theta}$ is suppresssed for clarity). We estimate the model parameters as follows. First we optimize θ with σ fixed at zero using a training data set (when $\sigma \to 0 \quad \text{the} \quad \theta_j \to \bar{\theta}$), then

$$\theta^* = \underset{\bar{\theta}}{\text{argmax}} \sum_i \ln p(y_i; x_i, \bar{\theta}) \ , \tag{7.7}$$

which is equivalent to maximum likelihood parameter estimation of the base model. Next, we optimize over σ, (using a validation data set), with θ fixed at the previous estimate, θ^*,

$$\sigma^* = \underset{\sigma}{\text{argmax}} \sum_i \ln \frac{1}{m} \sum_{\theta_j} p(y_i; x_i, \theta_j) \ \text{ where } \ \theta_{j=1}^m \underset{\text{IID}}{\leftarrow} N(\theta^*, \sigma^2 I) \ . \tag{7.8}$$

Then at test time the ensemble prediction is

$$p(y_i; x_i, \theta^*, \sigma^*) \approx \frac{1}{m} \sum_{\theta_j} p(y_i; x_i, \theta_j) \ \text{ where } \ \theta_{j=1}^m \underset{\text{IID}}{\leftarrow} N(\theta^*, \sigma^{*2} I) \ . \tag{7.9}$$

In our experiments, perhaps somewhat surprisingly, $\mathbb{L}(\sigma)$ has a well-defined maximum away from $\sigma = 0$ (which corresponds to the baseline model). As σ grows from 0, $\mathbb{L}(\sigma)$ rises to a well-defined peak value, then falls dramatically (Figure 7.1). Conveniently, the calibration quality tends to grow favorably until the $\mathbb{L}(\sigma)$ peak is reached. It may be that $\mathbb{L}(\sigma)$ initially grows because the classifiers corresponding to the ensemble parameters remain accurate, and the ensemble performs better as the classifiers become more independent [31]. Figure 7.1 shows $\mathbb{L}(\sigma)$ for experiments with InceptionV3 [175], along with the average log-likelihoods ($\overline{\ln(L)}$) of the individual ensemble members. Note that in the figures, in the current machine learning style, we have used averaged log-likelihoods, while in this section we use the estimation literature convention that log-likelihoods are summed rather than averaged. We can see that for several members, $\overline{\ln(L)}$ grows somewhat initially, this indicates that the optimal parameter from training is not optimal for the validation data. Interestingly, the ensemble has a more robust increase, which persists over scale substantially longer than for the individual networks. We have observed this $\mathbb{L}(\sigma)$ "increase to peak" phenomenon in many experiments with a wide variety of networks.

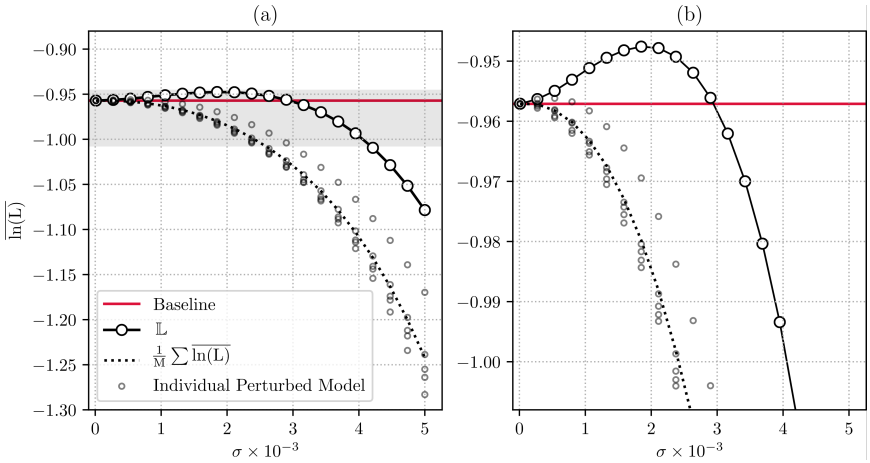


Figure 7.1: Parameter Ensembling by Perturbation (PEP) on pre-trained InceptionV3 [175]. The rectangle shaded in gray in (a) is shown in greater detail in (b). The average log-likelihood of the ensemble average, $\mathbb{L}(\sigma)$, has a well-defined maximum at $\sigma = 1.85 \times 10^{-3}$. The ensemble also has a noticeable increase in likelihood over the individual ensemble item average log-likelihoods, $\overline{\ln(L)}$ and over their average. In this experiment, an ensemble size of 5 (M = 5) was used for PEP and the experiments were run on 5000 validation images.

7.2.3 Local Analysis

In this section, we analyze the nature of the PEP effect in the neighborhood of θ^*. Returning to the log-likelihood of a PEP ensemble (Eq. 7.6), and "undoing" the approximation by sample average,

$$\mathbb{L}(\sigma) \approx \sum_i \ln \mathbb{E}_{\theta \sim N(\theta^*, \sigma^2 I)} \left[L_i(\theta) \right] \ . \tag{7.10}$$

Next we develop a local approximation to the expected value of the log-likelihood. Suppose $x \sim N(\mu, \Sigma)$. We seek a local approximation to $\mathbb{E}_x \left[f(x) \right]$. Using a second order Taylor expansion about μ,

$$\mathbb{E}_x \left[f(x) \right] \approx \mathbb{E}_x \left[f(\mu) + (x - \mu)^T \nabla f(\mu) + \frac{1}{2} (x - \mu)^T H f(\mu) (x - \mu) \right] \tag{7.11}$$

where $Hf(x)$ is the Hessian of $f(x)$. Then, as the gradient term vanishes,

$$\mathbb{E}_x \left[f(x) \right] \approx f(\mu) + \frac{1}{2} \mathbb{E}_x \left[(x - \mu)^T H f(\mu) (x - \mu) \right] \tag{7.12}$$

$$\mathbb{E}_x\left[f(x)\right] \approx f(\mu) + \frac{1}{2}\mathbb{E}_x\left[x^T Hf(\mu)x - 2x^T Hf(\mu)\mu + \mu^T Hf(\mu)\mu\right] \quad (7.13)$$

or

$$\mathbb{E}_x\left[f(x)\right] \approx f(\mu) + \frac{1}{2}\left[\mathbb{E}_x\left[x^T Hf(\mu)x\right] - \mu^T Hf(\mu)\mu\right] \ . \quad (7.14)$$

Now using $\mathbb{E}_x\left[x^T \Lambda x\right] = T_R(\Lambda\Sigma) + \mu^T\Lambda\mu$,($T_R$ is the trace, see [117]),

$$\mathbb{E}_x\left[f(x)\right] \approx f(\mu) + \frac{1}{2}T_R(Hf(\mu)\Sigma) \ . \quad (7.15)$$

For $x \sim N(\mu, \Sigma)$

$$\mathbb{E}_x\left[f(x)\right] \approx f(\mu) + \frac{1}{2}T_R(Hf(\mu)\Sigma) \ , \quad (7.16)$$

where $Hf(x)$ is the Hessian of $f(x)$ and T_R is the trace. In the special case that $\Sigma = \sigma^2 I$,

$$\mathbb{E}_x\left[f(x)\right] \approx f(\mu) + \frac{\sigma^2}{2} \triangle f(\mu) \quad (7.17)$$

where $\triangle$ is the Laplacian, or mean curvature. The appendix shows that the third Taylor term vanishes due to Gaussian properties, so that the approximation residual is $O(\sigma^4\partial^4 f(\mu))$ where ∂^4 is a specific fourth derivative operator.

Applying this to the log-likelihood in Eq. 7.10 yields

$$\mathbb{L}(\sigma) \approx \sum_i \ln\left[L_i(\theta^*) + \frac{\sigma^2}{2} \triangle L_i(\theta^*)\right] \approx \sum_i \left[\ln L_i(\theta^*) + \frac{\sigma^2}{2}\frac{\triangle L_i(\theta^*)}{L_i(\theta^*)}\right] \quad (7.18)$$

(to first order), or

$$\mathbb{L}(\sigma) \approx \mathcal{L}(\theta^*) + B_\sigma(\theta^*) \ , \quad (7.19)$$

where $\mathcal{L}(\theta)$ is the log-likelihood of the base model (Eq. 7.2) and

$$B_\sigma(\theta) \doteq \frac{\sigma^2}{2}\sum_i \frac{\triangle L_i(\theta)}{L_i(\theta)} \quad (7.20)$$

is the PEP effect. Note that the PEP effect value may be dominated by data items that have low likelihood, perhaps because they are difficult cases, or incorrectly labeled. Next we establish a relationship between the PEP effect

and the Laplacian of the log-likelihood of the base model. From Appendix (Eq 11) ,

$$\triangle\mathcal{L}(\theta) = \sum_i \left[\frac{\triangle L_i(\theta)}{L_i(\theta)} - (\nabla \ln L_i(\theta))^2 \right] \tag{7.21}$$

(here the square in the second term on the right is the dot product of two gradients) Then

$$\triangle\mathcal{L}(\theta) = \frac{2}{\sigma^2} B_\sigma(\theta) - \sum_i (\nabla \ln L_i(\theta))^2 \tag{7.22}$$

or

$$B_\sigma(\theta) = \frac{\sigma^2}{2} \left[\triangle\mathcal{L}(\theta) + \sum_i (\nabla \ln L_i(\theta))^2 \right] \ . \tag{7.23}$$

The empirical Fisher information (FI) is defined in terms of the outer product of gradients as

$$\widetilde{F}(\theta) \doteq \sum_i \nabla \ln L_i(\theta) \nabla \ln L_i(\theta)^T \tag{7.24}$$

(see [93]) . So, the second term above in Eq. 7.23 is the trace of the empirical FI. Then finally,

$$B_\sigma(\theta) = \frac{\sigma^2}{2} \left[\triangle\mathcal{L}(\theta) + T_R(\widetilde{F}(\theta)) \right] \ . \tag{7.25}$$

The first term of the PEP effect, the mean curvature of the log-likelihood can be positive or negative, (we expect it to be negative near the mode), while the second term, the trace of the empirical Fisher information, is non-negative. As the sum of squared gradients, we may expect the second term to grow as θ moves away from the mode.

The first term may also be seen as a (negative) trace of an empirical FI. If the sum is converted to an average it approximates an expectation that is equal to the negative of the trace of the Hessian form of the FI, while the second term is the trace of a different empirical FI. Empirical FI is said to be most accurate at the mode of the log-likelihood [93]. So, if θ^* is close to the log-likelihood mode on the new data, we may expect the terms to cancel. If θ^* is farther from the log-likelihood mode on the new data, they may no longer cancel.

Next, we discuss two cases, in both we examine the log-likelihood of the validation data, $\mathcal{L}(\theta)$, at θ^*, the result of optimization on the training data. In general, θ^* will not coincide with the mode of the log-likelihood of the

validation data. **Case 1:** θ^* is 'close' to the mode of the validation data, so we expect the mean curvature to be negative. **Case 2**: θ^* is 'not close' to the mode of the validation data, so the mean curvature may be positive. We conjecture that case 1 characterizes the likelihood landscape on new data when the baseline model is not overfitted, and that case 2 is characteristic of an overfitted model (where, empirically, we observe positive PEP effect).

As these are local characterizations, they are only valid near θ^*. While the analysis may predict PEP effect for small σ, as it grows, and the θ_j move farther from the mode, the log-likelihood will inevitably decrease dramatically (and there will be a peak value between the two regimes).

There has been a lot of work recently concerning the curvature properties of the log-likelihood landscape. Gorbani et al. point out that "Hessian of training loss ... is crucial in determining many behaviors of neural networks"; they provide tools to analyze the Hessian spectrum and point out characteristics associated with networks trained with BN [47]. Sagun et al. [151] point out that there is a 'bulk' of zero valued eigenvectors of the Hessian that can be used to analyze overparameterization , and in a related paper discuss implications that "shed light on the geometry of high-dimensional and non-convex spaces in modern applications" [152]. Fort et al. [40] analyze Deep Ensembles from the perspective of the loss landscape, discussing multiple modes and associated connectors among them. While the entire Hessian spectrum is of interest, some insights may be gained from the avenues to characterizing the mean curvature that PEP provides.

7.3 Experiments

This section reports performance of PEP, and compares it to temperature scaling [54], MCD [41], and Deep Ensembles [96], as appropriate. The first set of results are on ImageNet pre-trained networks where the only comparison is with temperature scaling (no training of the baselines was carried out so MCD and Deep Ensembles were not evaluated). Then we report performance on smaller networks, MNIST and CIFAR-10, where we compare to MCD and Deep Ensembles as well. We also show that the PEP effect is strongly related to the degree of overfitting of the baseline networks.

Evaluation metrics: Model calibration was evaluated with negative log-likelihood (NLL), Brier score [17] and reliability diagrams [127]. NLL and Brier score are proper scoring rules that are commonly used for measuring the quality of classification uncertainty [41, 54, 96, 140]. Reliability diagrams plot expected accuracy as a function of class probability (confidence), and

perfect calibration is achieved when confidence (x-axis) matches expected accuracy (y-axis) exactly [54, 127]. Expected Calibration Error (ECE) is used to summarize the results of the reliability diagram. Details of evaluation metrics are given in the Supplementary Material.

7.3.1 ImageNet Experiments

We evaluated the performance of PEP using large scale networks that were trained on ImageNet (ILSVRC2012) [150] dataset. We used the subset of 50,000 validation images and labels that is included in the development kit of ILSVRC2012. From the 50,000 images, 5,000 images were used as a validation set for optimizing σ in PEP, and temperature T in temperature scaling. The remaining 45,000 images were used as the test set. Golden section search [138] was used to find the σ^* that maximizes $\mathbb{L}(\sigma)$. The search range for σ was 5×10^{-5}–5×10^{-3}, ensemble size was 5 (M=5), and number of iterations was 7. On the test set with 45,000 images, PEP was evaluated using σ^* and with ensemble size of 10 (M=10). Single crop of the center of images was used for the experiments. Evaluation was performed on six pre-trained networks from the Keras library[26]: DenseNet121, DenseNet169 [68], InceptionV3 [175], ResNet50 [60], VGG16, and VGG19 [166]. For all pre-trained networks, Gaussian perturbations were added to the weights of all convolutional layers. Table 7.1 summarizes the optimized T and σ values, model calibration in terms of NLL, Brier score, and classification errors. For all the pre-trained networks, except VGG19, PEP achieves statistically significant improvements in calibration compared to the baseline and temperature scaling. Note the reduction in top-1 error of DenseNet169 by about 1.5 percentage points, and the reduction in all top-1 errors. Figure 7.2 shows the reliability diagram for DenseNet169, before and after calibration with PEP with some corrected misclassification examples.

Table 7.1: ImageNet results: For all models except VGG19 , PEP achieves statistically significant improvements in calibration compared to baseline (BL) and temperature scaling (TS), in terms of NLL and Brier score. PEP also reduces test errors, while TS does not have any effect on test errors. Although TS and PEP outperform baseline in terms of ECE% for DenseNet121, DenseNet169, ResNet, and VGG16, the improvements in ECE% is not consistent among the methods. T^* and σ^* denote optimized temperature for TS and optimized sigma for PEP, respectively. Boldfaced font indicates the best results for each metric of a model and shows that the differences are statistically significant (p-value<0.05).

		σ^*	Negative log-likelihood			Brier score			ECE%			Top-1 error %	
Model	T^*	$\times 10^{-3}$	BL	TS	PEP	BL	TS	PEP	BL	TS	PEP	BL	PEP
DenseNet121	1.10	1.94	1.030	1.018	**0.997**	0.357	0.356	**0.349**	3.47	**1.52**	2.03	25.73	**25.13**
DenseNet169	1.23	2.90	1.035	1.007	**0.940**	0.354	0.350	**0.331**	5.47	**1.75**	2.35	25.31	**23.74**
InceptionV3	0.91	1.94	0.994	0.975	**0.950**	0.328	0.328	**0.317**	**1.80**	4.19	2.46	22.96	**22.26**
ResNet50	1.19	2.60	1.084	1.057	**1.023**	0.365	0.362	**0.350**	5.08	**1.97**	2.94	26.09	**25.18**
VGG16	1.09	1.84	1.199	1.193	**1.164**	0.399	0.399	**0.391**	2.52	2.08	**1.64**	29.39	**28.83**
VGG19	1.09	1.03	1.176	1.171	1.165	0.394	0.394	0.391	4.77	4.50	4.48	28.99	28.75

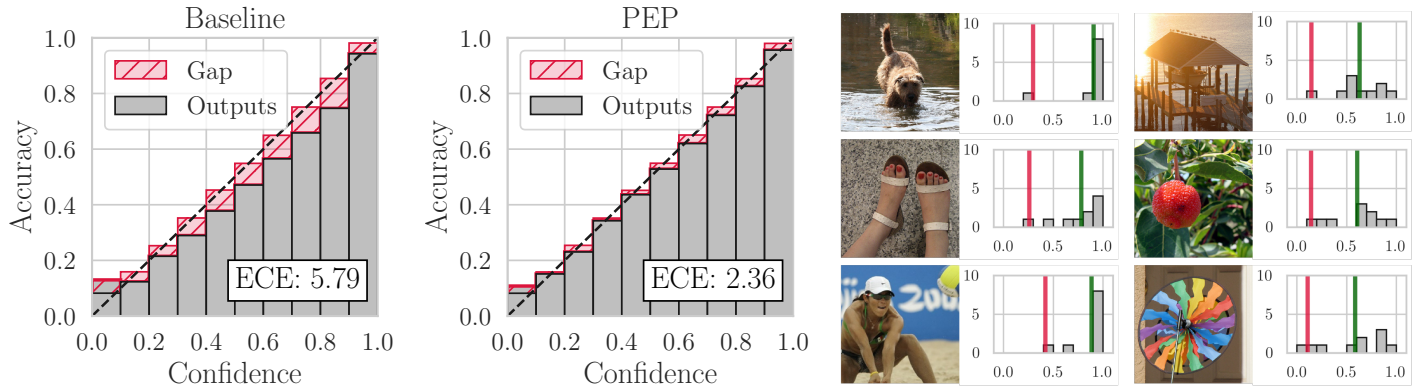


Figure 7.2: Improving pre-trained DenseNet169 with PEP (M=10). (a) and (b) show the reliability diagrams of the baseline and the PEP. (c) shows examples of misclassifications corrected by PEP. The examples were among those with the highest PEP effect on the correct class probability. (c) Top row: brown bear and lampshade changed into Irish terrier and boathouse; Middle row: band aid and pomegranate changed into sandal and strawberry; Bottom row: bathing cap and wall clock changed into volleyball and pinwheel. The histograms at the right of each image illustrate the probability distribution of ensemble. Vertical red and green lines show the predicted class probabilities of the baseline and the PEP for the correct class label. (For more reliability diagrams see Supplementary Material.)

7.3.2 MNIST and CIFAR-10 Experiments

The MNIST handwritten digits dataset [97] consists of 60,000 training images and 10,000 test images. The CIFAR-10 dataset [91] consists of 50,000 training images and 10,000 test images. We created validation sets by setting aside 10,000 and 5,000 training images from MNIST and CIFAR-10, respectively. For the MNIST dataset, the predictive uncertainty was evaluated for two different neural networks: a Multi-layer Perception (MLP) and a Convolutional Neural Network (CNN) similar to LeNet [99] but with smaller kernel sizes. The MLP is similar to the one used in [96] and has 3 hidden layers with 200 neurons each, ReLu non-linearities, and BN after each layer. For MCD experiments, dropout layers were added after each hidden layer with 0.5 dropout rate as was suggested in [41]. The CNN for MNIST experiments has two convolutional layers with 32 and 64 kernels of sizes 3×3 with stride size of 1 followed by two fully connected layers (with 128 and 64 neurons each) with BN after both types of layers. Here, again for MCD experiments, dropout was added after all layers with 0.5 dropout rate, except the first and last layers. For the CIFAR-10 dataset, the CNN architecture has 2 convolutional layers with 16 kernels of size 3×3 followed by a max-pooling of 2×2; another 2 convolutional layers with 32 kernels of size 3×3 followed by a max-pooling of size 2×2. And finally, two dense layers of size 128, and 10. BN was applied to all convolutional layers. For MCD experiments, dropout was added similar to CNN for MNIST experiments. Each network was trained and evaluated 25 times with different initializations of parameters (weights and biases) and random shuffling of the training data. For optimization, stochastic gradient descent with the Adam update rule [86] was used. Each baseline was trained for 15 epochs. Training was performed for another 25 rounds with dropout for MCD experiments. Models trained and evaluated with active dropout layers were used for MCD evaluation only, and baselines without dropout were used for the rest of the experiments. The Deep Ensembles method was tested by averaging the output of the 10 baseline models. MCD was tested on 25 models and the performance was averaged over all 25 models. Temperature scaling and PEP were tested on the 25 trained baseline models without dropout and the results were averaged.

Table 7.2: MNIST and CIFAR-10 results: The table summarizes experiments described in Section 7.3.2.

Experiment	Baseline	PEP	Temp. Scaling	MCD	Deep Ensembles
	NLL				
MNIST (MLP)	0.096 ± 0.01	0.079 ± 0.01	0.074 ± 0.01	0.094 ± 0.00	0.044 ± 0.00
MNIST (CNN)	0.036 ± 0.00	0.034 ± 0.00	0.032 ± 0.00	0.031 ± 0.00	0.021 ± 0.00
CIFAR-10	1.063 ± 0.03	0.982 ± 0.02	0.956 ± 0.02	0.798 ± 0.01	0.709 ± 0.00
	Brier				
MNIST (MLP)	0.037 ± 0.00	0.035 ± 0.00	0.035 ± 0.00	0.040 ± 0.00	0.020 ± 0.00
MNIST (CNN)	0.016 ± 0.00	0.015 ± 0.00	0.015 ± 0.00	0.014 ± 0.00	0.010 ± 0.00
CIFAR-10	0.469 ± 0.01	0.450 ± 0.01	0.447 ± 0.01	0.381 ± 0.01	0.335 ± 0.00
	ECE %				
MNIST (MLP)	1.324 ± 0.16	0.528 ± 0.12	0.415 ± 0.10	2.569 ± 0.17	0.839 ± 0.08
MNIST (CNN)	0.517 ± 0.07	0.366 ± 0.08	0.259 ± 0.06	0.832 ± 0.06	0.287 ± 0.05
CIFAR-10	11.718 ± 0.72	4.599 ± 0.82	1.318 ± 0.26	7.109 ± 0.62	8.867 ± 0.23
	Classification Error %				
MNIST (MLP)	2.264 ± 0.22	2.286 ± 0.24	2.264 ± 0.22	2.452 ± 0.14	1.285 ± 0.05
MNIST (CNN)	0.990 ± 0.13	0.990 ± 0.12	0.990 ± 0.13	0.842 ± 0.06	0.659 ± 0.03
CIFAR-10	33.023 ± 0.68	32.949 ± 0.74	33.023 ± 0.68	27.207 ± 0.66	22.880 ± 0.21

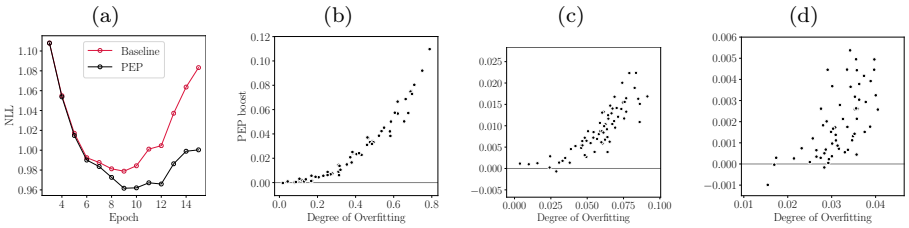


Figure 7.3: The relationship between overfitting and PEP effect. (a) shows the average of NLLs on test set for CIFAR-10 baselines (red line) and PEP $\mathbb{L}$ (black line). The baseline curve shows overfitting as a result of overtraining. The degree of overfitting was calculated by subtracting the training NLL (loss) from the test NLL (loss). PEP reduces overfitting and improves log-likelihood. PEP effect is more substantial as the overfitting grows. (b), (c), (d) shows scatter plots of overfitting vs PEP effect for CIFAR-10, MNIST(MLP), and MNIST(CNN), respectively.

Table 7.2 compares the calibration quality and test errors of baselines and PEP, temperature scaling [54], MCD [41], and Deep Ensembles [96]. The averages and standard deviation values for NLL, Brier score, and ECE% are provided. For all cases, it can be seen that PEP achieves better calibration in terms of lower NLL compared to the baseline. Deep Ensembles achieves the best NLL and classification errors in all the experiments. Compared to the baseline, temperature scaling and MCD improve calibration in terms of NLL for all three experiments.

Effect of Overfitting on PEP Effect: We ran experiments to quantify the effect of overfitting on PEP effect, and optimized σ values. For the MNIST and CIFAR-10 experiments, model checkpoints were saved at the end of each epoch. Different levels of overfitting as a result of over-training were observed for the three experiments. σ^* was calculated for each epoch and PEP was performed and the PEP effect was measured. Figure 7.3 (a), shows the effect of calibration on calibration and reducing NLL for CIFAR-10 models. Figures 7.3 (b-d) shows that PEP effect increases with overfitting. Furthermore, we observed that the σ^* values also increase with overfitting, meaning that larger perturbations are required for more overfitting.

7.4 Conclusion

We proposed PEP for improving calibration and performance in deep learning. PEP is computationally inexpensive and can be applied to any pre-trained

network. On classification problems, we show that PEP effectively improves probabilistic predictions in terms of log-likelihood, Brier score, and expected calibration error. It also nearly always provides small improvements in accuracy for pre-trained ImageNet networks. We observe that the optimal size of perturbation and the log-likelihood increase from the ensemble correlates with the amount of overfitting. Finally, PEP can be used as a tool to investigate the curvature properties of the likelihood landscape.

Chapter 8

Conclusion and Future Work

Prostate cancer is the leading cause of cancer death in North American men and the second most common cancer in men worldwide. The ultimate diagnosis of prostate cancer is through histopathology analysis of prostate biopsy or radical prostatectomy. MRI has shown promising results for detection and characterization of prostate cancer and in guiding biopsy needles to suspicious targets. Despite promising results in using MRI for prostate cancer management, open problems exist regarding detection and characterization of prostate cancer and image-guided interventions.

In this thesis, novel algorithms and methods were proposed with the ultimate goal of improving MRI-guided prostate cancer diagnosis and interventions. In Chapter 2, we proposed models to classify prostate cancer at a given biopsy location as clinically significant or not in diagnostic MRI images. We further proposed models to handle biopsy location uncertainty at training and inference times. In Chapter 3, we proposed models to automatically detect the tip of biopsy needles on intra-procedural MRI images. In Chapter 4, we investigated domain adaptation techniques to see if we could tune a CNN trained to perform a task on MRI images acquired with different acquisition parameters. In Chapter 5, we proposed a partial Dice loss function for weakly-supervised segmentation with single points and scribbles. In Chapter 6, we studied uncertainty estimation in semantic segmentation and proposed methods to improve confidence calibration using ensemble of models. Finally, in Chapter 7, we proposed a general methodology for uncertainty estimation of neural networks using parameter ensembling by perturbation.

8.1 Contributions

This thesis is an attempt to develop techniques that are essential for MRI-guided prostate cancer diagnosis and interventions. In the course of achieving this objective, the following contributions were made:

- A novel deep learning technique was proposed for diagnosing clinically significant prostate cancer in mpMRI. The method uses diffusion-

weighted imaging (DWI) and dynamic contrast-enhanced (DCE) MRI sequences and information about the location of the suspicious target to diagnose clinically significant cancer. The proposed method was tested on an unseen patient dataset of 206 findings from 140 patients and achieved an area under the curve of receiver operating characteristic (AUC) of 0.80. The performance is comparable with the AUC values of experienced human readers for PI-RADS.

- A novel probabilistic framework was proposed to include biopsy location uncertainty at the inference for diagnosis of clinically significant prostate cancer lesions with FCNs. Moreover, a Gaussian weighted loss was proposed as a label imputation mechanism for training FCNs with sparse biopsy data. The proposed loss function was compared with partial cross-entropy (CE) where biopsy locations are used for loss calculation in optimization. It was observed that using the updated biopsy location improves sensitivity significantly through detecting lesions where the biopsy location was displaced. The proposed method was trained and validated using a 6-fold cross validation scheme with 352 biopsy locations from 203 patients suspicious of prostate cancer.

- A novel asymmetric 3D deep CNN was developed to localize and visualize the tip and trajectory of biopsy needles in MRI. Needles were annotated on 583 T2-weighted intra-procedural MRI scans acquired after needle insertion for 71 patients. The accuracy of the proposed method, as tested on previously unseen data, was 2.80 mm average in needle tip detection, and 0.98° in needle trajectory angle. Additionally, an observer study was designed in which independent annotations by a second observer, blinded to the original observer, were compared to the output of the proposed method. The resultant error was comparable to the measured inter-observer concordance, reinforcing the clinical acceptability of the proposed method. To the best of our knowledge, this was the first report of a fully automatic system for biopsy needle segmentation and localization in MRI with deep convolutional neural networks.

- A novel technique was developed for domain adaptation of networks trained with one set of MRI acquisition parameters. The following questions regarding domain adaptation were investigated: Given a fitted model on a certain dataset domain, 1) How much data from the new domain is required for a decent adaptation of the original network?; and, 2) What portion of the pre-trained model parameters should be

retrained given a certain number of the new domain training samples? We trained a CNN on one set of images and evaluated the performance of the domain-adapted network on the same task with images from a different domain. We then compared the performance of the model to the surrogate scenarios where either the same trained network is used or a new network is trained from scratch on the new dataset. The proposed method is capable of tuning the deep network to the new domain.

- A novel technique was proposed for weakly-supervised semantic segmentation with point and scribble supervision in FCNs. A novel loss function, partial Dice loss, was proposed a variant of Dice loss [124] for deep weakly-supervised segmentation with sparse pixel-level annotations. Partial Dice loss was compared with partial cross-entropy [27, 177] in terms of segmentation quality. Finally, point and scribble-supervised segmentation were compared with fully-supervised on five different semantic segmentation tasks from medical images of the heart, the prostate, and the kidney. In a majority of these experiments, partial Dice loss provided statistically significant performance improvement over partial cross-entropy. The use of single point supervision results in $51\%-95\%$ of the performance of fully supervised training and the use of single scribble supervision achieves $86\%-97\%$ of the performance of fully supervised training.

- A novel technique was developed for confidence calibration and predictive uncertainty estimation for deep medical image segmentation. Despite of high quality segmentations, FCNs trained with batch normalization and Dice loss are poorly calibrated. We systematically compared cross-entropy loss with Dice loss in terms of segmentation quality and uncertainty estimation of FCNs; We proposed model ensembling for confidence calibration of the FCNs trained with BN and Dice loss; We further assessed the ability of calibrated FCNs to predict the segmentation quality of structures and detect out-of-distribution test examples. We consistently demonstrated that model ensembling is considerably effective for confidence calibration.

- A novel technique was developed for confidence calibration uncertainty estimation of neural networks. The proposed technique, parameter ensembling by perturbation (PEP) approach, prepares an ensemble of parameter values as perturbations of the optimal parameter set from training by a Gaussian with a single variance parameter. Experiments

on classification benchmarks such as MNIST and CIFAR-10 showed improved calibration and likelihood. To demonstrate the scalability of PEP on deep networks, experiments were conducted on ImageNet, these show that PEP can be used for uncertainty estimation and probability calibration on pre-trained networks.

8.2 Future Work

Novel methods have been presented in this thesis for MRI-guided prostate cancer diagnosis and interventions. In addition, we proposed methods for domain adaptation, confidence calibration and uncertainty estimation in medical images. A number of interesting areas of research can be suggested as follows:

- The proposed models for cancer diagnosis were developed and validated only on a cohort of patient from a single institution. Further experimental investigations are needed to be done using larger multi-institute datasets. It would be essential to determine the performance of the proposed prostate cancer diagnosis approaches across a wider range of patient populations.

- Future work requires to explore the use of the posterior probabilities on latent true biopsy coordinates for improving training procedures of CAD systems with noisy ground truth. Through an expectation-maximization (EM) framework, the posterior can be used as the $E-$step to re-estimate probability distribution on biopsy locations given the prior knowledge and classifier's output. Maximum likelihood estimation can be updated to include the current knowledge of the distribution.

- Further work needs to be carried out to embed the proposed automatic localization method in the workflow of the transperineal in-gantry MRI-targeted prostate biopsies. In order to do this, a study has to be designed to determine how the needle trajectory should be presented to the interventionalist to help them make the most efficacious decisions – e.g. should the insertion point or angle of a suboptimal trajectory be changed – during the procedure. On a wider level, research is also needed to transfer the framework and proposed methodology to other types of image-guided procedures that involve needle detection and localization.

- In Chapter 4, due to lack of access to multi-domain prostate cancer datasets, we studied transfer learning for the problem of brain white matter hyperintensities (WMH) segmentation. Further experiments are required to evaluate the proposed methods with multi-institutional prostate cancer MRI datasets.
- Additional work needs to be carried out to establish the effect of loss function on confidence calibration for deep FCNs that were proposed in Chapter 6. It would be interesting to investigate the calibration and segmentation quality of other loss functions such as combinations of Dice loss and cross-entropy loss, as well as the recently proposed Lovász-Softmax loss [14].
- The proposed parameter ensembling by perturbation (PEP) method was evaluated on computer vision benchmarks. Further evaluation of PEP on medical imaging benchmarks and applications including prostate cancer diagnosis in MRI would be interesting.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Amir H Abdi, Christina Luong, Teresa Tsang, Gregory Allan, Saman Nouranian, John Jue, Dale Hawley, Sarah Fleming, Ken Gin, Jody Swift, Robert Rohling, and Purang Abolmaesumi. Automatic quality assessment of echocardiograms using convolutional neural networks: Feasibility on the apical Four-Chamber view. *IEEE Transactions on Medical Imaging*, 36(6):1221–1230, 2017.

[3] M Aboofazeli, P Abolmaesumi, P Mousavi, and G Fichtinger. A new scheme for curved needle segmentation in three-dimensional ultrasound images. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1067–1070, 2009.

[4] 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.

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[6] Simon Andermatt, Simon Pezold, and Philippe Cattin. Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-Data. In *Deep Learning and Data Labeling for Medical Applications*, pages 142–151. Springer International Publishing, 2016.

[7] Samuel G Armato, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S Kirby, Nicholas Petrick, George Redmond, Maryellen L

Giger, Kenny Cha, Artem Mamonov, et al. PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging*, 5(4):044501, 2018.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[9] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017.

[10] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117, 2017.

[11] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019.

[12] Christoph Baur, Shadi Albarqouni, and Nassir Navab. Semi-supervised deep learning for fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 311–319. Springer, 2017.

[13] Parmida Beigi, Robert Rohling, Tim Salcudean, Victoria A Lessoway, and Gary C Ng. Needle trajectory and tip localization in Real-Time 3-D ultrasound using a moving stylus. *Ultrasound in Medicine and Biology*, 41(7):2057–2070, 2015.

[14] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pages 4413–4421, 2018.

[15] Jeroen Bertels, David Robben, Dirk Vandermeulen, and Paul Suetens. Optimization with soft Dice can lead to a volumetric bias. *arXiv preprint arXiv:1911.02278*, 2019.

[16] Andreas Bresser et al. Python pathfinding. `https://github.com/brean/python-pathfinding`.

[17] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[18] Jinzheng Cai, Youbao Tang, Le Lu, Adam P Harrison, Ke Yan, Jing Xiao, Lin Yang, and Ronald M Summers. Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2D RECIST . In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 396–404. Springer, 2018.

[19] Andrew Cameron, Farzad Khalvati, Masoom A Haider, and Alexander Wong. MAPS: A quantitative radiomics approach for prostate cancer detection. *IEEE Transactions on Biomedical Engineering*, 63(6):1145–1156, 2016.

[20] Yigit B Can, Krishna Chaitanya, Basil Mustafa, Lisa M Koch, Ender Konukoglu, and Christian F Baumgartner. Learning to segment medical images with scribble-supervision alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 236–244. Springer, 2018.

[21] Ian Chan, William Wells, 3rd, Robert V Mulkern, Steven Haker, Jianqing Zhang, Kelly H Zou, Stephan E Maier, and Clare M C Tempany. Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. *Medical physics*, 30(9):2390–2398, 2003.

[22] Quan Chen, Xiang Xu, Shiliang Hu, Xiao Li, Qing Zou, and Yunpeng Li. A transfer learning approach for classification of clinical significant prostate cancers from mpMRI scans. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101344F. International Society for Optics and Photonics, 2017.

[23] Shuai Chen, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs van Tulder, and Marleen de Bruijne. Multi-task attention-based semi-supervised learning for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–465. Springer, 2019.

[24] V. Cheplygina, I. P. Pena, J. H. Pedersen, D. A Lynch, L. Sørensen, and M. de Bruijne. Transfer learning for multi-center classification of chronic obstructive pulmonary disease. *arXiv preprint arXiv:1701.05013*, 2017.

[25] Eleni Chiou, Francesco Giganti, Elisenda Bonet-Carne, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Prostate cancer classification on verdict DW-MRI using convolutional neural networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 319–327. Springer, 2018.

[26] François Chollet et al. Keras. `https://keras.io`, 2015.

[27] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 424–432. Springer, 2016.

[28] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, pages 2898–2909, 2019.

[29] Bob D de Vos, Jelmer M Wolterink, Pim A de Jong, Tim Leiner, Max A Viergever, and Ivana Isgum. ConvNet-Based localization of anatomical structures in 3-D medical images. *IEEE Transactions on Medical Imaging*, 36(7):1470–1481, 2017.

[30] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[31] Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[32] SP DiMaio, DF Kacher, RE Ellis, G Fichtinger, N Hata, GP Zientara, LP Panych, R Kikinis, and FA Jolesz. Needle artifact localization in 3T MR images. *Studies in Health Technology and Informatics*, 119:120, 2005.

[33] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. CT Mok, L. Shi, and P. A. Heng. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Transactions on Medical Imaging*, 35(5):1182–1195, 2016.

[34] Timothy Dozat. Incorporating nesterov momentum into adam. In *ICLR Workshop*, 2016.

[35] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[36] 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.

[37] A. Esteva, B. Kuprel, R. A Novoa, J. Ko, S. M Swetter, H. M Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[38] Andriy Fedorov, Kemal Tuncali, Fiona M Fennessy, Junichi Tokuda, Nobuhiko Hata, William M Wells, Ron Kikinis, and Clare M Tempany. Image registration for targeted MRI-guided transperineal prostate biopsy. *Journal of Magnetic Resonance Imaging*, 36(4):987–992, 2012.

[39] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sébastien Ourselin, and Tom Vercauteren. Generalised Wasserstein Dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *International MICCAI Brainlesion Workshop*, pages 64–76. Springer, 2017.

[40] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[41] 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.

[42] Pierre-Antoine Ganaye, Michaël Sdika, and Hugues Benoit-Cattin. Semi-supervised learning for segmentation under semantic constraint. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–602. Springer, 2018.

[43] M. Ghafoorian, N. Karssemeijer, I. WM van Uden, F.E de Leeuw, T. Heskes, E. Marchiori, and B. Platel. Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Medical Physics*, 43(12):6246–6258, 2016.

[44] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Mayra Bergkamp, Joost Wissink, Jiri Obels, Karlijn Keizer, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, and Bram Platel. Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *Neuroimage Clinical*, 14:391–399, 2017.

[45] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Inge WM van Uden, Clara I Sanchez, Geert Litjens, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, and Bram Platel. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):1–12, 2017.

[46] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 516–524. Springer, 2017.

[47] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241, 2019.

[48] Valentina Giannini, Simone Mazzetti, Enrico Armando, Silvia Carabalona, Filippo Russo, Alessandro Giacobbe, Giovanni Muto, and Daniele Regge. Multiparametric magnetic resonance imaging of the prostate with computer-aided detection: experienced observer performance study. *European Radiology*, 27(10):4200–4208, 2017.

[49] Shoshana B Ginsburg, Ahmad Algohary, Shivani Pahwa, Vikas Gulani, Lee Ponsky, Hannu J Aronen, Peter J Boström, Maret Böhm, Anne-Maree Haynes, Phillip Brenner, Warick Delprado, James Thompson, Marley Pulbrock, Pekka Taimen, Robert Villani, Phillip Stricker, Ardeshir R Rastinehad, Ivan Jambor, and Anant Madabhushi. 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.

[50] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[52] Joseph Görres, Michael Brehler, Jochen Franke, Karl Barth, Sven Y Vetter, Andrés Córdova, Paul A Grützner, Hans-Peter Meinzer, Ivo Wolf, and Diana Nabers. Intraoperative detection and localization of cylindrical implants in cone-beam CT image data. *International Journal of Computer Assisted Radiology and Surgery*, 9(6):1045–1057, 2014.

[53] Henry Gray. *Anatomy of the human body*, volume 8. Lea & Febiger, 1878.

[54] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

[55] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.

[56] Freddie C Hamdy, Jenny L Donovan, J Athene Lane, Malcolm Mason, Chris Metcalfe, Peter Holding, Michael Davis, Tim J Peters, Emma L Turner, Richard M Martin, et al. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New England Journal of Medicine*, 375(15):1415–1424, 2016.

[57] Xiang Hao, Kristen Zygmunt, Ross T Whitaker, and P Thomas Fletcher. Improved segmentation of white matter tracts with adaptive Riemannian metrics. *Medical Image Analysis*, 18(1):161–175, 2014.

[58] Elmira Hassanzadeh, Daniel I Glazer, Ruth M Dunne, Fiona M Fennessy, Mukesh G Harisinghani, and Clare M Tempany. Prostate imaging reporting and data system version 2 (PI-RADS v2): a pictorial review. *Abdominal Radiology*, 42(1):278–289, 2017.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[61] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[62] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[63] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, 2019.

[64] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. In *Advances in Neural Information Processing Systems*, pages 909–918, 2018.

[65] William Thomas Hrinivich, Douglas A Hoover, Kathleen Surry, Chandima Edirisinghe, Jacques Montreuil, David D'Souza, Aaron Fenster, and Eugene Wong. Simultaneous automatic segmentation of multiple needles using 3D ultrasound for high-dose-rate prostate brachytherapy. *Medical Physics*, 44(4):1234–1245, 2017.

[66] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2019.

[67] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[68] 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.

[69] Abhaya Indrayan. *Medical biostatistics*. Chapman and Hall/CRC, 2012.

[70] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[71] Junichiro Ishioka, Yoh Matsuoka, Sho Uehara, Yosuke Yasuda, Toshiki Kijima, Soichiro Yoshida, Minato Yokoyama, Kazutaka Saito, Kazunori Kihara, Noboru Numao, et al. Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. *BJU International*, 122(3):411–417, 2018.

[72] Ahmadreza Jeddi, Mohammad Javad Shafiee, Michelle Karg, Christian Scharfenberger, and Alexander Wong. Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness. *arXiv preprint arXiv:2003.01090*, 2020.

[73] Zhanghexuan Ji, Yan Shen, Chunwei Ma, and Mingchen Gao. Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 175–183. Springer, 2019.

[74] Hongsheng Jin, Zongyao Li, Ruofeng Tong, and Lanfen Lin. A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. *Medical physics*, 45(5):2097–2107, 2018.

[75] Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019.

[76] Kaggle. TGS salt identification challenge, segment salt deposits beneath the earth's surface. `https://www.kaggle.com/c/tgs-salt-identification-challenge`, 2018.

[77] K. Kamnitsas, C. Ledig, V. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[78] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 450–462. Springer, 2017.

[79] Davood Karimi, Qi Zeng, Prateek Mathur, Apeksha Avinash, Sara Mahdavi, Ingrid Spadinger, Purang Abolmaesumi, and Septimiu E Salcudean. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Medical Image Analysis*, 57:186–196, 2019.

[80] Moritz Kasel-Seibert, Thomas Lehmann, René Aschenbach, Felix V Guettler, Mohamed Abubrig, Marc-Oliver Grimm, Ulf Teichgraeber, and Tobias Franiel. Assessment of PI-RADS v2 for the detection of prostate cancer. *European Journal of Radiology*, 85(4):726–731, 2016.

[81] Veeru Kasivisvanathan, Antti S Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A Mynderse, Markku H Vaarala, Alberto Briganti, Lars Budäus, Giles Hellawell, Richard G Hindley, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*, 378(19):1767–1777, 2018.

[82] 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.

[83] 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 5574–5584, 2017.

[84] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis*, 54:88–99, 2019.

[85] Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. *arXiv preprint arXiv:1806.04854*, 2018.

[86] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[87] Atilla P Kiraly, Clement Abi Nader, Ahmet Tuysuzoglu, Robert Grimm, Berthold Kiefer, Noha El-Zehiry, and Ali Kamen. Deep convolutional encoder-decoders for prostate cancer detection and classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 489–497. Springer, 2017.

[88] 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*, 2017.

[89] Simon Kohl et al. A probabilistic U-Net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.

[90] Axel Krieger, Sang-Eun Song, Nathan Bongjoon Cho, Iulian I Iordachita, Peter Guion, Gabor Fichtinger, and Louis L Whitcomb. Development and evaluation of an actuated MRI-compatible robotic system for MRI-guided prostate intervention. *IEEE/ASME Transactions on Mechatronics*, 18(1):273–284, 2013.

[91] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[92] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of

automatic segmentation of white matter hyperintensities; results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568, 2019.

[93] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems 32*, pages 4156–4167. Curran Associates, Inc., 2019.

[94] Jin Tae Kwak, Sheng Xu, Bradford J Wood, Baris Turkbey, Peter L Choyke, Peter A Pinto, Shijun Wang, and Ronald M Summers. Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging. *Medical physics*, 42(5):2368–2378, 2015.

[95] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using Bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. *Medical Imaging with Deep Learning*, 2018.

[96] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[97] Yann LeCun. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998.

[98] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[99] 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.

[100] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[101] 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.

[102] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):17816, 2017.

[103] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[104] Zhibin Liao, Hany Girgis, Amir Abdi, Hooman Vaseli, Jorden Hetherington, Robert Rohling, Ken Gin, Teresa Tsang, and Purang Abolmaesumi. On modelling label uncertainty in deep neural networks: Automatic estimation of intra-observer variability in 2D echocardiography quality assessment. *IEEE Transactions on Medical Imaging*, 39(6):1868–1883, 2019.

[105] Paweł Liskowski and Krzysztof Krawiec. Segmenting retinal blood vessels with deep neural networks. *IEEE Transactions on Medical Imaging*, 35(11):2369–2380, 2016.

[106] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

[107] Geert Litjens et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[108] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014.

[109] Lizhi Liu, Zhiqiang Tian, Zhenfeng Zhang, and Baowei Fei. Computer-aided Detection of Prostate Cancer with MRI: Technology and Applications. *Academic Radiology*, 23(8):1024–1046, 2016.

[110] Saifeng Liu, Huaixiu Zheng, Yesu Feng, and Wei Li. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013428. International Society for Optics and Photonics, 2017.

[111] Stacy Loeb, Marc A Bjurlin, Joseph Nicholson, Teuvo L Tammela, David F Penson, H Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *European Urology*, 65(6):1046–1055, 2014.

[112] 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.

[113] Bradley Christopher Lowekamp, David T Chen, Luis Ibáñez, and Daniel Blezek. The design of SimpleITK. *Frontiers in Neuroinformatics*, 7:45, 2013.

[114] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, volume 30, 2013.

[115] David JC MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[116] Andre Mastmeyer, Guillaume Pernelle, Ruibin Ma, Lauren Barber, and Tina Kapur. Accurate model-based segmentation of gynecologic brachytherapy catheter collections in MRI-images. *Medical Image Analysis*, 42:173–188, 2017.

[117] Arakaparampil M Mathai and Serge B Provost. *Quadratic forms in random variables: theory and applications*. Dekker, 1992.

[118] Alireza Mehrtash, Mohsen Ghafoorian, Guillaume Pernelle, Alireza Ziaei, Friso G Heslinga, Kemal Tuncali, Andriy Fedorov, Ron Kikinis, Clare M Tempany, William M Wells, et al. Automatic needle segmentation and localization in MRI with 3-D convolutional neural networks: Application to MRI-targeted prostate biopsy. *IEEE Transactions on Medical Imaging*, 38(4):1026–1036, 2018.

[119] Alireza Mehrtash, Mehran Pesteie, Jorden Hetherington, Peter A. Behringer, Tina Kapur, William M. Wells III, Robert Rohling, Andriy Fedorov, and Purang Abolmaesumi. DeepInfer: Open-source deep learning deployment toolkitfor image-guided therapy. In *SPIE Medical Imaging*. International Society for Optics and Photonics, 2017.

[120] Alireza Mehrtash, Alireza Sedghi, Mohsen Ghafoorian, Mehdi Taghipour, Clare M Tempany, William M Wells III, Tina Kapur, Parvin Mousavi, Purang Abolmaesumi, and Andriy Fedorov. Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101342A. International Society for Optics and Photonics, 2017.

[121] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 2020.

[122] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[123] Anneke Meyer, Alireza Mehrtash, Marko Rak, Daniel Schindele, Martin Schostak, Clare Tempany, Tina Kapur, Purang Abolmaesumi, Andriy Fedorov, and Christian Hansen. Automatic high resolution segmentation of the prostate from multi-planar MRI. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 177–181, 2018.

[124] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[125] Mehdi Moradi, Septimiu E Salcudean, Silvia D Chang, Edward C Jones, Nicholas Buchan, Rowan G Casey, S Larry Goldenberg, and Piotr Kozlowski. Multiparametric MRI maps for detection and grading of dominant prostate tumors. *Journal of Magnetic Resonance Imaging*, 35(6):1403–1413, 2012.

[126] MRBrainS18. Grand challenge on MR brain segmentation at MICCAI 2018. `https://mrbrains18.isi.uu.nl/`, 2018.

[127] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2901–2907, 2015.

[128] Huu-Giao Nguyen, Céline Fouard, and Jocelyne Troccaz. Segmentation, separation and pose estimation of prostate brachytherapy seeds in CT images. *IEEE Transactions on Biomedical Engineering*, 62(8):2012–2024, 2015.

[129] Emilie Niaf, Olivier Rouvière, Florence Mège-Lechevallier, Flavie Bratan, and Carole Lartizien. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. *Physics in Medicine & Biology*, 57(12):3833–3851, 2012.

[130] Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen. ASDNET: Attention based semi-supervised deep networks for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 370–378. Springer, 2018.

[131] Marc Niethammer, Kilian M Pohl, Firdaus Janoos, and William M Wells III. Active mean fields for probabilistic image segmentation: Connections with Chan–Vese and Rudin–Osher–Fatemi models. *SIAM Journal on Imaging Sciences*, 10(3):1069–1103, 2017.

[132] Paul M Novotny, Jeff A Stoll, Nikolay V Vasilyev, Pedro J del Nido, Pierre E Dupont, Todd E Zickler, and Robert D Howe. GPU based real-time instrument tracking with three-dimensional ultrasound. *Medical Image Analysis*, 11(5):458–464, 2007.

[133] Lauren J O'Donnell and Carl-Fredrik Westin. Automatic tractography segmentation using a high-dimensional white matter atlas. *IEEE Transactions on Medical Imaging*, 26(11):1562–1575, 2007.

[134] Xi Ouyang, Zhong Xue, Yiqiang Zhan, Xiang Sean Zhou, Qingfeng Wang, Ying Zhou, Qian Wang, and Jie-Zhi Cheng. Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest X-ray. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 613–621. Springer, 2019.

[135] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[136] Tobias Penzkofer, Kemal Tuncali, Andriy Fedorov, Sang-Eun Song, Junichi Tokuda, Fiona M Fennessy, Mark G Vangel, Adam S Kibel, Robert V Mulkern, William M Wells, Nobuhiko Hata, and Clare M C Tempany. Transperineal in-bore 3-T MR imaging-guided prostate

biopsy: a prospective clinical observational study. *Radiology*, 274(1):170–180, 2015.

[137] Guillaume Pernelle, Alireza Mehrtash, Lauren Barber, Antonio Damato, Wei Wang, Ravi Teja Seethamraju, Ehud Schmidt, Robert A Cormack, Williams Wells, Akila Viswanathan, and Tina Kapur. Validation of catheter segmentation for MR-Guided gynecologic cancer brachytherapy. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, Lecture Notes in Computer Science, pages 380–387. Springer, Berlin, Heidelberg, 2013.

[138] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[139] Michael Quentin, Dirk Blondin, Christian Arsov, Lars Schimmöller, Andreas Hiester, Erhard Godehardt, Peter Albers, Gerald Antoch, and Robert Rabenalt. Prospective evaluation of magnetic resonance imaging guided in-bore prostate biopsy versus systematic transrectal ultrasound guided prostate biopsy in biopsy naïve men with elevated prostate specific antigen. *The Journal of Urology*, 192(5):1374–1379, 2014.

[140] Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer, 2005.

[141] Khashayar Rafat Zand, Caroline Reinhold, Masoom A Haider, Asako Nakai, Laurian Rohoman, and Sharad Maheshwari. Artifacts and pitfalls in MR imaging of the pelvis. *Journal of Magnetic Resonance Imaging*, 26(3):480–497, 2007.

[142] 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.

[143] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 36(2):674–683, 2016.

[144] Prashanth Rawla. Epidemiology of prostate cancer. *World Journal of Oncology*, 10(2):63, 2019.

[145] Mark Renfrew, Mark Griswold, and M Cenk Çavuşoğlu. Active localization and tracking of needle and target in robotic image-guided intervention systems. *Autonomous Robots*, 42(1):83–97, 2018.

[146] Raanan Yehezkel Rohekar, Yaniv Gurwicz, Shami Nisimov, and Gal Novik. Modeling uncertainty by learning a hierarchy of deep neural connections. In *Advances in Neural Information Processing Systems*, pages 4246–4256, 2019.

[147] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[148] Holger Roth, Ling Zhang, Dong Yang, Fausto Milletari, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Weakly supervised segmentation from extreme points. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, pages 42–50. Springer, 2019.

[149] Matthias Rottmann and Marius Schubert. Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[150] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[151] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[152] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[153] Jörg Sander, Bob D de Vos, Jelmer M Wolterink, and Ivana Išgum. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094919. International Society for Optics and Photonics, 2019.

[154] Patrick Schelb, Simon Kohl, Jan Philipp Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Sebastian Bickelhaupt, Tristan Anselm Kuder, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Klaus H Maier-Hein, and David Bonekamp. Classification of cancer at prostate MRI: Deep learning versus clinical PI-RADS assessment. *Radiology*, page 190938, 2019.

[155] Ivo G Schoots, Monique J Roobol, Daan Nieboer, Chris H Bangma, Ewout W Steyerberg, and MG Myriam Hunink. Magnetic resonance imaging–targeted biopsy may enhance the diagnostic accuracy of significant prostate cancer detection compared to standard transrectal ultrasound-guided biopsy: a systematic review and meta-analysis. *European Urology*, 68(3):438–450, 2015.

[156] Jarrel CY Seah, Jennifer SN Tang, and Andy Kitchen. Detection of prostate cancer on multiparametric MRI. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013429. International Society for Optics and Photonics, 2017.

[157] Suman Sedai, Bhavna Antony, Dwarikanath Mahapatra, and Rahil Garnavi. Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using Bayesian deep learning. In *Computational Pathology and Ophthalmic Medical Image Analysis*, pages 219–227. Springer, 2018.

[158] Suman Sedai, Bhavna Antony, Ravneet Rai, Katie Jones, Hiroshi Ishikawa, Joel Schuman, Wollstein Gadi, and Rahil Garnavi. Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 282–290. Springer, 2019.

[159] Alireza Sedghi, Alireza Mehrtash, Amoon Jamzad, Amel Amalou, William M Wells III, Tina Kapur, Jin Tae Kwak, Baris Turkbey, Peter Choyke, Peter Pinto, et al. Improving detection of prostate cancer foci via information fusion of MRI and temporal enhanced ultrasound.

International Journal of Computer Assisted Radiology and Surgery, 2020.

[160] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9030–9038, 2019.

[161] Vijay Shah, Baris Turkbey, Haresh Mani, Yuxi Pang, Thomas Pohida, Maria J Merino, Peter A Pinto, Peter L Choyke, and Marcelino Bernardo. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. *Medical physics*, 39(7):4093–4103, 2012.

[162] Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. In *Advances in Neural Information Processing Systems*, pages 7375–7385, 2018.

[163] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[164] M Minhaj Siddiqui, Soroush Rais-Bahrami, Baris Turkbey, Arvin K George, Jason Rothwax, Nabeel Shakir, Chinonyerem Okoro, Dima Raskolnikov, Howard L Parnes, W Marston Linehan, et al. Comparison of MR/ultrasound fusion–guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer. *JAMA*, 313(4):390–397, 2015.

[165] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30, 2020.

[166] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[167] SE Song, NB Cho, II Iordachita, P Guion, Fichtinger G, A Kaushal, K Camphausen, and LL Whitcomb. Biopsy catheter artifact localization in MRI-guided robotic transrectal prostate intervention. *IEEE Transactions on Biomedical Engineering*, 59(7):1902–11, 2012.

[168] Yang Song, Yu-Dong Zhang, Xu Yan, Hui Liu, Minxiong Zhou, Bingwen Hu, and Guang Yang. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. *Journal of Magnetic Resonance Imaging*, 2018.

[169] 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.

[170] Thomas A Stamey, Fuad S Freiha, John E McNeal, Elise A Redwine, Alice S Whittemore, and Hans-Peter Schmid. Localized prostate cancer. relationship of tumor volume to clinical significance for treatment of prostate cancer. *Cancer*, 71(S3):933–938, 1993.

[171] Susan Standring. *Gray's Anatomy : The Anatomical Basis of Clinical Practice*. Gray's Anatomy. Elsevier Health Sciences, 41 edition, 2016.

[172] Carole H Sudre, Beatriz Gomez Anson, Silvia Ingala, Chris D Lane, Daniel Jimenez, Lukas Haider, Thomas Varsavsky, Ryutaro Tanno, Lorna Smith, Sébastien Ourselin, et al. Let's agree to disagree: Learning highly debatable multirater labelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 665–673. Springer, 2019.

[173] Carole H Sudre et al. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer, 2017.

[174] Sharmin Sultana, Jason Blatt, Benjamin Gilles, Tanweer Rashid, and Michel Audette. MRI-based medial axis extraction and boundary segmentation of cranial nerves through discrete deformable 3D contour and surface models. *IEEE Transactions on Medical Imaging*, 2017.

[175] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. arxiv 2015. *arXiv preprint arXiv:1512.00567*, 1512, 2015.

[176] N. Tajbakhsh, J. Y Shin, S. R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[177] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised CNN segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2018.

[178] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019.

[179] Clare Tempany, Jagadeesan Jayender, Tina Kapur, Raphael Bueno, Alexandra Golby, Nathalie Agar, and Ferenc A Jolesz. Multimodal imaging for improved diagnosis and treatment of cancers. *Cancer*, 121(6):817–827, 2015.

[180] Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. In *International Conference on Machine Learning*, pages 4914–4923, 2018.

[181] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 13888–13899, 2019.

[182] Gaurie Tilak, Kemal Tuncali, Sang-Eun Song, Junichi Tokuda, Olutayo Olubiyi, Fiona Fennessy, Andriy Fedorov, Tobias Penzkofer, Clare Tempany, and Nobuhiko Hata. 3T MR-guided in-bore transperineal prostate biopsy: A comparison of robotic and manual needle-guidance templates. *Journal of Magnetic Resonance Imaging*, 42(1):63–71, 2015.

[183] Baris Turkbey, Andrew B. Rosenkrantz, Masoom A. Haider, Anwar R. Padhani, Geert Villeirs, Katarzyna J. Macura, Clare M. Tempany, Peter L. Choyke, François Cornud, Daniel J. A. Margolis, Harriet C. Thoeny, Sadhna Verma, Jelle O. Barentsz, and Jeffrey C Weinreb. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European Urology*, 2019.

[184] M Uherčík, J Kybic, H Liebgott, and C Cachard. Model fitting using RANSAC for surgical tool localization in 3-D ultrasound images. *IEEE Transactions on Biomedical Engineering*, 57(8):1907–1916, 2010.

[185] A. G. van Norden, K. F. de Laat, R. A. Gons, I. W. van Uden, E. J. van Dijk, L. J. van Oudheusden, R. A. Esselink, B. R. Bloem, B. G. van Engelen, M. J. Zwarts, I. Tendolkar, M. G. Olde-Rikkert, M. J. van der Vlugt, M. P. Zwiers, D. G. Norris, and F. E. de Leeuw. Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol. *BMC Neurology*, 11:29, 2011.

[186] A. Van Opbroek, M A. Ikram, M. W Vernooij, and M. De Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Transactions on Medical Imaging*, 34(5):1018–1030, 2015.

[187] Sadhna Verma, Peter L Choyke, Steven C Eberhardt, Aytekin Oto, Clare M Tempany, Baris Turkbey, and Andrew B Rosenkrantz. The current state of MR imaging-targeted biopsy techniques for detection of prostate cancer. *Radiology*, 285(2):343–356, 2017.

[188] P C Vos, J O Barentsz, N Karssemeijer, and H J Huisman. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. *Physics in Medicine & Biology*, 57(6):1527–1542, 2012.

[189] Juan Wang, Huanjun Ding, FateMeh Azamian, Brian Zhou, Carlos Iribarren, Sabee Molloi, and Pierre Baldi. Detecting cardiovascular disease from mammograms with deep learning. *IEEE Transactions on Medical Imaging*, 2017.

[190] Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang, and Kwang-Ting Cheng. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. *IEEE Transactions on Medical Imaging*, 37(5):1127–1139, 2018.

[191] William M Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, 1996.

[192] Rogier R Wildeboer, Ruud JG van Sloun, Hessel Wijkstra, and Massimo Mischi. Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods. *Computer Methods and Programs in Biomedicine*, 189:105316, 2020.

[193] Onno Wink, Wiro J Niessen, and Max A Viergever. Multiscale vessel tracking. *IEEE Transactions on Medical Imaging*, 23(1):130–133, 2004.

[194] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Automatic segmentation and disease classification using cardiac cine MR images. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 101–110. Springer, 2017.

[195] Tineke Wolters, Monique J Roobol, Pim J van Leeuwen, Roderick CN van den Bergh, Robert F Hoedemaeker, Geert JLH van Leenders, Fritz H Schröder, and Theodorus H van der Kwast. A critical analysis of the tumor volume threshold for clinically insignificant prostate cancer using a data set of a randomized screening trial. *The Journal of Urology*, 185(1):121–125, 2011.

[196] David A Woodrum, Akira Kawashima, Krzysztof R Gorny, and Lance A Mynderse. Targeted prostate biopsy and MR-guided therapy for prostate cancer. *Abdominal Radiology*, 41(5):877–888, 2016.

[197] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

[198] Han Zheng, Lanfen Lin, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Ruofeng Tong, and Jian Wu. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 148–156. Springer, 2019.

[199] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 593–602, 2019.