Exploiting Supplementary Data and Knowledge for Improved CNN-based Segmentation of Prostate Structures in T2-weighted MRI

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieurin (Dr.-Ing.)

angenommen durch die Fakultät für Informatik der
Otto-von-Guericke-Universität Magdeburg

von Anneke Meyer, M. Sc.
geb. am 18.03.1990 in Nienburg/Weser

Gutachterinnen/Gutachter
Prof. Dr. Christian Hansen
Prof. Dr.-Ing. Dorit Merhof
Prof. Andrey Fedorov, Ph.D.

Magdeburg, den 31. März 2022

ANNEKE MEYER

EXPLOITING SUPPLEMENTARY DATA AND KNOWLEDGE FOR IMPROVED CNN-BASED SEGMENTATION OF PROSTATE STRUCTURES IN T2-WEIGHTED MRI

DISSERTATION

Supervisors

Prof. Dr. Christian Hansen

Dr.-Ing. Marko Rak

ACKNOWLEDGEMENTS

This dissertation bases on research that has been conducted at the Otto-von-Guericke University Magdeburg and the Research Campus STIMULATE in Magdeburg, Germany.

The work has been funded by the EU and the federal state of Saxony-Anhalt, Germany, and by the European Regional Development Fund under the operation numbers ZS/2016/04/78123 and ZS/2016/08/80388. The Titan X Pascal GPU used for this research was donated by the NVIDIA Corporation. Moreover, the author of this thesis received a travel grant from the Deutscher Akademischer Austauschdienst (DAAD) 'Programm Kongressreisen' and a 3-months stipend from the Brigham and Women's Hospital (BWH) Boston, MA, USA.

ABSTRACT

Magnetic resonance imaging (MRI) is gaining increasing importance for the diagnosis and treatment of prostate cancer (PCa). One integral part in the analysis of MRI scans is the segmentation of prostate structures, which are needed for multiple tasks in clinical assessment of PCa, and for the planning and monitoring of therapeutic interventions.

Convolutional neural networks (CNNs) have proven to be the top choice for many computer vision tasks, including medical image analysis. Consequently, a large body of research has been carried out on CNN-based segmentation of the prostate whole gland and its subdivision into two anatomical zones: the peripheral zone (PZ) and the transition zone (TZ). Far less research has been conducted on the segmentation of other structures that are relevant in PCa assessment and treatment planning. In this thesis, we set out to close this gap by investigating not only an improved segmentation of the whole gland, but extending the automatic segmentation to a more detailed division of the interior prostate gland, and to adjacent structures that are relevant for reducing the risks of adverse therapy side effects.

In this context, we contribute novel methods that leverage supplementary data from different levels of clinical datasets to improve the accuracy and robustness of CNN algorithms for prostate structure segmentation. With our work, we aim to mitigate challenges in their development with respect to prostate structures segmentation in specific, and CNN-based methods in general. These challenges include the quality of underlying images, the necessity of a large amount of labeled training data, and the performance drop due to domain shift.

To overcome the lower image quality in parts of the prostate on axial MRI scan directions, we propose a 3D anisotropic multi-stream CNN. Our method improves the segmentation performance for the prostate by allowing for incorporation of multiple scan directions. Moreover, we contribute a novel, semi-supervised learning algorithm to leverage unlabeled data for improving the segmentation outcomes and reducing the CNN's demand for labeled data. Lastly, we exploit that, although the CNN's performance drops on data from different distributions, its knowledge can be used to improve in the new domain. We introduce a simple yet effective semi-supervised domain adaptation technique that improves the segmentation quality in the new domain with only small amounts of labelled data. With our proposed methods, this thesis takes a further step towards reliable automatic segmentation of prostate structures. Thereby, we do not only focus on the improvement of the CNN algorithms, but we also introduce means to make the methods more applicable in practice.

ZUSAMMENFASSUNG

Die Magnetresonanztomographie (MRT) gewinnt zunehmend an Bedeutung für die Diagnose und Behandlung von Prostatakarzinomen. Ein wesentlicher Bestandteil der Analyse von MRT-Bildern ist die Segmentierung von Prostatastrukturen. Diese werden für verschiedene Aufgaben bei der klinischen Beurteilung von Prostatakarzinomen sowie für die Planung und Überwachung fokaler und lokoregionaler therapeutischer Eingriffe benötigt.

Convolutional Neural Networks (CNNs) haben sich als primäre Lösung für viele Aufgaben im Bereich der Computer Vision erwiesen. Dies schließt auch die medizinische Bildanalyse mit ein. Folglich wurden zahlreiche Forschungsarbeiten zur CNN-basierten Segmentierung der gesamten Prostata und ihre Unterteilung in zwei anatomische Zonen (periphere Zone (PZ) und Übergangszone (TZ)) entwickelt. Weit weniger erforscht wurde die Segmentierung anderer Strukturen, die für die Beurteilung und die Behandlungsplanung von Prostatakrebs relevant sind. In dieser Arbeit beabsichtigen wir diese Lücke zu schließen, indem wir nicht nur eine verbesserte Segmentierung der gesamten Prostata anvisieren, sondern die automatische Segmentierung auf eine detailliertere Unterteilung der inneren Prostata ausweiten. Darüber hinaus weiten wir die Segmentierung auf benachbarte Strukturen aus, die für die Reduktion von Therapienebenwirkungen relevant sind.

In diesem Zusammenhang stellen wir neue Methoden vor, die zusätzliche Daten von verschiedenen Ebenen klinischer Datensätze nutzen, um die Genauigkeit und Robustheit von CNN-Algorithmen zur Segmentierung der Prostatastruktur zu verbessern. Mit unserer Arbeit zielen wir darauf ab, Herausforderungen bei der Entwicklung und Anwendung von CNN-Algorithmen in Hinblick auf Prostatasegmentierung im Speziellen und CNN-Methoden im Allgemeinen, zu verringern. Zu diesen Herausforderungen gehören die Qualität der zugrundeliegenden Bilder, die Notwendigkeit großer Mengen an gelabelten Trainingsdaten sowie Performanceeinbußen aufgrund des sogenannten Domain-Shifts.

Um die geringere Bildqualität in Teilen der Prostata bei axialen MRT-Scanrichtungen zu kompensieren, stellen wir ein anisotropes 3D-Multistream-CNN vor. Unsere Methode verbessert die Segmentierungsqualität für die Prostata, indem es die Einbeziehung mehrerer Scanrichtungen ermöglicht. Darüber hinaus führen wir einen neuartigen semi-supervised Algorithmus ein, der ungelabelte Daten zur Verbesserung der Segmentierungsergebnisse nutzt und somit den Bedarf des CNN an annotierten Daten reduziert. Des Weiteren machen wir uns zunutze, dass die Performance des CNN bei Daten aus verschiedenen Verteilungen zwar abnimmt, dessen Wissen aber genutzt werden kann,

um die Ergebnisse auf neuartigen Daten zu verbessern. Wir stellen ein einfaches aber effektives semi-supervised Verfahren zur sogenannten Domain Adaptation vor, das die Segmentierungsqualität in der neuen Domain mit einer kleinen Menge an gelabelten Daten verbessert.

Mit den von uns entwickelten Methoden leistet diese Arbeit einen Beitrag für eine zuverlässigere automatische Segmentierung von Prostatastrukturen. Dabei haben wir unseren Fokus nicht nur auf die Verbesserung von quantitativen Ergebnisse der CNN-Algorithmen gelegt, sondern auch auf die Verbesserung ihrer Anwendung in der Praxis.

CONTENTS

LIST OF ACRONYMS

ABD	average boundary distance
ADC	apparent diffusion coefficient
AFS	anterior fibromuscular stroma
BPH	benign prostatic hyperplasia
CAD	computer-aided diagnosis
CNN	convolutional neural network
CT	computer tomography
CZ	central zone
DA	domain adaptation
DCE	dynamic contrast-enhanced
DL	deep learning
DPU	distal prostatic urethra
DSC	Dice similarity coefficient
DWI	diffusion-weighted imaging
EUS	external urethral sphincter
FCN	fully convolutional neural network
GPU	graphical processing unit
HD	Hausdorff distance
MC	Monte Carlo
mpMRI	multi-parametric magnetic resonance imaging
MRI	magnetic resonance imaging
NVB	neurovascular bundles
PCa	prostate cancer
PI-RADS	Prostate Imaging - Reporting and Data System
PSA	prostate-specific antigen
PZ	peripheral zone
SSL	semi-supervised learning
TRUS	transrectal ultrasound
T1w	T1-weighted
T2w	T2-weighted
TL	transfer learning
TZ	transition zone
UATS	uncertainty aware temporal self-learning

1 INTRODUCTION

1.1 MOTIVATION

In many Western countries, such as the United States of America and Germany, prostate cancer (PCa) is the cancer most frequently diagnosed in men (Siegel et al., 2020). The widespread application of screening methods in industrialized nations has led to an increased detection rate of early-stage cancer and thus decreased the mortality rate of PCa. On the other hand, this screening also enhances the detection of less aggressive and slow-growing tumors which may not cause any harm. Consequently, the early detection rates have led to a discussion about overdiagnosis and overtreatment, which can come with risks and complications for the patient who can then die of causes other than the cancer itself (Loeb et al., 2014).

Thanks to its high soft-tissue contrast, clinical workflows of prostate cancer increasingly involve multi-parametric magnetic resonance imaging (mpMRI) to enhance the diagnosis, localization, staging and therapy of PCa. For example, tissue biopsies are still the standard of care for diagnosis, and mpMRI as a planning and guiding tool can increase the diagnostic accuracy and reduce unnecessary biopsies (Verma et al., 2017; Leest et al., 2019). By supporting a more precise characterization of the disease, the employment of mpMRI leads to an improved risk stratification of patients. Therefore, instead of treating the whole gland aggressively with adverse side effects negatively impacting the quality of life, therapy alternatives like active surveillance or focal therapy can be considered (Litwin and Tan, 2017).

With the widespread use of magnetic resonance imaging (MRI) in the clinical routine, the robust and reliable automatic analysis of MRI images gains increasing importance. Deep learning (DL) techniques, and convolutional neural networks (CNNs) in particular, are nowadays the top performers in the medical image analysis field (Litjens et al., 2017b) and have the potential to improve, accelerate and automate different tasks in the clinical routine, for example in diagnosis (Esteva et al., 2017), treatment planning and monitoring (Wang et al., 2020a; Laukamp et al., 2019), as well as patient and physician education (Seok et al., 2021; Engelhardt et al., 2018). One integral step in several clinical and research workflows for PCa is the segmentation of the prostate, and its interior and adjacent structures on T2-weighted (T2w) MRI scans. Segmenting the structures manually is very time-consuming and requires medical expertise. Moreover, it is subject to variations

among different annotators. Therefore, methods that obtain accurate and reliable automatic segmentation results are highly desired.

LIMITATIONS OF CURRENT METHODS While the research on automatic DL-based segmentation of prostate structures is an active research field, and has yielded results in the range of human readers, the proposed algorithms have limitations and challenges regarding prostate structures segmentation in specific, and CNN-based methods in general. The limitations that we set out to overcome in our work include (1) previous segmentation methods' neglect of a more fine-grained interior and adjacent anatomy of the prostate, (2) their input relying only on image data that suffers from lower quality for the extreme parts of the prostate, (3) CNNs' general demand for large quantities of labeled training data and (4) their performance drop on unseen datasets.

There exist further challenges and limitations of CNNs, such as (uncalibrated) overconfident predictions (Mehrtash et al., 2020) and the incapability of life-long learning without forgetting when moving from one task to another (Singh et al., 2020). However, in this thesis, we target the four shortcomings listed above:

1. Previous works on prostate segmentation have mainly focused on the segmentation of the whole gland and its two major zones: the *transition zone* (TZ) (central gland) and the *peripheral zone* (PZ). However, with an increasing use of MRI in various applications, the consideration of other structures becomes relevant. A more detailed analysis of the inner anatomy of the prostate could, for example, provide better landmarks for correlating MRI data with other imaging modalities as histopathology (Kwak et al., 2016), and enable a more standardized reporting of prostate exams (Turkbey et al., 2019). Also, automatic segmentations of critical structures for PCa treatment can potentially automate a better planning of surgery or radiation therapy (Wake et al., 2020) that could reduce the side effects and risks of PCa therapy (Nguyen et al., 2017; Mungovan et al., 2017).

2. Methods have thus far mainly relied on the axial T2w scan of the prostate. The T2w acquisition allows for a good distinction of anatomy, but suffers from partial volume effects due to its high slice thickness. Consequently, relying only on one scan direction prevents the exact distinction of the gland boundaries in parts of the prostate.

3. As is the case for all DL-based methods, the CNNs need to be trained with a large amount of labeled data. While researchers working on the analysis of other image types (e.g., street scenes or text recognition) can resort to crowd-sourcing tools (Kovashka et al., 2016), it is at least unclear how this can be carried out in

the medical domain. Here, good-quality labeled data is costly to obtain and needs to involve medical experts.

4. One other major drawback of modern neural networks is their problem with the so-called domain shift, which causes models trained on a dataset from one domain, to substantially degrade in performance on data from another domain. In the medical context, a domain can be another scanner, site or imaging protocol. To circumvent the need to create a new training dataset for each new domain, different domain adaptation methods or training variants have been proposed to achieve higher robustness of the models. However, they largely require either collecting data from a variety of domains, or the data from the original domain needs to be available, which is often impractical due to the costly data annotation and privacy concerns of medical data.

1.2 SCOPE AND CONTRIBUTIONS

In this thesis, we address the limitations described above. Besides considering the whole gland and its two major zones, we additionally targeted other structures in the course of this thesis that have not yet been investigated for automatic segmentation.

Specifically, we examined the feasibility of segmenting a more detailed anatomy of the prostate, extending the two-zones segmentation by the *anterior fibromuscular stroma* (AFS) and the *distal prostatic urethra* (DPU), which is enclosed by the prostate. Moreover, we apply CNN-based segmentation to critical structures for PCa treatment, namely the *external urethral sphincter* (EUS) and the *neurovascular bundles* (NVB), whose damage is correlated with urinary complications and sexual dysfunction.

In this context, we introduce novel methods that exploit different types of data, that are easily available in clinical workflows to improve the methods' performance. Basically, we investigated using data originating from three different levels of clinical datasets:

- *Patient-level:* According to the standard protocol for mpMRI acquisitions (Turkbey et al., 2019), it is essential to acquire not only the axial T2w scan, but at least one additional scan direction. We studied the incorporation of additional scan directions (i.e., *multi-planar data*) to reduce segmentation errors in regions with high partial volume effect.

- *Intra-domain-level:* As labelling data is expensive and tedious, unlabeled data is easier to obtain and more often available from clinical partners. Therefore, we developed a semi-supervised segmentation method, and investigated how leveraging extra *unlabeled data* of the same domain can support an improved segmentation.

- *Inter-domain-level*: Although performance of CNN methods degrades due to the domain shift, there is still valuable *knowledge from the external domain* available in the models (that were trained on the original domain data). We studied how this information can be exploited to improve segmentation in a new domain without the necessity to access the original domain data.

Although our focus for all methods was on the application of prostate MRI, we investigated the generalization capabilities for the algorithms developed to exploit unlabeled data and external domain knowledge on other tasks and types of data. We could demonstrate that the methods can be beneficial to other problems as other segmentation targets and other modalities.

STRUCTURE OF THE THESIS In this work, we investigated the usage of supplementary data and knowledge for an improved segmentation of different prostate structures. To describe the context and our methods, this thesis is structured as follows:

- Chapter 2 introduces the medical background and technical fundamentals for this work.
- Chapter 3 outlines our proposed multi-stream CNN architecture to incorporate multi-planar data in order to improve segmentation performance for the prostate gland.
- Chapter 4 presents a semi-supervised segmentation method, that adds unlabeled data for the task of a detailed zonal anatomy segmentation for the prostate.
- Chapter 5 describes a domain-adaptation method for the segmentation of critical structures for PCa therapy that relaxes the requirement of original domain data being available.
- Chapter 6 concludes this thesis by giving a brief recap of the methods and contributions of our work and discussing their limitations and potential future work.

2

PRELIMINARIES

In this chapter, we provide the fundamentals of the medical and technical background for a better framing of the proposed methods and their clinical context. We cover the medical background by including details about the prostate anatomy, PCa, and prostate MRI in Section 2.1.

With respect to the technical fundamentals, in Section 2.2, we first outline main CNN architectures that are encountered for medical image segmentation, and cover methods that are used to measure the uncertainty of network predictions. We continue with basics of semi-supervised learning (SSL) techniques (Section 2.3), which we employ in the methods of Chapters 4 and 5. Lastly, in Section 2.4 we provide details about methodological concepts used throughout this work, which include evaluation measures and implementation details about software and packages used.

2.1 MEDICAL BACKGROUND

In the following, we cover the medical background for our work. This section starts with an introduction into the anatomy of the prostate and its surrounding structures in Section 2.1.1, and continues with information on PCa and its diagnosis and therapies (Section 2.1.2). The medical background is concluded with a summary of the characteristics of MRI scans that are acquired for PCa detection and therapy planning (Section 2.1.3).

2.1.1 *Prostate Anatomy*

The prostate is a walnut-sized fibromuscular gland in the male reproductive system. The secretion produced in the prostatic gland makes up together with the secretion from the seminal vesicles the main part of seminal fluid (Standring et al., 2016). The gland surrounds the prostatic urethra and is located below the bladder and anterior to the rectum (see Figure 2.1), through which it can be palpated. The superior part of the prostate (*base*) is contiguous with the bladder neck. The inferior part (*apex*) encloses the connection of the prostatic and the membranous urethra (Standring et al., 2016) and is contiguous with the pelvic diaphragm (Hautmann and Gschwend, 2014). The prostate lacks a true histological capsule, but it is enclosed by an outer band of fibromuscular tissue that is incomplete anteriorly and at the apex (Turkbey et al., 2019).

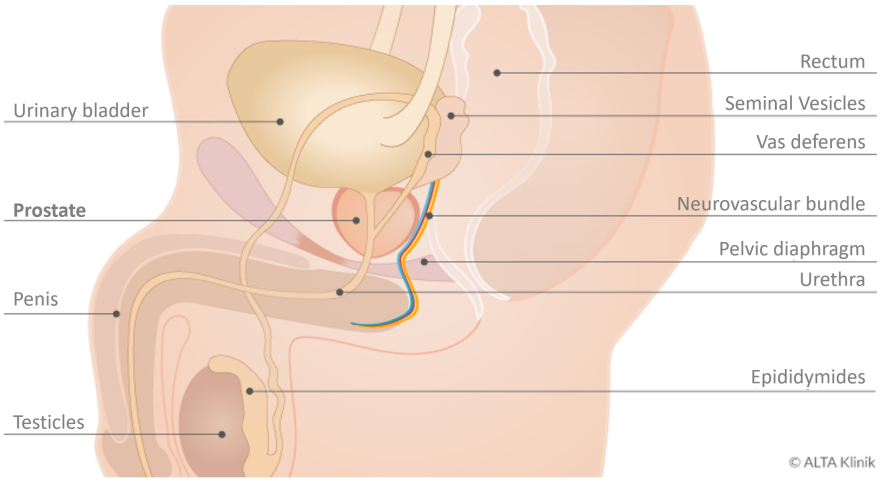


Figure 2.1: Location of the prostate in the male reproductive system. Image courtesy of ALTA Klinik GmbH, Bielefeld, Germany.

The prostatic urethra is dilated in the center of the prostate. This part of the urethra is named verumontanum (or seminal colliculus). Here, the ejaculatory ducts, which drain seminal fluid from the seminal vesicles and spermatozoa from the testis (via vas deferens), join the urethra. The prostatic urethra is enclosed by two different sphincter mechanisms (see Figure 2.2). The internal urethral sphincter is located at the bladder neck. Its main role is to prevent retrograde ejaculation into the bladder (Jacob, 2008). The EUS is located below the apex of the prostate within the pelvic diaphragm. It may also extend into the apex, depending on the individual shape of the men's prostate (Lee et al., 2006). The EUS has a key role in maintaining urinary continence (Jacob, 2008).

Nerves and vessels, which supply the prostate and neighboring structures, run posterolateral to the prostate. However, this NVB is not an anatomically defined cord, but rather a complex network ('veil') that 'embraces' the gland posteroranterior, thinning out in the anterior direction (Hautmann and Gschwend, 2014). As the NVB continues to be the cavernous nerves that facilitate penile erection, the damage of the bundles during surgery can cause impotence (Standring et al., 2016).

Different concepts for the schematic division of the prostate exist. We follow the histological division by McNeal (1981). In this, the prostate is comprised of four histological zones: the TZ, the central zone (CZ), the PZ and the AFS (see Figure 2.3). The CZ encloses the ejaculatory ducts, and makes up approximately 20 % of prostate's volume in normal anatomy of adult men aged younger than 40 years. The TZ surrounds the urethra proximal to the verumontanum and accounts for only 5% of the volume. With 70% of prostate's volume, the PZ is the largest zone. It is a cup-shaped structure that defines the apex of the prostate and encloses the transition zone (Standring et al., 2016). The anterior part of the prostate is covered by the AFS, which is a non-glandular

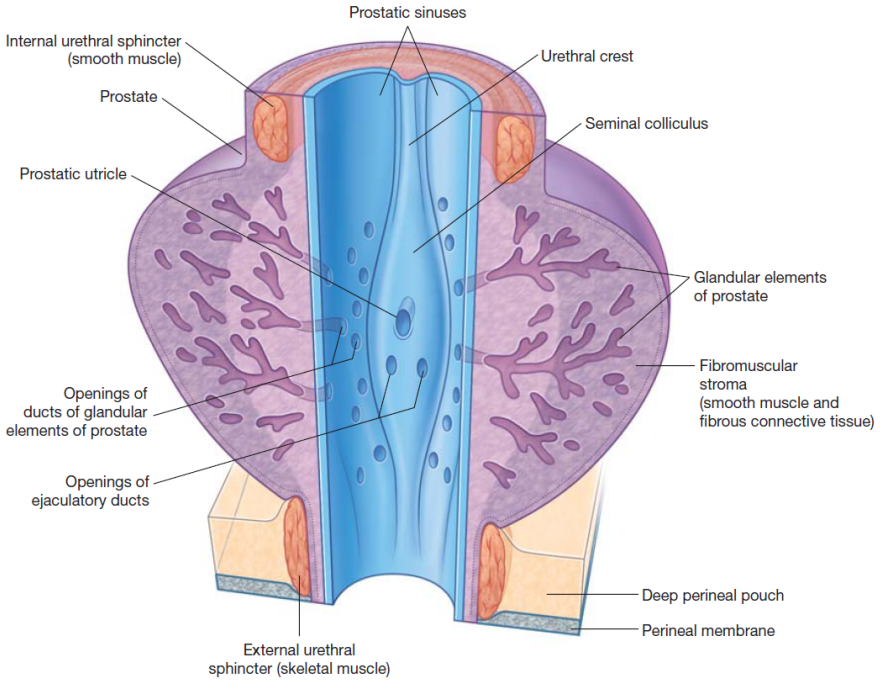


Figure 2.2: Prostate in coronal section with internal and external urethral sphincter. Figure from Drake et al. (2010), reprinted with permission from Elsevier.

fibromuscular structure that extends from the apex to the bladder neck (Standring et al., 2016).

With the age of 45 to 50 years, the size of the prostate extends due to benign prostatic hyperplasia (BPH). While the gland weighs approximately 8 g in youth, its weight may range between 40 g and 150 g when the prostate enlarges in the course of BPH (Standring et al., 2016). Since this age-related condition is caused by expansion of the TZ, BPH affects the overall ratio of the zone's volume. When the benign enlargement develops, the TZ will make up for an increasing amount of the gland (Turkbey et al., 2019). Consequently, the other zones get compressed and the CZ can often not be distinguished anymore from the PZ in MRI scans.

2.1.2 *Prostate Cancer*

With 191,930 new cases and more than 33,000 estimated deaths in the United States of America, PCa is the most common type of cancer and is the second most deadly cancer among men (Siegel et al., 2020). However, thanks to the recent advances in cancer treatment and diagnosis, the death rate decreased by more than 50% since 1993, and the 5-year survival rate for all stages combined is currently 98% for men in the United States of America (Siegel et al., 2020).

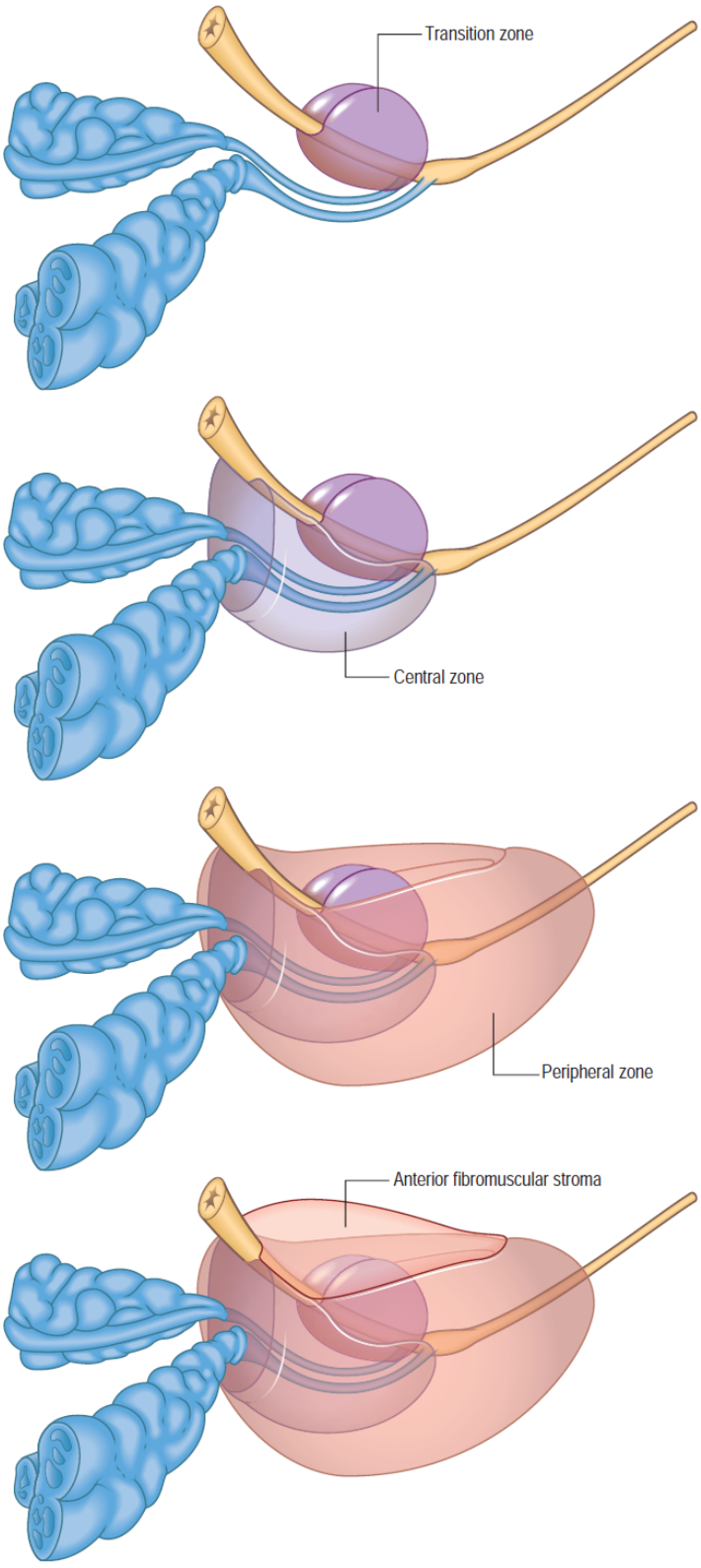


Figure 2.3: The four zones of the prostate: transition zone, central zone, peripheral zone and anterior fibromuscular stroma. Figure from Standring et al. (2016) and Wein et al. (2012), reprinted with permission from Elsevier.

Among the different factors that are discussed in the literature to elevate the risk of PCa, age is considered as the main risk factor for developing PCa. For example, the average age of diagnosis for prostate cancer worldwide is 66 years (Rawla, 2019).

Approximately 70 - 75 % of PCa arises in the PZ, and 20 - 30 % in the TZ (Turkbey et al., 2019). It occurs in a majorly multifocal manner, often with varying grades of malignancy in the different tumors (Hautmann and Gschwend, 2014). In the course of growing, the cancer often penetrates through the prostate capsule in the area where small nerve branches enter the prostate in the base and apex (Turkbey et al., 2019). In higher stages, the tumor may infiltrate proximal structures such as seminal vesicles, bladder neck, EUS, rectum, levator ani muscle and/or pelvic wall (Hautmann and Gschwend, 2014). In metastatic stages of the disease, the tumor can spread into the proximal lymph nodes, and into the skeletal system (Hautmann and Gschwend, 2014).

The increasing detection rate of early stage tumors makes it crucial to differentiate between *significant* and *insignificant* PCa. Epidemiologically, insignificant PCa is defined as harmless, when, based on the lifetime risk estimates, no symptomatic or clinical PCa will develop (Van der Kwast and Roobol, 2013). Although guidelines for the classification of insignificant PCa exist, the specific parameters and thresholds for its determination are widely discussed regarding overdiagnosis and overtreatment reduction (Van der Kwast and Roobol, 2013).

DIAGNOSIS In advanced stages, patients with PCa may develop symptoms as problems with urination (because the tumor may obstruct the urethra) or skeletal pains from metastases (Hautmann and Gschwend, 2014). If interpreted correctly, they can indicate the presence of PCa. Also, some tumors can be palpable during digital rectal examination (Hautmann and Gschwend, 2014).

The consideration of the prostate-specific antigen (PSA) value as an indicator is becoming more and more common and is the most important parameter for the detection of early stage PCa (Gasser, 2015). On the other hand it is widely discussed with respect to overdiagnosis (Litwin and Tan, 2017). The PSA value can also elevate due to inflammation or BPH. As its level correlates with the size of the prostate, the volume of the prostate or the transition zone should be taken into account for an improved interpretation of the PSA value (Gasser, 2015).

In recent years, mpMRI has gained increasing importance in the enhanced diagnosis of prostate cancer. The final diagnosis of PCa, however, can only be made based on the histological evaluation of prostate tissue gathered during needle biopsy (Litwin and Tan, 2017). Conventionally, schematic biopsies with transrectal ultrasound (TRUS) guidance is carried out as standard of care (Litwin and Tan, 2017). The precision of biopsies can be improved by targeted biopsies that rely on the guidance of MRI, such as MRI-TRUS fused biopsy, cognitive

biopsy after a visual review of MRI or in-bore MRI percutaneous biopsy (Litwin and Tan, 2017).

TREATMENT Depending on the grade of PCa and other factors such as life expectancy, different approaches are subsequent to staging. Men with lower life expectancy or low-risk PCa, can be candidates for watchful waiting (relieving symptoms with palliative intent) or active surveillance (recurrent imaging and biopsy with curative intent for men developing significant disease) (Litwin and Tan, 2017).

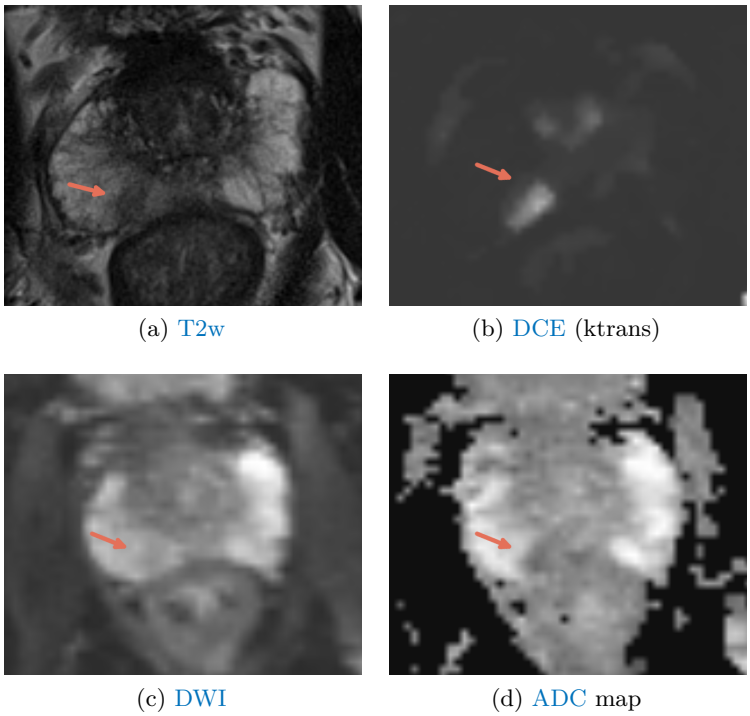
Radical prostatectomy is recommended for patients with localized and non-metastasized PCa that is limited to the prostate gland (Hautmann and Gschwend, 2014). Prostatectomy includes the surgical removal of the prostate and the seminal vesicles. To reduce the risk of complications as erectile dysfunction or incontinence, nerve-sparing or EUS-sparing surgeries can be considered for patients with non-extraprostatic extending tumors (Hautmann and Gschwend, 2014).

An alternative to prostatectomy for patients at this stage is radiation therapy that has the potential to reduce the side effect on sexual function and urinary control. On the other hand, it is often accompanied with nocturial and bowel dysfunctions (Litwin and Tan, 2017). Focal procedures like brachytherapy, high-intensity-focused ultrasound or cryotherapy have less side effects than prostatectomy or radiation therapy (Gasser, 2015; Litwin and Tan, 2017). However, they are not all part of a clinical routine yet (Litwin and Tan, 2017). For more advanced (metastatic) stages of PCa, hormon deprivation therapy and chemotherapy are the main treatment options (Litwin and Tan, 2017).

2.1.3 *Prostate MRI*

To enhance the diagnosis, localization, and therapy planning or guidance of PCa, mpMRI that combines anatomic T2w scans with functional and physiological assessment is gaining increasing importance. The current guidelines for standardizing imaging protocols and the diagnosis of PCa on the basis of mpMRI, are condensed in the Prostate Imaging - Reporting and Data System (PI-RADS) version 2.1 document (Turkbey et al., 2019). According to PI-RADS v2.1, mpMRI for the prostate should include T1-weighted (T1w) and T2w sequences as well as diffusion-weighted imaging (DWI) and dynamic contrast-enhanced (DCE) images (see Figure 2.4). The images should be obtained by 1.5 or 3 tesla scanners with endorectal or external (surface) phased array coils or the concurrent use of both coil types.

The purpose of T1w sequences is the detection of hemorrhage in the prostate or seminal vesicles, and skeletal or nodal metastases when combined with a contrast agent (Turkbey et al., 2019). T2w scans provide anatomic details with high soft tissue contrast which enables the distinction of prostate zones (see Figure 2.5 for examples of T2w scans).

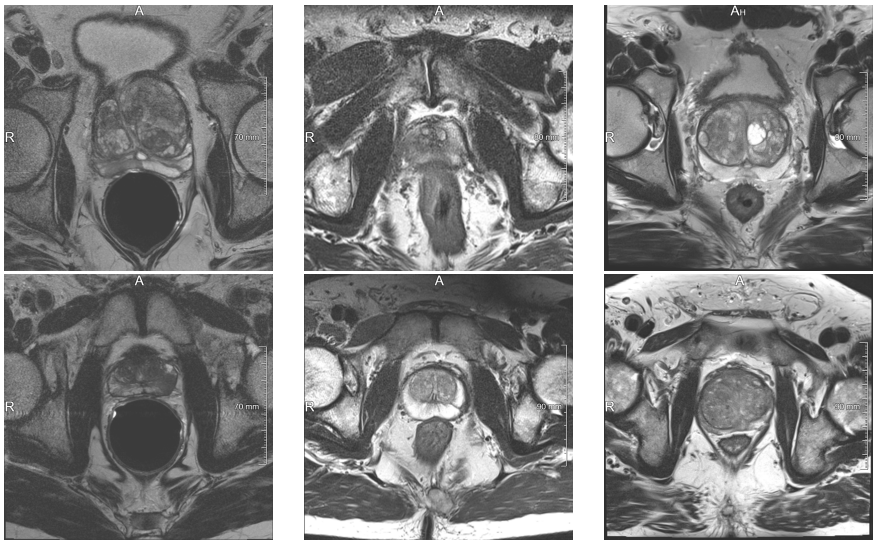


(a) T2w (b) DCE (ktrans)

(c) DWI (d) ADC map

Figure 2.4: mpMRI of the prostate with clinical significant lesion in the peripheral zone (orange arrow).

Moreover, extraprostatic extension of tumors and abnormalities in the prostate tissue may be determined (Turkbey et al., 2019). DWI measures the diffusion motion of water molecules in the tissue. Tumors restrict the motion of the molecules and can thus be differentiated in DWI scans. Apparent diffusion coefficient (ADC) maps that may enhance diagnostic performance, are calculated from DWI scans by acquiring the images with different gradient amplitudes (b-values). DCE MRI requires the injection of a contrast agent. Its temporal acquisition allows for the computation of different dynamic parameters of the agent within the tissue (Somford et al., 2008), e.g. uptake and wash-out parameters of the contrast agent. Because tumors cause increased vascularization (angiogenesis) during their growth, they may be distinguished from healthy tissue in DCE images (Somford et al., 2008). However, as the vascularization of prostate tumors is heterogeneous, DCE is rather considered as a 'back-up' modality and may support the radiologist in detecting smaller lesions (Turkbey et al., 2019). The different scan modalities cover the whole prostate and consist of multiple axial 2D slices, which are acquired gap-free with higher in-plane resolution when compared to the slice-thickness. For T2w images, there should be additional acquisitions in at least one other orthogonal plane (coronal and/or sagittal) (Turkbey et al., 2019). MRI scans of the prostate do not have standardized intensity units as, for example, the Hounsfield unit in computer tomography (CT) scans.



Philips 1.5T, combined endorectal and surface coil

Siemens TrioTim 3T, surface coil

Siemens Skyra 3T, surface coil

Figure 2.5: Example T2w scans. For all examples, an axial slice of the midgland is depicted. One can see the high variation in the scans caused by different scanners, acquistion protocols and medical conditions.

PI-RADS 2.1 assessment includes a five-point scale that expresses the likelihood of clinically significant PCa being present. It takes into account the anatomical zones of the prostate. Depending on the zone, the MRI modalities are interpreted differently for assigning PI-RADS scores. Different benign diseases, such as prostatitis, cysts, BPH, fibrosis, etc., change the appearance of the prostate on MRI and make PCa diagnosis and zonal distinction more challenging (Stabile et al., 2020).

Although diagnostic capabilities in detection and biopsy planning/ execution can be increased through mpMRI, this imaging technique is more often encountered in academic centers and less often in medical practices in the US and Germany (Stabile et al., 2020; Saar et al., 2020). Moreover, its widespread application is impeded by its high cost (Hutchinson and Lotan, 2017). One common problem with mpMRI is the high inter-reader variability despite systems as PI-RADS v2.1 (Stabile et al., 2020). Within the last years, the performance of computer-aided diagnosis (CAD) has increased systematically and can be considered as essential enhancement to human diagnosis (Stabile et al., 2020), having the potential to decrease the inter-reader variability.

2.2 CONVOLUTIONAL NEURAL NETWORKS

Having covered the medical background for this thesis, the following sections provide details on the technical fundamentals, beginning with CNNs and the main CNN architectures applied to medical segmentation.

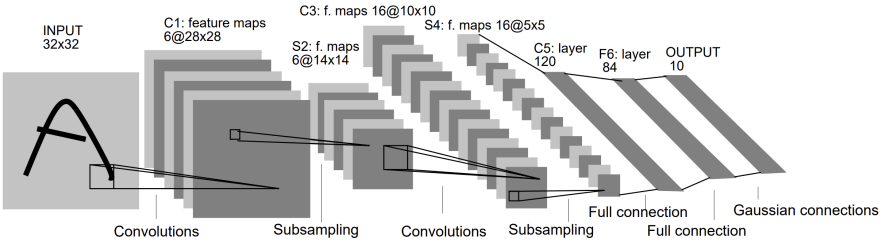


Figure 2.6: Design of a CNN proposed by LeCun et al. (1998). The convolutional kernels are slid along the image in the input layer. The pooling operation ('subsampling') decreases the resolution of the image or feature maps and increases the spatial context for subsequent layers. The final layers in this architecture are fully connected to allow for a classification output. Figure from LeCun et al. (1998), ©1998 IEEE.

CNNs are a class of artificial neural networks that dominate the current field of machine learning for images. They have been originally proposed by LeCun et al. (1989) for handwritten zip code recognition, but have not found their breakthrough until the successful introduction of the AlexNet in 2012 by Krizhevsky et al. (2012) for the ImageNet challenge (Deng et al., 2009). Since then, the top image analysis methods where progressively based on CNNs.

CNNs are conceptually inspired by the visual cortex of humans and are designed to be more efficient for data with grid-like topology, such as images, than other *fully-connected* architectures, where each node in one layer is connected to every node in the immediate previous and next layer. Having individual connections for each pixel in the image is computationally very expensive and makes the network prone to overfitting. Moreover, considering the individual pixel per connection, discards the spatial information of the pixel's neighborhood.

CNNs have two key components: *convolutional* and *pooling* operations (see Figure 2.6). Convolutional operations allow sharing of parameters for different regions of the image, because the convolutional filters, or kernels, are slid along the image and apply the same transformation in different locations, which leads to a drastic reduction of parameters compared to the fully-connected networks. Moreover, it makes CNNs equivariant to translation, which describes that the output changes in the same way as the input (Goodfellow et al., 2016).

The architecture of a CNN is organized in a layer-wise manner. Each layer in the CNN consists of a set of convolutional kernels (filters) K with weights $\mathcal{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_K\}$ and biases $\mathcal{B} = \{b_1, b_2, ..., b_K\}$. Each kernel generates a feature map of its input which is then transformed nonlinearly in an element-wise manner. In a CNN, the kernels of layer l take as input the feature maps $\boldsymbol{X}_{l-1}$ of its previous layer to generate the features for layer l:

$$\boldsymbol{X}_k^l = \sigma(\boldsymbol{W}_k^{l-1} * \boldsymbol{X}_k^{l-1} + b_k^{l-1}) \,, \tag{2.1}$$

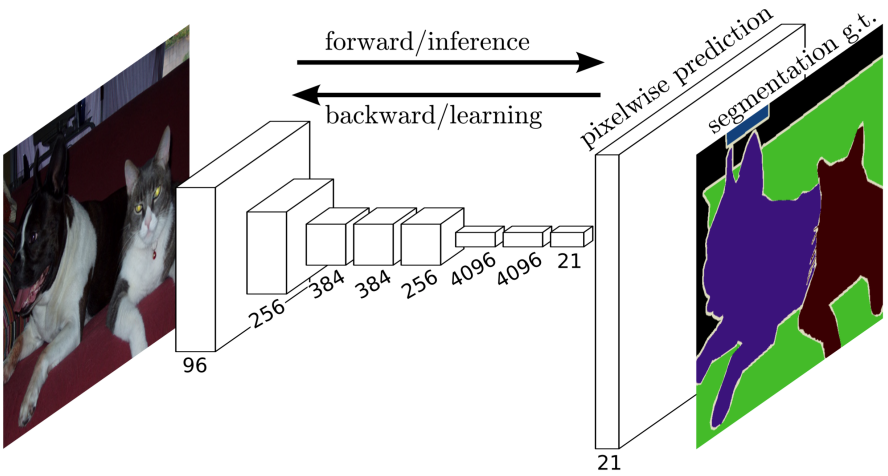


Figure 2.7: Original architecture of the FCN proposed. Image from Long et al. (2015), ©2015 IEEE.

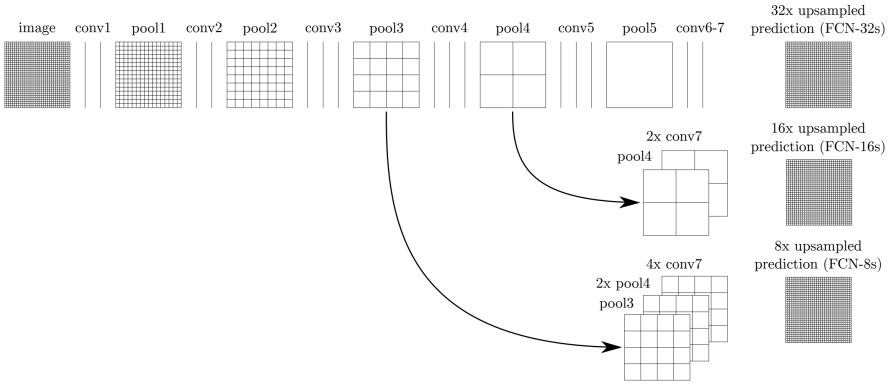


Figure 2.8: Upsampling mechanism for a more detailed output of the FCN. Image from Long et al. (2015), ©2015 IEEE.

where $*$ denotes the convolution operation and σ the nonlinear function.

Pooling operations are applied in CNNs to increase the spatial context throughout the layers. These operations summarize their input (cluster of feature values) into one single value, commonly by extracting the maximum (*max pooling*) or the average (*average pooling*).

The parameters $\mathcal{W}$ and $\mathcal{B}$ in a CNN are learned through optimizing the objective function, the *loss*, by means of the gradient descent optimization. For this, the gradient for the loss with respect to every parameter in the CNN needs to be calculated, which is performed by the *backpropagation* algorithm (Rumelhart et al., 1986).

In the following sections, we outline the most prominent CNN architectures designed for segmentation tasks (Section 2.2.1). Moreover, in Section 2.2.2, we describe means to estimate prediction uncertainty of neural networks that are used in our methods described in Chapters 4 and 5.

2.2.1 *Segmentation Architectures*

For segmentation, early approaches solved the problem of pixel-wise labeling by pixel-wise classification using one patch per pixel. Those methods employed classification networks with convolutional layers in the beginning and fully connected layers in the last part of the network. Due to the fully connected layers, the networks could not exploit the size invariance of the convolutional layers and therefore the images had to be resized to specific dimensions. Furthermore, the methods were rather slow because multiple forward-passes were required per prediction of a dense segmentation map.

To account for these shortcomings, Long et al. (2015) proposed the *fully convolutional neural network* (FCN) in 2015 (see Figure 2.7). For the FCN, Long et al. (2015) exchanged the fully connected layers with convolutional layers, which were subsequently upsampled (32× by fractionally strided convolutions (deconvolution) in one step) to obtain a dense prediction map of the input image size. To forward lower-level but higher-resolution information to the 32× upsampled map, links were introduced that upsample the output of earlier convolution layers (see Figure 2.8). These finer predictions are then added to the final output.

Later in 2015, Ronneberger et al. (2015) proposed the U-Net architecture as an extension to the FCN. The U-Net and variants thereof have so far been the most commonly used network architecture for image segmentation, which is reflected in its high Google citation score ($> 29{,}500$[1]). As this architecture forms the basis for our proposed methods, we explain it in more detail in the following paragraphs.

U-NET The name U-Net is inspired by its U-shaped architecture that originates from the two symmetric paths (see Figure 2.9): the *encoder* (contracting path) which extracts features from the input image and compresses it to less spatial extent and the *decoder* (expansive path) which incrementally upsamples the features back to the input image size. In contrast to the FCN, the decoder has a larger number of filters (feature channels) that should allow the network to "propagate context information to higher resolution layers" (Ronneberger et al., 2015). The encoder and decoder are connected through a bottom layer which holds the latent features of the network.

Both paths consist of different stages that act on different resolution levels. Each stage is a convolutional block which has two convolutional layers with a 3×3 kernel and a stride of 1. The convolutional layers are each followed by a rectified linear unit (ReLU) function introducing nonlinearity into the network:

$$\sigma(x) = max(0, x)\,. \tag{2.2}$$

1 as of August 11th, 2021

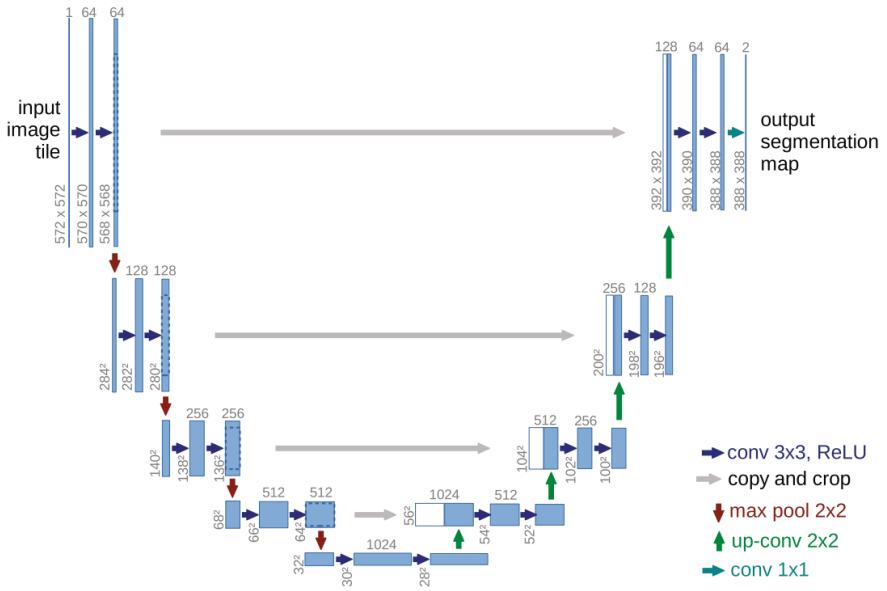


Figure 2.9: Original architecture of the U-Net. The blue boxes represent feature maps with the number of their channels noted on top and their in-plane dimensions noted at their side. Figure from Ronneberger et al. (2015), reprinted with permission from Springer, ©2015.

Each stage in the encoder is followed by a downsampling operation, which increases the receptive field of the network. In the original U-Net architecture, the dowsampling is carried out via 2×2 max pooling with a stride of 2. In the literature U-Net implementations exist that perform the downsampling by means of convolutions with strides larger than one (strided convolutions) instead of max pooling. Furthermore, the number of feature channels is doubled with each stage on the network's encoder and at the bottom most layer to increase the capacity.

Symmetrically, each stage of the decoder begins with an upsampling operation and the number of feature channels is halved in each convolutional block. The upsampling is carried out with deconvolutions in the original U-Net method (Ronneberger et al., 2015) which allows learning a nonlinear upsampling, but it is also common to apply plain bilinear interpolation. The upsampled feature channels are concatenated with the features from the encoder's same resolution level right before feeding them into the convolutional block. Similar to the FCN, this allows to propagate the finer details from earlier stages of the network and enabling a finer segmentation map as output.

The network concludes with a final 1×1 convolution layer that outputs the c channels, with c being the number of classes of the segmentation task. It is followed by a task-specific activation function, typically a sigmoid function for binary segmentation:

$$\text{sigmoid}(z) = \frac{1}{1 + e^{-z}} \,. \tag{2.3}$$

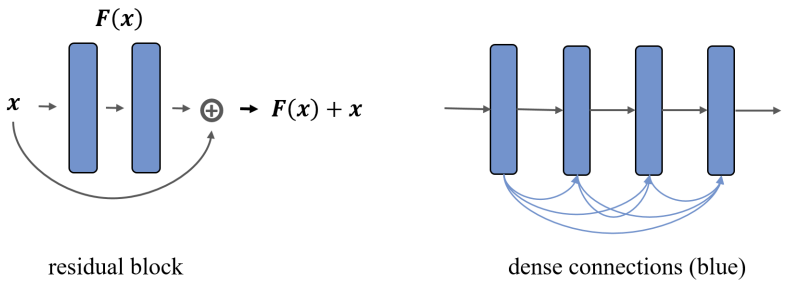


Figure 2.10: Schematic illustration of a residual (He et al., 2016) and a dense block (Huang et al., 2017b).

In multi-class settings, the softmax-function is used:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_i e^{-z_i}} \text{ for } i = 1..c\,. \tag{2.4}$$

EXTENSIONS OF THE U-NET The basic architecture of the U-Net has been modified to a diverse collection of variants over the years. We now give a brief overview about most commonly observed variants or extensions that have also been reused and applied in the context of prostate segmentation.

- *3D variant*: The U-Net is mostly applied in its (original) 2D variant. However, for volumetric medical images, a 3D version was proposed by Çiçek et al. (2016) and Milletari et al. (2016). The main difference is that instead of 2D operations, 3D operations (convolutions, max pooling and upsampling) are applied. This allows an end-to-end processing and prediction of volumetric scans. Because the 3D operations are computationally more expensive, Milletari et al. (2016) used less feature channels than the 2D U-Net (Ronneberger et al., 2015).

- *residual blocks*: Residual learning has been introduced by He et al. (2016) in their ResNet to enable better learning for deeper networks. Instead of learning the desired feature mapping $\mathcal{H}(x)$ of the input x in the stacked layers directly, He *et al.* suggested to learn the residual mapping $\mathcal{F}(x) = \mathcal{H}(x) - x$. This is realized by implementing shortcut connections that add the input x to the output of the stacked layers (see Figure 2.10). The authors hypothesize that it is easier for the network to optimize such a residual mapping than the direct mapping itself. Furthermore, residual connections enable a better gradient flow and mitigate the vanishing gradient problem (Nielsen, 2015) encountered in deep models.

- *dense blocks*: Huang et al. (2017b) evolved the residual connection concept within their proposed method to all subsequent layers. Thus, within a block of layers (with same feature map sizes), every

layer is connected by all its preceding layers and furthermore passes its feature to subsequent layers (Figure 2.10). In contrast to the residual connections, the previous layer's information is not added but concatenated. Networks that incorporate dense blocks are supposed to require less parameters, as the network does not need to relearn redundant features. Similar to the ResNet (He et al., 2016), this concept enables a better gradient flow, allowing much wider and deeper networks.

- *dilated convolutions*: Pooling operations are included in CNN architectures to increase the network's receptive field. However, they also reduce the resolution of the feature maps. Dilated convolutions have been proposed by Yu and Koltun (2016) to increase the receptive field without compromising the resolution. This is realized by widening the convolutional kernel by inserting "holes".

- *multi-scale mechanisms*: Different concepts have been proposed, that exploit information on multiple scales for medical image segmentation. One example are cascaded networks, for example, used by Pan et al. (2019), who employed two U-Nets. Their second network obtains as input a cropped image, whereas the cropping information is derived from the first network output. On a more detailed level, multi-scale information can be obtained by combining convolutions with different kernel sizes in one block, as for example proposed by Jia et al. (2020) with their pyramid convolutional architecture.

- *deep supervision*: In the original U-Net setting, the loss of the network is computed only based on last layer's prediction. Deeply supervised approaches introduce multiple auxiliary outputs at different resolution levels of the decoder (Figure 2.11). The overall loss of the network is then obtained by additionally incorporating the auxiliary outputs. This allows to inject gradients directly into deeper layers of the network (Isensee et al., 2021).

2.2.2 *Uncertainty Measures*

Modern neural networks have achieved state-of-the-art results for various tasks. Yet, despite their high accuracy, their output was found to be overconfident in their predictions and not well-calibrated regarding their predictive uncertainty (Guo et al., 2017) (see Figure 2.12 for an example). Thus, even when the softmax output of the model is high, its prediction can be uncertain (Gal and Ghahramani, 2016). This is particularly problematic in settings when downstream decisions, as in autonomous driving or computer-aided diagnosis, rely on the network's confident faulty predictions instead of asking for (manual) interventions in the case of uncertainty. Moreover, uncertainty measures play an

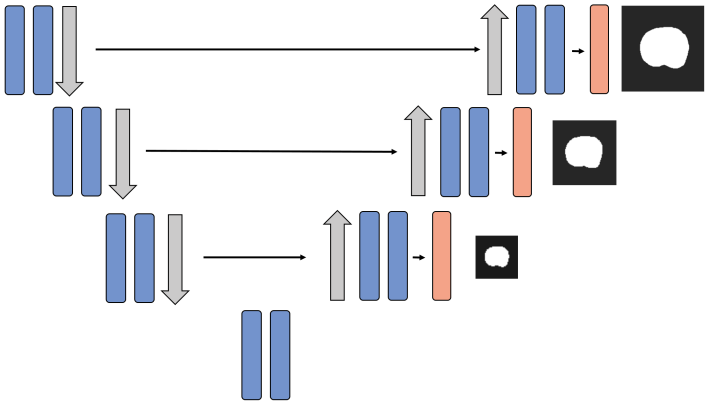


Figure 2.11: Illustrative example for a deeply supervised U-Net. The red boxes in the network represent the output layers.

important role in the field of SSL, where the uncertainty information is leveraged to reduce the incorporation of false predictions (of unlabeled data) into the model refinement (Sedai et al., 2019; Li et al., 2020b; Nie et al., 2018).

There are basically two different kinds of sources for uncertainty in machine learning: *aleatoric* and *epistemic* uncertainty (see Figure 2.13). Aleatoric uncertainty is due to ambiguities or noise inherent in the data, for example, caused by sensor noise (Kendall and Gal, 2017). Epistemic uncertainty represents uncertainty in the model parameters (Kendall and Gal, 2017). In contrast to aleatoric uncertainty, epistemic uncertainty is reducible, for instance, by including more data into the training of the network (Hüllermeier and Waegeman, 2021).

Several works have been proposed in the literature, that aim to capture uncertainty for neural networks. The research line of *Bayesian neural networks* proposes mathematically grounded solutions that output the uncertainty of the methods (Gal and Ghahramani, 2016). Instead of learning deterministic network parameters, Bayesian neural networks learn the posterior distributions for their weights, given the training data. This allows for the inference of a predictive distribution (of label probabilities) instead of a single point estimate, as in traditional neural networks. The methods are trained using Bayesian inference to find the posterior distribution of the model parameters. As Bayesian inference is computationally intractable (it requires integration over the whole model parameter space), various approximating methods have been proposed, such as Markov Chain Monte Carlo concepts (Neal, 2008) or variational Bayesian methods (Blundell et al., 2015). The output quality of Bayesian neural networks, however, depends on defining the right prior distributions for the model parameters and the quality of approximation (Lakshminarayanan et al., 2017). Moreover, they are hard to implement and slow to train (Lakshminarayanan et al., 2017).

In the context of medical image segmentation, several concepts have been included in order to model the uncertainty of the outcome, for

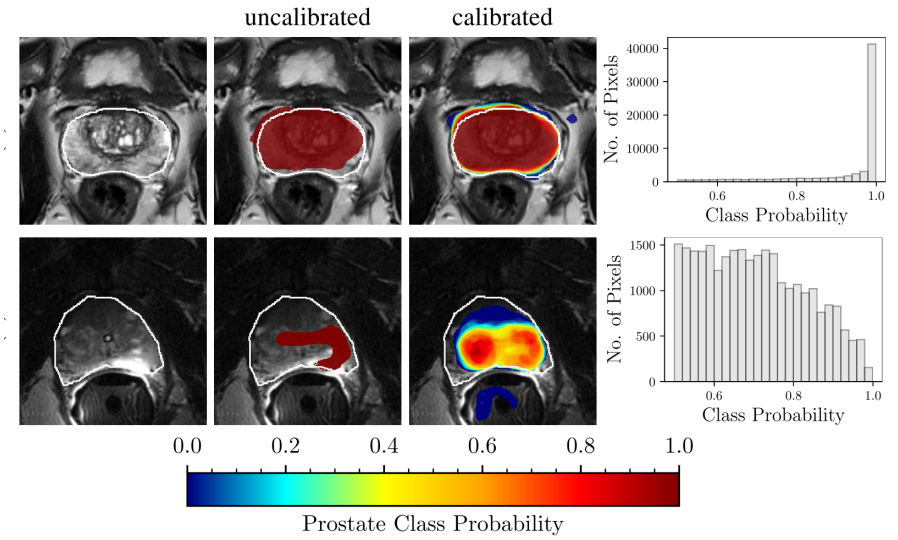


Figure 2.12: Examples for uncalibrated and calibrated (via deep ensembles) predictions. The model was trained on MRI acquired with a surface coil. The first row illustrates outcomes for an image from the training domain (top row) and the second row shows an outcome for an out-of-distribution image (acquired with endorectal coil). The histograms on the right show the calibrated class probabilities for both example images. The wide distribution in the bottom histogram indicate that it is an out-of-distribution example. Image from Mehrtash et al. (2020), ©2020 IEEE.

example, based on conditional variational autoencoders (Sohn et al., 2015; Baumgartner et al., 2019; Kohl et al., 2018; Bian et al., 2020) discriminative networks (Nie et al., 2018) or on sampling different augmentations of the test input (Venturini et al., 2020). Nonetheless, the two majorly encountered concepts in medical image segmentation are *Monte Carlo* (MC) *dropout* (Gal and Ghahramani, 2016) and *deep ensembles* (Lakshminarayanan et al., 2017), which will be described below. For a broader overview about uncertainty estimation techniques for DL applications, we refer to the recent review in (Abdar et al., 2021).

MONTE CARLO DROPOUT Dropout is usually employed in neural network architectures to reduce overfitting (Srivastava et al., 2014). By "dropping out" a random fraction of units (neurons) of the network at each training stage, the network is temporarily trained on a sub-network, inducing a regularization effect on the overall network. For this purpose, a hyperparameter p is introduced that determines the probability of the individual neuron's drop-out.

As Bayesian approximation to estimate uncertainty, Gal and Ghahramani (2016) proposed to apply dropout during inference. This allows to obtain predictions from F forward passes with different dropout masks, which can be seen as MC samples from the space of all available (sub-)

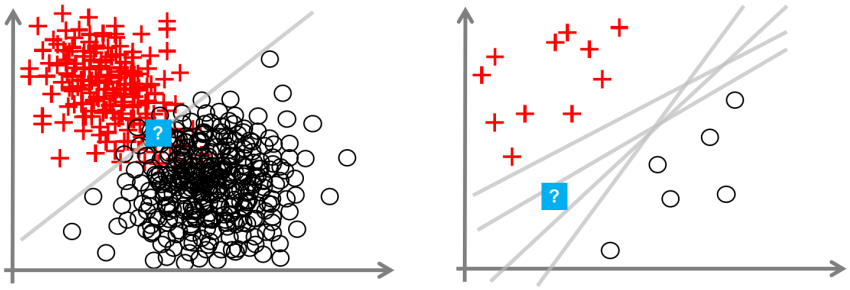


Figure 2.13: Examples for aleatoric and epistemic uncertainty. The aleatoric uncertainty (left image) arises from an overlap between the data distributions. The uncertainty in the right image is epistemic, where the correct model hypothesis can not be made due to a lack of data. Figure from Hüllermeier and Waegeman (2021), licensed under CC BY 4.0.

networks. Different quantities can then be obtained from this predictive distribution, e.g. the mean, variance or entropy.

Due to its simplicity, the MC dropout method is frequently applied in DL applications. Kendall et al. (2017) were the first to integrate MC dropout uncertainty estimates for a segmentation network and found that dropping out the deepest half of encoder and decoder layers prior to the down- and up-sampling operation lead to the best results.

DEEP ENSEMBLES Lakshminarayanan et al. (2017) proposed a non-Bayesian method that does not require any specific training paradigms. The authors investigated the uncertainty estimation of deep ensembles by training M models and incorporating adversarial examples (Szegedy et al., 2015) in the training procedure. They found that deep ensembles achieved high quality predictive uncertainty estimates (performing similarly or even better as Bayesian approximates). Moreover, the deep ensembles resulted in higher uncertainties for *out-of-distribution* samples that are far from the training datasets. Similarly, Mehrtash et al. (2020) confirmed the suitability of deep ensembles for better calibrated predictive uncertainty and out-of-distribution detection for the task of medical image segmentation. They also demonstrated that deep ensembles outperformed MC dropout for estimating the uncertainty of the output for three different MRI segmentation tasks, including prostate segmentation (see Figure 2.12).

2.3 SEMI-SUPERVISED LEARNING

For the training of CNNs, different strategies exist that can be categorized into *supervised*, *semi-supervised* and *unsupervised* methods. Typically, CNNs are trained in a supervised manner. This means that for every training sample x_i in our dataset, we have access to a label y_i, such that $D = \{(x_1, y_1), .., (x_n, y_n)\}$, and the method tries to learn

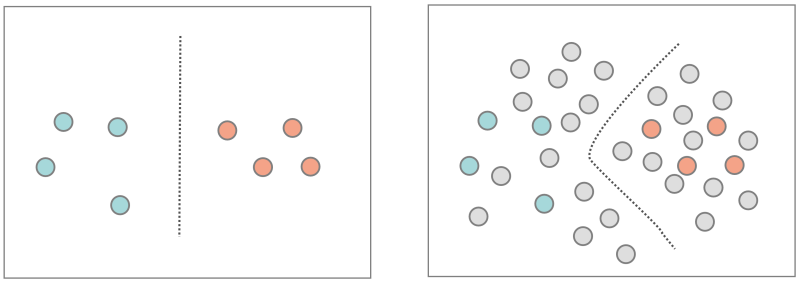


Figure 2.14: Illustration of the modification in the hypothesis or the decision boundary under the influence of unlabeled data. The left image portrays the decision boundary when using only labeled data, where turquoise circles belong to class A and red circles to class B. The right image portrays the change in decision boundary under the influence of unlabeled data denoted by grey circles.

a mapping from x_i to y_i. In unsupervised methods, the data available for learning consists only of unlabeled samples $D = \{x_1, .., x_n\}$, and the objective of these methods is to learn about the relevant structure in the data. This includes estimating the density $p(x)$ underlying the samples X, but also other forms, such as dimensionality reduction and clustering, fall into the unsupervised learning category (Chapelle et al., 2006). The definition of semi-supervised learning SSL lies between unsupervised and supervised learning. For SSL, we have a labeled dataset $D_L = \{(x_1, y_1), .., (x_n, y_n)\}$ and additionally an unlabeled dataset $D_U = x_{l+1}, .., x_{l+u}$, whereas $|D_L|$ is typically much smaller than $|D_U|$. By using the additional unlabeled data, semi-supervised learning can improve the CNN's decision boundary and therefore improve over its supervised counterpart (see Figure 2.14).

This, however, can only be achieved when the knowledge that is obtained from D_U, can provide information to define a better decision boundary for the task at hand (Ouali et al., 2020). If this is not the case, the inclusion of D_U will not improve or can even lead to a decline in model performance. Chapelle et al. (2006) define three assumptions about the data structure, that have to hold for SSL being beneficial for the learning.

- *Smoothness assumption:* If a sample x_i is close to another sample x_j in a high-density region, then their respective labels y_i, y_j should also be close. On the other hand, if two samples are lying in a low-density region, their labels need not be close. The smoothness assumption thus implies that the mapping functions should be smooth in high-density regions.

- *Cluster assumption:* This assumption can be considered as a special case of the smoothness assumption. It says that samples, which lie in one cluster, share the same label. However, it does not imply, that each class is only represented in a single cluster. The cluster assumption can be equivalently formulated as *low*

density separation, which states that decision boundaries should run through low-density regions.

- *Manifold assumption:* It states that the (high-dimensional) data lie on a low-dimensional manifold. This assumption relates to the problem that with higher dimensional data, the pair-wise distances between the data points become less expressive (for discriminative tasks). If we can learn the lower-dimensional manifold, for instance, by leveraging the unlabeled data, the task will be simpler to solve (Ouali et al., 2020).

These assumptions were originally defined for traditional machine learning methods that rely on linear models. But they are also considered as assumptions for deep SSL techniques.

According to Ouali et al. (2020), deep SSL concepts can be categorized into consistency regularization, pseudo-labeling approaches, graph-based methods and entropy-minimization. As *pseudo-labeling* approaches (Section 2.3.1) and *consistency regularization* (Section 2.3.2) are underlying concepts for our methods presented in the course of this thesis, we describe them in more detail in the following sections. The methods have been proposed originally for classification, but most of them have also found their way into the segmentation application.

2.3.1 *Pseudo-Labeling*

Pseudo-label methods incorporate the unlabeled data by inferring predictions (pseudo labels) on it with a classifier, which was originally trained on the labeled data. These pseudo-labels are then fed into the model's training. Different variants exist that vary in the way how these pseudo-labels are produced.

SELF-LEARNING *Self-Learning*, also known as self-training or self-labeling, is the oldest and most straight-forward SSL paradigm, and dates back to the 1960's (Scudder, 1965; Agrawala, 1970). The core mechanism is to use a trained model (initially trained on labeled data $D_L = (X_L, Y_L)$) to predict pseudo labels $\hat{Y}_u$ for unlabeled data $D_U = X_U$. This way the labeled set for training is enlarged to $(D_L \cup D_U)$, where $D_U = (X_U, \hat{Y}_U)$. The enlarged training dataset is then extended by the pseudo-labeled data and fed back into the network's training to improve the model. To iteratively improve the model, the process of pseudo-labeling and model refinement with expanded dataset is repeated until a stopping criterion is met (e.g. fixed number of iterations or the model converged).

As simply reusing all pseudo-labels for training can degenerate the performance when prediction errors get amplified during the self-learning cycle, methods have been proposed, which filter the pseudo-labels to contain only confident predictions. This sample confidence can be either

absolute (above a threshold) or relative (include only the top n confident samples). By incorporating only high confidence predictions, the self-learning methods rely on the cluster assumption (Cheplygina et al., 2019).

MULTI-VIEW TRAINING Multi-View training can be considered as a variant of pseudo-labeling. The algorithms rely on multiple views that complement each other, such as views obtained by different sensors or by generating limited views of the original data (Ouali et al., 2020). For each view, a different model is trained, whose predictions on D_U can then be exploited to increase performance of the other view's models.

One variant of multi-view-training is *co-training*, proposed by Blum and Mitchell (1998). Two models are trained individually on one view, whereas each view could be sufficient for learning. In this training setting, one model provides the pseudo-label on unlabeled data for the other model, if its prediction is above a certain confidence threshold.

As another variant of multi-view training, Zhou and Li (2005) introduced *tri-training*. Contrary to co-training, tri-training does not require different views of the data from the instance space. As the name already implies, three models are used, which are initialized via training on different datasets generated by bootstrap sampling from the original D_L. Samples from D_U are then iteratively added to the training data of one model, if the other two models agree on its label.

2.3.2 *Consistency Regularization*

Besides pseudo-labeling approaches, consistency regularization is one of the most popular concepts applied to SSL algorithms in the field of medical image segmentation. Rather than directly including the pseudo-labeled samples to the labeled training set and treating them as ground truth, consistency regularization penalizes the deviation of a model's prediction for one input sample, that was subject to different perturbations. Both D_L and D_U data can be used for this regularization, as no ground truth label is required. Because consistency regularization forces slightly perturbed samples to have the same label, it bases on a weak variant of the smoothness assumptions (Engelen and Hoos, 2020). The learned function is smoothed in the vicinity of the data points, which can also be seen as pushing the decision boundary further to low-density regions, complying with the low-density separation of the cluster assumption. (Ouali et al., 2020). In the following passages, the basic and most popular methods are described.

PI-MODEL The *π-model* is a deep learning SSL paradigm described by Laine and Aila (2017) (see Figure 2.15). For the π-model, the input sample is evaluated twice with different random perturbations (augmentation and dropout), resulting in two outputs z and $\tilde{z}$.

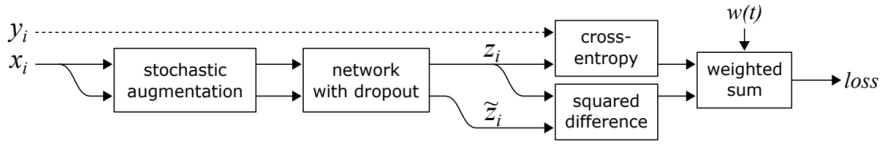


Figure 2.15: Training pass of the π-model. Figure from Laine and Aila (2017).

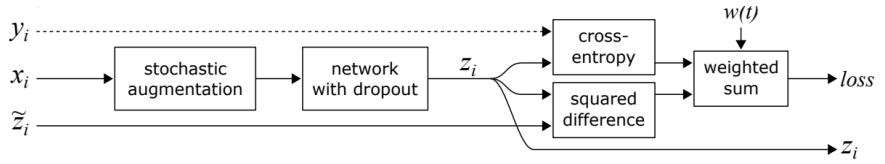


Figure 2.16: Training pass of the temporal ensembling concept. Figure from Laine and Aila (2017).

In addition to the supervised (task) loss, another weighted component is added, that enforces a consistency between p and $\tilde{p}$:

$$L = L_{\text{Task}} + w(t) L_{\text{Cons}} \,. \tag{2.5}$$

The weighting coefficient is defined by a ramp-up function $w(t)$ dependent on the current epoch t. This hyperparameter controls that the loss is governed by the labeled data in the beginning of training, when the model is not very stable.

TEMPORAL ENSEMBLING *Temporal ensembling* is an extension of the π-model and was also proposed by Laine and Aila (2017). The algorithm builds upon the π-model (see Figure 2.16), and is extended by a temporal ensemble of predictions Z which is used instead of the second network evaluation for consistency calculation. Z is updated epoch-wise as a weighted moving average where more recent updates have higher impact on the ensemble:

$$Z_i \leftarrow \alpha Z_i + (1 - \alpha) z_i \,, \tag{2.6}$$

with z being the current epoch's output and α a momentum term, which influences how far the ensemble reaches into historical predictions. As the ensemble Z_i is initialized as zero vector, the training target vectors $\hat{z}_i$ are determined by $\hat{z}_i \leftarrow Z_i / (1 - \alpha)^t$ to correct for the start-up bias.

Next to the reduced training time (due to the omission of the second evaluation pass), the ensemble has the advantage that it is less noisy and is most probably closer to the real ground truth and therefore a better estimate than a single epoch's prediction.

MEAN TEACHER An alternative to aggregate the predictions over time is to ensemble the model weights as proposed in the *mean teacher* algorithm by Tarvainen and Valpola (2017). The method employs two

models: a student and a teacher. The student model's weights θ_s are aggregated into the teacher model weights θ_t over the training steps t:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s\,. \tag{2.7}$$

The consistency loss is then calculated as the dissimilarity of the teacher's and student's prediction

The authors argue that the mean teacher has two advantages over temporal ensembling: (1) it allows for an update of the ensemble at each training step and not only at every epoch, which is an appreciated property for large datasets and speeds up training pace. (2) The weight averages affect all layers of the method and not just the output, such that the teacher model is supposed to have better representations.

VIRTUAL ADVERSARIAL TRAINING The inputs in the proposed methods so far, are perturbed in a random fashion. In *virtual adversarial training*, proposed by Miyato et al. (2018), the unlabeled input is perturbed in a direction such that the model's output distribution diverges the most from current model's output distribution.

2.4 METHODOLOGICAL PRELIMINARIES

Having described the medical background and the technical fundamentals for our methods, we now conclude this chapter with general information about methodological preliminaries, which we use throughout this thesis. This includes details about evaluation measures (Section 2.4.1) and algorithm implementations (Section 2.4.3).

2.4.1 *Evaluation Measures*

In this work, we use different measures to evaluate the performance of the proposed methods. We assess the performance of our proposed methods with well-established measures, which have been extensively applied in related work, including the PROMISE12 challenge on prostate segmentation (Litjens et al., 2014b). Specifically, we use the Dice similarity coefficient (DSC) (Dice, 1945) as well as the average boundary distance (ABD) and the 95th percentile Hausdorff-Distance (95-HD) between surface points of both volumes. All evaluation measures are computed in 3D. The Dice similarity coefficient is defined as

$$\mathrm{DSC}(X,Y) = \frac{2\,|X \cap Y|}{|X| + |Y|}\,, \tag{2.8}$$

with X being the predicted and Y being the ground truth voxels. It measures the ratio of overlap between two segmentations and ranges between [0,1] or [0%, 100%], respectively.

The ABD is defined as:

$$\begin{aligned}\mathrm{ABD}(X_S, Y_S) = \frac{1}{|X_S| + |Y_S|} \Bigg(&\sum_{x \in X_S} \min_{y \in Y_S} \mathrm{ED}(x, y) \\ &+ \sum_{y \in Y_S} \min_{x \in X_S} \mathrm{ED}(y, x) \Bigg) ,\end{aligned} \tag{2.9}$$

where X_S and Y_S are the sets of surface points of the predicted and ground truth segmentation. ED is the Euclidean distance operator.

Lastly, the Hausdorff distance (HD) is defined as

$$\begin{aligned}\mathrm{HD}(X_S, Y_S) &= \max\left(\mathrm{HD}'(X_S, Y_S), \mathrm{HD}'(Y_S, X_S)\right) \\ \text{with } \mathrm{HD}'(X_S, Y_S) &= \max_{x \in X_S} \left(\min_{y \in Y_S} \mathrm{ED}(x, y) \right) .\end{aligned} \tag{2.10}$$

We use the 95th percentile for implementation of HD (the so-called 95-HD), as this measure is more often applied in related work (Litjens et al., 2014b), leveraging comparability with previous works.

2.4.2 *Statistical Evaluation*

There exist different strategies on how to design the experimental setup for statistical evaluation of CNN methods. The most-established variant is the k-fold cross-validation, which is carried out in two different manners in the literature: *without* (e.g., Qin et al., 2020) or *with a hold-out test dataset* (e.g., Ghavami et al., 2019; Isensee et al., 2021). We apply the variant with the hold-out test set, because this guarantees that the hyperparameters of the network are not overfitted to the data from the actual test set.

Thus, the k folds are obtained from the training data only, and they determine the training and validation split for the model development, resulting in k models for evaluation. The evaluation is then carried out on the hold-out test dataset for each of the k folds (see Figure 2.17). By having k predictions for each test case available, there are again two strategies for a final evaluation: either averaging the predictions

Figure 2.17: Exemplary k-fold cross validation ($k = 5$) where the turquoise boxes represent the validation samples for the network training.

into an ensemble (as in Ghavami et al. (2019)) and thus having one performance estimate for each case, or using the k predictions for k estimates. Although model ensembling is an important strategy to improve the robustness of the methods, we decided to rely on the k performance estimates for the final evaluation, unless stated otherwise, because this allows for direct investigation of the method's performance.

To evaluate our models quantitatively, we report the mean and the standard deviations. Furthermore, we quantify whether there exist statistically significant differences between the distributions of different methods. We apply the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1992) for this task, because we could not assume Gaussianity for all distributions (evaluated for randomly selected experiments with the Kolmogorov-Smirnov test (Massey Jr, 1951)).

2.4.3 *Implementation Details*

The work in this thesis was implemented with Python and developed on mainly two different machines. The methods introduced in Chapter 3 and 4 were trained on a Linux machine with an Intel® Core™ i7-6850K CPU @ 3.60 GHz and 11 GB RAM NVIDIA GeForce GTX 1080 Ti graphical processing units (GPUs) inside. The algorithm introduced in Chapter 5 were mainly developed and trained on a Linux machine equipped with an Intel® Core™ i7-6700 CPU @ 3.40GHz and a 12 GB NVIDIA Titan X Pascal GPU.

For the deep learning components of our work, we used the *Keras* framework[2], which is now integrated into the *Tensorflow*[3] library. Operations on the medical images, as loading, pre-processing, augmentation etc., were conducted with the *SimpleITK*[4] and *numpy* Python packages.

2 https://keras.io/
3 https://www.tensorflow.org/
4 https://simpleitk.org/

3

EXPLOITING MULTI-PLANAR DATA

The general performance gain through the development of various CNN methods for image analysis tasks can also be observed in the context of prostate structures segmentation. However, the methods' performance is naturally bounded by the quality of the underlying image. The majority of methods proposed for prostate segmentation take the axial T2w MRI volume as input, which suffers from characteristically lower quality for some parts of the prostate. Therefore, with our work covered in this chapter, we investigate whether additional volumes acquired from different scan directions can compensate for the lower quality, and consequently improve the overall segmentation performance for the whole gland. We design an anisotropic 3D multi-stream CNN architecture, which can process this patient-level data simultaneously, allowing for a direct assessment of the supplementary data's impact on the overall outcome. The results indicate that the input of additional, patient-level data improves, in particular, in regions where the axial plane suffers from partial volume effects.

The content of this chapter builds upon the following publications:

- A. Meyer, A. Mehrtash, M. Rak, D. Schindele, M. Schostak, C.-M. Tempany, T. Kapur, P. Abolmaesumi, A. Fedorov, C. Hansen, "Automatic high resolution segmentation of the prostate from multi-planar MRI," in *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 177-181, © 2018 IEEE.

- D. Schindele, A. Meyer, D. F. von Reibnitz, V. Kiesswetter, M. Schostak, M. Rak, C. Hansen (2020). "High resolution prostate segmentations for the ProstateX-Challenge" [Dataset]. The Cancer Imaging Archive.

- A. Meyer[1], G. Chlebus[1], M. Rak, D. Schindele, M. Schostak, B. van Ginneken, A. Schenk, H. Meine, H.K. Hahn, A. Schreiber and C. Hansen, 2021. "Anisotropic 3D multi-stream CNN for accurate prostate segmentation from multi-planar MRI," *Computer Methods and Programs in Biomedicine*, 200, p.105821.

1 Joint primary authorship. The design of the network and organization of the datasets was contributed by the thesis author. Grzegorz Chlebus conceptualized and set up the hyperparameter search. Both authors had equal contributions in implementation, experimental setup design and evaluation.

STRUCTURE OF THE CHAPTER We begin this chapter by introducing the clinical motivation for automatic prostate segmentations and provide more details on the technical motivation and contribution of our work (Section 3.1). To put our method into the context of existing literature, we provide a thorough overview about the state-of-the-art of prostate segmentation algorithms together with their limitations in Section 3.2. Then, we describe our proposed anisotropic 3D multi-stream architecture in Section 3.3 and outline the experimental setup to evaluate our method and the impact of multi-planar data in Section 3.4. Details on the experimental results are reported in Section 3.5. Lastly, we discuss the results and our method's and experiments' limitations, as well as future research directions in Section 3.6. This chapter is then concluded with a brief recap of of our work in Section 3.7.

3.1 INTRODUCTION

A precise and automatic segmentation of the whole gland on T2w MRI is commonly desired for a variety of tasks in research and clinical practice. Automating prostate segmentation has the potential to (1) reduce the time for diagnosis and therapy planning, (2) to create more reliable and reproducible segmentations, and (3) to improve the outcome of PCa detection and interventions.

In PCa diagnosis, prostate segmentations may be a pre-processing step in computer-aided detection and assessment of prostate tumors (Sun et al., 2019). Segmentations can also be used to correlate MRI with histological images to obtain insights about the origin of MRI features (Shah et al., 2009; Kwak et al., 2016). Automatic segmentations facilitate the measurement of gland volume and can potentially make the measurement more accurate and reproducible. The gland volume is, for example, needed to calculate the PSA density, which is considered as a superior indicator for clinically significant PCa over PSA value alone (Yusim et al., 2020; Turkbey et al., 2019).

Moreover, gland segmentations are required in the planning of MRI-based dose calculation of radiation therapy (Siversson et al., 2015; Greer et al., 2019). For radical prostatectomy procedures, a 3D printed or virtual model of the prostate and its neighboring structures supports the planning, clinical education and patient counseling (Porpiglia et al., 2018; Wake et al., 2020). Furthermore, prostate segmentations are used to propagate high detailed image information of the T2w MRI on intraoperative TRUS images via segmentation-based registration, to guide, for example, prostate biopsy (Fedorov et al., 2015; Bashkanov et al., 2021), robot-assisted laparoscopic prostatectomy (Mohareri et al., 2015) or needle insertion in brachytherapy (Chen et al., 2021).

MOTIVATION AND CONTRIBUTIONS With the advance of CNNs, a new performance standard has been achieved for medical image analysis.

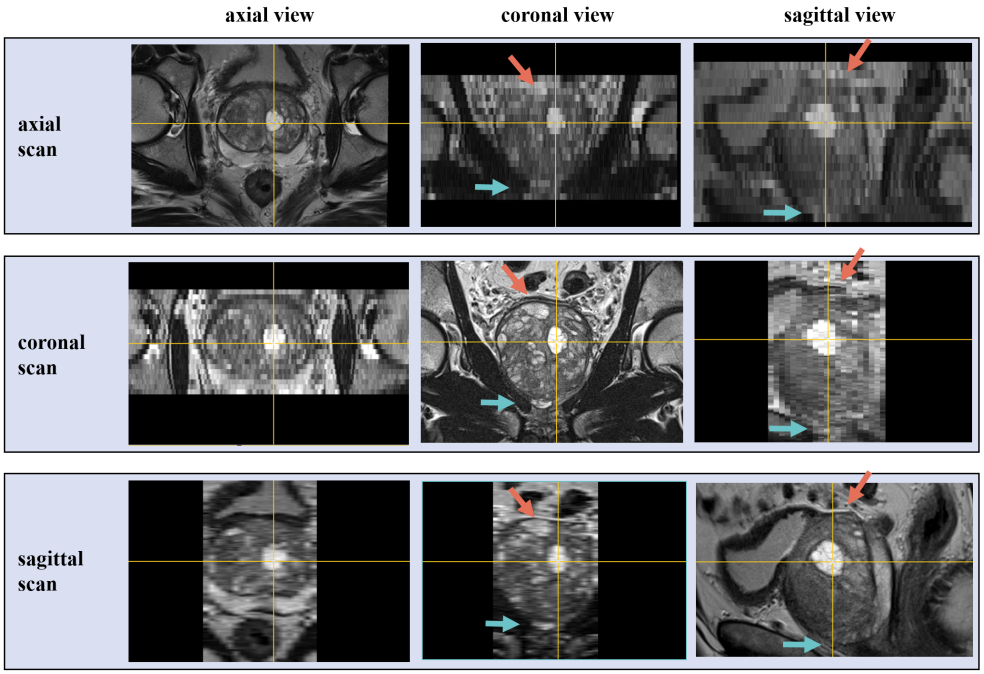


Figure 3.1: Visualization of the independent orthogonal scans of one patient illustrating their anisotropic nature. The first row depicts the axial scan that is normally used for segmentation. As can be seen in the sagittal and coronal view of that axial scan, the apical (turquoise arrow) and base (orange arrow) region lack clear boundaries of the prostate due to partial volume effect. In the sagittal and coronal scans, the prostate tissue in these regions can be distinguished more clearly from non-prostate tissue.

This also applies to the task of prostate gland segmentation on T2w MRI, where records for the public PROMISE12 challenge (Litjens et al., 2014c) are set on a regular basis. As we will learn in the subsequent Section 3.2, several methods with complex and elaborated architectures or loss functions have been proposed to enhance the performance of CNNs. However, with only few exceptions, these algorithms all consider solely the axial scans of the prostate.

Prostate MRI is highly anisotropic for the typical acquisition protocols, resulting in a factor of 6-10 difference between the out-of-plane and in-plane resolution. As can be seen in the top row of Figure 3.1, this leads to strong partial volume effects. Consequently, the prostate gland boundary can be challenging to accurately localize in the axial image in the apex and base regions, where the prostate cannot be clearly distinguished from surrounding structures like seminal vesicles, neurovascular bundles or muscular tissue. Although in PCa imaging protocols as in Turkbey et al. (2019), it is mandatory to acquire at least one additional scan direction (sagittal or coronal), the majority of proposed methods rely purely on the axial T2w MRI scan of the prostate. In multiple clinical routines, even all three scan directions are acquired for better interpretation. These multi-planar scans complement each other and could guide the

segmentation algorithm in the apex and base regions, where the single axial scan does not provide sufficient information. Thus, these additional scans could be used to improve the prostate segmentation quality.

In this chapter, we propose an anisotropic 3D multi-stream CNN architecture, that allows for simultaneous multi-planar input from T2w MRI, to aim for a high resolution and precise segmentation of the prostate gland. The proposed network design fuses information from anisotropic images, alleviating the need for image resampling to isotropic voxel size, and thus reducing the network's memory requirements.

We quantified the influence of information from additional image orientations on segmentation quality by comparing performance of a baseline single-plane model (processing only axial images) with dual-plane (axial + sagittal) and triple-plane (axial + sagittal + coronal) models. Quantitative results based on different compositions of the image data from two datasets and multiple sites demonstrate that the exploitation of this patient-level data is beneficial for the overall prostate segmentation outcome.

3.2 RELATED WORK

In the following, we give an overview about existing literature for prostate whole gland segmentation. We begin by outlining the majorly employed approaches that use only the axial T2w volume as input (Section 3.2.1). Subsequently, existing methods that incorporate multi-planar information are covered in Section 3.2.2. A summary of the CNN-based approaches and their performance is given in Table 3.1. We conclude with outlining their limitations (3.2.3), which motivated us to propose our method described in the following Section 3.3.

3.2.1 *Axial Plane Prostate Segmentation*

Before the advance of deep learning, prostate segmentation was mainly performed with atlas-based segmentation or deformable models based on hand-crafted features. A comprehensive summary of those methods is given in Ghose et al. (2012). Early approaches incorporating deep learning used voxel-wise classification to yield a segmentation mask. For instance, Liao et al. (2013) learned deep features with a stacked independent subspace analysis network in an unsupervised fashion and performed segmentation with label propagation from atlases. Guo et al. (2016) also used deep features but generated by a supervised stacked sparse autoencoder, yielding a prostate likelihood map, which is then segmented by a deformable model. Jia et al. (2017) performed patch-based voxel-wise prediction with an ensemble of four deep CNNs.

During the last five years, CNNs were increasingly introduced into the context of medical image segmentation. The FCN and more notably, the U-Net, have been adapted in various manners for prostate segmentation

Author	Method	Dataset	n_{train}	n_{test}	DSC [%]	ABD [mm]	95-HD [mm]
Milletari et al. (2016)	3D U-Net - residual connections	PROMISE12	50	30	86.9 ± 3.3	2.23 ± ?	5.71 ± 1.20
Cheng et al. (2017)	VGG Net (2D) with side-outputs - weighted fusion of side-outputs - tri-planar (indiv. networks)	internal	4-fold	100	88.6 ± ?	-	14.53 ± ?
Jia et al. (2017)	2D CNN-based voxel-wise classification - network ensemble	PROMISE12*	5-fold	50	88.0 ± 4.0	1.74 ± 0.42	5.00 ± 1.25
Yu et al. (2017)	3D - long and short residual connections - deep supervision	PROMISE12	50	30	89.4 ± ?	1.95 ± ?	5.54 ± ?
Zhu et al. (2017b)	2D U-Net - deep supervision	internal (ERC)	77	4	88.5 ± ?	-	-
Brosch et al. (2018)	regression network for prediction of distance from 3D mesh to boundary points	PROMISE12	50	30	90.5 ± ?	1.71 ± ?	4.94 ± ?
Karimi et al. (2018)	regression network for prediction of shape model parameters	internal (surface)	49	26	88.0 ± ?	2.02 ± ?	-
Lozoya et al. (2018)	dual planar - individual axial + sagittal 2D U-Nets	PROSTATEx	60	20	81.0 ± ?	-	-
To et al. (2018)	3D U-Net - residual and dense connections	PROMISE12	50	30	89.4 ± ?	-	-
Zhu et al. (2018)	3D U-Net - dense blocks	internal (ERC)	65	16	82.1 ± ?	-	-
Chen et al. (2019)	two stacked 2D U-Nets (feature fusion at multiple levels)	PROMISE12	50	30	90.0 ± ?	1.59 ± ?	5.58 ± ?
Hassanzadeh et al. (2019)	2D U-Net - dense blocks - residual blocks	PROMISE12*	10-fold	50	87.3 ± ?	-	-
Jia et al. (2019)	3D ResNet + 3D & 2D (boundary) decoder - pyramid conv. block - deep supervision	PROMISE12	50	30	91.4 ± ?	1.36 ± ?	3.93 ± ?
Pan et al. (2019)	cascaded 3D U-Net - dilated convolutions	PROMISE12	50	30	90.5 ± ?	-	4.47 ± ?
Wang et al. (2019a)	3D U-Net - residual connections - group dilated convolution - deep supervision	PROMISE12	5-fold	50	88.0 ± 5.0	1.02 ± 0.35	9.50 ± 5.11
Yuan et al. (2019)	2D encoder-decoder network - dense block	PROSTATEx	218	24	87.1 ± 6.6	2.23 ± 1.06	6.12 ± 2.16
Grall et al. (2019)	adversarial setting - 2D U-Net generator - conditional discriminator	internal (mpMRI)	30	10	73.0 ± 16.0	-	9.59 ± 4.37
Jia et al. (2020).	3D ResNet - 3D anisotropic decoder - pyramid conv. blocks	PROMISE12	50	30	90.6 ± ?	1.45 ± ?	4.13 ± ?
Riepe et al. (2020)	anisotropic 3D multi-stream U-Net, anisotropic conolutions, tri-planar	PROSTATEx (surface)	5-fold	40	90.0 ± 1.0	-	-
Umapathy et al. (2020)	cascaded 2D U-Net - residual blocks and residual skip connections	internal (mpMRI)	67	8	91.0 ± 0.02	-	-
Zhu et al. (2020)	3D U-Net - residual and dense blocks - boundary-weighted loss - transfer learning	PROMISE12 +int.	141	30	91.4 ± ?	1.35 ± ?	4.27 ± ?
Isensee et al. (2021)	3D anisotropic U-Net - self-adapting training pipeline - deep supervision	PROMISE12	50	30	91.9 ± 2.7	1.24 ± 0.29	3.95 ± 1.02
Meyer et al. (2021a)	anisotropic 3D multi-stream U-Net, anisotropic max pooling, tri-planar	PROSTATEx + int. (surf.)	47	19	93.1 ± 3.0	0.88 ± 0.48	3.07 ± 2.15
Meyer et al. (2021a)	anisotropic 3D multi-stream U-Net, anisotropic max pooling, dual-planar (axial + sag.)	PROSTATEx + int. (surf.)	47	19	93.3 ± 2.8	0.84 ± 0.51	3.00 ± 2.58

Table 3.1: State-of-the art supervised CNN-based methods for prostate segmentation and their reported performance. If not stated otherwise, methods were run on T2w data. "ERC" or "surface" in the dataset column denote endorectal or surface coil. PROMISE12 contains both ERC and surface coil data. PROMISE12* states that the method was not evaluated on the official test data. If methods were additionally evaluated on other but the PROMISE12 challenge set, we report the PROMISE12 results only for reasons of clarity. "k-fold" in the "n_{train}" column specifies a k-fold cross validation on the number of cases noted in column "n_{test}". Our proposed method's results are included in the last rows.

on T2w MRI. For instance, Milletari et al. (2016) proposed to apply a 3D U-Net with strided convolutions (for downsampling) that could be applied in an end-to-end-manner to the whole volume without any patch-based strategy. In their work, they were also the first to incorporate the DSC as loss function in the training.

Learning and segmentation performance can benefit from different aspects regarding network design to retain fine-detailed information and alleviate the vanishing gradient problem. For example, deep supervision was employed in several prostate segmentation works to feed gradient information directly into deeper layers, e.g. as in Zhu et al. (2017b). Isensee et al. (2021) also included deep supervision in their self-adapting pipeline for different segmentation tasks and applied anisotropic convolutions ($3 \times 3 \times 1$) in their 3D U-Net to reduce information loss due to the high slice thickness along the craniocaudal axis.

While the U-Net architecture employs skip connections from the encoder to the decoder part of the network, Yu et al. (2017) analyzed the effect of short and long residual connections and showed that a combination thereof is beneficial in a 3D CNN for prostate segmentation. Wang et al. (2019b) observed improvements with residual connections between neighboring blocks in combination with group dilated convolutions for multi-scale features and deep supervision.

The use of dense connections that enhance feature reuse and propagation has been shown to improve performance additionally. Therefore, Hassanzadeh et al. (2019) evaluated the use of various residual and dense connection setups. Yuan et al. (2019) made use of densely connected blocks in both encoder and decoder and trained with a joint loss function that incorporates the DSC and the reconstruction error from the dense blocks' output. Also, Zhu et al. (2018), To et al. (2018), Liu et al. (2020a), and Zhu et al. (2020) harnessed dense blocks in their architectures.

A very different approach from the end-to-end prediction of a dense segmentation map was proposed in two works that formulated the segmentation problem as a regression task. Karimi et al. (2018) let a convolutional regression network predict rotation, position and shape parameters to fit a shape model to the prostate contours of an input image. Brosch et al. (2018) combined a 3D shape model with a convolutional regression network which obtains the distance from a surface mesh to the corresponding boundary point of the prostate.

Multi-scale approaches have been proposed by several researchers to derive features from different scales of context. Pan et al. (2019) and Umapathy et al. (2020) employed cascaded U-Nets. In their works, the first network acts as a global estimator, and a subsequent network refines the segmentation mask with more local information. Jia et al. (2020) accounted for the anisotropic resolution of the prostate scans in their segmentation network, in which they combined a ResNet (He et al., 2016) encoder with an anisotropic convolutional decoder. They

introduced skip connections, that are built up from multi-scale (spatial pyramid) convolutional blocks with parallel convolutions of different kernel size to jointly compute local and global image features. Their 3D APA-Net was employed as a generator into an adversarial setting, in which a discriminator was trained to refine the output segmentation.

In another work, Jia et al. (2019) proposed to combine a ResNet encoder with a 3D segmentation decoder for intra-class consistency and an auxiliary 2D boundary decoder for inter-class discrimination. In the 3D segmentation decoder, they included a channel attention mechanism, deep supervision and similar to Jia et al. (2020) anisotropic convolution and pyramid convolutional blocks. In both decoders, dense connections were inserted between different stages. Currently[2], this work is placed 3rd in the PROMISE12 challenge, where prostate segmentation algorithms are evaluated on MRI acquired at different sites with varying acquisition protocols and scanner vendors. The first place is an extension of Jia et al. (2020), which has not been published properly, yet. Similarly, the current second place of PROMISE12 reused concepts from Jia et al. (2020) and Jia et al. (2019) as for example edge attention (in 3D), pyramid convolutional block with attention mechanism, anisotropic convolutional blocks and a 3D ResNet encoder, which they pre-trained on another MRI dataset.

3.2.2 *Multi-Planar Prostate Segmentation*

While the algorithms described above all base on the axial scans only, there have been some proposals for methods incorporating additional scan directions, which are required to be obtained in the clinical routine. Cheng et al. (2017) leveraged multi-planar data by automatically segmenting the prostate with multiple so-called holistically-nested edge detector networks. For each orthogonal scan (axial, sagittal and coronal), a 2D VGG network (Simonyan and Zisserman, 2015) with multiple side-outputs is trained separately and subsequently the three segmentation outcomes with low out-of-plane resolution are used for surface extraction with ball pivoting, followed by Poisson surface reconstruction to obtain a hole-free and smooth surface as algorithm output.

Furthermore, Lozoya et al. (2018) assessed the effect of single and dual plane segmentation by training an ensemble of two deeply supervised 2D CNNs independently on axial and sagittal volumes. The models process three consecutive image slices to segment the middle one. The axial and sagittal segmentations are then fused by assigning each voxel the foreground label, where either the sagittal or the axial network predicted a foreground label. Compared to a single-plane baseline, the results showed an improvement for the dual plane approach.

2 as of June, 17th 2021

3.2.3 *Limitation of Current Approaches*

The works summarized in Section 3.2.1 demonstrate the suitability of deep learning approaches for the task of automatic prostate segmentation. However, as stated before, these methods only rely on the axial scans, such that the methods need to extrapolate for regions with diminished details due to the partial volume effect.

The methods outlined in Section 3.2.2 addressed this issue by using multiple scan directions as input. While these multi-planar approaches show that the exploitation of additional scan directions is beneficial for the segmentation quality, they have some limitations. Firstly, both approaches train independent CNNs per MRI orientation, which prevents the models from learning how to combine the information coming from different orientations. Secondly, only 2D neural networks are employed which cannot capture the inherent volumetric information of MRI scans. Being able to analyze the 3D image context is important for prostate segmentations as demonstrated by the top performing methods in the current leaderboard of the PROMISE12 challenge[3]. And thirdly, all works that have leveraged multi-planar data before, used different methods and datasets. Consequently a thorough investigation of the difference between two and three planes has not been possible yet.

In this work, we target these limitations by proposing a novel[4] multi-stream CNN architecture that processes simultaneously anisotropic multi-planar 3D MRI scans to produce a high-resolution prostate segmentation (Section 3.3). We evaluated our method and the effect of exploiting one, two and three input scan directions as network inputs on the overall performance. For this purpose, we used different compositions of two prostate MRI datasets from multiple sites (Section 3.3).

3.3 TECHNICAL METHODS

We can basically define two variants of combining multiple planes as input for CNNs. The first way is to train multiple networks separately with each network taking one orthogonal scan as input. The output of the three networks is then fused in a postprocessing step as in (Cheng et al., 2017; Lozoya et al., 2018). The alternative is to input all planes into one (multi-stream) network, which allows to process them simultaneously.

3 https://promise12.grand-challenge.org/evaluation/challenge/leaderboard/, last accessed June 2021.

4 In 2020, an abstract was published by Riepe et al. (2020) that also proposed an anisotropic multi-stream 3D U-Net for prostate segmentation. Their work and our journal paper arose independently and at a similar time (our work (Meyer et al., 2021a) has been submitted to Elsevier Computer Methods and Programs in Biomedicine in January 2020). The main difference between both works is the scope, as the method by Riepe et al. (2020) compared merely single axial plane vs. triple-plane (axial+sagittal+coronal) input on a single dataset ($n = 40$). Moreover, we introduced the idea of a multi-stream architecture for multi-planar segmentation already in our ISBI paper (Meyer et al., 2018).

Due to its simplicity in deployment, we focused our work on the multi-stream architecture. This has the additional benefit, that we can investigate the influence of additional planes directly, as the ensembling of network outputs has a benefit on performance in general (Goodfellow et al., 2016). The following sections describe our proposed anisotropic 3D multi-stream CNN (Section 3.3.1), and provide details on its training (Section 3.3.2).

3.3.1 *Anisotropic 3D Multi-Stream CNN*

We base our architecture on the 3D U-Net proposed by Çiçek et al. (2016) with four resolution levels. We extended this architecture as a multi-stream architecture, whose number of encoder branches corresponds to the number of inputs. Therfore, we propose three instances of this architecture: the *single-*, *dual-* and *triple-plane* network. Figure 3.2 illustrates the triple-plane model variant with three branches in the encoder processing axial, coronal, and sagittal acquisitions.

All current top five approaches in the PROMISE12 challenge (Litjens et al., 2014c) consider the anisotropy of the axial T2w MRI scan in their architecture. This is in line with our work on prostate zone segmentation where we found the anisotropic U-Net variant to perform better than its isotropic counterpart (see Chapter 4). We adopt this design for our multi-stream architecture, where our encoder branches reflect the anisotropy of their input's volume (e.g. 144×144×36 for the axial volume and 144×36×144 for the coronal volume). For this purpose, our network performs max pooling operations with an anisotropic pool size (e.g., 2×2×1 for the axial stream) after the first two convolutional blocks, resulting in equally sized outputs in the third resolution level. The input to the third convolutional block is then the concatenation of the individual stream's outputs.

In the decoder, the feature map sizes are increased via tri-linear upsampling or transposed convolution (the determination of the upsampling mode is subject to the hyperparameter search). In the bottom-most and each decoder's convolutional block, we employed a dropout layer (Srivastava et al., 2014) in between the two convolutional layers to regularize the network training and prevent overfitting. The first convolutional layer of the network has 8 feature maps, which are doubled in each resolution level. This results in a maximum of 128 channels at the bottom of the network in the latent feature space. To pass detailed features from the earlier network levels, we employed skip connections from each resolution level in the different encoders to the corresponding level in the decoder. For this, the anisotropic dimension of the feature maps of the first two resolution levels is upsampled to fit the feature map dimensions in the isotropic decoder. The final layer of the network is a convolutional layer with 1×1×1 kernel and sigmoid activation function

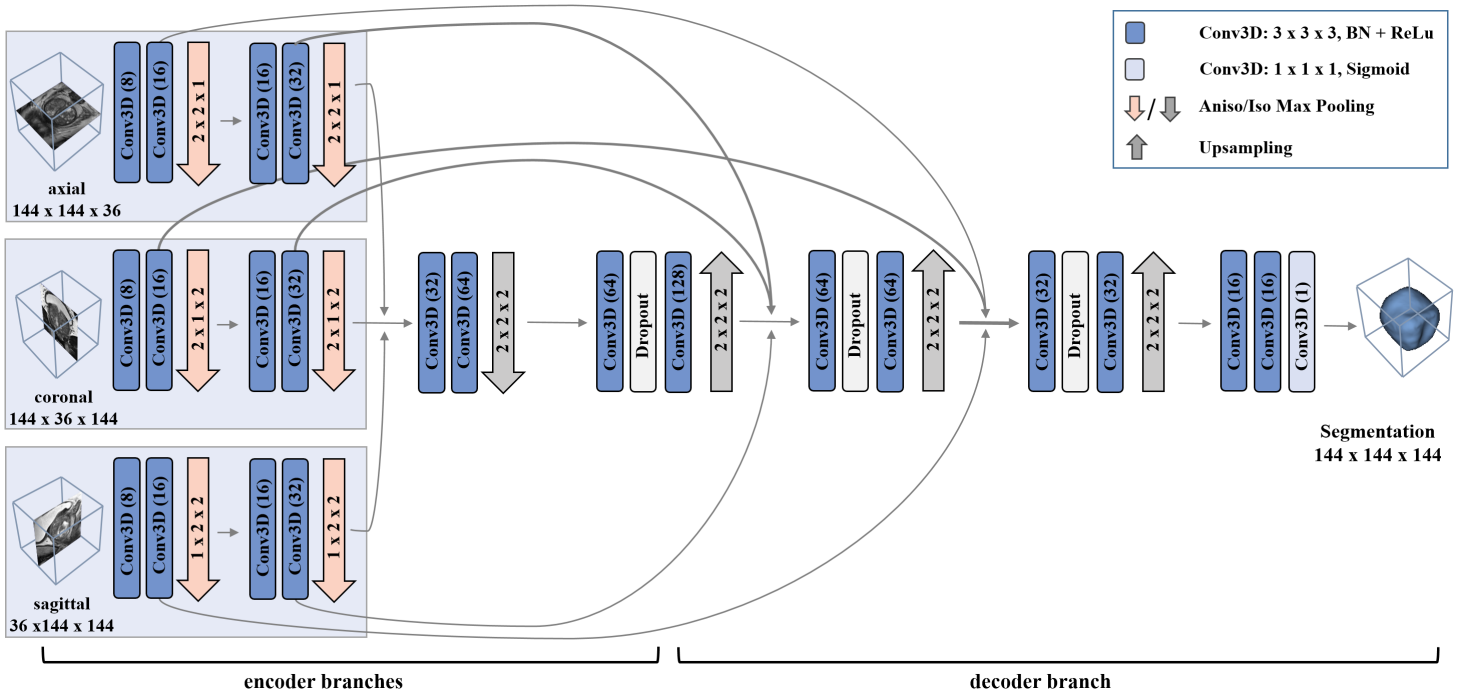


Figure 3.2: Triple-planar multi-stream 3D network processing axial, coronal, and sagittal MR volumes. The number in parentheses for the convolution layers denote the feature map count. The numbers in the max pooling and upsampling arrows denote the down- and upsampling factor for each dimension. The upsampling is performed either by tri-linear upsampling or 3D transposed convolution. Convolutions are all applied with a $3 \times 3 \times 3$ kernel. For reasons of clarity, we omit the anisotropic upsampling for the skip-connections that start from the encoders' two lowest resolution levels.

that outputs the predicted segmentation mask as one channel. The resulting high-resolution mask has a size of 144×144×144 voxels.

We employed optional batch normalization (Ioffe and Szegedy, 2015) between the convolutional layer and the ReLU activation function to improve the network learning. Analogous to the upsampling mode, the usage of batch normalization is subject to hyperparameter optimization and thus not set for every network instance. We give more details on the network's training and hyperparameter search in the following.

3.3.2 *Training Details*

The three network instances (single, dual, and triple) are individually trained with the negative soft DSC loss function (Milletari et al., 2016):

$$\text{loss}(\hat{y}, y) = -\frac{2\sum_{i=1}^{N}\hat{y}_i y_i + \epsilon}{\sum_{i=1}^{N}\hat{y}_i + \sum_{i=1}^{N} y_i + \epsilon}\,, \tag{3.1}$$

with N being the total number of voxels, $\hat{y}$ and y the predicted and manual reference segmentations, respectively, and ϵ a small constant to ensure numerical stability. We ran the training with the Adam optimizer (Kingma and Ba, 2015) for a maximum of 270 epochs, with an early stop criterion if the validation loss does not improve by at least $\delta = 0.001$ for 100 iterations. The mini-batch size was set to one due to GPU memory capacity (NVIDIA GeForce GTX 1080 Ti 11GB).

We applied random geometric transformations to augment the training set and increase the method's robustness. The augmentation includes axial flips, elastic deformations, translations and rotations. Unnatural transformations such as top-bottom and front-back flips were not used.

In order to determine the best configurations for the different network instances, we carried out a hyperparameter optimization. We applied the method from Falkner et al. (2018) that involves a combination of Hyperband (Li et al., 2017) with Bayesian optimization (Shahriari et al., 2015) to achieve fast convergence to optimal configurations. We focused on those hyperparameters, which were empirically found to have substantial influence on model performance: learning rate, dropout rate, upsampling mode (tri-linear or transposed convolutions) and batch normalization. The hyperparameter search was carried out on the concatenation of the first folds from both datasets. The best performing hyperparameters for each approach, selected based on the validation loss, are summarized in Table 3.2. The total numbers of trainable parameters for the single-, dual, and triple-plane of the proposed network architectures are 1.4, 1.6, and 1.7 million, respectively. Consequently, the proposed strategies are using a similar network capacity.

Hyperarameter	Single-Plane	Dual-Plane	Triple-Plane
learning rate	1.28×10^{-4}	1.31×10^{-4}	2.99×10^{-4}
dropout rate	0.6	0.2	0.2
batch normalization	no	no	yes
upsampling mode	tri-linear	transposed conv.	transposed conv.

Table 3.2: Best performing hyperparameters for each of the investigated network architecture instances.

3.4 EXPERIMENTAL SETUP

Following the description of our proposed multi-stream CNN and its training, we now provide details on the evaluation thereof. We begin with details on the two datasets that we used for the performance investigation (Section 3.4.1). Section 3.4.2 then outlines the experiments that we have designed to assess the performance of our anisotropic 3D multi-stream CNN, and to investigate the impact of the inclusion of additional orthogonal planes into the prostate segmentation.

3.4.1 *Datasets*

Methods targeting the segmentation of the prostate glands are often benchmarked in the PROMISE12 challenge (Litjens et al., 2014b). As this challenge dataset only consists of axial T2w scans (see Table 3.3), we were not able to make this comparison for our proposed method. Instead, we used two other datasets for the evaluation of our approaches: the PROSTATEx dataset and an in-house dataset. Both datasets contain axial, sagittal and coronal T2w scans acquired without an endorectal coil. The scans represent prostates with clinical variability such as tumors, cysts, and benign prostatic hyperplasia. Details on the resolution of the orthogonal scans can be found in Table 3.3. The following paragraphs cover more details on the PROSTATEx and the in-house-dataset with their corresponding reference segmentation creation, as well as their pre-processing and training/testing split.

PROSTATEX DATASET The PROSTATEx dataset is publicly available through the SPIE-AAPM-NCI Prostate MRI Classification Challenge (Litjens et al., 2017a; Litjens et al., 2014a; Clark et al., 2013), which was designed for predicting the clinical significance of prostate lesions. The dataset comprises multiparametric MRI acquired by two different types of Siemens 3T scanners: the MAGNETOM Trio and Skyra. For this dataset, we selected only cases, where the entire prostate was acquired by all orthogonal volumes. Cases in which, for example, the axial scan missed parts of the prostate's apex, were excluded to fairly

Dataset	Scan	Resolution [mm]
PROSTATEx	axial	[0.5-0.6] x [0.5-0.6] x [3-5]
	sagittal	0.56 x 0.56 x [3-4]
	coronal	[0.56-0.6] x [0.56-0.6] x [3-4.5]
In-House	axial	0.5 x 0.5 x 2.75
	sagittal	0.5 x 0.5 x 3.25
	coronal	0.5 x 0.5 x 2.76
PROMISE12	axial	[0.27-0.63] x [0.27-0.63] x [2.2-3.6]
	sagittal	not available
	coronal	not available

Table 3.3: Resolution details for Prostate MRI datasets. For our study, we used the PROSTATEx and the in-house dataset. We could not use the PROMISE12 dataset because it did not contain the sagittal and coronal scans.

evaluate the impact of the different inputs. For the in-house dataset, no such cases were found.

Since there are no reference segmentations of the glands available in the challenge dataset, we created 66 segmentations for randomly chosen T2w volumes. We took considerable care in this ground truth creation process, since delineating the organ contours only on the axial scan could bear the risk of missing important details in apex and base, where prostate boundaries are not clearly distinguishable in the axial scans.

Therefore, the segmentations were obtained manually for each individual scan direction by a medical student with 3D Slicer (Fedorov et al., 2012), followed by a review and, if necessary, corrections of an expert urologist. The final isotropic high-resolution prostate mask was extracted by taking the average of linearly resampled distance transformations of the individual segmentations and thresholding the result at zero (similar to the shape-based interpolation by Herman et al. (1992)). The final masks were again reviewed by an expert under consideration of all three orthogonal scans and corrected if necessary.

IN-HOUSE DATASET The second dataset is an in-house dataset containing 89 axial, sagittal and coronal T2w MR scans acquired on a Philips Achieva 3T imager. In the clinical routine, gland segmentations have been obtained with commercial software (DynaCAD, Philips Invivo) in a semi-supervised manner. As the software only considers the axial T2w volumes, we resampled the segmentations to an isotropic resolution via shape-based interpolation (Herman et al., 1992). Subsequently, the medical student and expert urologist reviewed and corrected the

isotropic segmentations with 3D Slicer by simultaneously considering all three orthogonal scans.

PRE-PROCESSING Because of the subsequent acquisition of the different orthogonal volumes, their alignment in the patient space may have been corrupted through motion of the patient or the patient's bowel. Therefore, the alignment of the orthogonal scans was checked visually using 3D Slicer. We observed, that in about 10% of the cases, one or multiple scans were misaligned. For these cases, we performed a manual rigid registration of affected images.

For network training and prediction, the three scans were pre-processed by resampling (linear interpolation) them into a common coordinate system. The resulting resolution was $0.5\times0.5\times2.0$ mm for axial scans, $0.5\times2.0\times0.5$ mm for coronal scans, and $2.0\times0.5\times0.5$ mm for the sagittal scans corresponding to their anisotropic acquisition. Next, the images were cropped, such that the resulting volume is the intersection of the three scans which has an in-plane size of 184×184 and an out-of-plane size of 46 (if necessary, the volume was cropped or padded to obtain the desired in-plane and out-of-plane size). As intensity normalization, the gray values were clipped to the 1st and 99th percentiles and afterwards normalized to an intensity range of [0,1].

DATA SPLIT We followed the k-fold evaluation concept described in Section 2.4.1. As hold-out test set, we set aside 19 randomly chosen cases for each dataset that were not considered for training. The remaining images were split into four folds for cross-validation. Thus, the folds of the in-house dataset consist of 52 training and 18 validation images each, while the PROSTATEx folds contains 35 training and 12 validation images. Based on these data splits, we carried out our evaluation experiments that are described in the next section.

3.4.2 *Evaluation Design*

To obtain quantitative evaluation measures, we first applied connected components analysis to the automatic segmentation outcomes, removing every component except for the largest. Then, the post-processed predictions were evaluated individually with respect to the DSC, the 95-HD and the ABD as described in Section 2.4.1. The evaluation is carried out globally on the whole volume and regionally on the base, mid-gland and apex. For the regional assessment, we divided the volume into thirds based on the manual reference segmentation (along the craniocaudal axis in a slice-wise manner).

With our experiments we aimed to investigate our major research question: do additional scans directions as input help in obtaining an improved segmentation outcome? Moreover, we evaluated whether there are any performance differences between our multi-stream network

and an ensemble of networks. Finally, we compared our method to the average manual performances among expert readers (inter-reader-variance). Details on the design of the evaluation experiments are described in the following paragraphs.

MULTI-PLANAR VS. AXIAL NETWORK In order to investigate whether multi-planar input is beneficial for the overall segmentation outcome, we compared the results for the triple-plane (axial + sagittal + coronal) and dual-plane (axial + sagittal) network to the single axial network, which serves as a baseline. As for the dual-plane network, there would also be other compositions possible, namely axial + sagittal and sagittal + coronal. We focused our experiments solely on the axial + sagittal variant of the dual-plane network, because the axial scan is the main orthogonal direction in PCa MRI and we observed, that our clinical partners worked preferably with the sagittal scan as additional plane when manually segmenting prostate structures.

Having two different datasets (PROSTATEx and in-house) available, there exist different strategies on how to train and evaluate the methods to investigate the impact that the additional scan directions have on the overall outcome. Therefore, we implemented two different scenarios:

- *Scenario I* - train and evaluate the model on a merged dataset.
- *Scenario II* - train and evaluate the models on individual datasets.

For Scenario I, we concatenated the respective training and validation sets of the two datasets for each fold. By comparing models resulting from both scenarios, we can verify whether segmentation quality for any of the network instances can benefit from training on multi-site data. Additionally, we compared our method's performance with those reported in the literature.

MULTI-STREAM VS. ENSEMBLE As already introduced before, another variant to exploit multi-planar data for the segmentation is to train an ensemble of networks. We directly compared our multi-stream architecture with such an approach. For this purpose, we trained three independent 3D single-plane models for each image orientation, whose outputs were combined in a post-processing step via majority voting. We also evaluated output combination using shape-based interpolation, but the results degraded in comparison to the majority vote. This experiment was carried out on the PROSTATEx dataset.

INTER-READER VARIANCE To provide a comparison of our proposed automatic multi-plane method to the general performance of manual (expert) segmentations, we indirectly compared the results of our method to the average manual performance reported in the PROMISE12 challenge and the one that we assessed for the PROSTATEx data within another project. In the PROMISE12 challenge (Litjens et al., 2014c),

the automatic results submissions were also compared to the inter-reader variance between an experienced clinical reader and a relatively inexperienced nonclinical reader (two years research of prostate MRI).

For our work on prostate zone segmentation that is covered in the next chapter, we evaluated the inter-reader-performance between two expert urologists (Section 4.4.1). The two urologists (each one supported by a medical student) were asked to outline the glandular structures in the axial scans of 20 cases from the PROSTATEx challenge. It has to be noted that those cases do not cover the hold-out test cases of this work. Nevertheless, we can still get a notion of how much two expert segmentations can vary for this dataset that we incorporated in our method.

3.5 RESULTS

In the following sections, we report the results for the experiments described above. In Section 3.5.1, we assess the outcomes of our multi-plane network instances to the axial single-plane state-of-the-art methods. The ensemble of single-plane networks (one for each orientation) is compared to the tri-planar multi-stream network in Section 3.5.2. And a comparison to the inter-reader-variance is made in Section 3.5.3 based on results of other works and on other datasets.

3.5.1 *Multi-planar vs. Axial Network*

We compared the outcomes of our multi-plane networks (triple and dual) to the axial single-plane network within the two training scenarios (train one model on the merged dataset and train one individual model per dataset). Moreover, we indirectly rank our methods' results according to the performance measures reported in the state-of-the-art described in Section 3.2.

TRAINING ON MERGED DATASETS In the training Scenario I, we trained and evaluated the network instances on the concatenation of the PROSTATEx and the in-house dataset. Table 3.4 summarizes the outcomes for this experiment. Furthermore, the results for this scenario are visualized in the boxplots in Figure 3.3. Visual examples of the segmentation results from the single-, dual- and triple plane network variants are depicted in Figure 3.4. In general, we can see that the additional scan directions used by the multi-plane models improved the segmentation quality when compared to the single-plane model, whereas the dual-plane network obtained the overall best results.

With respect to the DSC, the dual-plane approach that incorporates axial and sagittal volumes, works significantly better ($p < 0.01$) than the single-plane approach in every region. The dual-plane method achieved an average DSC of 93.3% for the whole gland (vs. 92.7% for single-

		Single	Dual	Triple
DSC[%]	Whole	92.7 ± 3.0	$\mathbf{93.3 \pm 2.8}^{***}$	$93.1 \pm 3.0^{***}$
	Apex	88.8 ± 8.7	$\mathbf{90.1 \pm 7.7}^{***}$	$89.6 \pm 9.5^{**}$
	Mid	95.6 ± 2.0	$\mathbf{95.8 \pm 2.1}^{***}$	95.4 ± 2.4
	Base	89.8 ± 5.3	$90.4 \pm 7.5^{***}$	$\mathbf{90.6 \pm 5.6}^{***}$
ABD[mm]	Whole	0.90 ± 0.05	$\mathbf{0.84 \pm 0.51}^{***}$	$0.88 \pm 0.48^{***}$
	Apex	0.99 ± 0.69	$\mathbf{0.86 \pm 0.54}^{***}$	$0.92 \pm 0.70^{*}$
	Mid	$\mathbf{0.76 \pm 0.30}$	$0.78 \pm 0.66^{***}$	0.83 ± 0.58
	Base	1.01 ± 0.63	$\mathbf{0.95 \pm 0.91}^{***}$	$0.95 \pm 0.71^{**}$
95-HD[mm]	Whole	3.10 ± 1.68	$\mathbf{3.00 \pm 2.58}^{***}$	$3.07 \pm 2.15^{**}$
	Apex	2.99 ± 1.89	$\mathbf{2.65 \pm 1.40}^{**}$	$2.81 \pm 1.75^{*}$
	Mid	$\mathbf{2.48 \pm 1.41}$	$2.69 \pm 3.45^{***}$	2.74 ± 2.51
	Base	3.10 ± 1.96	$2.93 \pm 2.56^{**}$	$\mathbf{2.90 \pm 2.20}^{**}$

Best results are marked bold. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 3.4: Average evaluation measures for **Scenario I** (training and evaluation on merged datasets). Asterisks mark significant differences to the results of the single-plane model.

plane), 90.1% (vs. 88.8%) in the apex and 95.8% (vs. 95.6%) and 90.4% (vs. 89.8%) for mid-gland and base, respectively. It has to be noted that the average ABD and 95-HD for the mid-region in Table 3.4 are worse for dual-plane than single-plane, but we found that the dual-plane model performs better when the median as well as the distribution of results is considered (see Figure 3.3).

The triple-plane model performed significantly better ($p < 0.05$) than the single-plane model regarding the DSC, ABD, and 95-HD for all regions except the mid-gland, too. Incorporating all three planes achieved an average DSC of 93.1% for the whole gland and DSCs of 89.6%, 95.4% and 90.6% for apex, mid-gland and base, respectively.

For the mid-gland, the average evaluation metrics for the single-, dual- and triple-plane approaches are very close to each other. However, for the other regions, the difference in performance between dual and triple-plane is less than between single-plane and triple- or dual-plane for Scenario I.

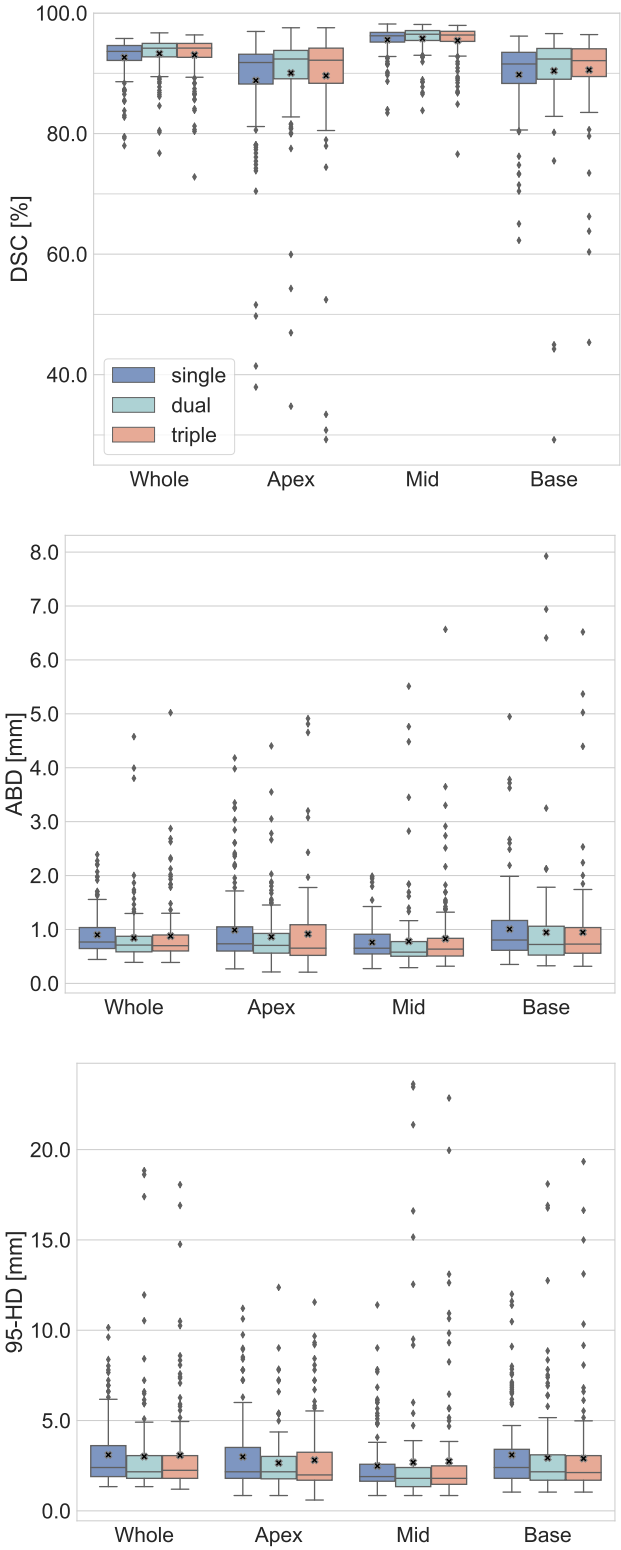
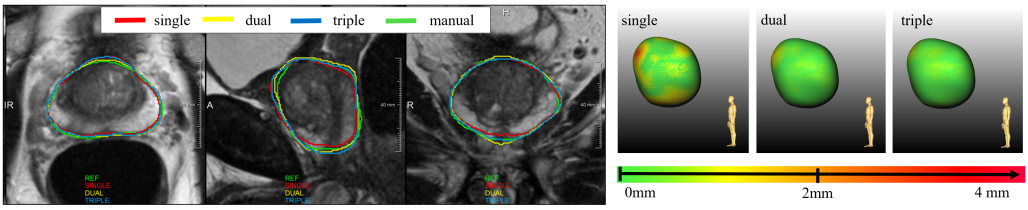
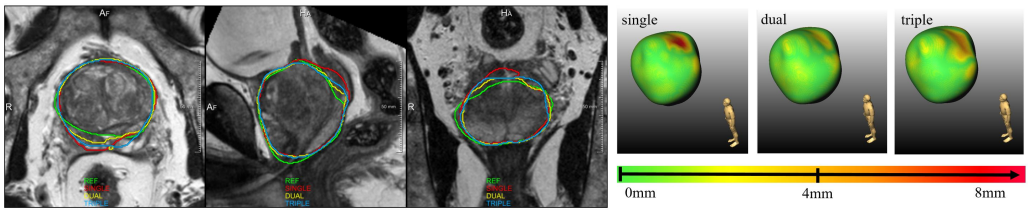


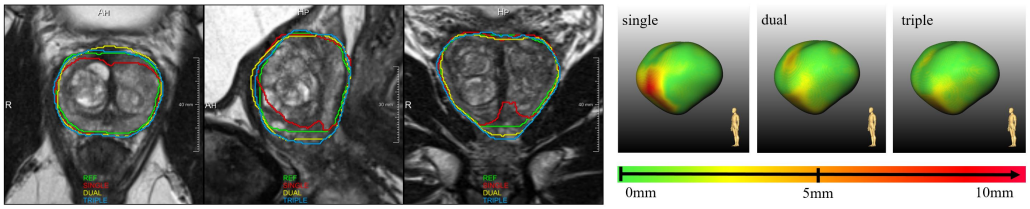
Figure 3.3: Boxplots showing the DSC, ABD, and 95-HD for the whole gland and its subregions for single- dual, and triple-plane models. Models were trained and evaluated on merged datasets (**Scenario I**). The marker 'x' indicates the mean of the distributions.



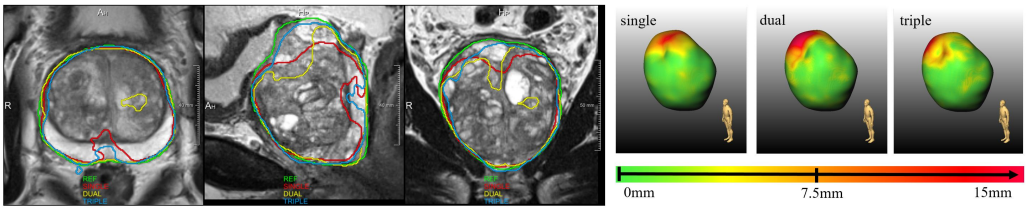
(a) Simple case where all approaches perform about equally well.



(b) Challenging case where dual-/triple- plane approaches are beneficial. When considering only the axial plane, we yield overestimation in the base region.



(c) Another challenging case where dual-/triple- plane approaches are beneficial. Segmentation in apical region of the prostate is improved.



(d) Challenging case, where all approaches fail, presumably due to strong heterogeneity in the prostate gland.

Figure 3.4: Four examples with different characteristics. On the left, segmentations in the image plane are depicted. Left column is the axial view, central column is sagittal view, and right column depicts the coronal view. On the right the surface distance between ground truth and prediction are shown for each approach.

TRAINING ON INDIVIDUAL DATASETS In Scenario II, we trained models separately on the individual datasets. The results for this scenario are summarized in Table 3.5. A visual presentation of the evaluation measures' distributions for each dataset is furthermore provided in the boxplots in Figure 3.5.

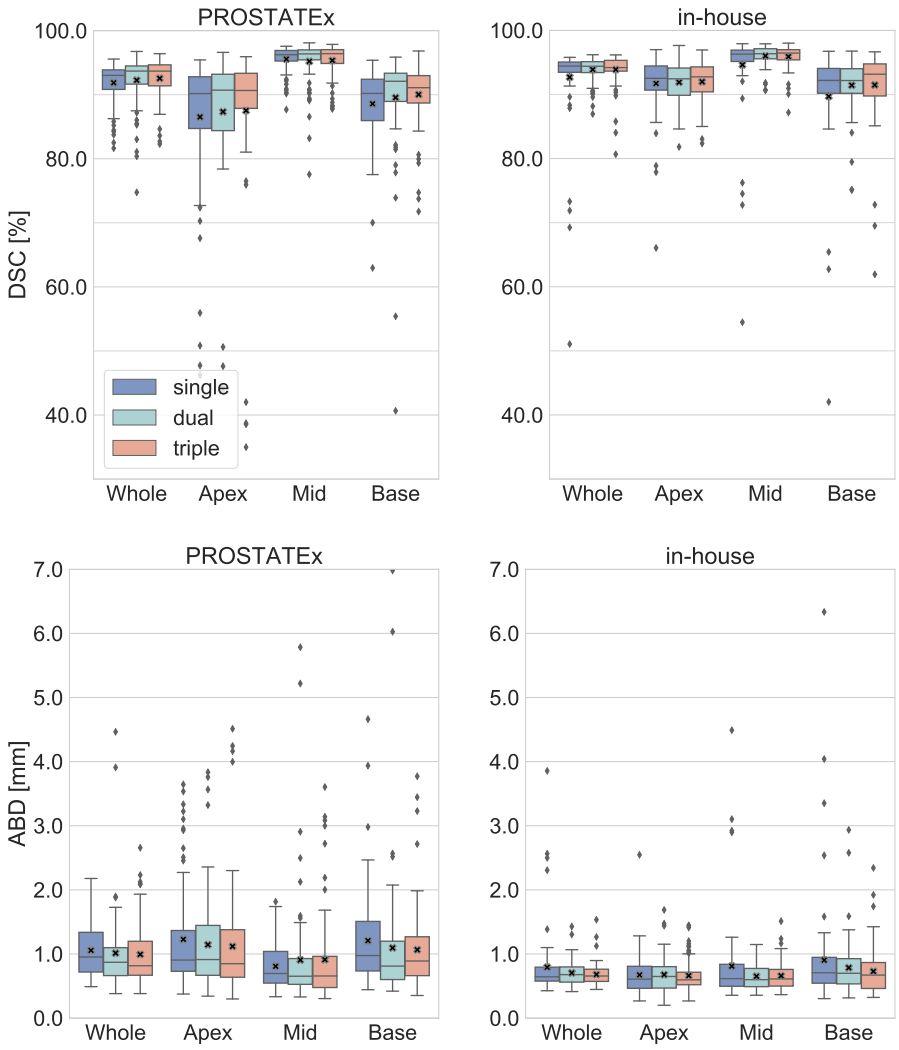
For the PROSTATEx dataset, the average DSC for the whole gland could be improved with each additional input plane from 91.9% (only axial) and 92.3% (axial + sagittal) to 92.6% (axial + sagittal + coronal). The improvement over the pure axial input was statistically significant for both dual-plane ($p < 0.001$) and triple-plane ($p < 0.01$). Statistical significant differences could also be found for the base, but not for mid-gland and apex. The average 95-HD for training on single-plane, dual-plane and triple-plane was 3.73 mm, 3.60 mm and 3.67 mm for the whole gland, respectively.

Performance for the in-house dataset was generally higher than for the PROSTATEx dataset. The average DSC for the whole gland was 92.7% for single-plane and 93.9% for both dual- and triple plane. The mean 95-HD for the in-house dataset was 2.60 mm, 2.41 mm and 2.16 mm for single-, dual- and triple-plane, respectively. These distances are smaller than the slice thickness which ranges from 2.75 mm to 3.25 mm for the individual planes.

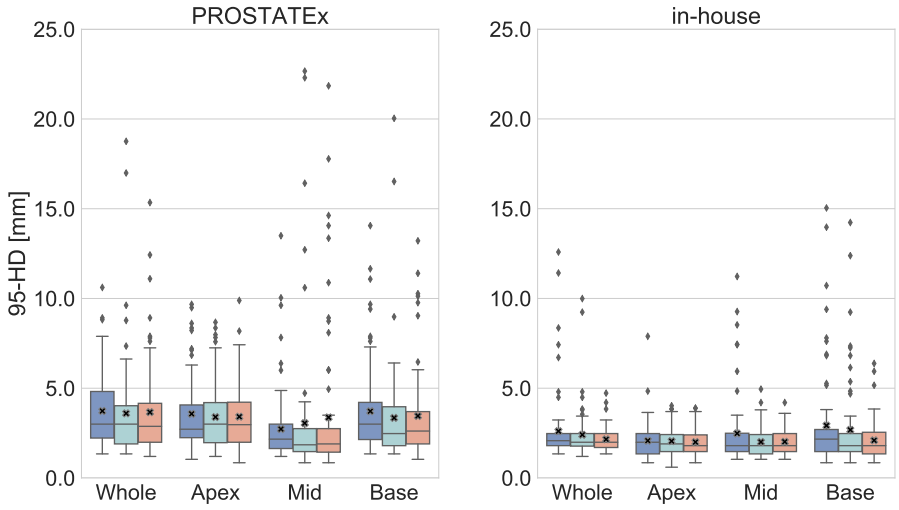
COMPARISON OF TRAINING SCENARIOS Quantitative differences between both scenarios can be examined when comparing the results in Table 3.5 to those in Table 3.4. Opposed to Scenario I (Table 3.4), where the dual-plane approach achieved the best performance for the evaluation measures in general, the triple-plane approach generally performs better than dual-plane for both datasets and the majority of regions and evaluation measures in Scenario II (Table 3.5). Thus, dual-plane seems to be more robust to variations in the training data if multiple data sources are used. However, the quantitative differences between dual- and triple-plane in both training scenarios are not statistically significant.

Moreover, we can find less significant differences between the single- and multi-plane approaches than in Scenario I. This may be caused by the fact that less training data was available for the experiments in Scenario II. While we had approximately 117 cases available in the network training for Scenario I, the training set size was effectively reduced to 47 and 70 training samples in Scenario II, for the PROSTATEx and in-house experiments, respectively. Additionally, only half of the number of test samples was available for statistical testing when the methods were evaluated on the individual datasets instead of the concatenation (merge) of the two test sets, which may further reduce statistical differences.

We also investigated whether any of the proposed network instances benefits from a specific training scenario. For this, we have split the test set for Scenario I into the two dataset-specific test sets from Scenario II



(b) ABD



(c) 95-HD

Figure 3.5: Boxplots showing the DSC scores for the whole gland and its subregions for single- dual- and triple-plane models. Models were trained and evaluated on individual datasets (**Scenario II**).

		PROSTATEx			**In-House**		
		Single	Dual	Triple	Single	Dual	Triple
DSC[%]	Whole	91.9 ± 3.6	$92.3 \pm 3.1^{***}$	$\mathbf{92.6 \pm 2.9}^{**}$	92.7 ± 6.3	$\mathbf{93.9 \pm 1.7}$	$\mathbf{93.9 \pm 2.4}^{*}$
	Apex	86.5 ± 10.1	87.3 ± 10.6	$\mathbf{87.5 \pm 12.2}$	91.7 ± 3.5	91.9 ± 2.9	$\mathbf{92.0 \pm 2.9}$
	Mid	$\mathbf{95.6 \pm 1.8}$	95.2 ± 2.4	95.3 ± 2.5	94.6 ± 6.1	$\mathbf{96.0 \pm 1.5}^{*}$	95.9 ± 1.7
	Base	88.6 ± 4.9	$89.6 \pm 6.1^{***}$	$\mathbf{90.0 \pm 4.4}^{**}$	89.7 ± 10.2	91.4 ± 3.5	$\mathbf{91.5 \pm 5.1}^{**}$
ABD[mm]	Whole	1.06 ± 0.39	$1.01 \pm 0.47^{***}$	$\mathbf{0.99 \pm 0.46}^{*}$	0.79 ± 0.49	0.70 ± 0.15	$\mathbf{0.68 \pm 0.15}^{*}$
	Apex	1.23 ± 0.76	1.14 ± 0.72	$\mathbf{1.12 \pm 0.83}$	0.67 ± 0.24	0.68 ± 0.23	$\mathbf{0.66 \pm 0.24}$
	Mid	$\mathbf{0.81 \pm 0.31}$	0.91 ± 0.61	0.91 ± 0.63	0.81 ± 0.63	$\mathbf{0.65 \pm 0.20}$	0.66 ± 0.20
	Base	1.21 ± 0.66	$1.09 \pm 0.77^{***}$	$\mathbf{1.07 \pm 0.60}^{**}$	0.90 ± 0.78	0.78 ± 0.36	$\mathbf{0.73 \pm 0.32}^{**}$
95-HD[mm]	Whole	3.73 ± 1.81	$\mathbf{3.60 \pm 2.13}^{*}$	3.67 ± 2.23	2.60 ± 1.80	$2.41 \pm 1.10\pm$	$\mathbf{2.16 \pm 0.57}$
	Apex	3.57 ± 1.80	$\mathbf{3.39 \pm 1.76}$	3.41 ± 1.78	2.08 ± 0.72	2.05 ± 0.62	$\mathbf{2.00 \pm 0.60}$
	Mid	$\mathbf{2.73 \pm 1.79}$	3.05 ± 2.71	3.37 ± 3.26	2.48 ± 1.70	$\mathbf{2.00 \pm 0.64}$	2.02 ± 0.62
	Base	3.72 ± 2.39	$\mathbf{3.35 \pm 2.32}^{*}$	$3.46 \pm 2.17^{**}$	2.92 ± 2.53	2.68 ± 2.05	$\mathbf{2.10 \pm 0.89}^{***}$

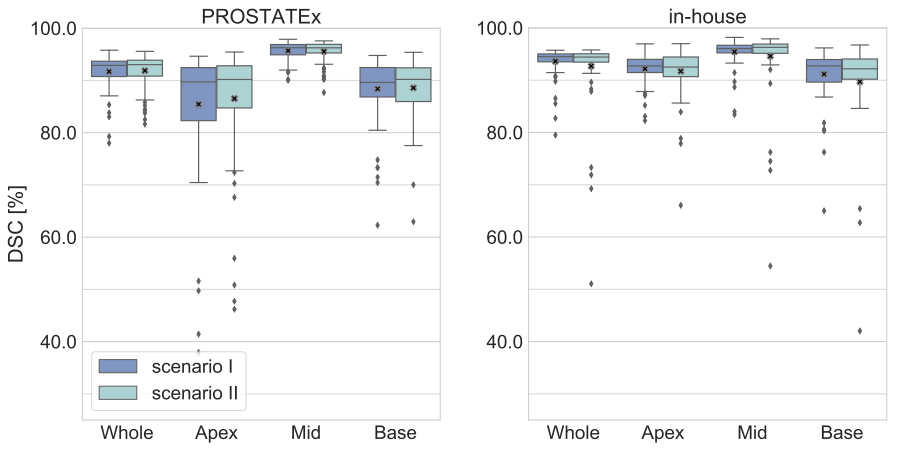
Best results are marked bold. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 3.5: Average evaluation measures for **Scenario II** (models are trained and evaluated on each dataset individually). Asterisks mark significant differences to the results of the single-plane model.

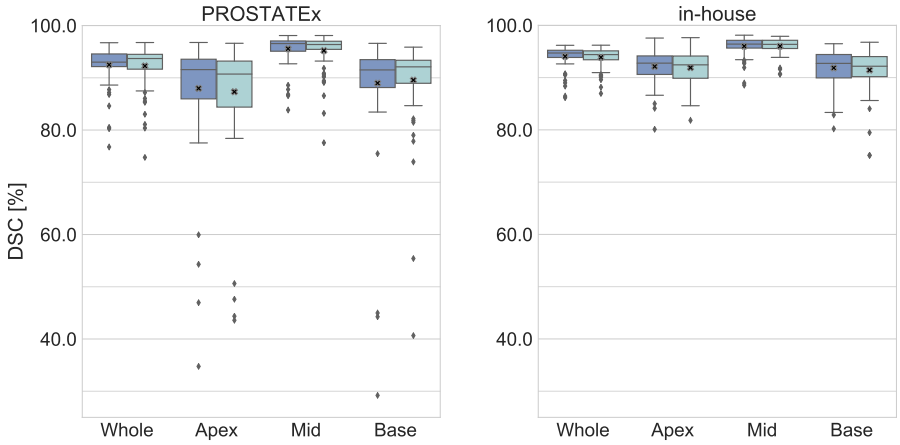
to allow direct comparison of the scenario-based results for each method (see boxplots in Figure 3.6). We could not find any consistent differences for any of the proposed methods. However, we could again observe that the performance for the in-house dataset is generally better than for the PROSTATEx dataset in both scenarios.

COMPARISON TO STATE-OF-THE-ART Table 3.1 summarizes the works that have been so far proposed in the literature for the whole gland segmentation. Highest quantitative performance scores were achieved by Isensee et al. (2021) on the official PROMISE12 test dataset with an average DSC of 91.9 % and minimum boundary distances (ABD: 1.24 mm, 95-HD: 3.95 mm).

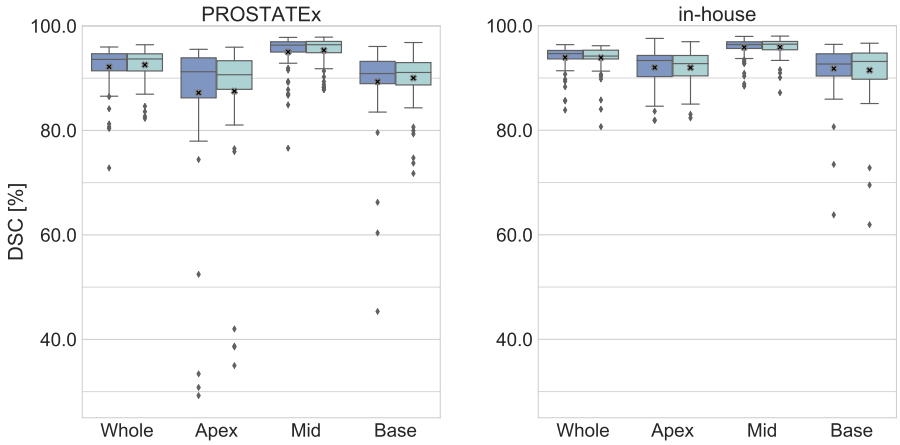
A direct comparison to our method's results can not be drawn because the image quality, dataset aggregation and size of training dataset for the PROMISE12 challenge are very different to the datasets that we have used in our work (see Table 3.3). However, we can still see that the obtained results quality of our method with a DSC of up to 93.3 %, an ABD of 0.84 mm and a 95-HD of 3.00 mm ranks in the top performances obtained for prostate segmentation methods. This is also the case when we trained the multi-plane models with less samples (n=47 and n=70) on different datasets (training Scenario II). Here we achieved average DSCs of 92.6 % and 93.9 %, respectively, for the triple plane network.



(a) single-plane



(b) dual-plane



(c) triple-plane

Figure 3.6: Boxplots comparing DSC scores for the whole gland and its sub-regions for models trained on the merged datasets (Scenario I) and the individual dataset (Scenario II). The models obtained from both training scenarios were benchmarked on the same test samples for the PROSTATEx and in-house dataset. The marker 'x' indicates the mean of the distributions.

		ensemble	multi-stream
DSC [%]	Whole	92.6 ± 0.3	92.6 ± 0.3
	Apex	87.1 ± 1.3	87.5 ± 1.2
	Mid	95.5 ± 0.2	95.3 ± 0.3
	Base	90.1 ± 0.5	90.0 ± 0.5
ABD[mm]	Whole	0.95 ± 0.40	1.00 ± 0.49
	Apex	1.14 ± 0.87	1.12 ± 0.84
	Mid	0.79 ± 0.34	0.91 ± 0.70
	Base	1.03 ± 0.66	1.07 ± 0.65
95-HD[mm]	Whole	3.10 ± 1.61	3.67 ± 2.60
	Apex	3.13 ± 1.77	3.41 ± 1.91
	Mid	2.29 ± 1.22	3.38 ± 3.99
	Base	3.01 ± 1.99	3.46 ± 2.44

Table 3.6: Comparison of two approaches for generating segmentations from triplanar input (ensemble of networks and the multi-stream network). No significant differences were found.

3.5.2 *Multi-Stream vs. Ensemble*

We compared our triple-plane architecture processing all orthogonal images simultaneously, which directly outputs a prostate segmentation, with an ensemble approach from the literature (Cheng et al., 2017; Lozoya et al., 2018). The results are listed in Table 3.6. The differences between both approaches are only minor and could not be confirmed statistically for any region and evaluation measure. No clear winning method could be extracted from these results. However, we have to point out that the ensembling method generally benefits from aggregating multiple model predictions, while the multi-stream architecture results rely only on a single output. Nevertheless, the results for the ensemble technique are in line with the outcome of our study that the input of multiple planes improves over a single-plane input.

3.5.3 *Inter-Reader Variance*

To put our automatic segmentation results into perspective, we were interested to see in what range the inter-reader variability of prostate segmentation is (see Table 3.7). In the literature, second reader segmentation evaluation has been investigated within the scope of the PROMISE12 challenge (Litjens et al., 2014b). The authors report a mean DSC of 90.0 % between two expert segmentations for the whole gland and 80.0 % and 86.0 % for the apex and base, respectively. For

		PROMISE12 (n=30)	**PROSTATEx (n=20)**
DSC	Whole	90.0 ± 3.0	93.2 ± 2.0
	Apex	80.0 ± 11.0	90.4 ± 3.9
	Mid	n.a.	96.1 ± 1.7
	Base	86.0 ± 6.0	89.3 ± 4.0
ABD	Whole	1.82 ± 0.36	0.66 ± 0.27
	Apex	2.55 ± 1.08	0.63 ± 0.29
	Mid	n.a	0.49 ± 0.23
	Base	2.21 ± 0.80	0.86 ± 0.52
95-HD	Whole	5.64 ± 1.73	3.16 ± 1.27
	Apex	6.36 ± 2.40	2.84 ± 1.15
	Mid	n.a	2.02 ± 1.05
	Base	6.28 ± 2.95	3.56 ± 1.64

Table 3.7: Evaluation measures for inter-reader variability for the PROMISE and our PROSTATEx test dataset

the whole gland, they reported an inter-reader variability of 5.64 mm for 95-HD.

Moreover, we assessed the manual performance among two expert readers for 20 cases of the PROSTATEx dataset. The average inter-reader DSC for the whole gland, apex and base for these cases were 93.0 %, 90.0 % and 89.0 %, respectively. The 95-HD was 3.15 mm for the whole gland, which corresponds approximately to the thickness of one slice. Comparing these results to the overall DSC of 93.0 % for the dual- and triple-plane model (with respect to the whole gland), we are clearly in the range of inter-reader variability.

3.6 DISCUSSION

Within this study, we assessed whether patient-level data could be leveraged to obtain a more reliable automatic prostate segmentation. Our results demonstrate that incorporating multi-planar data into the automatic segmentation does improve the segmentation outcome compared with including only the axial scan as network input. The improvements are effective in the apex and base of the gland. These regions are more challenging to segment because the partial volume effect has the highest impact on image quality here. For the mid-gland, the segmentation quality could not be improved. However, this finding is not surprising, because the axial scans already provide good contrast in the mid-gland region.

The quantitative differences between the three proposed models may not be large, but depending on the clinical application, the improved accuracy can be critical. For example, to determine the extraprostatic extension to proximal structures as seminal vesicles, a very precise definition of the boundary would be important. Moreover, it has been shown that the shape of the apex impacts the recovery of urinary continence after radical prostatectomy (Lee et al., 2006). Therefore, providing the surgeon with the most reliable delineation of the prostate gland is crucial.

Having two different datasets, we evaluated our methods within two training scenarios: (1) training on the merged dataset, and (2) training separately on the individual datasets. We observed that the quantitative evaluation measures in both scenarios are considerably better for the in-house datasets than for the PROSTATEx data. We assume that the reason for these results is two-fold: First, the number of cases in the datasets are not balanced. The in-house dataset had almost 50% more cases available for training (n=70) than the PROSTATEx dataset (n=47). Second, the reference annotations were created with different methods. While the annotations for the PROSTATEx dataset were created entirely manually, the in-house dataset was segmented semi-automatically in the first stage and later refined manually. Even when experts review and correct the semi-automatically generated segmentations, there may still be a potential bias towards the semi-automatic segmentations, which could result in more consistent segmentations than with manual delineation. One might also argue that the image quality is another factor for performance quality. However, we could not confirm this visually.

Comparing the dual- with the triple-plane network, we did not notice any large differences in either training scenarios. Consequently, questions about preference for the dual- or triple-plane variant could not be answered unequivocally. However, the dual-plane (axial+sagittal) approach seems to be a good trade-off between computational costs and segmentation quality.

Although no differences were found when comparing the results of the ensemble network and our multi-stream architecture, we think that the multi-stream approach is superior to the ensemble because it requires less parameters (factor of 2.7) and therefore is easier to deploy in production. Moreover, using a common decoder for all image orientations (as in the multi-stream architecture) can be seen as a regularizer, which can help in minimizing the generalization error on other datasets/tasks.

LIMITATIONS AND FUTURE WORK We could show that the automatic segmentation performance of the methods is in the range of expert performance. However, individual cases, as shown in Figure 3.4d, still indicate that automatic segmentations need to be further improved in the future. Approaches that can detect anatomically incorrect predic-

tions of the network, for example with shape priors (Liu et al., 2020b), could improve the method with respect to this aspect.

Future work should also focus on the architecture of the model. So far, we have only investigated the concatenation of the individual encoder branches of our network at one specific level. It would be interesting to examine, how and whether the outcome changes if the branches were fused at another location. Second, we have used only a slightly modified version of the 3D U-Net. Although the effectiveness of various architecture extensions for medical segmentation has been questioned by Isensee et al. (2021), they found that deep supervision seems to be beneficial in general. Thus, it should be incorporated in future experiments. Third, our network architecture needs to be adapted and retrained if the number of orthogonal scans changes. For the practical application, it will be helpful if the network can accept a variable number of inputs. A training paradigm, where different inputs are randomly set to empty (zero-) arrays, should be examined to make the remaining network robust to all different compositions of input scans.

In cases with multiple inputs, it is necessary to ensure that they are well aligned. Some cases required manual registration of the orthogonal scans. Therefore, automatic registration algorithms should be investigated to compensate for potential transformations among the inputs (Haskins et al., 2020). This could lead to an increased performance of the multi-planar approaches, as the manual registration may have not compensated for all motion artifacts and may be less precise than an automatic method. Lastly, it would be interesting to apply our method to other clinical use cases where multi-planar imaging is acquired (e.g., cardiac MRI).

3.7 SUMMARY

The objective of our work was to determine whether prostate segmentation performance could be increased by incorporation of patient-level data, specifically, sagittal and coronal T2w volumes. These volumes are required to be obtained in typical PCa MRI acquisition protocols and are thus easily available in clinical practice. In our work, we developed an anisotropic 3D multi-stream CNN for whole gland segmentation that allows incorporating different numbers of orthogonal input volumes.

We assessed different input compositions for our network on the basis of two datasets from multiple sites, for which we applied different training and evaluation set aggregations. The most important finding of our study is that the use of multi-planar strategies significantly improves segmentation performance when compared to using only axial volumes. The improvements could be found in particular for the apex and base regions, where the axial scans suffer from lower image quality due to partial volume effects. The improvement was consistent for all datasets and dataset-aggregations. Moreover, our methods obtained

segmentation quality of clinical experts with respect to the inter-reader variance, and could be ranked within the top performing methods of the state-of-the-art.

For implementing the method in clinical practice, future work needs to incorporate an automatic registration algorithm that aligns network inputs to each other. Moreover, other architecture variants, including different concatenation levels of the individual encoders, should be assessed to further improve the segmentation outcome.

4

EXPLOITING UNLABELED DATA

The remarkable performance of most CNN methods builds upon the availability of a large amount of labeled training data. However, obtaining the necessary amount of annotated data poses a challenge for medical image segmentation tasks, as voxel-wise labeling is very time-consuming and requires medical expertise. To this end, we propose a novel SSL method that exploits the unlabeled data from the intra-domain level to decrease the workload of manual annotations and improve the method's performance. With our method, we aim for a reliable automatic segmentation of the gland into four different structures: the PZ, TZ, AFS, and DPU. This extends the state-of-the art works that have so far only focused on the two-class segmentation of PZ and TZ. We base our method on two established algorithms from the SSL literature: uncertainty-aware self-learning and consistency-based regularization. Our results demonstrate the effectiveness of our method, which achieved results on the level of human performance and furthermore outperformed other state-of-the-art methods.

This chapter is based on the following publications:

> A. Meyer, M. Rak, D. Schindele, S. Blaschke, M. Schostak, A. Fedorov, C. Hansen, "Towards patient-individual PI-Rads v2 sector map: CNN for automatic segmentation of prostatic zones from T2-weighted MRI," in *Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 696-700, © 2019 IEEE.
>
> A. Meyer, D. Schindele, D. F. von Reibnitz, M. Rak, M. Schostak, C. Hansen (2020). "PROSTATEx zone segmentations" [Dataset]. The Cancer Imaging Archive.
>
> A. Meyer[1], S. Ghosh[1], D. Schindele, M. Schostak, S. Stober, C. Hansen, and M. Rak, 2021. "Uncertainty-aware temporal self-learning (UATS): Semi-supervised learning for segmentation of prostate zones and beyond," *Artificial Intelligence in Medicine*, 116, p. 102073.

1 Joint primary authorship. The paper builds upon the method developed within the master's thesis by Suhita Ghosh (2019), which has been directed and supervised by the thesis author.

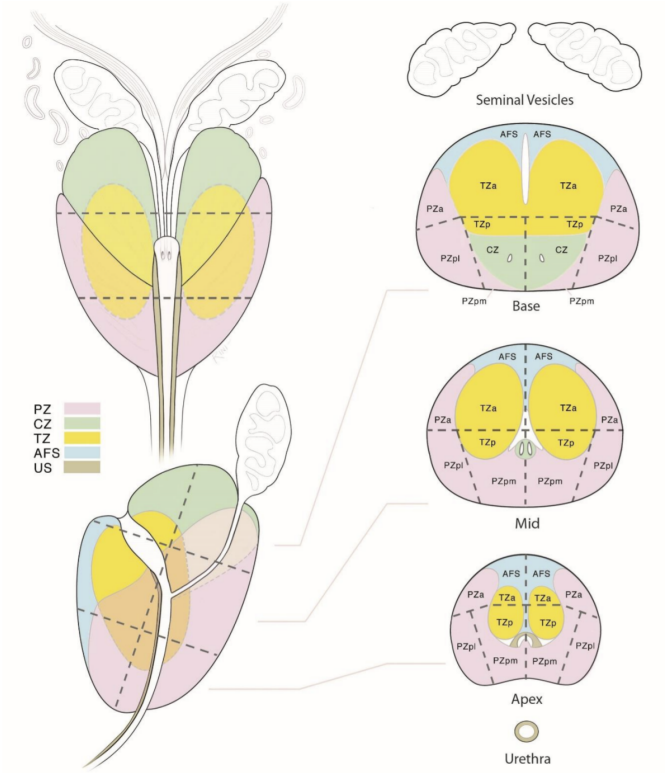
STRUCTURE OF THE CHAPTER The remainder of this chapter is organized as follows. We will first introduce the clinical context for the automatic prostate zone segmentation and motivate our technical contribution in Section 4.1. In Section 4.2, we report the current state-of-the-art on related methods to our work. After motivating our work with the limitation of previously published approaches, the subsequent Section 4.3 covers our proposed anisotropic 3D U-Net and our uncertainty aware temporal self-learning (UATS) technique. Section 4.4 describes the data and experimental design to evaluate our methods, and Section 4.5 reports the obtained results. The results, and our method's limitations, are discussed in Section 4.6. Section 4.7 concludes this chapter with a summary of our study.

4.1 INTRODUCTION

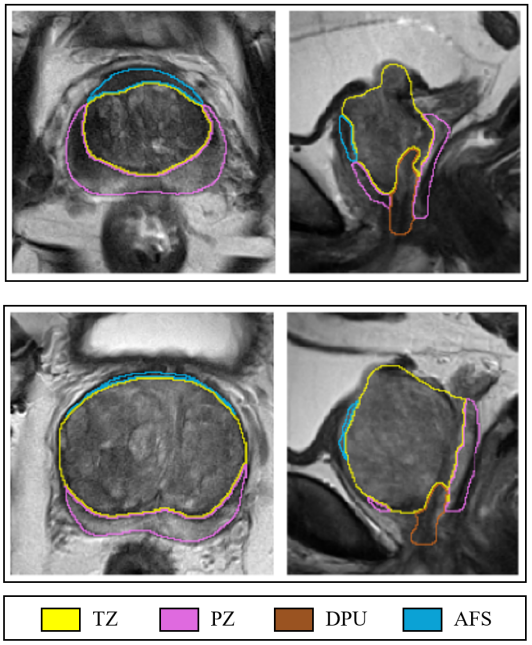
In the previous Chapter 3, we targeted the segmentation of the whole prostate. However, a more detailed segmentation of the prostate that accounts for the different anatomical zones and other interior structures can be valuable for several applications. Therefore, in the work covered in this chapter, we aim for a reliable and automatic segmentation of the PZ, TZ, AFS and DPU in T2w MRI scans. Such a detailed anatomical segmentation of the prostate can be leveraged, for example, in the context of lesion assessment with the PI-RADS v2.1 guidelines (Turkbey et al., 2019) (see Section 2.1.3).

PI-RADS v2.1 includes a so-called sector map for a more standardized lesion location assignment (see Figure 4.1). The sector map consists of 41 sectors of which 38 are related to prostate zones. The remaining three sectors are the seminal vesicles and the external urethral sphincter. The sector map should support standardized reporting and "facilitate precise localization for MR-targeted prostate biopsy and therapy, pathological correlation, and research" (Turkbey et al., 2019). Furthermore, it could also provide a "roadmap for surgical dissection at the time of radical prostatectomy" (Turkbey et al., 2019). However, having only one fixed atlas (the sector map) for the prostate that does not reflect the enlargements and different shapes of the prostate in real patients, limits its applicability. Radiologists must transfer the lesion location of the current case to the one in the sector map (Greer et al., 2018). Consequently, the sector map has not been found to be effective for the standardized communication of lesion locations, as variability among radiologists was shown to be high (Greer et al., 2018). Therefore, an automatic segmentation of the zones would be a step towards a patient-individual sector map and could increase the repeatability and consistency of reporting, having an impact on all the above-mentioned applications.

Moreover, the zonal information can be included into automatic PCa detection algorithms, which has been demonstrated to increase the accuracy of the methods (Mehrtash et al., 2017a; de Vente et al., 2020)



(a) PI-RADS v2.1 sector map. Figure from Turkbey et al. (2019), reprinted with permission from Elsevier.



(b) Manual segmentations on two T2w MRIs in axial (left) and sagittal (right) view. The sagittal T2w scan was included and segmentation masks were upsampled for better visualization.

Figure 4.1: Schematic division of the prostate according to the PI-RADS v2.1 sector map (a) which is transferred on two MRI scans with manual segmentations (b). The CZ is omitted in the manual segmentations and our work in general, because it could not be distinguished for the majority of cases.

and the volume of PZ and TZ can contribute to an automatic detection of BPH. Lastly, all the target structures could potentially be used to obtain a better registration of MRI scans to histopathological scans, as in Kwak et al. (2016).

MOTIVATION AND CONTRIBUTIONS The success of supervised deep learning approaches heavily depends on the amount of labeled examples used in training. Unfortunately, it is very challenging and expensive to get a considerable amount of high quality annotated data in the medical domain, as the annotations cannot be obtained by crowdsourcing and need a medical expert's involvement.

The dearth of good quality annotated data motivated exploring techniques that require limited supervision (Tajbakhsh et al., 2020), such as weakly supervised methods, transfer learning, and SSL. In this part of the thesis, we propose a SSL method to improve the CNN's performance for segmentation of prostate zones. To this end, we combine two state-of-the-art SSL techniques for segmentation: uncertainty-aware self-learning and temporal-ensembling. Accordingly, we name our method *uncertainty aware temporal self-learning* (UATS).

For the task of prostate zone segmentation, we designed an adapted 3D U-Net architecture (Çiçek et al., 2016) which served as backbone architecture and considers the anisotropy of the axial T2w MRI scans. We demonstrate that with the incorporation of data from the intra-domain level, our SSL method improves over the fully supervised training. We also show that our method improves the segmentation performance compared to other SSL techniques and achieved higher robustness against different levels of noise on the input data than the supervised training. Furthermore, we confirmed UATS' potential to improve upon supervised baselines on two other important biomedical datasets.

With our work, we are the first to automatically segment a more detailed anatomy of the prostate consisting of DPU and AFS, in addition to the usually segmented PZ and TZ on T2w MRI scans. Furthermore, to our best knowledge, no SSL method has been proposed so far for prostate zone segmentation.

4.2 RELATED WORK

In this section the state-of-the-art for automatic prostate zone segmentation (Section 4.2.1) and semi-supervised techniques in deep biomedical image segmentation (Section 4.2.2) is outlined. The section concludes with a summary of these methods' limitations that encouraged us to develop our method.

4.2.1 *Zone Segmentation*

The reported performance of state-of-the-art methods for prostate zone segmentation presented in this section is summarized in Table 4.1. Because the methods were developed and evaluated on different datasets, a direct performance comparison between them is impossible, but the overview can still give an impression of how the segmentation performance improved over the years.

The input to the proposed state-of-the-art methods for prostate zone segmentation is generally either multiparametric MRI or only T2w MRI scans. Similar to prostate whole gland segmentation, zonal segmentation of the prostate has been performed with deformable models, atlas-based segmentation and traditional machine learning. Early approaches by Makni et al. (2011), Litjens et al. (2012) and Toth et al. (2013) required a manual input of the whole gland contour as algorithm initialization. Qiu et al. (2014) relaxed this requirement to a couple of boundary points as manual input and proposed to use a spatial prior with an appearance model in a graph-based optimization. Later methods that have been published since 2016, were fully automatic methods that did not require any manual input. For example, Chilali et al. (2016) performed segmentation by using atlas images and evidential C-Means clustering.

The majority of automatic methods, however, employs variants or extensions of the U-Net which will be described in the following. Clark et al. (2017) proposed an architecture with four consecutive 2D CNNs. The networks are responsible for detection (classification) and subsequent segmentation of the prostate in a first and second step which is followed by detection and segmentation of the TZ in a third and fourth step. Also Zabihollahy et al. (2019) used separate 2D U-Nets to first segment the whole gland which is followed by segmentation of the TZ with a second network. Mooij et al. (2018) segmented PZ and TZ by means of a 3D U-Net based architecture that considers the anisotropic resolution of MRI scans for the architecture design: instead of overall 3D convolutions and 3D max poolings, the authors employ 2D convolutions and 2D max pooling in the high resolution directions and use 3D operations only in the last resolution layer.

Other works concentrated on an improved feature representation and propagation in their architectures. Rundo et al. (2019) added squeeze-and-excitation modules to every resolution stage of the encoder and decoder to increase representational power of the feature maps. They also showed that training on multiple datasets improved intra- and cross-dataset generalization of the network. Aldoj et al. (2020) implemented dense blocks to the U-Net architecture to improve the CNN's performance. Liu et al. (2019) developed an encoder-decoder architecture with a ResNet encoder and a (multi-scale) feature pyramid block. They extended this method with spatial attention for the input

Method	Input	Method	n_{train}	n_{test}	PZ	TZ
Semi-Automatic						
Makni et al. (2011)	mpMRI	manual gland contours & modified evid. C-means clustering	-	31	76.0 ± 6.0	87.0 ± 4.0
Litjens et al. (2012)	mpMRI	manual gland contours & linear discriminant classifier	-	48	75.0 ± 7.0	89.0 ± 3.0
Toth et al. (2013)	T2w	manual gland contours & coupled levelsets	-	40	68.0 ± ?	79.0 ± ?
Qiu et al. (2014)	T2w	manual boundary points & graph-based approach	43		69.1 ± 6.9	82.2 ± 3.0
Automatic						
Chilali et al. (2016)	T2w	probabilistic atlases & evidential C-means-clustering	-	22	62.0 ± 7.3	70.2 ± 12.1
Clark et al. (2017)	mpMRI	multiple 2D CNNs for detection and segmentation (U-Net)	78	26	-	84.7 ± ?
Mooij et al. (2018)	T2w	3D U-Net - aniso conv.	5-fold	53	60.0 ± ?	85.0 ± ?
Liu et al. (2019)	T2w	(2D) ResNet encoder, multi-scale attention block	250	63	74.0 ± 8.0	86.0 ± 7.0
Rundo et al. (2019)	T2w	2D U-Net with squeeze and excitation blocks	4-fold	40	76.6 ± 7.8	90.7 ± 3.1
Zabihollahy et al. (2019)	ADC	separate 2D U-Nets for whole gland and TZ segmentation	100	125	83.3 ± 9.6	86.3 ± 10.7
Zabihollahy et al. (2019)	T2w	separate 2D U-Nets for whole gland and TZ segmentation	100	125	86.2 ± 3.7	91.1 ± 8.9
Aldoj et al. (2020)	T2w	2D U-Net with dense blocks	141	47	78.1 ± 2.5	89.5 ± 2.0
Liu et al. (2020d)	T2w	similar to Liu et al. (2019), but with additional spatial attention	259	45	80.0 ± 5.0	89.0 ± 4.0
Qin et al. (2020)	T2w + ADC	3D ResNet - multi-scale block - channel attention - multi-directional edge loss	10-fold	202	78.5 ± 1.7	90.8 ± 1.1
Meyer et al. (2021b)	T2w	3D U-Net - anisotropic max pooling - semi-supervised (UATS) - 4 zones	78	20	79.3 ± 4.9	87.1 ± 6.5

Table 4.1: Overview of the performance (DSC [%]) of different approaches for zonal prostate segmentation in the literature. The size of the training (n_{train}) and test dataset (n_{test}) is included when information was provided in the paper. "k-fold" in the "n_{train}" column specifies a k-fold cross validation on the number of cases noted in column "n_{test}". Meyer et al. (2021b) specifies our work presented in this chapter. Please note that we targeted a four-zone segmentation, which impairs direct comparison of quantitative results.

image in Liu et al. (2020d) and measured the uncertainties via MC dropout (Gal and Ghahramani, 2016) for the automatic predictions of TZ and PZ. They found that automatic segmentations were most uncertain at the junction of PZ, TZ and AFS.

Lastly, Qin et al. (2020) integrated multi-scale pyramid convolution blocks into their network. The output of the pyramid block's channels gets weighted by an attention module. Moreover, they proposed a multi-directional edge loss, which is based on wavelet decompositions and was shown to be effective with several network architectures for the prostate zone segmentation on bi-parametric input.

4.2.2 *Semi-Supervised Segmentation*

To reduce the amount of expensive manual labels, multiple deep SSL techniques have been proposed to medical image segmentation. In the following, we focus our state-of-the-art summary to methods incorporating the concepts of pseudo-labeling (self-learning and multi-view training) as well as consistency regularization. Pseudo-labeling and consistency regularization are currently two of the most common techniques encountered in the SSL literature and have also been used in our proposed method. The mechanisms behind these techniques are described in our preliminaries in Section 2.3.

Besides these two concepts, there are several other techniques that have been employed for semi-supervised learning. For example, self-supervision can be applied to pre-train the network with the help of unlabeled data. This is realized by introducing proxy labels that do not require real ground truth, such as the prediction of image orientation (flipping or rotation angle) (Tajbakhsh et al., 2019). Moreover, other works integrated techniques such as contrastive learning (e.g., Chaitanya et al., 2020; Hu et al., 2021), shape priors (as in Chen et al., 2020b; Lu et al., 2021), or graph-based methods (e.g., Huang et al., 2021; Sun et al., 2021) to leverage unlabeled data. For a more detailed overview on other SSL concepts and methods, we refer to the surveys from Tajbakhsh et al. (2020) and Peng and Wang (2021).

SELF-LEARNING Bai et al. (2017) were the first to introduce self-learning for biomedical segmentation in the context of heart chambers segmentation. Although they refine their pseudo labels through conditional random fields, their method brought only limited performance gain. Therefore, in the context of pelvic organ segmentation, Nie et al. (2018) proposed a discriminative network for voxel-wise confidence guidance, that allows to augment the training data by only the most reliable regions of the pseudo labels. Similarly, Sedai et al. (2019) included MC dropout (Gal and Ghahramani, 2016, see Section 2.2.2) to weigh down presumably unreliable pseudo labels for retinal fundus image segmentation.

Li et al. (2020b) combined self-learning with a self-supervised strategy to improve the pseudo label quality for skin and histopathological image segmentation, that combines self-learning with self-supervision. The encoder of their network has the auxiliary task of solving a jigsaw puzzle (selecting the right permutations causing a disarranged jigsaw image). Uncertainty and an ensemble of segmentations are inferred from the recurrent optimization of the encoder, which leads to better pseudo labels and an increased performance.

MULTI-VIEW TRAINING In multi-view learning approaches, multiple networks, that are supposed to generate different predictions, are leveraged to create more reliable pseudo labels. The different views can be generated, for example, by obtaining axial, sagittal and coronal 2D planes of an input volume and train one network for each plane which has been proposed by Zhou et al. (2019) for multi-organ segmentation and by Zhao et al. (2019) for brain segmentation. The three network predictions for unlabeled data are then fused per majority voting into a consensus volume, that can be reused to create more reliable pseudo labels for the extended training data.

Xia et al. (2020) proposed a co-training strategy with N views generated by rotations and permutations, where each model is trained by an uncertainty-weighted ensemble of the predictions of the remaining $N-1$ view's models. Peng et al. (2020) combined co-training with virtual adversarial training (Miyato et al., 2018), to obtain a more diverse set of models.

CONSISTENCY REGULARIZATION A large body of research in the field of deep semi-supervised segmentation incorporates some form of consistency regularization. The unlabeled images are leveraged in this SSL concept to induce a regularization by penalizing the deviation of a model's prediction for one input sample that was subject to different perturbations (see Section 2.3.2).

Bortsova et al. (2019) measure the consistency with the two outputs of a siamese network, where each branch receives a differently transformed input per elastic deformation. Fang and Li (2020) obtained the consistency loss from the predictions of two decoders that share one encoder. Their algorithm also includes entropy minimization to push the decision boundary into the low-density regions.

However, most approaches that enforce consistency, employ the mean teacher concept (Tarvainen and Valpola, 2017). The teacher model is a historical ensemble of past iterations' model weights from the so-called student model. The purpose of this teacher model is then to provide more robust targets for the consistency loss. Cui et al. (2019) employed the mean teacher model with perturbations induced by applying noise to the images. For the task of ischemic stroke lesion segmentation, they showed improved performance of their network by incorporation of

the mean teacher strategy. Li et al. (2020a) applied a mean teacher variant that enforces consistency regarding geometric transformations for multiple segmentation tasks. Fotedar et al. (2020) could improve the performance of the mean teacher model by applying a stronger and more diverse set of transformations. They used strong intensity, geometric and image-mixture transformations to the student model's input while applying only mild transformations to the teacher model to keep its predictions reliable. They demonstrated their method's effectiveness for several medical image analysis tasks.

Other researchers proposed to constrain the consistency loss to only the less uncertain regions. For example, Yu et al. (2019) restricted the consistency loss to the most reliable regions of the teacher prediction via an MC dropout uncertainty estimate. Yang et al. (2020) implemented two U-Nets for catheter segmentation. For both networks, inter-network consistency and the intra-network consistency (similar to the πmodel) is measured. The consistency losses are limited to the less uncertain regions per MC dropout and an adversarial loss enhances contextual similarity between the outputs for labeled and unlabeled data.

Furthermore, strategies have been proposed, that evaluate the consistency at multiple levels of the network. Li et al. (2021b) improved the quality of standard mean teacher method for left atrium segmentation by applying the consistency regularization at multiple scales of the decoder either between two networks or within one network (Luo et al., 2021). Wang et al. (2020c) weighted the consistency loss with double-uncertainty: uncertainty on the prediction level and on the feature channel level, both under dropout and random noise during inference. Alternatively, the consistency can be calculated on a local (patch-based) and a global structural level as Hang et al. (2020) demonstrated for left atrium segmentation. Entropy minimization was further added to the unlabeled data predictions to motivate more confident outputs.

Very recently, combinations of the here presented methods have been proposed that further improve the segmentation performance over the other SSL methods. Wang et al. (2021b) combine the mean teacher algorithm with self-learning in a cascaded framework in the context of whole heart segmentation. In an initial stage, a mean teacher method is trained on both labeled and unlabeled data. The resulting model is supposed to be superior to a supervised model and can thus provide more reliable pseudo labels. Afterwards, a self-learning routine is applied, where the pseudo label's reliability is quantified by a shape prior obtained from an autoencoder reconstruction network. In another work by Wang et al. (2021a), the mean teacher model was combined with multi-view-training. In their method, K views are trained per mean teacher, thus $K \times 2$ models need to be updated during training.

4.2.3 *Limitation of Current Approaches*

The current state-of-the art demonstrates that CNNs can be successfully employed for automatic segmentation of the prostate zones from MRI. However, as mentioned above, these approaches targeted only the coarse division into PZ and TZ and do not consider other interior structures as AFS and DPU. Moreover, most methods that have a top performance, need a large quantity (≥ 100) of training samples.

To reduce the number of labeled samples, SSL strategies have been proposed for different biomedical segmentation tasks, that leverage the frequently available unlabeled data. Pseudo-labeling methods have shown their potential for various applications, but their benefit is limited when the pseudo label quality is not sufficient. As described above, several strategies have been proposed for selecting only the reliable pseudo labels or samples of unlabeled data, such as introducing adversarial training (Nie et al., 2018) or multi-view strategies (Xia et al., 2020).

Consistency regularization methods, and especially the mean teacher approach, are currently one of the most popular SSL techniques proposed for biomedical image segmentation. However, mean teacher methods require at least two models that need to be available during training.

Although the methods summarized in this related work section have proven to be very effective, one major downside of the successful mean teacher and the more elaborated pseudo-labeling methods is that they make the training very resource-intensive for 3D data. Therefore, patch-based (Cui et al., 2019) or sliding window (Yu et al., 2019) strategies are required that limit the receptive field of the neural network and introduce another algorithm component that needs careful optimization for training and inference (Madesta et al., 2020). This motivated us to rely on less expensive methods with respect to the GPU memory requirements, allowing to process the whole gland during training and inference.

4.3 TECHNICAL METHODS

Encouraged by the potential that SSL has demonstrated in the past for various medical imaging problems, we developed a novel method that combines pseudo-labeling and consistency regularization. To be precise, we extend uncertainty-aware self-learning by the concept of temporal ensembling (see Section 2.3.2). With the combination of these techniques, we address the challenges of computational complexity and insufficient pseudo label quality. Instead of relying on multiple models during training (as the mean teacher method or multi-view training), temporal ensembling uses a historical ensemble of model predictions that can be stored offline and thus does not increase computational ressources on the limited GPU. Furthermore, this historical ensemble has

the potential to provide more reliable pseudo labels for the self-learning routine than a single network inference.

Our SSL method uses two stages of training: (I) In a warm-up phase, a supervised model $f(\cdot)$ is trained with only labeled data $D_L = \{(x_i, y_i)\}_{i=1}^{l}$ until convergence. For each image x_i from $X \in \mathbb{R}^{H \times W \times D}$, there exists a ground truth segmentation map y_i from $Y_L \in \{0,1\}^{H \times W \times D \times C}$, where W, H, D are the dimensions of the volume and C defines the number of class labels. (II) semi-supervision is then added to improve the performance of $f(\cdot)$ by leveraging unlabeled data $D_U = \{x_i\}_{i=l+1}^{l+u}$. At this stage, we extend self-learning by temporal ensembling. The idea behind self-learning is to get an improved model iteratively through an expanded dataset $D_L \cup D_U$ comprising the labeled dataset D_L along with the unlabeled images X_U and their pseudo labels $\hat{y}_i \in \hat{Y}_U$ with $\hat{y}_i = f(x_i, \theta)$, such that $D_U = (X_U, \hat{Y}_U)$. To constrain the influence of wrong pseudo labels, we base on the most confident predictions using uncertainty measures. We combine this idea with concepts derived from temporal ensembling, where the pseudo-labels are updated with the ensemble predictions rather than the current epoch's prediction. Also, a consistency loss is calculated between the current and ensemble predictions, which enforces consistency between the current and the previous epochs' predictions, preventing huge gradient updates.

Prior to the development of our method, no algorithm has been proposed to incorporate the concept of temporal ensembling for biomedical segmentation. However, concurrently with our method, Cao et al. (2020) developed an uncertainty aware temporal ensembling method for the mass segmentation in breast ultrasound images[2]. The major difference to our method is, that Cao et al. (2020) relied only on the temporal ensembling method and restricted the consistency loss to the most certain ensemble regions. Contrary, we propose to use temporal ensembling *and* self-learning, where we consider all voxels for the consistency loss but restrict the input for the self-learning to the most confident regions.

In the following Section 4.3.1, we first describe our backbone network architecture, for which we designed an anisotropic U-Net variant and the post-processing. Hereby, we focus our methods on the multi-class segmentation of prostate zones which was our main objective in this study. We will then continue with the details on our SSL technique in Section 4.3.2, which can be applied to all types of segmentation tasks as we will see in our evaluation study in Section 4.5.3.

4.3.1 *Anisotropic 3D U-Net*

We implemented a variant of the 3D U-Net as backbone architecture that considers the anisotropic nature of the T2w prostate MRI scans for its design (see Figure 4.2).

2 Their method was published in the beginning of October, while our draft for the manuscript was circulated to the co-authors end of September.

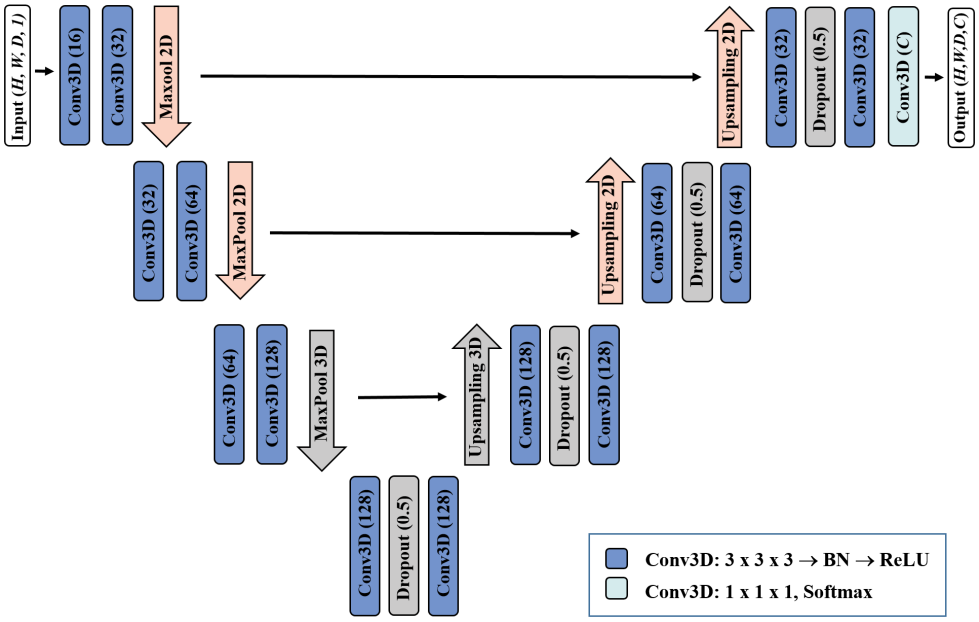


Figure 4.2: Proposed anisotropic architecture of the network for zonal segmentation. The architecture is based on the 3D U-Net (Çiçek et al., 2016). The orange arrows highlight the location of 2D max pooling and 2D upsampling operations.

In the encoder, the image is downsampled by means of three resolution steps. Each layer in one resolution step consists of two $3 \times 3 \times 3$ convolution filters with ReLU activation and a successive max pooling operation. On the way down, the number of filters increases from 16 for the first layer to 256 in the bottom layer. In contrast to the original architecture, we used anisotropic $2 \times 2 \times 1$ max pooling to take the highly anisotropic input data into account. Only the last max pooling is isotropic with $2 \times 2 \times 2$. Similarly, the decoder path with transposed convolution ($3 \times 3 \times 3$ kernel) employs a stride of 2 in each dimension for its first layer. It is followed by two $3 \times 3 \times 3$ convolution layers with decreasing number of filters. With respect to the anisotropic downsampling, we used transposed convolution with a stride of $2 \times 2 \times 1$ for the last two decoder resolution steps to maintain the symmetrical design of the U-Net.

As in the original architecture, we employed skip connections to transfer high resolution information from the encoder path to the same level of the decoder path. Batch normalization after every convolution was added for faster learning. As regularization, dropout with a rate of 0.5 was included in the bottom most layer and in the decoding layers. The last layer of the network is a $1 \times 1 \times 1$ convolution with softmax activation function and a resulting output of 5 channels: one each for TZ, PZ, DPU, AFS and background. Due to its 'winner-takes-it-all'-nature, the softmax function is optimal for creating a preferably non-overlapping multi-class segmentation.

POST-PROCESSING To obtain a final topologically correct multi-class segmentation for the zone segmentation, we post-processed the prediction of the network $\hat{y}_i$ as follows. First, the prediction gets thresholded. Second, for every class in the thresholded prediction, a connected components analysis is applied and only the largest component is kept. Voxels resulting in a label-free state after connected components were now assigned a new label with:

$$\hat{y}_{c,i} = \begin{cases} 1, & \text{if } \hat{y}_{c,i} = \max_{c \in C} \mathrm{SDF}(\hat{y}_{c,i}) \,, \\ 0, & \text{otherwise} \,, \end{cases} \tag{4.1}$$

with $\mathrm{SDF}(\cdot)$ being a signed euclidean distance function that assigns positive values inside and negative values outside the segmentation. Consequently, every voxel that had a label-free state, gets assigned to the zone of the nearest labeled voxel according to the shape-based distance measure.

4.3.2 *Semi-supervised Learning*

Instead of starting our SSL-based training from scratch, it is intuitive to first train a model in supervised fashion on (X_L, Y_L) only, leveraging better pseudo label predictions for X_U for the initialization of our SSL technique and thus avoiding degenerated models. As a second step, our semi-supervised method is applied, for which an overview is depicted in Figure 4.3. A pseudo-code for the method can be found in Algorithm 1 at the end of this section. We start the training with the expanded dataset comprising the ground truth (manual) labels and the pseudo labels, which are derived from the pre-trained model. Thus, the pseudo labels act as targets for the unlabeled samples. The overall loss function L_{Total} in the SSL stage contains two components: *task* and *consistency* loss and is defined as weighted combination

$$L_{Total} = L_{Task} + \lambda L_{Cons} \,, \tag{4.2}$$

where λ is the consistency loss weighting coefficient.

TASK LOSS We chose the continuous DSC (cDSC) (Shamir et al., 2019) to implement the task loss L_{Task}. It can handle probabilistic segmentation better than the regular DSC that requires at least one binary input. This exempts us from defining any thresholds on the

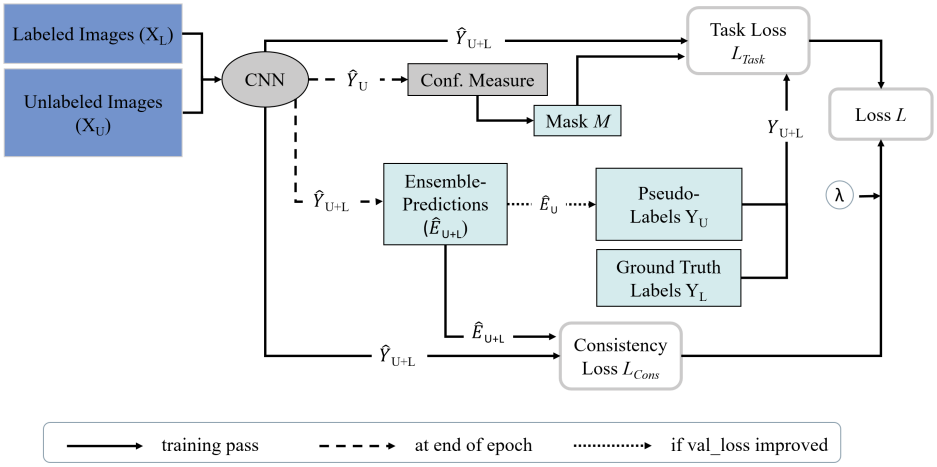


Figure 4.3: Concept of our uncertainty aware temporal self-learning (UATS).

network predictions and from losing information the network provides with its probabilistic output. Our task loss is defined as:

$$L_{Task} = L_{cDSC}(y, \hat{y}, m) = \\ - \sum_{c \in C} \frac{2 \sum_{i=1}^{N} m_i \hat{y}_{c,i} y_{c,i} + \epsilon}{a_c \sum_{i=1}^{N} m_i y_{c,i} + \sum_{i=1}^{N} m_i \hat{y}_{c,i} + \epsilon}, \quad (4.3)$$

where $\hat{y}$ is the model's prediction and $y \in Y_{L \cup U}$ is the corresponding ground truth or pseudo label and ϵ a small constant to ensure numerical stability. N is the total number of voxels with i being an index for a specific voxel. C is the set of different classes and a_c is a specific coefficient for the cDSC (see Shamir et al. (2019) for details). Mask $m \in \{0,1\}^{H \times W \times D}$ is one for all voxels of Y_L and for the n most confident voxels of Y_U, and zero otherwise. We will provide details on the definition of the n most confident voxels after describing the temporal ensembling component in the next paragraph.

TEMPORAL ENSEMBLE As a novelty, we propose to use the temporal ensemble of predictions $\hat{E}$ used in the consistency loss (Laine and Aila, 2017) as the pseudo labels. With an ensemble of predictions, we consider multiple hypotheses rather than a probable noisy single hypothesis. Therefore, this strategy reduces the effect of noisy labels generated for the unlabeled images. $\hat{E}$ is an exponential moving average over epochs and its update at each epoch t is defined as in the original work from Laine and Aila (2017):

$$\hat{E} \leftarrow \alpha \hat{E} + (1 - \alpha) \hat{y}, \quad (4.4)$$

with α being defined as the momentum term controlling the contribution of historical data to the ensemble (Laine and Aila, 2017). In the original temporal ensembling approach, L_{cons} is included right in

the beginning of the network training and the ensemble is initialized as zero vector. Because we start our SSL method with a previously trained supervised model, we can initialize $\hat{E}$ prior to the first epoch with the supervised model's prediction. This spares us from applying a ramp-up weight for L_{cons}. Theoretically, this exempts us as well from applying the start-up bias correction on the ensemble (see Section 2.3 or (Laine and Aila, 2017)), but we found its inclusion to be beneficial for the overall performance due to experiments (see Appendix A.3).

To increase the reliability of the ensemble, we propose to update $\hat{E}$ only during those epochs where the validation loss decreases, i.e., when the model performs well on the unseen data. To be precise, $\hat{E}$ is updated class-wise, such that the labels for the classes are updated only when the class-specific loss improved on the validation loss.

UNCERTAINTY MEASURES We constrain the influence of wrong pseudo labels on the task loss by including only the n most confident voxels per batch as pseudo labels Y_U. We evaluated MC dropout (Gal and Ghahramani, 2016) (see Section 2.2.2) and the historically averaged softmax prediction of the temporal ensemble as confidence measures for the pseudo-label selection process.

For the MC dropout uncertainty estimation, we used the entropy of the resulting multiple predictions that are obtained by activating dropout at inference. With f denoting a forward pass, MC entropy H for every voxel is defined as:

$$H(F) = -\sum_{c \in C} \left(\frac{1}{F} \sum_{f=1}^{F} \hat{y}_{c,f} \right) log \left(\frac{1}{F} \sum_{f=1}^{F} \hat{y}_{c,f} \right). \tag{4.5}$$

To extract the most confident voxels per class, we first got the indices of those voxels belonging to class c and having the maximum softmax prediction (or the minimum entropy H) across classes. Then, for each class, the indices of the top n_c confident voxels are returned. An example for a resulting confidence mask can be found in Appendix A.4.

CONSISTENCY LOSS Consistency loss L_{Cons} is obtained by calculating the dissimilarity between the (bias corrected) ensemble predictions $E_{L \cup U}$ and the current network predictions $\hat{Y}_{L \cup U}$. Hence, L_{Cons} acts as a regularizer enforcing a smoother gradient update. In the original temporal ensembling method designed for classification, the dissimilarity is measured with mean squared error. For our segmentation task, we base L_{Cons} on the cDSC (Shamir et al., 2019) of the two segmentations (Equation 4.3) as it is less sensitive to class imbalance and has been found to work best in preliminary experiments on the validation set. We define the consistency loss as the dissimilarity (Cha, 2007) between the current prediction $\hat{y}$ and E as:

$$L_{Cons} = 1 - L_{cDSC}(\hat{y}, E, m = \mathbf{1}). \tag{4.6}$$

Algorithm 1: Uncertainty-aware Temporal Self-Learning (UATS)

Input: supervised model f_θ, input images X, set of indices for labeled images L, set of indices for unlabeled images U, ground truth labels Y_L, number of confident voxels per class n, *num_epochs*, α, *early_stop_condition*

Output: semi-supervised model f_θ

/* Initialize semi-supervised training. For reasons clarity, we differentiate between confidence masks for the task (M_{Task}) and the consistency loss (M_{Cons}). */

$M_{\text{Task},i\in U} \leftarrow \mathbf{0}$ // initialize voxels of task mask with zeros

$M_{\text{Task},i\in L} \leftarrow \mathbf{1}$ // set task mask to 1 for all labeled images

$M_{\text{Cons},i\in L\cup U} \leftarrow \mathbf{1}$ // set cons. mask to 1 for all voxels

$\hat{E}_{L\cup U} \leftarrow f_\theta(X_{L\cup U})$ // initialize ensemble prediction

$E_{L\cup U} \leftarrow \hat{E}_{L\cup U}$ // initialize target vectors

$v_{conf} = \emptyset$ // initialize confident voxels

$cur\_val\_loss \leftarrow \infty$, $min\_val\_loss \leftarrow \infty$

/* semi-supervised training */

for *t in [1, num_epochs]* **do**

  while *not early_stop_condition* **do**

    $M_{\text{Task},i\in U}[v_{\text{conf}}] = 1$ // update mask for unlabeled images

    $Y_U \leftarrow E_U$ // assign ensemble to pseudo labels

    for *each minibatch* $b \in \{L \cup U\}$ **do**

      $\hat{Y}_{i\in b} \leftarrow f_\theta(X_{i\in b})$

      $loss \leftarrow L_{Task}(\hat{Y}_{i\in b}, Y_{i\in b}, M_{\text{Task},i\in b})$
      $+\lambda L_{Cons}(\hat{Y}_{i\in b}, E_{i\in b}, M_{\text{Cons},i\in b})$

      update θ with Adam optimizer

    update *cur_val_loss*

    /* update ensemble class-wise if class-specific validation loss improves */

    for *each class c* **do**

      if *cur_val_loss[c] < min_val_loss[c]* **then**

        $\hat{E}[c] \leftarrow \alpha\hat{E}[c] + (1-\alpha)\hat{Y}[c]$ // update ensemble

        $E \leftarrow \text{bias\_correction}(\hat{E})$ // update target vectors

        $min\_val\_loss[c] \leftarrow cur\_val\_loss[c]$

    $v_{\text{conf}} \leftarrow \text{get\_conf\_voxels}(f_\theta, \hat{E}, n)$ // update conf. voxels

Mask $m \in \{0,1\}^{H\times W\times D}$ is one everywhere, as all the voxels are considered irrespective of their confidence.

4.4 EXPERIMENTAL SETUP

The following sections cover the experimental setup to assess our methods described in the previous Section 4.3. Our experiments are mainly carried out on a prostate MRI dataset, because our main objective was to achieve robust and reliable segmentation outcomes of the prostate's interior anatomy. However, as our SSL strategy is applicable much more

	Prostate Zones	**Hippocampus**	**Skin Lesions**
Image Type	T2w MRI	T1w MRI	color photographs
Dimensions	3D	3D	2D
Classes	4	2	1
Size $D_{L,\text{train}}$	78	210	2494
Size $D_{U,\text{train}}$	236	130	1000
Size $D_{L,\text{test}}$	20	50	500

Table 4.2: Overview of the datasets used for the evaluation of our method.

widely, we decided to investigate its ability to generalize on other challenging tasks, too. To this end, we evaluated our method additionally on a skin lesion and hippocampus segmentation dataset. We first describe the three datasets used in our study in Section 4.4.1 and summarize the training details, which are partly task-dependent, in Section 4.4.2. Then the design of our experiments is outlined in Section 4.4.3.

4.4.1 *Data*

In the following, we provide more details on our benchmark datasets for prostate zones, hippocampus and skin lesion segmentation, as well as their respective pre-processing and augmentation strategies. We refer to Table 4.2 for a quick overview about the datasets. The table summarizes the main dataset characteristics and the ratio of labeled/unlabeled data as well as the number of training and hold-out test images for each task.

PROSTATE ZONE SEGMENTATION We used the publicly available SPIE-AAPM-NCI PROSTATEx challenge dataset (Litjens et al., 2017a; Litjens et al., 2014a; Clark et al., 2013) for the evaluation of our method for the task of prostate zone segmentation. The images were acquired by two different types of Siemens 3T MRI scanners (MAGNETOM Trio and Skyra) with a pelvic phased array coil at the Radboud University Medical Centre (Radboudumc). Overall, we extracted 334 T2w volumes (axial, sagittal and coronal volumes) and created manual ground truth (PZ, TZ, AFS and DPU zones segmentation) for randomly selected 98 cases. The manual segmentations of the prostate zones were created on the axial volume with additional consideration of the sagittal volume with 3D Slicer (Fedorov et al., 2012) by a medical student and subsequently corrected by an expert urologist (Reader 1). As hold-out test data, we used 20 randomly selected labeled cases.

For evaluating the inter-reader variability, a second ground truth was created by a second expert urologist (Reader 2) with the help of another medical student for these 20 test cases. Additionally, for 10 out of the 20 test cases, a third expert segmentation was generated by

an assistant radiologist (Reader 3). Segmentations from Reader 1 (the same who segmented the training data) was set as our ground truth. Segmentations from Reader 2 and Reader 3 were evaluated against this ground truth for the inter-reader variance. The average performance of Reader 2 & 3 forms the inter-reader level.

Pre-processing and Augmentation: The original volumes had varying resolution of $[0.3 - 0.6] \times [0.3 - 0.6] \times [3.0 - 5.0]$ mm. Therefore, we resampled the volumes' resolution to a common spacing of $0.5 \times 0.5 \times 3$ mm. The volumes were cropped to a unified size of $168 \times 168 \times 32$ voxels by considering the intersecting volume of the axial, sagittal and coronal T2w sequence. To be clear, we only used the cropped and resampled axial T2w volume as network input. The other volumes were only considered for automatically determining the region of interest.

The intensities were cropped to the first and 99th percentile and subsequently normalized to an interval of [0,1]. We applied online augmentation of the training data by random application of the following transformations: left-right flipping, 3D rotation, scaling and 3D translation. Instead of nearest neighbor interpolation, we applied a shape-based interpolation as proposed in (Herman et al., 1992) for the augmentation which produced smoothly transformed segments despite the anisotropic resolution.

HIPPOCAMPUS SEGMENTATION For hippocampus segmentation, we used Task04 of the Medical Decathlon Challenge (Simpson et al., 2019). The dataset consists of two classes: the hippocampus proper (CA1-4 and dentate gyrus) and parts of the subiculum, which together are more frequently named as hippocampal formation. The T1w sagittal volumes were acquired with a Philips Achieva scanner at the Vanderbild University Medical Center (Nashville, TN, USA). The dataset contains 390 T1w images of healthy people and patients with non-affective psychotic disorders. For 260 out of the 390 images, labels are provided as training data. The remaining 130 samples are originally used as test data in the challenge with labels withheld from the public. Thus, we randomly set aside 50 labeled samples from the original training data for our testing purposes and used the original test data (n=130) as unlabeled (training) data for our UATS method.

Pre-processing and augmentation: The dataset's volumes are provided with uniform spacing of $1.0 \times 1.0 \times 1.0$ mm for all volumes. We standardized the sizes of the volumes to $48 \times 64 \times 48$ voxels and normalized the intensities to an interval of [0,1]. As data augmentation, we applied 3D rotation, scaling and 3D translation.

SKIN LESION SEGMENTATION For skin lesion segmentation we used the ISIC 2018 challenge data (Codella et al., 2018; Tschandl et al., 2018). The dataset consists of high-resolution color photographs of the

skin from all anatomic sites. It contains both benign and malignant lesions with a higher percentage of the first. Images were acquired with a variety of dermatoscope types and from different institutions. The size of the images ranges from 566×679 to 4499×6748 pixels. The original challenge dataset provides 2594 labeled training and 1000 unlabeled testing samples. For our experiments, we used 500 randomly selected labeled samples from the original training dataset as our test data. We used the original unlabeled test data as unlabeled training data for our UATS method. In summary, we included 2094 labeled and 1000 unlabeled images for training and 500 labelled images for testing.

Pre-processing and augmentation: We employed intensity normalization and resized the images to 256×192 pixels during pre-processing. Data augmentation included rotation, translation, scaling as well as left-right and top-down flipping.

4.4.2 *Training Details*

We used the same training strategy for all the named tasks, but the underlying model differs slightly due to the nature of the data. For our main task prostate zone segmentation, we applied the anisotropic 3D U-Net that we developed with respect to the nature of the 3D MRI scans (see Section 4.3.1). For the other two tasks we employed dataset-specific variants derived from this architecture as our aim was not to achieve state-of-the-art results but to fairly compare the proposed UATS approaches with the supervised baseline. Specifically, for hippocampus segmentation, we employed a similar architecture as backbone, but with isotropic max pooling in all resolution layers and starting with 32 filters in the first convolutional layer. Similarly, we used a 2D U-Net variant for skin segmentation with a starting filter size of 64 in the first layer. The probability of the dropout in each resolution layer of the decoder path is 0.5 for all models.

The models were trained using ADAM (Kingma and Ba, 2015) optimizer based on 75% of the training data. The remaining 25% were used as validation data during training. The supervised baseline models were trained with the multi-class DSC loss (see Appendix A.1 for details). All the models were trained for a maximum of 300 epochs. We used early stopping in all the experiments, where training was terminated when no improvement could be detected for the validation loss within 30 epochs. The model with the overall lowest validation loss was selected and used for the evaluation on test data. For the prostate segmentation, we post-processed the predictions for the supervised baseline and the UATS method as described in Section 4.3.1.

The model hyperparameters were selected empirically. In all the temporal ensembling and UATS experiments α was set to 0.6. The consistency coefficient was set to $\lambda = 1$ unless the consistency loss dominated the task loss. In this case, we set $\lambda = 0$ epoch-wise to ensure

the task loss is the main driver. To evaluate the MC entropy as uncertainty measure, we implemented 10 forward passes of the network. The remaining hyperparameters as confident pseudo label voxels, learning rate and batch size were set task-specific and can be found in Table A.1 in Appendix A.2.

4.4.3 *Evaluation Design*

We carried out several experiments to investigate the performance and effectiveness of our methods. In this section, we describe the design of the experiments, the results of which are then presented in Section 4.5.

For our evaluation, we report the DSC and the ABD as evaluation measures, and refer to the appendix for additional insights of the 95-HD for our prostate zone experiments.

As stated before, we focus our evaluation on the task of prostate segmentation. To put the quantitative results of our methods for this task into context, we compared them with the respective inter-reader variability (Section 4.5.1) that was assessed as described in Section 4.4.1.

For evaluating the automatic deep learning methods, we conducted a randomly sampled four-fold cross validation for different training and validation splits if not stated otherwise. For each method, the four trained models were then evaluated individually as described in Section 2.4.1.

In Section 4.3, we proposed an anisotropic 3D U-Net for the prostate segmentation, that should better account for the characteristics of the prostate MRI data than its isotropic counterpart. We compared the performance of both architecture variants for the fully supervised segmentation on the labeled dataset D_L to assess whether the model benefits from the anisotropic architecture (Section 4.5.2).

Subsequently, we investigated the performance of our semi-supervised UATS method (both the MC entropy and the softmax variant), which we trained on $D_{L \cup U}$ (Section 4.5.3). Furthermore, we conducted experiments that evaluated (1) the performance of other state-of-the-art SSL methods, (2) the impact of the different components of our UATS method in an ablation study, (3) UATS's robustness against noise in the input images, and (4) the effectiveness of UATS for other biomedical segmentation tasks (generalizability). We will give more details on these experiments' design in the following passages.

COMPARISON TO THE SSL STATE-OF-THE-ART To compare our UATS method to other state-of-the-art SSL techniques, we evaluated the performance of temporal ensembling as in Laine and Aila (2017) with L_{Cons} as in Equation 4.6 and self-learning similar to Bai et al. (2017). For this self-learning version, we used continuous predictions as pseudo labels in combination with the cDSC, which worked better than thresholding the predictions. Bai et al. (2017) updated the pseudo labels

only after a specific interval of 50 epochs regardless of the quality of the model concerning the validation data. Consequently, pseudo labels quality may decrease and the method may suffer from these low-quality labels. Thus, we apply a variant that updates the pseudo labels always and only when the validation loss improves (Self-Learning Update).

ABLATION STUDY We conducted an ablation study to investigate the individual effect of different components on the overall performance. For this, we employ the UATS softmax variant and make the following changes to its implementation: First, we omitted the consistency loss L_{Cons} and only optimized the network parameters during training with L_{Task}. Second, we set the parameter for the percentage of confident voxels to $n = 100\%$. This way, *all* pseudo label voxels are considered as confident and not only the top n voxels (with highest probability) per class. Thus, no confidence measure was included in the UATS method. And third, instead of an ensemble of predictions $\hat{E}$, we used the current prediction of the network $\hat{y}$ as pseudo labels Y_U.

ROBUSTNESS AGAINST NOISE In this experiment, we investigated the robustness of the supervised baseline and both UATS methods against varying levels of noise. For this, we applied different levels of noise on the test data and evaluated the performance of the methods. We applied Gaussian additive noise with $\mu = 0$ and varying σ to our test images. Because MRI images have different intensity ranges, we applied the noise to the normalized test image. Then, we normalized the noisy images again to obtain the same intensity range from [0,1] as for the training images. We applied noise with $\sigma = \{0.01, 0.025, 0.05, 0.1, 0.2\}$ which corresponds to an average signal-to-noise-ratio ($SNR = \frac{\mu_{img}}{\sigma_{noise}}$) of $SNR = \{26.8, 10.7, 5.4, 2.7, 1.3\}$. An example case with increasing noise is displayed in Figure 4.4.

GENERALIZABILITY To show that UATS is generally applicable, we additionally benchmarked it on a hippocampus and skin lesion segmentation dataset and for varying amounts of labeled data. We set up the experimental scheme for all datasets as follows. We randomly sampled 10 %, 25 %, 50 % and 100 % of the available labeled training samples and applied the plain supervised baseline and UATS. Hereby, we used always the same amount of unlabeled samples (Size $D_{U,train}$ in Table 4.2) and the full validation sets. For the hippocampus and skin experiments, we included additionally 5 % of labeled data. The random sampling was repeated three times for different train/validation splits, averaging result qualities to get reasonable grounds for comparison.

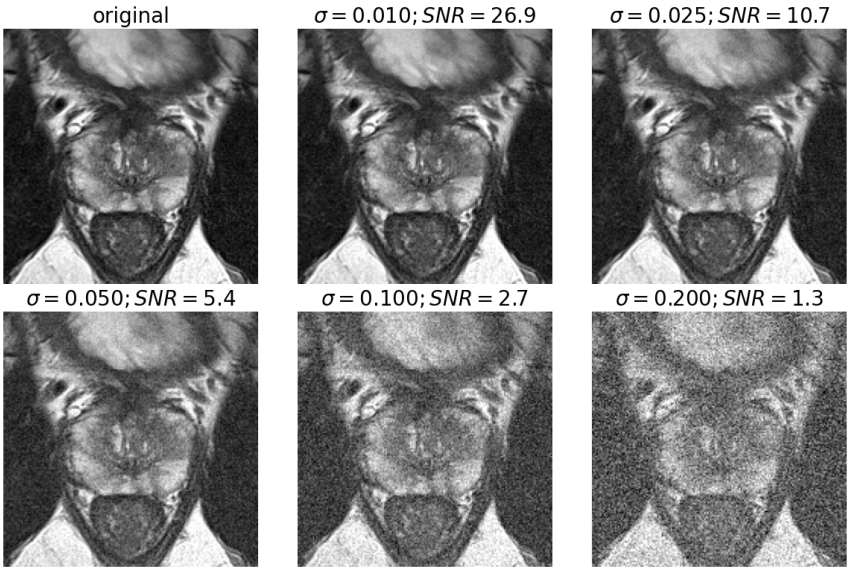


Figure 4.4: Application of additive Gaussian noise with varying σ to one example test case.

4.5 RESULTS

The following sections report the results for our experiments described in the previous Section 4.4. Before assessing the impact of our proposed SSL method, we begin with details on the inter-reader variance (Section 4.5.1) and continue with the comparison of the supervised anisotropic and isotropic 3D U-Nets (Section 4.5.2). The results for the assessment of our UATS method are then reported in Section 4.5.3 regarding

- the effectiveness of MC entropy and softmax uncertainty measures,
- the performance of other state-of-the-art SSL techniques,
- an ablation study for different components of the UATS algorithm,
- UATS's robustness against noise in the input images,
- the effectiveness of UATS for other biomedical segmentation tasks (generalizability) and with varying size of D_L.

Lastly, we rank our UATS results with respect to other segmentation results reported in the literature for the prostate zone, hippocampus and skin lesions.

4.5.1 *Inter-reader Variance*

We evaluated the performance of different clinical experts to obtain an estimate for the inter-reader variance. Details on the results can be found in the top rows of Table 4.3. On average, the inter-reader

Algorithm	**PZ**		**TZ**		**DPU**		**AFS**	
	DSC (%)	ABD (mm)	DSC (%)	ABD (mm)	DSC (%)	ABD (mm)	DSC (%)	ABD (mm)
Manual								
Expert1 vs. Expert2	81.8 ± 3.4	0.70 ± 0.35	87.8 ± 5.8	0.86 ± 0.31	60.6 ± 8.9	1.33 ± 0.49	51.0 ± 11.1	1.91 ± 1.10
Expert1 vs. Expert3	78.0 ± 5.4	1.02 ± 0.60	82.8 ± 5.7	1.07 ± 0.34	64.1 ± 4.9	1.26 ± 0.39	46.8 ± 15.1	2.42 ± 1.24
∅ Inter-reader-level	79.9 ± 4.2	0.86 ± 0.45	85.3 ± 5.8	0.97 ± 0.32	62.4 ± 7.8	1.30 ± 0.46	48.9 ± 12.6	2.16 ± 1.15
Supervised								
Supervised ANISO	77.4 ± 5.7	1.08 ± 0.67	86.7 ± 7.2	0.91 ± 0.36	70.6 ± 14.7	1.08 ± 1.32	46.1 ± 13.6	3.55 ± 3.18
Supervised ISO	75.4 ± 6.4***	1.23 ± 0.87***	85.9 ± 6.7***	1.01 ± 0.42***	68.4 ± 18.7	1.74 ± 5.35	42.0 ± 14.8**	4.00 ± 3.57**
Proposed Methods								
UATS Entropy	78.9 ± 5.0***	**0.97 ± 0.59*****	**87.3 ± 6.5****	**0.89 ± 0.36**	73.6 ± 9.9***	0.87 ± 0.46**	**50.1 ± 10.3*****	**2.95 ± 2.05***
UATS Softmax	**79.3 ± 4.9*****	1.00 ± 0.65***	87.1 ± 6.5	0.92 ± 0.38	**74.9 ± 9.7*****	**0.83 ± 0.43*****	49.5 ± 11.9***	2.96 ± 2.29***
Semi-Supervised								
Temporal Ensembling	77.4 ± 6.1	1.18 ± 0.89	87.0 ± 6.5	0.93 ± 0.42	71.6 ± 12.0	0.95 ± 0.56	44.2 ± 11.9	3.28 ± 2.77
Self-Learning as in Bai et al. (2017)	77.1 ± 5.9	1.13 ± 0.82	84.4 ± 6.4***	1.14 ± 0.46***	68.8 ± 13.2**	1.07 ± 0.70***	43.4 ± 11.2**	3.41 ± 2.30
Self-Learning Update	77.0 ± 5.6	1.29 ± 0.98**	85.4 ± 6.0***	1.05 ± 0.37***	71.2 ± 11.0	0.97 ± 0.63	41.2 ± 12.9***	4.42 ± 3.38***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. p-values according to Wilcoxon signed-rank test.

Table 4.3: DSC and ABD of different prostate zone segmentation strategies. Best results are marked bold. Asterisks indicate significant differences in the distributions of the marked and the supervised ANISO method.

performance was 79.9 %, 85.3 %, 62.4 % and 48.9 % for PZ, TZ, DPU and AFS, respectively. We can see that there exist clear differences in the manual segmentations. For example, Reader 2 was much closer to Reader 1 than Reader 3. This highlights that the delineation of the zones is challenging. The clinicians confirmed that the AFS is the most difficult structure to segment as boundaries are not clearly visible for a large part and the structure has high variety of shape and appearance. Thus, the inter-reader variability is very high. We also expect the intra-reader variability to be high, but need to confirm this with further experiments in future work.

4.5.2 *Supervised Baseline*

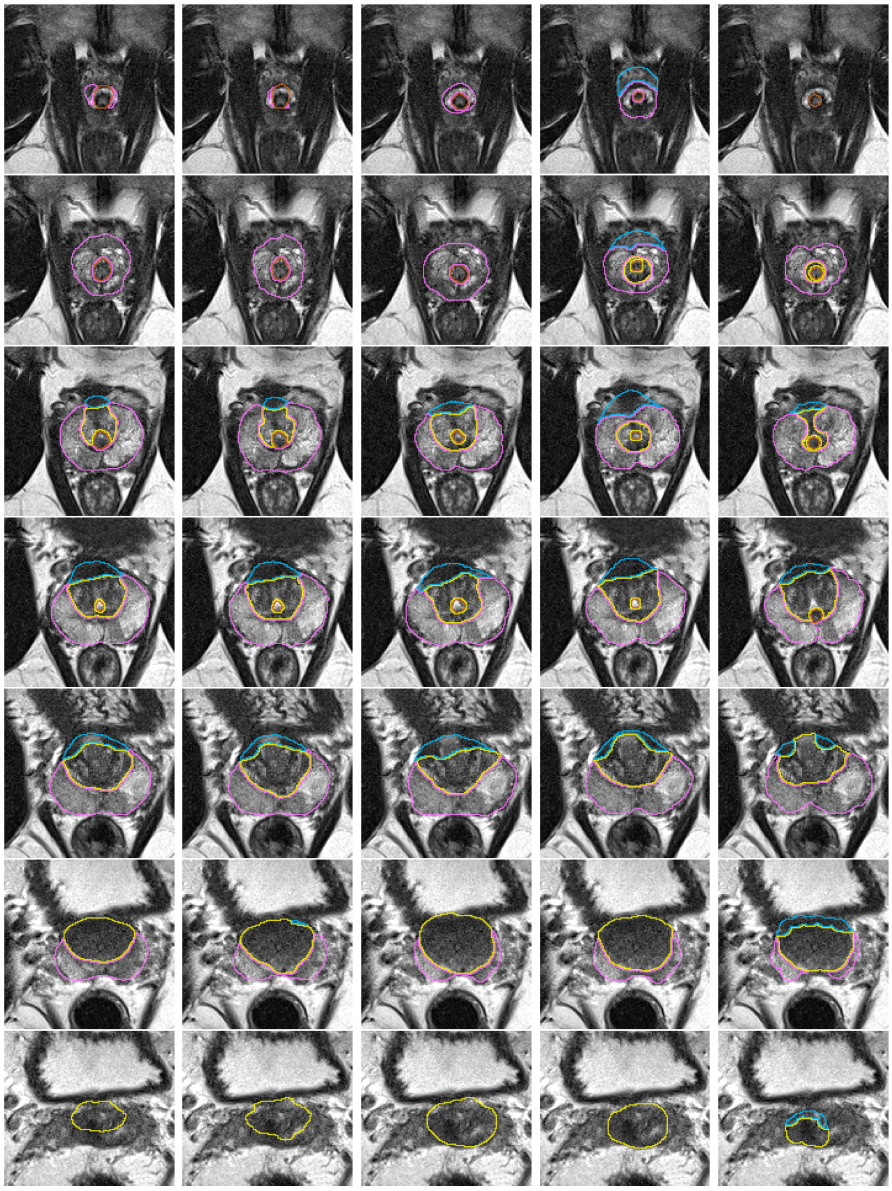
The quantitative results for the anisotropic and isotropic U-Net variant are summarized in Table 4.3. The anisotropic variant consistently improved performance over the isotropic standard architecture and achieved DSCs of 77.4 % vs. 75.4 % for PZ, 86.7 % vs. 85.9 %for TZ, while DPU and AFS resulted in DSCs of 70.6 % vs 68.4 % and 46.1 % vs. 42.0 % respectively.[3] Wilcoxon signed rank test confirmed these improvement findings for PZ, TZ and AFS ($p < 0.01$). Our results of the anisotropic supervised method are in the range of average inter-reader variability for TZ and even above for DPU. For the other two zones, however, the manual segmentations are of considerably better quality than the automatic segmentations, indicating the need for further improvement. Smaller volumes generally have the tendency to obtain lower accuracy for region-based measures, such as the DSC, because smaller errors have a larger weight on the overall measure. Consequently, it is not surprising that DPU and AFS obtained worse results than TZ and PZ. Distance-based measures such as ABD show that the quality of automatic DPU segmentation is still good. On the other hand, regarding the AFS, the supervised automatic method clearly needs improvement. The high standard variance demonstrates that some cases are nearly as good as manual segmentations but many cases are not.

4.5.3 *Semi-supervised Learning*

Qualitative results for our UATS method are visualized in Figure 4.5. Results of our experiments on UATS performance are summarized in Table 4.3 and in the boxplots in Figure 4.6.

As summarized in Table 4.3, UATS Entropy and UATS Softmax significantly outperformed the supervised baseline for all zones with most performance gain for the minority classes DPU and AFS. TZ

3 We have to point out that the results for both variants of the 3D U-Net differ from our previous work (Meyer et al., 2019) because in that work we did not use any validation data for the final model. Thus, the effective training data size was larger in (Meyer et al., 2019). However, the method is the same for both training procedures.



(a) Softmax (b) Entropy (c) Reader 1 (d) Reader 2 (e) Reader 3

Figure 4.5: Example segmentation results of one test case for the UATS softmax and entropy approaches and the three readers. The four structures PZ (pink), TZ (yellow), DPU (brown) and AFS (blue) are depicted.

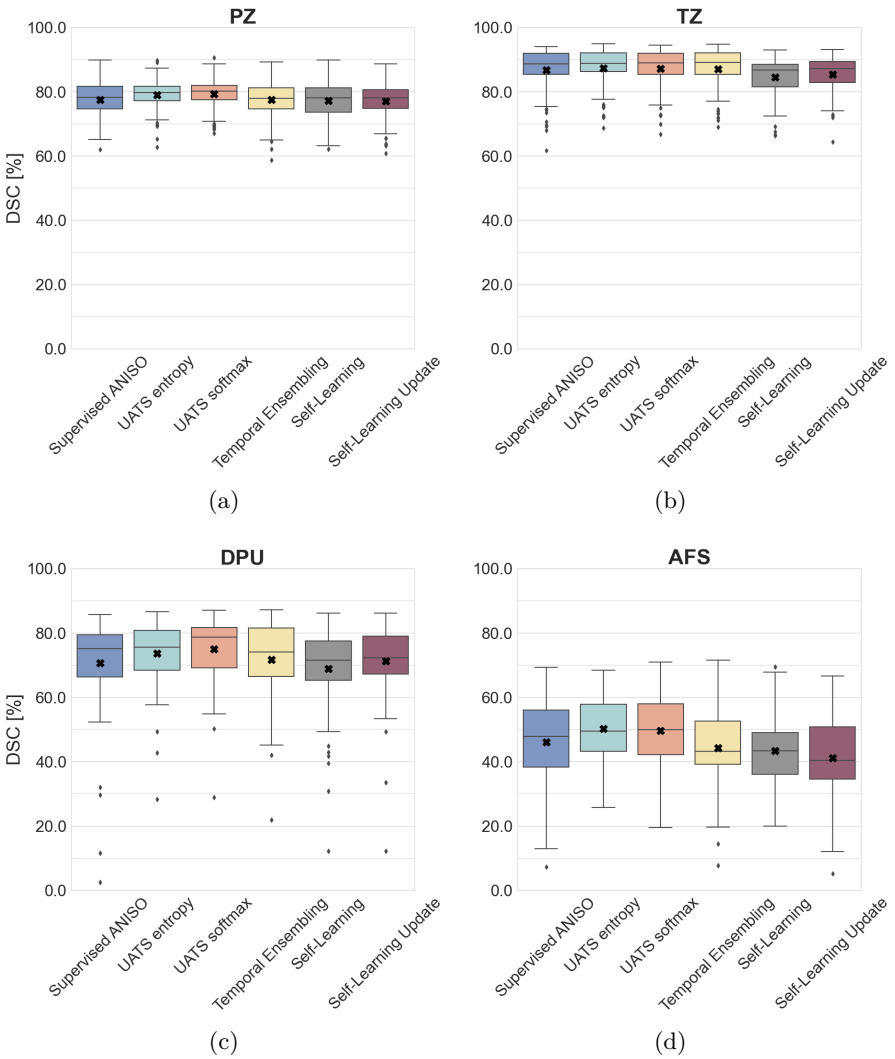


Figure 4.6: Boxplots for the segmentation results of the four zones of the prostate. The marker ('x') represents the mean DSC value. Results are given as the DSC of the ground truth and automatic segmentation. The state-of-the-art methods are shown in comparison the supervised baseline and to our proposed UATS method.

gained only little from semi-supervision, irrespective of the method considered.

With respect to the average inter-reader performance, our UATS' performance is on the level of human expert performance for all structures. The UATS segmentations yield even higher DSC for all structures except PZ. We carried out further analysis of the model results for PZ and examined in which region most of the automatic and expert segmentation disagreement occur. We found that most deviations from the inter-reader level can be encountered in the upper (cranial) third of the PZ, which is located in the prostate's base in proximity to the seminal vesicles. This region suffers from partial volume effect due to the high slice thickness and the intensity similarity of PZ and seminal vesicles. While the automatic segmentation only processes the (anisotropic) axial scan, the human readers could additionally verify their segmentations in the sagittal scans, in which the seminal vesicles can be better distinguished from PZ tissue. We assume that this is one reason why automatic performance is lower than the inter-reader-level for PZ. Another reason could be that, for PZ, the amount of labeled data needs to be increased to better cover the structure's variety.

The above statements are valid for both UATS confidence measures: softmax probability and MC dropout entropy. However, we would recommend the softmax probability for this task. This is because both performances are about equal and the softmax probability is cheaper to compute because it does not require several forward passes.

COMPARISON TO SSL STATE-OF-THE-ART The performance of other SSL state-of-the-art methods can be found in Table 4.3 and boxplots with more details are illustrated in Figure 4.6. Although temporal ensembling showed promising results for TZ and DPU, it could not achieve consistent and significant improvement over the supervised variant across all classes.

It is interesting to see that the two relatively simple pseudo labeling approaches generally lead to a performance decline compared to the supervised baseline for most structures. This demonstrates that for the prostate dataset, naive self-learning setups potentially suffer from the false predictions they produce during the self-learning cycle. However, in combination with the uncertainty awareness and consistency loss, we see that UATS improves significantly over the supervised baseline and the other state-of-the-art methods.

ABLATION STUDY In our ablation study, we analyzed the effect of different components of the UATS method, which was trained with the softmax uncertainty. Quantitative results are visualized in the boxplots in Figure 4.7.

Considering the results of the experiment that omits L_{Cons}, we found decreased performance for the smaller structures DPU and AFS when

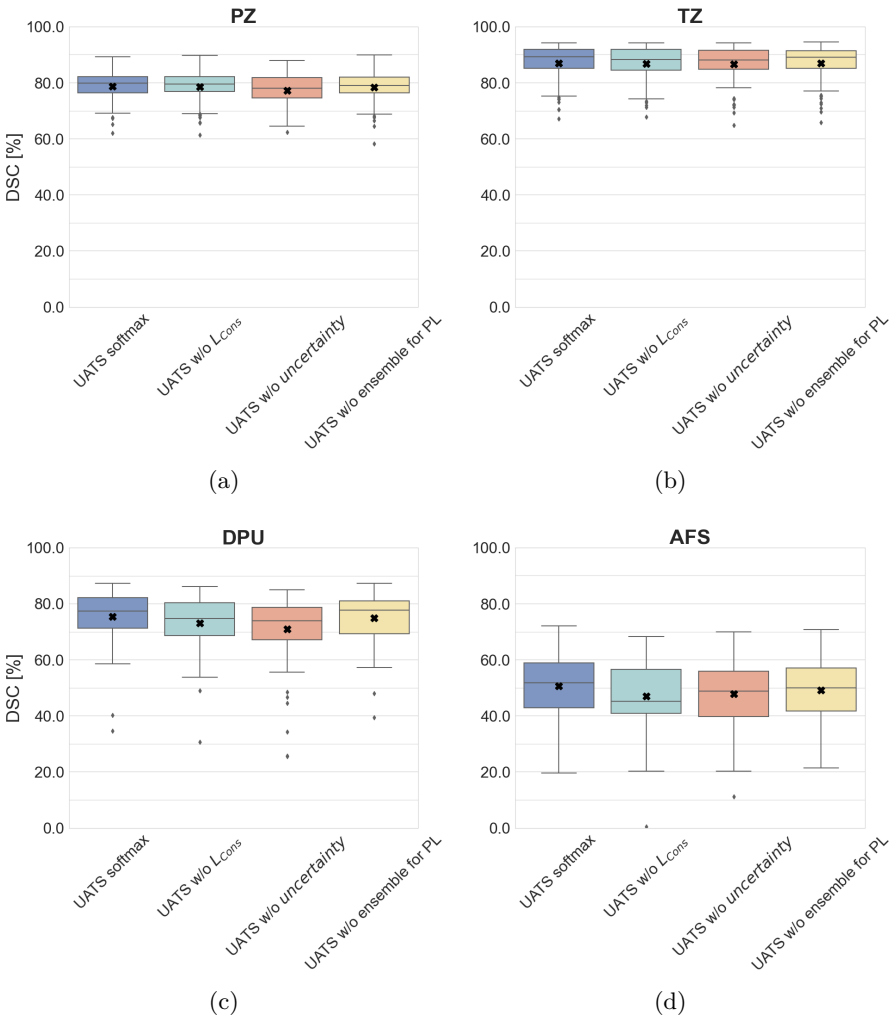


Figure 4.7: Boxplots for the segmentation results of the ablation study in comparison to the UATS softmax variant. Results are given as the DSC of the ground truth and automatic segmentation. The marker ('x') represents the mean DSC value.

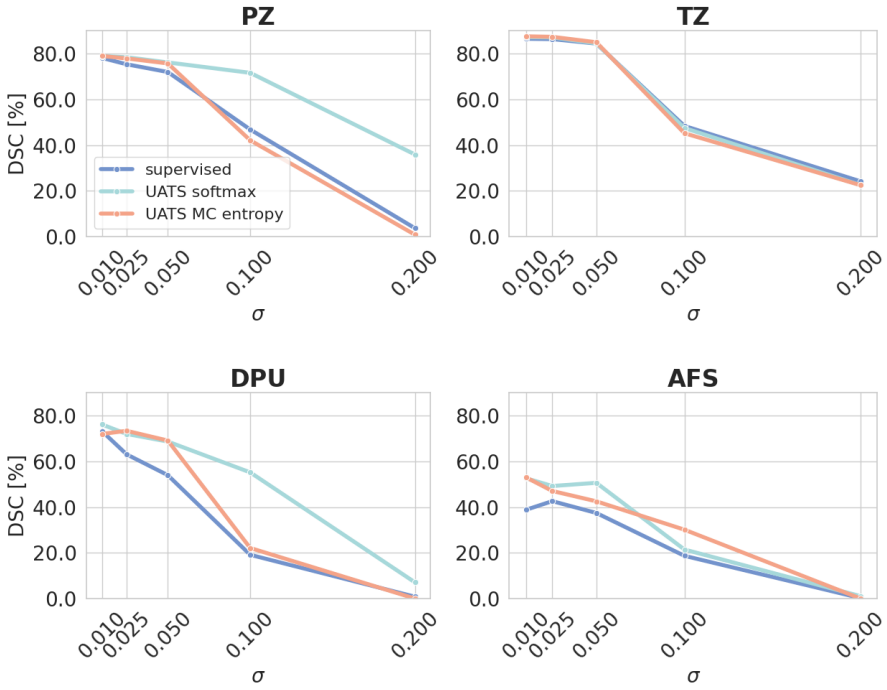


Figure 4.8: Performance of the supervised baseline with UATS softmax and UATS entropy on the test dataset with added Gaussian noise with varying σ. Performance measure is DSC.

the network was trained only with L_{Task}. Laine and Aila (2017) demonstrated that temporal ensembling could better cope with noisy labels than a simply supervised baseline. We can assume that this has a more massive effect on smaller labels because the ratio of falsely labeled voxels is larger when the structure's total size is relatively small.

By only considering the n most confident voxels of each class, we mitigate the likelihood of falsely labeled voxels in our pseudo labels. If we consider all pseudo voxels for the labels, we see an evident performance decline for PZ, DPU and AFS in comparison to our proposed UATS softmax. This demonstrates that for a beneficial effect of self-learning, the choice of labels incorporated into the task loss is essential. This finding is in line with many studies that implement self-learning and some form of uncertainty awareness (e.g., Nie et al. (2018)).

In contrast to the consistency loss and confidence guidance, the ensemble of predictions for pseudo labels does not contribute evidently to the improvement as the confidence and the consistency loss. We hypothesize that selecting the most confident voxels already reduces the false voxel label's contribution, and the consistency loss compensates further for the false voxels.

ROBUSTNESS AGAINST NOISE We evaluated the performance of the supervised baseline, the UATS softmax and the UATS entropy approach under varying additive Gaussian noise. The results are plotted

for DSC in Figure 4.8. The general performance drops for all approaches with noise strength above $\sigma = 0.05$. For TZ, all three approaches perform similar. For PZ, AFS, and DPU, however, the UATS softmax suffers much less from noise, indicating its increased robustness against noise compared to the other two approaches.

We did not apply any intensity augmentation in our training. Consequently, we assume that all methods would generally improve their performance with increasing noise, if they had seen it during training. On the other hand, we can still infer from this experiment that seeing more data during training, even when no labels are available, can lead to potential robustness of the method.

GENERALIZABILITY ACROSS TASKS To evaluate the generalization capability of our method, we assessed its performance for other datasets (hippocampus and skin) and for varying numbers of labeled data. Quantitative results of our generalizability experiment are presented in the diagrams in Figure 4.9. We included visual comparisons of the supervised and UATS predictions for the different datasets used for this experiment in Appendix A.5. Reviewing the results, we can conclude that UATS outperformed the supervised baseline irrespective of the dataset and the number of labeled samples. Only for the hippocampus segmentation, the UATS entropy variant lead to decreased results for 100 % of labeled data. However, the supervised baseline and the UATS entropy outcome are still very close with less than 1 % difference in their average DSC (87.2 % vs. 86.7 %).

The general tendency is that the smaller the amount of labeled samples, the larger the gain from unlabeled samples. This is a common observation in SSL methods, also mentioned in Bai et al. (2017). The rationale is quite intuitive, at some point, the variability in appearance and shape are well-covered by the labeled samples, yielding diminishing returns from unlabeled samples. Additionally, for 100 % of labeled data, there were more labeled than unlabeled samples available for the skin and hippocampus tasks, which reduces the effect of the unlabeled data further. The only exception from this intuitive finding is the DPU segmentation with 10 % of labeled samples, where the gain from UATS is small. Presumably, this is caused by the fact that 10 % equals 6 labeled samples, which might be generally insufficient for a reasonable DPU segmentation.

The above findings are true for both UATS confidence measures. Although there exist some difference in their respective performance, we could also find that these were generally rather small.

COMPARISON TO STATE-OF-THE-ART As summarized in Table 4.1, various approaches have been proposed for the segmentation of only PZ and TZ. When comparing our UATS method with other approaches, UATS performs in the mid-range. But a direct comparison

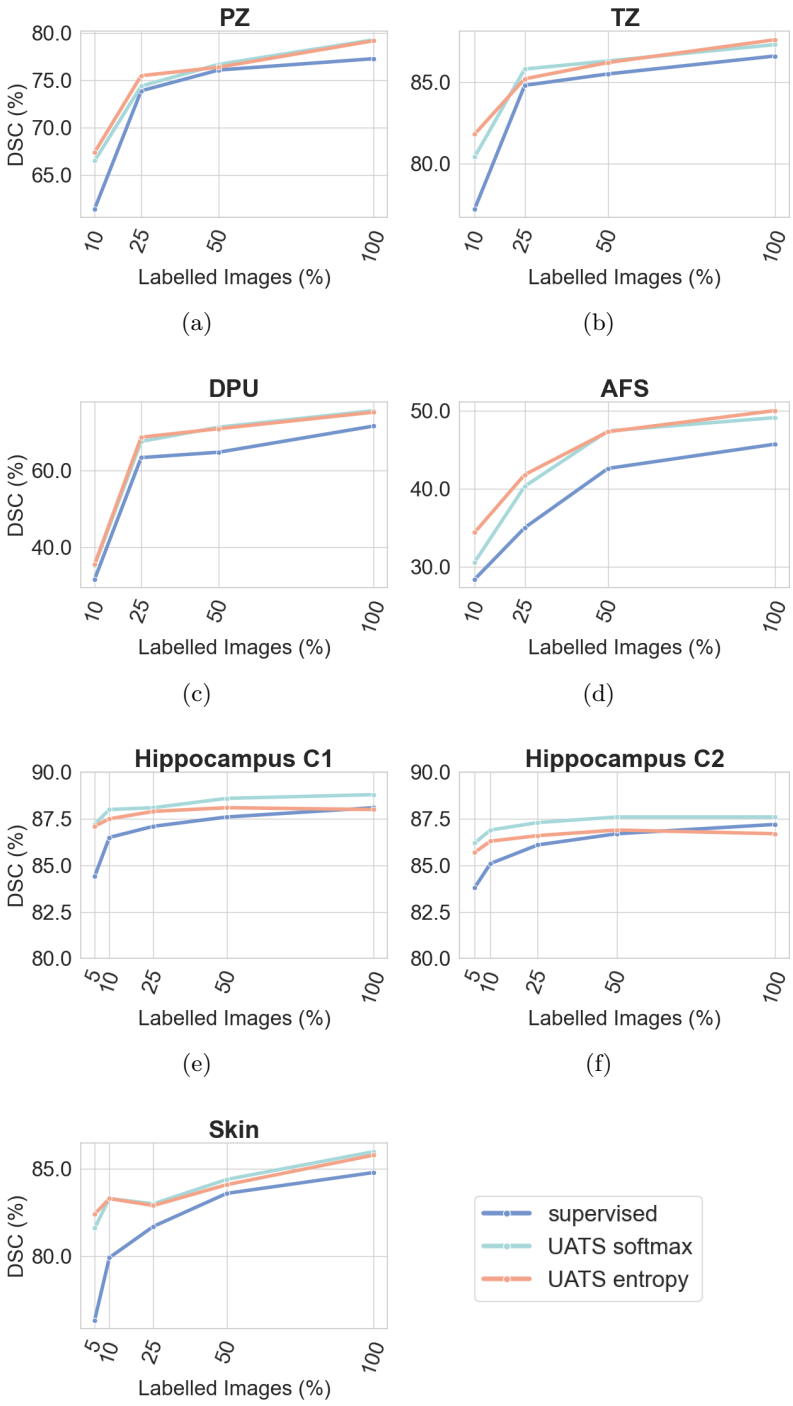


Figure 4.9: Supervised and UATS methods' performance (DSC) depending on the amount of labeled samples for the tasks of prostate zone, skin lesion and hippocampus segmentation.

to other methods can not be drawn for two reasons. First, we target a more detailed anatomy inside the prostate such that our problem definition is more difficult than the two class segmentation of PZ and TZ. And second, the underlying datasets are different. They differ in sample size, acquisition protocols, included sequences and image quality.

For the ISIC2018 skin lesion segmentation challenge, hundreds of methods are listed in the live leader-board, and the currently best one[4] achieves a DSC of 91.5 % while our UATS method obtained a DSC of 86.0%. We can only assume why there is such a performance gap between our method and the top participants. One reason might be the varying amounts of training data. As we set 500 labeled samples for test purposes aside, we have 500 labeled samples less for training. Another reason could be that we did not evaluate on the same and a smaller test dataset. Furthermore, the models for the challenge are maximally fine-tuned to this specific task. We, on the other hand, used a very basic backbone architecture. Additionally, an ensemble of models, more sophisticated pre-processing and augmentation strategies, and more complex network architectures (e.g. multi-scale or attention-based) have been used. Our method does support such techniques, but it is beyond the scope of our work to implement them.

The Medical Decathlon challenge focuses on developing methods that have high generalization capacity on different segmentation tasks. We followed a similar strategy, and thus our results are more comparable to this challenge than the ISIC2018. In the ongoing Medical Decathlon challenge, the best method's[4] result achieved on hippocampus is a DSC of 90.0% for class 1 and DSC of 89.0% for class 2. With UATS, we achieved a DSC of 88.8% and 87.6%, respectively. This places our method in the range of the twelfth-best result (88% and 88%), whereas we had less labeled data for training (with the same reason we have given for ISIC2018) [4].

4.6 DISCUSSION

In this work, we proposed a novel semi-supervised learning strategy in the context of a detailed prostate zone segmentation that has not been targeted before by other researchers. We found that adapting the network architecture to the anisotropic nature of the MRI scans lead to a performance increase for all structures. Moreover, in our experiments, we could show that our proposed semi-supervised method has the potential to leverage additional unlabeled data to increase the outcome quality for prostate zone segmentation and other biomedical segmentation tasks. This demonstrates the benefit of exploiting data from the intra-domain level of the clinical data structure.

Our method incorporates a temporal ensemble (of predictions) that is used for the consistency loss and to create higher quality pseudo

4 as of July, 9th 2021

labels. Other SSL methods (e.g., Yu et al., 2019; Li et al., 2021b; Wang et al., 2020c), on the other hand, employ the mean teacher method (Tarvainen and Valpola, 2017) to create another type of ensemble based on model weights. The authors of the mean teacher method argue that ensembling of the model weights has two advantages over prediction ensembling: (1) the ensemble update can be carried out at every iteration (instead of every epoch) and scales better to large datasets, (2) better representations can be obtained because the averaging affects all layers of the model and not just the output. We decided on the temporal ensemble of predictions because it is more GPU memory efficient. However, the prediction ensemble could be easily exchanged with the model ensemble. It would be of value for future work to evaluate whether there exist difference in performance between these two variants.

The proposed method is based upon a rather simple U-Net architecture. We chose this architecture because of its successful application to a multitude of segmentation tasks. However, we want to point out that UATS is network-independent and could also be applied to any other network architecture.

The segmentation quality for the AFS zone is rather low for both the manual and automatic approach. This highlights that more care should be taken for the ground truth segmentations, for example with a consensus segmentation among multiple readers. This could produce more consistent labels and could potentially improve the automatic segmentation result, too. However, even if the segmentation quality is not yet perfect, we believe that it is of sufficient quality to be of valuable information for the consistent lesion location assignment (via the PI-RADS sector map) and for clinical studies.

We compared softmax probability and MC dropout entropy as measures to select the most confident voxels. In general, we could not find consistent and significant differences among these two approaches. We assume that this is because we applied the relative selection of confident voxels instead of an absolute one (only considering those above a threshold). Thus, for both methods, the uncertain boundary region is usually avoided in the pseudo label selection, irrespective of the confidence measure's absolute value, which causes the method to generate similar confidence masks.

In another experiment, we investigated the robustness of the methods against noise. Although both UATS variants demonstrated more robustness against noise than the supervised baseline, we could also observe that this is the only experiment where UATS softmax performed clearly better than UATS entropy (for DPU and AFS). As this contrasts with our findings that both variants do not differ for other experiments, it will be of high interest to investigate whether this observation is consistent for other tasks and datasets.

Besides noise, there are various other factors that impact the quality of images. For example, images can be blurry due to motion of the

patient during acquisition, or they can be corrupted by bias field. Additionally, the images may have been obtained by another acquisition protocol or scanner, introducing a domain-shift (see Section 5.1). More intensity-based data augmentation and more task-specific pre-processing could, but does not guarantee, an increase to the robustness of all algorithms. Most state-of-the art approaches use established pre-processing consisting of intensity normalization and spatial normalization (image resolution and size). We followed this style for our work, as our focus was to investigate how additional unlabeled data can improve over a supervised baseline.

Although the CZ is another anatomical zone of the prostate, we did not account for it, because this zone is frequently compressed in elderly men due to increased growth of the TZ (a benign medical condition known as BPH) (Standring et al., 2016). Therefore our TZ segmentation encompasses both the TZ and CZ, a compromise that all other related works on automatic zone segmentation have made, too. Nevertheless, even without the distinction of the CZ, the detailed anatomical segmentation of the prostate can be leveraged for several clinical tasks.

LIMITATIONS AND FUTURE WORK In our generalization experiment, we examined whether improvement gains through SSL could also be obtained with lower amounts of labeled data. However, we did not vary the amount of unlabeled data. It should be of future work to analyze the effect different ratios of labeled *and* unlabeled data have on the outcome. This would be necessary to give guidelines on the application of the method to other tasks.

Another limitation of our evaluation for different number of labeled training samples is that we did not reduce the number of validation data. Consequently, the ratio of labeled validation to labeled training data is artificially high for the experiments with decreased sizes of labeled training datasets (Oliver et al., 2018). On the other hand, the supervised baseline receives the same data aggregation. Consequently, our findings that UATS exploits the additional unlabeled data for improved performance and that less labeled data is required with our UATS method are still valid. However, what effect the validation dataset size has on the outcome and whether the validation set can be omitted should also be evaluated in future work.

Furthermore, we want to investigate whether a relation can be quantified automatically for a given task to answer the following two questions. First, can we estimate beforehand how much semi-supervision will help? Second, how do we select the optimal samples that should be labeled? For example, approaches that measure the model uncertainty (Gal and Ghahramani, 2016; Mehrtash et al., 2020) and estimate the segmentation quality (Robinson et al., 2018) could be investigated to address these research questions. Moreover, we would be interested to examine

whether the approach could gain improvement on other biomedical imaging tasks, for example classification.

4.7 SUMMARY

We proposed a semi-supervised method for prostate zone segmentation from T2w MRI with the aim to exploit unlabeled data from the intra-domain level to make the model more accurate and robust. To the best of our knowledge, we are the first to address simultaneous segmentation of the TZ, PZ, DPU and AFS in an attempt to reproduce the anatomical prostate division according to the PI-RADS v2.1 sector map in a patient-specific manner. Our method combines uncertainty-aware self-learning and temporal ensembling into a novel framework to improve supervised deep learning models by commonly available unlabeled data.

Regarding prostate zone segmentation, our method yields results from the quality on a clinical expert level. The improved segmentation quality of the prostate zones may enable more precise and consistent lesion location assignment, as well as improved cancer therapy planning and it could increase the accuracy of automatic lesion detection and assessment methods. We showed that our method increases robustness against noise compared to the supervised baseline. Moreover, we demonstrated that UATS generalizes to other tasks by evaluating our method on additional biomedical challenge datasets. Our experiments demonstrated that our method improves upon the supervised baselines for different ratios of labeled samples and different tasks.

We found that, when gains from semi-supervision are larger, the higher the variability in appearance and shape and the smaller the amount of labeled samples. We also found that when enough labeled samples with sufficient quality become available, gains from semi-supervision will diminish at some point.

We used standard U-Nets as supervised grounds for comparison because these have demonstrated their potential for a wide range of tasks. However, our semi-supervised strategy also applies to network architectures beyond U-Nets. Therefore, our approach can have an impact on many different (biomedical) segmentation tasks by reducing the amount of necessary labeled images.

5

EXPLOITING EXTERNAL DOMAINS

One major challenge in the application of CNN methods is their lack of robustness on data that originates from another distribution than the data seen during training. In the medical field, this may cause a considerable performance drop when the CNN is applied to images that are, for example, acquired from another scanner or protocol. Re-training the network in the new domain is impractical, as it requires a large amount of labeled data. To this end, in this chapter, we design a method that exploits the knowledge from the external domain to improve segmentation in the new domain with only small amounts of labeled data. In the context of segmenting critical structures for PCa therapy, we propose a semi-supervised domain adaptation method that relaxes the common requirement of (labeled) data from the original domain being available. We demonstrate that our method's performance approaches the level of inter-reader variability for the majority of structures in the new domain.

This chapter is based on the following publication:

> A. Meyer, A. Mehrtash, M. Rak, O. Bashkanov, B. Langbein, A. Ziaei, A. S. Kibel, C. M. Tempany, C. Hansen, J. Tokuda, 2021. "Domain adaptation for segmentation of critical structures for prostate cancer therapy," *Scientific Reports*, 11, p. 11480.

Manual reference segmentations, that were created for a publicly available dataset within this work, were published as supplementary material of the paper.

STRUCTURE OF THE CHAPTER We begin the remainder of this chapter by introducing the clinical purpose for the targeted prostate structures, as well as our technical motivation and contribution (Section 5.1). Preliminaries that are relevant for this chapter follow in Section 5.2. We review the related work that has been carried out regarding the segmentation of the EUS and NVB, as well as the domain adaptation for medical image segmentation (Section 5.3). The designed technical methods and the experimental setup is described in Sections 5.4 and 5.5. The results follow in Section 5.6. This chapter is concluded with a discussion of the results (Section 5.7) and a brief summary in Section 5.8.

5.1 INTRODUCTION

The primary choice for the treatment of localized PCa is radical prostatectomy (Hautmann and Gschwend, 2014), which is carried out as either open surgery, laparoscopic or robot-assisted laparoscopic intervention. During this procedure, the prostate gland, seminal vesicles, and tumor are removed altogether, irrespective of the tumor size and location. Although radical prostatectomy is oncologically effective, it is frequently followed by sexual or urinary dysfunction. Multiple studies have shown that sparing the EUS and the NVBs could lead to an improved outcome of the surgery, with faster recovery of the patient regarding these functions (Nguyen et al., 2017; Mungovan et al., 2017).

MRI techniques allow for a more precise study of the tumor's involvement into these critical structures. Therefore, methods have been proposed to include virtual or 3D printed patient-specific models based on MRI data (see Figure 5.1) into the treatment planning, for instance, in Wake et al. (2020) and Wang et al. (2020b). Those models commonly comprise the prostate, tumor, NVB, EUS and other surrounding structures. The incorporation of 3D models makes the understanding of the tumor's location more intuitive for physicians and patients and can thus improve the treatment decision.

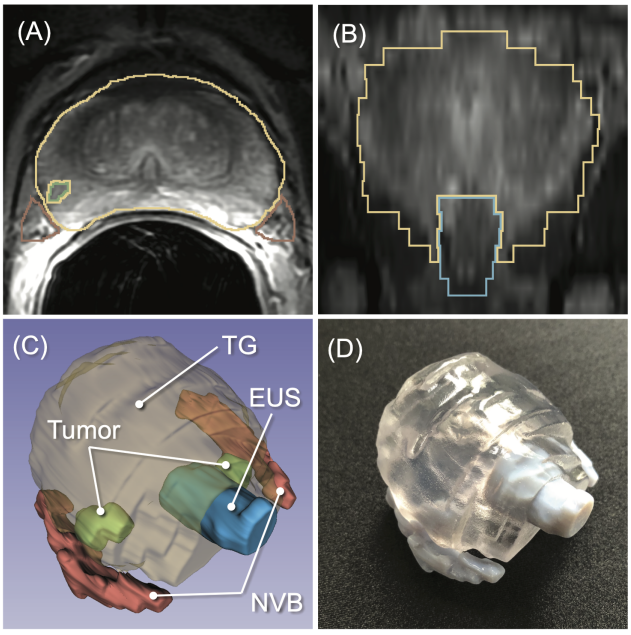


Figure 5.1: An example application of 3D segmentation of the prostate and adjacent structures for surgical planning. The prostate gland, NVB, EUS, and tumor are manually segmented on the preoperative T2w MRI (A, B) by a radiologist, and then converted to a 3D surface model (C). The model can also be 3D-printed (D) for surgical planning, and preoperative communication with the patient. Image was created by Junichi Todukda.

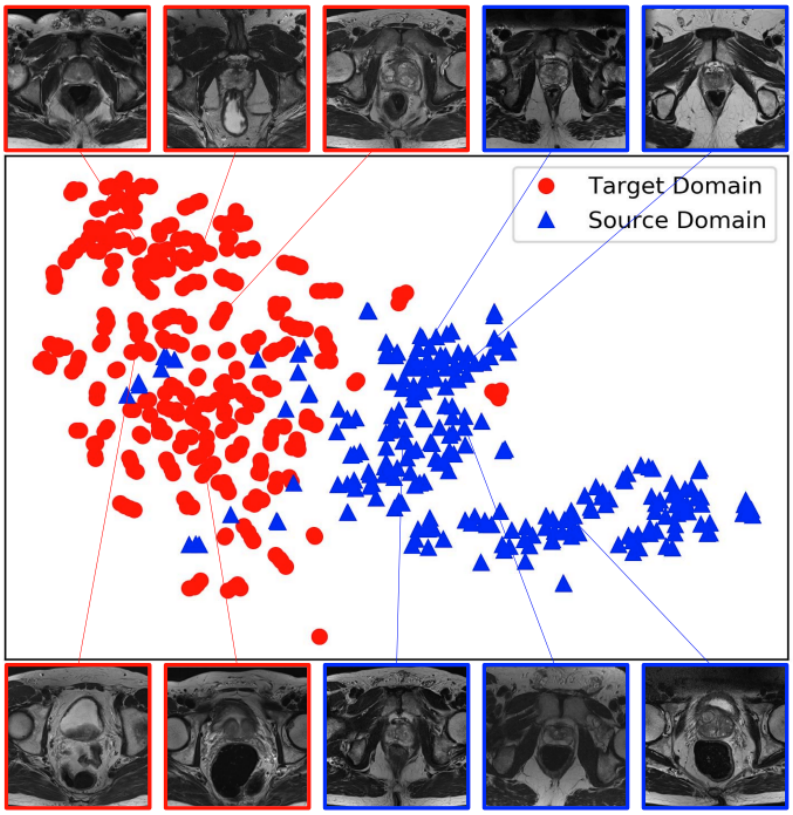


Figure 5.2: Visulaization of the domain shift for two prostate datasets: one private dataset acquired by a Philips 3 T scanner with endorectal coil (source), and the multi-vendor multi-center target dataset PROMISE12 (Litjens et al., 2014c). The images are mapped to a feature space (via a pre-trained VGG-16 network (Szegedy et al., 2015)) and t-SNE (Maaten and Hinton, 2008) is applied on these features to visualize the distribution of the two datasets. Figure from Zhu et al. (2020), © 2020 IEEE.

However, because segmenting individual structures manually is labor-intensive, its use in clinical routine is rather restricted. Therefore, an automatic and reliable segmentation of the critical structures may facilitate the use of 3D-models for the treatment planning and reduce the risks of overtreatment and complications. Furthermore, it may standardize PCa reports (Turkbey et al., 2019) and be employed in retrospective analysis (Inoue et al., 2009).

MOTIVATION AND CONTRIBUTION In the previous chapters, we have seen that a CNN model's performance can be improved by exploiting additional data that is often available in the clinical routine. For these methods, we assumed that training and test data share the same underlying distribution. This assumption holds if data from only one clinic is incorporated, but the distribution of medical datasets from another clinic will likely differ due to different scanner manufacturers, scanning parameters, subject cohorts, and other factors. This effect,

called *domain shift* (Quiñonero-Candela et al., 2009) (see Figure 5.2), leads to a performance drop of the models trained on data from the original *source domain* when applied to data from a new *target domain*. The effect of domain shift has been observed for DL models for several medical image analysis problems, including prostate segmentation (Gibson et al., 2018a).

Retraining the model from scratch in the new domain would require large amounts of annotated data, which is costly and often impractical. Therefore, various domain adaptation (DA) methods have been investigated and proposed in the last years to improve the model's performance on out-of-distribution data. However, most of them require that both source and target data are available. This requirement often becomes a burden when the model is deployed among multiple institutions, while the access to the source data is limited due to privacy concerns. Therefore, methods need to be explored that relax the requirement of source data. A trained model is less restrictive and easier to share compared to data from the source domain. Several deployment services exist that allow sharing off-the-shelf models (without the training data) for further reuse (Mehrtash et al., 2017b; Hosny et al., 2019). The concept of federated learning (Rieke et al., 2020) also exploits the fact that DL models are easier to share than their training data.

In this chapter, we propose a semi-supervised DA pipeline based on *transfer learning* (TL) (i.e., fine-tuning) that overcomes the necessity of the source data to be available. While TL is easy to apply and proven effective, a gap between the actual and desired performance remains, especially when only a few labeled target samples are available. To this end, we propose to combine TL with uncertainty-aware self-learning to exploit the information the additional unlabeled images offer. The combination of TL and self-learning has been investigated before in Zhou et al. (2018), who found that self-learning is the preferred choice of SSL techniques for TL for classification tasks. However, to our best knowledge, no such strategy has been used to address a segmentation or a DA task.

In summary, our main contributions are the following: Firstly, we investigate the automatic segmentation of the prostate, the EUS and NVB for the planning of prostate interventions on preoperative MRI. To the best of our knowledge, the EUS and NVB have not been segmented automatically yet. Secondly, we address the problem of domain shift for this task by proposing a semi-supervised DA pipeline that leverages knowledge from the inter-domain-level. This allows us to perform robust segmentation of the prostate and the critical structures on MRIs acquired outside the institution in which source training data were acquired. The proposed pipeline is simple yet effective and requires neither the source images and labels, nor any specific network architecture or training procedure in the source domain. Lastly, we demonstrate

that our method can be easily adapted to other problems and data in additional experiments on pancreas segmentation in CT scans.

5.2 PRELIMINARIES

In the following we define terminologies and notations used within this chapter (Section 5.2.1). Then, for the purpose of structuring related work on DA, we provide an overview about different DA problem settings that can be found in the field of medical image analysis (Section 5.2.2).

5.2.1 *Notations and Terminology*

This chapter covers a DA method that combines TL and semi-supervised self-learning to improve the performance of our CNN on data from a new clinic, scanner or acquisition protocol. In this section we define the terminologies TL and DA, as well as the notations used throughout this chapter. For this, we follow largely the definition given in the survey from Pan and Yang (2009).

To differentiate between the concepts of TL and DA, the terms *domain* and *task* are relevant. A domain $\mathcal{D}$ comprises a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ with $X = \{(x_1, ..., x_n)\} \in \mathcal{X}$. A task T comprises a label space $\mathcal{Y}$ and an objective predictive function $f_T(\cdot)$, which is learned from the training data, e.g. by a CNN. In TL, one aims to improve the learning of the predictive function $f_T(\cdot)$ in the target domain $\mathcal{T}$ by exploiting the knowledge from the source domain $\mathcal{S}$ and source task $T_\mathcal{S}$, where $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$ or $T_\mathcal{S} \neq T_\mathcal{T}$. The concept of DA is defined as a specific case of TL, where $T_\mathcal{S} = T_\mathcal{T}$, but the domains $\mathcal{T}$ and $\mathcal{S}$ are slightly different (Goodfellow et al., 2016).

The term TL is used interchangeably in the DL literature and can refer to either the concept described above or to *fine-tuning* of the model's weights. Fine-tuning is a specific method of TL in which a network is initialized with weights obtained from another domain or task and subsequently fine-tuned for the new task or data at hand. Using pre-trained models commonly reduces the amount of data necessary for the new task or domain (Tajbakhsh et al., 2016). To be clear, throughout the remainder of this thesis, we use the term TL to describe the method of fine-tuning.

Moreover, we will denote images from the source domain $\mathcal{S}$ as $X_\mathcal{S}$ and images from the target domain $\mathcal{T}$ as $X_\mathcal{T}$. Similarly, we denote labels from the source domain as $Y_\mathcal{S}$ and from the target domain as $Y_\mathcal{T}$.

5.2.2 *Problem Settings*

Deep learning methods are sensitive to domain shifts and DA is required to achieve improved performance for the target domain $\mathcal{T}$. There exist a large variety of methods in the literature proposed for the medical

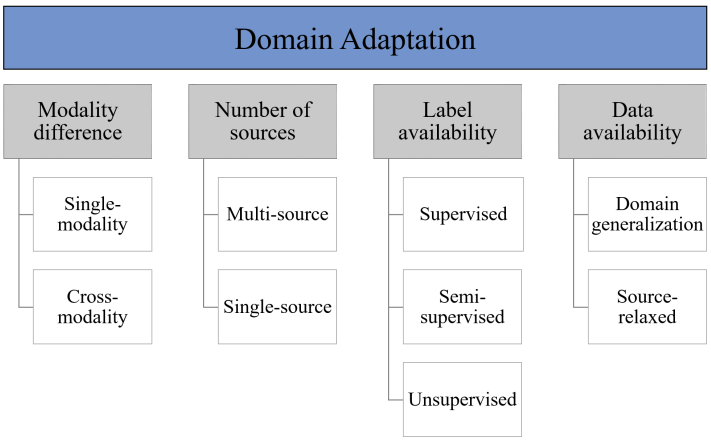


Figure 5.3: Categories of DA seen for medical image analysis (based on the DA categorization introduced by Guan and Liu (2021)).

image analysis field, which can not only be classified by the methods used, but also by their problem settings.

In the recent survey by Guan and Liu (2021), a classification of DA problem settings in medical image analysis is introduced. It has been adopted for this thesis for the most part and is illustrated in Figure 5.3. Following Guan and Liu (2021), we categorize deep learning-based DA problem settings with respect to differences in modalities, number of sources and availability of labeled samples in the target domain. Furthermore, we introduce the availability of data from either the source or the target domain as an additional class of DA problems. These categories are not exclusive - methods often root in multiple of these problem settings.

- MODALITY DIFFERENCE: There exist *single-modality* and *cross-modality* approaches in the literature. In single-modality DA, images from both domains are acquired by the same modality and the domain discrepancy is usually based on different scanners, protocols and sites. In cross-modality DA, the domain discrepancy originates from datasets being acquired by different modalities, as for example, CT scans in the source domain and MRI scans in the target domain.

- NUMBER OF SOURCES: A further differentiation is made as to whether the DA technique uses data from a *single source* domain or from *multiple source* domains. The usage of multiple sources has usually the advantage that models which have seen data from different domains during training, tend to a more robust performance in the target domain (Gibson et al., 2018a; Mårtensson et al., 2020). On the other hand, multi-source incorporation can also hold challenges in their training due to data heterogenity (Guan and Liu, 2021; Liu et al., 2020c).

- LABEL AVAILABILITY: DA methods can be split into *supervised*, *semi-supervised* and *unsupervised* problems, depending on availability of labels from the target domain $Y_{\mathcal{T}}$. Supervised DA techniques require various image/label pairs of $(X_{\mathcal{T}}, Y_{\mathcal{T}})$. Semi-supervised methods relax this requirement to a small number of $(X_{\mathcal{T}}, Y_{\mathcal{T}})$. Additionally, various samples without labels should be available. Unsupervised methods do not need any labels from the target domain, and rely only on $X_{\mathcal{T}}$.

- DATA AVAILABILITY: The last class of DA problem settings discussed here considers the data availability in the domains. Methods, that do not require any data at all from the target domain can be seen as an extreme case of unsupervised DA (Guan and Liu, 2021) and have been proposed within the research field of *domain generalization*. Techniques that do not require any data from the source domain are considered as *source-relaxed*, following the naming in the work of Bateson et al. (2020).

5.3 RELATED WORK

Following the definition of different problem settings for DA in medical image analysis, we now give an overview about existing domain adaptation techniques for medical image segmentation in general (Section 5.3.1). Subsequently, we summarize the related work that targets the segmentation of critical structures for radical prostatectomy procedures, namely the EUS and NVB (Section 5.3.2). For an overview about prostate gland segmentation methods, we refer to Section 3.2.

5.3.1 *Domain Adaptation*

The most straightforward way to achieve robustness in an unseen target domain is to include more heterogeneous data from multiple sources into the training as for example in Gibson et al. (2018b) and Mårtensson et al. (2020). But especially in the medical context, this is an impractical procedure, because it is often not feasible to aggregate data from different clinics due to privacy restrictions and limited availability. This is particularly true for studies that require highly-specialized labeled data that is only available in small portions. Therefore, the research on DA, which offers various other ways to improve model performance in the target domain, has received growing interest in the past years.

In the previous Section 5.2.2, we gave an overview about DA problem setting categories. In the following, we provide an overview about existing work on DA for medical image segmentation, which are grouped into supervised, unsupervised and semi-supervised as well as source-relaxed DA methods. As DA is a very large and rapidly evolving field of research, it would be out of scope for this thesis to summarize all

methods that have been published thus far. Instead, we aim to give an overview about main research directions and works that are relevant for our method. We refer the interested reader to the recent survey Guan and Liu (2021) for a more detailed summaray of the developments in this field.

SUPERVISED DOMAIN ADAPTATION A well-established supervised DA strategy is TL, which reuses the learned weights from training in other domains. The most widespread form of fine-tuning reuses CNN models pre-trained on natural images from the ImageNet challenge (Deng et al., 2009) as in Shin et al. (2016) and Tajbakhsh et al. (2016). However, it has also been successfully applied with models pre-trained on other medical datasets as a supervised DA strategy. TL has been shown to be very effective when only a small number of $(X_{\mathcal{T}}, Y_{\mathcal{T}})$-pairs are available for the DA of brain lesion segmentation (Ghafoorian et al., 2017; Valverde et al., 2019; Karimi et al., 2021) and pathological structure segmentation (Kaur et al., 2019). Valindria et al. (2018) demonstrated that fine-tuning the model with n annotated samples, that are most valuable samples for DA based on reverse classification accuracy, was more effective than fine-tuning with random samples. Karani et al. (2018) presented another variant of fine-tuning: their network learned domain-specific batch-normalization on a multi-source dataset. And for DA, only the batch-normalization parameters were fine-tuned with very few samples.

Bermúdez-Chacón et al. (2018) could improve electron microscopy segmentation in the target domain with only few labeled samples. They applied a coupled two-stream U-Net, where one stream is trained on source data and the other on target data. Feature sharing and regularization between both streams were applied for domain alignment.

For the task of prostate segmentation, Zhu et al. (2020) proposed a boundary-weighted DA strategy. They implemented two segmentation networks: one for the source domain and the other for the target domain, respectively. A discriminator tries to distinguish the decoders' features of both networks in an adversarial manner and drives the domain alignment of both networks. This adversarial loss is moreover weighted to focus on the boundary regions. With several labeled samples in the target domain available, they could show that incorporating the knowledge of the source domain improved performance in the target domain.

UNSUPERVISED DOMAIN ADAPTATION Unsupervised DA has gained growing attention in recent years with the advance of adversarial learning (Goodfellow et al., 2014) and is one of the most popular lines of research in DA. Frequently, unsupervised DA aligns the source and target domain distributions by enforcing similarity of (1) the input space at image level, (2) the output space (segmentation), or (3) the feature space during the DA process.

DA at the input space aims to transfer, for example, the source domain images into appearance of the target domain (also known as style transfer, e.g. via CycleGANs (Zhu et al., 2017a)) whereas the anatomical structure information is retained. Then, the source labels can be used to train a network that uses the style-transferred source images with target appearance (Huo et al., 2018; Chen et al., 2018).

DA at the output level assumes that segmentations from different domains have high similarity in the output space. In some methods, an adversarial loss is included on the network's output that drives the domain alignment by enforcing the CNNs to produce outputs that have consistent topology among different domains (Tsai et al., 2018; Yan et al., 2019). Another variant is to constrain the DA in the output space with specific shape or weak label priors as suggested by Bateson et al. (2021) for intervertebral discs and whole heart segmentation.

Most approaches concentrate on enforcing a domain-invariant (latent) feature space. This can, for example, be enforced by adversarial learning with a discriminator that aims to distinguish whether the encoder's feature maps originate from the source or the target domain (Kamnitsas et al., 2017; Dou et al., 2018). A combination of feature and input space alignment has been proposed by Chen et al. (2020a) for unsupervised cross-modality DA. Alternatively, Yang et al. (2019) proposed to use disentangled representations that decompose the input to a content- and a style-space. The segmentation is then learned on the domain-invariant content space.

Another popular line of research for DA, is to employ techniques, that have been originally introduced for SSL (see Section 2.3), into the context of unsupervised DA. The settings for both learning concepts are similar, except that for DA, the unlabeled data originates from another distribution. However, the application of SSL for DA methods has lead to promising results. For example, teacher-student models have been used to apply a consistency loss on unlabeled data of the target domain for spinal cord gray matter segmentation on MRI (Perone et al., 2019) and vessel segmentation on retinal fundus images (Fotedar et al., 2020). Another approach by Bian et al. (2020) combines uncertainty-aware self-learning with an adversarial loss that minimizes discrepancies between feature spaces of $X_{\mathcal{S}}$ and $X_{\mathcal{T}}$ for different medical segmentation tasks. The segmentation loss and the self-learning curriculum are furthermore guided using uncertainty information (via a conditional variational auto-encoder (Kohl et al., 2018)).

Ideally, there is no need at all to apply any DA method, because the source model is robust enough against the domain shift. To this end, works have been proposed that are categorized as domain generalization. One way to achieve such a generalization across domains is to apply extensive data augmentation as in Sheikh and Schultz (2020), Hesse et al. (2020), and Zhang et al. (2020). Zhang et al. (2020) for example applied stacked data augmentation transforms of $X_{\mathcal{S}}$ and $Y_{\mathcal{S}}$. For prostate

segmentation, a performance close to the state-of-the-art fully-supervised methods on the target domain was achieved when data augmentation was applied to a large source set with $|(X_{\mathcal{S}}, Y_{\mathcal{S}})| > 450$. Another way to achieve domain generalization is to train on multi-source data in a shape-aware meta-learning setting as Liu et al. (2020b) proposed for prostate segmentation. The authors trained a domain-robust model by using virtual 'meta-train' and 'meta-test' sets that simulate domain shift during training.

SEMI-SUPERVISED DOMAIN ADAPTATION In semi-supervised DA, there is limited labeled data and additional unlabeled data from the target domain available. Roels et al. (2019) proposed a Y-net shaped architecture with one encoder and two decoders (for segmentation and reconstruction, respectively) for the task of electron microscopic imaging segmentation. In a first stage, the segmentation decoder is trained on the labeled source data while the decoder reconstructs images from both source and target domain in an unsupervised manner to obtain more domain-invariant features. In a second stage, the reconstruction decoder is removed, and the labeled target data is used to fine-tune the segmentation decoder.

A more complex method is proposed by Li et al. (2021a) for cross-modality semi-supervised DA (MR to CT). They employed an intra-domain mean teacher model for consistency in the target domain, and an inter-domain mean teacher model with an appearance alignment via CycleGANs (Zhu et al., 2017a) that can map MR to CT images and vice versa. The knowledge transfer between the teacher and student models is designed in an uncertainty-aware manner induced by MC dropout (Kendall et al., 2017).

SOURCE-RELAXED DOMAIN ADAPTATION In contrast to other problem settings, only few methods have been proposed in the source-relaxed setting so far. For the application of lung segmentation, Venkataramani et al. (2019) proposed to condition the inference for new inputs in the target domain with features from a cluster of similar images of a support set from the target domain. These so-called context features are passed into the latent space of the encoder-decoder segmentation network and should allow for a life-long DA that can handle incremental changes in the dataset distributions. Bateson et al. (2020) proposed another unsupervised source-relaxed DA for spine segmentation with entropy-minimization in the target domain, which is regularized by a shape prior learned from the source data. A recent study by Xia et al. (2020) applied multi-view training (Section 2.3.1) to multi-organ segmentations in CT datasets. Their method showed to be effective even when no source data was included.

Karani et al. (2021) proposed a test-time adaptable unsupervised DA technique. Their method incorporates a shallow image normalization

network, a deep segmentation network and a denoising autoencoder trained on the source data. By means of the autoencoder's output (which serves as segmentation correction), the shallow normalization network can be fine-tuned with gradient information from the difference of autoencoder and segmentation network outputs. Lastly, the supervised TL (fine-tuning) approaches summarized above (for the supervised DA), can also be considered as source-relaxed methods (Ghafoorian et al., 2017; Valverde et al., 2019; Karimi et al., 2021; Kaur et al., 2019; Karani et al., 2018).

5.3.2 *Critical Structure Segmentation*

The literature overviews in the previous chapters presented a variety of approaches for the segmentation of the prostate gland and its substructures. Far less research has focused on the prostate's adjacent structures EUS and NVB so far which are critical for the outcome of prostatectomy interventions. NVB has only been segmented manually on MRI for registration of MRI and TRUS images (Yang et al., 2015). Our work in Chapter 4 addressed the segmentation of the distal prostatic urethra in a multi-class segmentation with the zonal anatomy of the prostate (Meyer et al., 2019; Meyer et al., 2021b). Another study used radiomics features to segment the peripheral zone and the prostatic urethra (Hambarde et al., 2019). However, no research has been carried out on the automatic segmentation of the EUS that we are aware of currently.

5.3.3 *Limitation of Current Approaches*

Contrary to the variety of methods proposed for whole gland (Section 3.2) and zones (TZ and PZ, Section 4.2.1) segmentation, there are no works yet on the automatic segmentation of NVB and EUS. Therefore, we aim to investigate whether a reliable segmentation can be obtained by applying CNNs for this task.

A common challenge for medical image segmentation is that the source data - $X_{\mathcal{S}}$ and $Y_{\mathcal{S}}$ - are not always available due to regulations and/or institutional policies on protected health information, despite the majority of DA techniques described above require them. Only few works exist, that target this limitation and do not require any images or labels from the source domain. On the other hand, these methods either require a multi-source setting for their training (Karani et al., 2018; Xia et al., 2020) which is commonly limited because of the scarcity of labeled medical data. Or they rely on a specific training paradigm (Venkataramani et al., 2019; Xia et al., 2020) which makes it impossible to reuse off-the-shelf models that did not consider specific requirements in their training procedure.

Lastly, other source-relaxed methods rely on knowledge about the shape of the structure to segment (Bateson et al., 2020; Karani et al., 2021), which is also gathered in the source domain in their implementations. Furthermore, the shape assumptions do not hold anymore in problems, where the domain-shift is induced by pathological causes. These limitations motivated us to develop an easy-to-apply but effective DA pipeline that does not require access to the source data nor any other knowledge about the structure-to-segment or any specific training paradigm. Our method is described in the following section.

5.4 TECHNICAL METHODS

Considering the limitations of the related work, the objective of our technical methods is two-fold: (1) to investigate the feasibility of automatic segmentation of the critical structures for the radical prostatectomy, including the prostate, NVB and EUS, and (2) to develop a source-relaxed domain-adaptation technique for this method that requires only few labeled training samples from the target domain $\mathcal{T}$.

Analogously, we divide this section into two parts. Firstly, our supervised training strategy for the critical structure segmentation in the source domain is described (Section 5.4.1). This will be the basis for investigating the feasibility of CNNs for NVB and EUS segmentation. Then, the proposed semi-supervised DA method is outlined (Section 5.4.2), which aims to improve the performance of the source model in the target domain.

5.4.1 *Supervised Learning (Source Domain)*

The supervised leaning uses a labeled dataset $D_L = \{x_i, y_i\}_{i=1}^{n}$. For each image x_i from $X \in \mathbb{R}^{H \times W \times D}$, there exists a ground truth segmentation map y_i from $Y \in \{0,1\}^{H \times W \times D \times C}$, where W, H, D are the dimensions of the volume and C defines the number of class labels. In our case, $C = 4$ due to the classes prostate, EUS, NVB and background. The network $f(\cdot)$ described in this section makes a prediction $\hat{y}_i$ for an input sample x_i, given the learned parameters θ, such that $\hat{y}_i = f(x_i, \theta)$ with $\hat{y}_{\in}[0,1]^{H \times W \times D \times C}$. Similar to the zone segmentation in Chapter 4, we used the adapted 3D U-Net that takes the anisotropic nature of the axial MRI scans into account. We employed the same architecture as described in Section 4.3.1, with the exception that we did not include any dropout layers, as we could not find improvements through their employment in preliminary experiments on the validation set.

DEEP ENSEMBLES To further improve the segmentation outcome, we used an ensemble of networks (deep ensemble). Deep ensembles have been shown to create more robust results than single networks (de Vente et al., 2020; Mehrtash et al., 2020). Model averaging commonly improves

performance, because an ensemble may compensate the different errors that were made by the different models on the test set (Goodfellow et al., 2016). Deep ensembles leverage different minima that CNNs can obtain because networks are subject to randomness during training. In our training setting, we employed random parameter initialization, random mini-batches generation during training and different random training/validation splits from a k-fold cross-validation to increase the local minima variability. We used an ensemble of k models and obtain a mean prediction μ of them:

$$\mu_i = \frac{1}{k} \sum_{i=1}^{k} f(x_i, \theta_k) \,. \tag{5.1}$$

POST-PROCESSING To obtain the final segmentation outcome, we post-process the (ensemble) prediction. In the first post-processing step, the prediction is thresholded to create a binary segmentation. To ensure topological correctness, the output is further post-processed with connected components analysis for the EUS and the prostate, for which only the largest component is kept. The connected component analysis is not applied to the NVB because NVB voxels are not always adjacent in neighboring slices due to the high slice thickness. A connected component analysis would, therefore, risk discarding actual NVB segments.

5.4.2 *Domain Adaptation*

We propose a source-relaxed DA technique. This means that we have only the k source models $f(\theta_{\mathcal{S}})$ and our target dataset $D_{\mathcal{T}}$ available. Due to our semi-supervised DA strategy, our target dataset consists of l labeled volumes $D_{\mathcal{T},L} = \{x_i, y_i\}_{i=1}^{l}$ and u unlabeled volumes $D_{\mathcal{T},U} = \{x_i\}_{i=l+1}^{l+u}$. Our DA method comprises two learning concepts: (1) TL as the first stage of DA, and (2) uncertainty-aware self-learning as a second stage to obtain more information about the distribution of the target domain $\mathcal{T}$. Our proposed semi-supervised DA method is depicted in Figure 5.4 and a summary is provided in Algorithm 2 at the end of this section.

STAGE I: TRANSFER LEARNING In our scenario, we find large differences in the shape and appearance of the structures between the source and target datasets due to using an endorectal coil in the source dataset (see Figure 5.5). The shape, location, and appearance of the structures-to-segment, particularly the NVB, are changed substantially because of the pressure from the endorectal coil in the source dataset. To compensate for this severe domain shift, we propose to have a small amount of labeled pairs $D_{\mathcal{T},L}$ with $l \leq 10$ in the target domain available.

With $D_{\mathcal{T},L}$, we fine-tune our source model $f(\theta_{\mathcal{S}})$ to a model adapted to the target domain $f(\theta_{\mathcal{T}})$. As we only have a minimal amount of

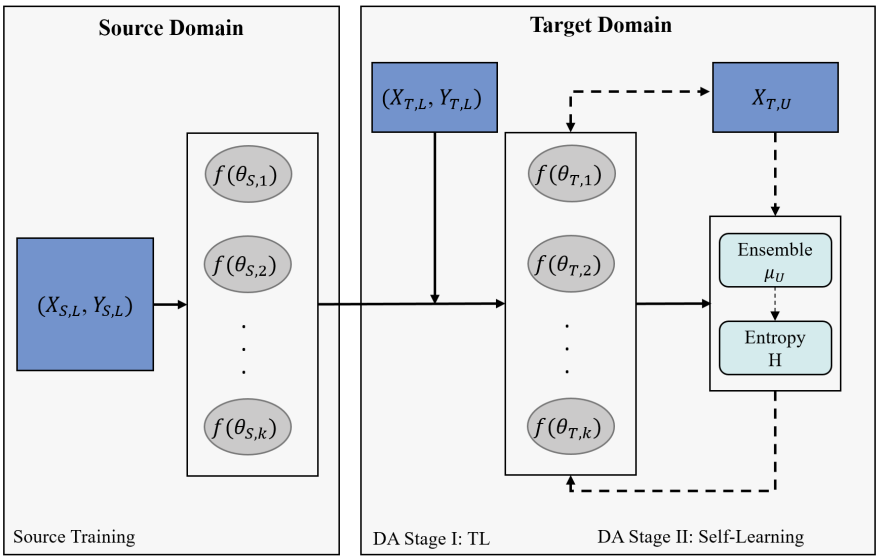


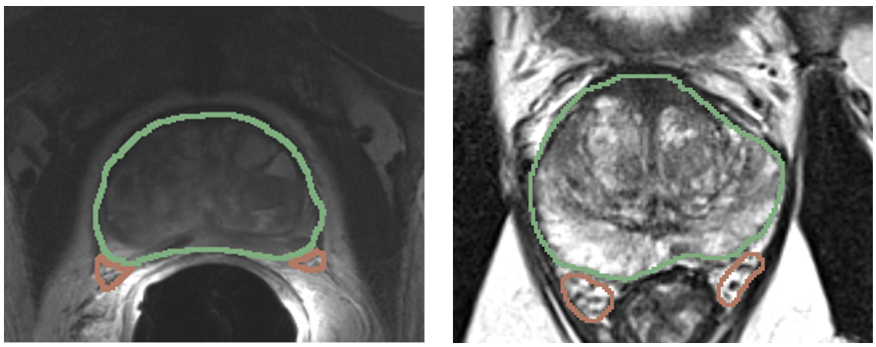
Figure 5.4: Proposed pipeline for the DA. The ensemble of k models is trained in the source domain with the labeled source data. Subsequently, these models are domain adapted by TL with the few labeled data from the target domain and furthermore refined with the self learning routine (dashed arrows) that includes ensemble-based pseudo labels and uncertainty (entropy) estimation.

labeled images, we fix the weights of the decoder and only fine-tune the encoder and the bottom layer weights of the source model. In preliminary experiments on the validation set, this has been working best for a small training dataset.

STAGE II: UNCERTAINTY-AWARE SELF-LEARNING The TL can be considered as a warm-up phase for the self-learning routine. This is followed by the second stage of our DA pipeline, which is the uncertainty-aware self-learning. At this stage, we used the labeled and unlabeled $D_{\mathcal{T},L}$ and $D_{\mathcal{T},U}$ for training. To reduce the negative predictions in the self-learning stage, we propose to use deep ensembles for better segmentation candidates and uncertainty estimation.

The self-learning routine (see Section 2.3.1) is a cycle consisting of label propagation to obtain pseudo labels Y_U, and retraining the model weights $\theta_{\mathcal{T}}$ with $D_{\mathcal{T},L\cup U}$ until the performance on the validation data does not improve any further. The fine-tuned model $f(\theta_{\mathcal{T}})$ from Stage I is used to obtain initial pseudo labels for the unlabeled data $X_U \in D_{\mathcal{T},U}$. Typically, three to five iterations have to be carried out until the model does not improve any further. In contrast to TL, in the self-learning training procedure, all weights are trained. The training objective at this stage is a weighted and masked DSC loss to train our network:

$$\text{loss}(y, \hat{y}, w, m) = -\frac{1}{|C|} \sum_{c \in C} \frac{w \cdot 2 \sum_{i=1}^{N} \hat{y}_{c,i} y_{c,i} m_i + \epsilon}{\sum_{i=1}^{N} \hat{y}_{c,i} m_i + \sum_{i=1}^{N} y_{c,i} m_i + \epsilon}, \quad (5.2)$$



(a) (b)

Figure 5.5: Example images for the prostate datasets: a) endorectal coil acquisition from the source dataset and b) pelvic coil acquisition from the target dataset. Segmentations of the prostate (green), NVB (brown) are overlayed.

where N is the number of the volume's voxels, $\hat{y}$ is the network's prediction, $y \in \{Y_U \cup Y_L\}$ are the pseudo and real ground truth labels, and ϵ ensures numerical stability as a small constant. The purpose of mask m and weight w is to reduce the impact of false predictions on the voxel- and subject-level, respectively. We describe how they are obtained, in the following.

For more reliable pseudo labels, we propose to use an ensemble of k target models $f(\theta_{\mathcal{T}})$ (see Equation 5.1), because deep ensembles have been shown to have better calibrated predictions (Section 2.2.2). The average predictions μ_U of this ensemble are post-processed as described in Section 5.4.1 (thresholding and connected components analysis) to obtain the binary pseudo labels Y_U and to remove uncertain and noisy predictions. The post-processing results in some voxels of the pseudo labels having no label given in any class channel c of Y_U (prostate, NVB, EUS and background). This is because either none of the classes is above the threshold, or the label has been removed through connected components analysis. With our **mask m**, we account only for voxels in our loss function that have any label given:

$$m_i = \begin{cases} 0, & \sum_C y_{c,i} = 0\,, \\ 1, & \sum_C y_{c,i} > 0\,. \end{cases} \tag{5.3}$$

Furthermore, deep ensembling can not only be used to improve segmentation accuracy, but has also been shown to be an appropriate means to estimate the uncertainty of prostate segmentation maps (Mehrtash et al., 2020). Hence, we utilized the entropy of ensemble predictions

for the subject-level uncertainty **weight w** to reduce the impact of low quality pseudo labels. The entropy is computed as:

$$H_i = -\frac{1}{N}\sum_{1}^{N}\sum_{c=1}^{C}\mu_c \log \mu_c \,. \tag{5.4}$$

The entropy is then normalized as:

$$H_i = \frac{H_i}{\max_i H_i} \,. \tag{5.5}$$

We exploit this uncertainty estimation for the pseudo labels by weighting their contribution to the overall loss with $w = 1 - H_i$. The weights for the labeled data remain unchanged ($w = 1$).

This uncertainty-based weight should balance the trade-off that one usually have to make for selecting the right weights for pseudo labels. Too high values of w for pseudo label samples can lead to a degeneration of model performance, if too many pseudo label voxels are misclassified. On the other hand, too small w for pseudo label samples may overemphasize the influence of the real ground truth samples, resulting in too little information from the unlabeled data for the gradient update. In this case, the model potentially overfits on the small amount of ground truth labels.

5.5 EXPERIMENTAL SETUP

Having described our technical methods in the previous section, we now turn to the details on the experimental setup for evaluating them. These details include a description of the datasets (Section 5.5.1), the training of the methods (Section 5.5.3), and the design of the experiments (Section 5.5.2).

5.5.1 *Data*

For the evaluation of our method, we used multiple datasets. Because our main objective for this chapter is the investigation of CNN-based segmentation of critical structures for prostatectomy, we evaluated the supervised CNN (Section 5.4.1) and our proposed DA method (Section 5.4.2) on prostate MRI. Moreover, we investigated the performance of our DA technique on pancreas CT data to demonstrate the generalization capability of our method.

For both types of segmentation tasks, prostate and pancreas segmentation, we used a source and target datasets, which we describe in the following paragraphs. Moreover, for each segmentation task, we cover details on the pre-processing and augmentation technique. A summary of the details for the prostate datasets can be found in Table 5.1 and information on the pancreas dataset is provided in Table 5.2

Algorithm 2: Semi-supervised Domain Adaptation

Input: set of k source models $f(\theta_{\mathcal{S}})$, input images $X \in D_{\mathcal{T}}$, set of indices for labeled images L, set of indices for unlabeled images U, ground truth labels $Y_L \in D_{\mathcal{T}}$

Output: set of k target models $f(\theta_{\mathcal{T}})$

```
/* Initialize algorithm                                          */
M ← 0                                  // initialize masks M with zeros
cur_val_loss ← ∞                       // current validation loss
min_val_loss ← ∞                       // minimum validation loss
w_U ← 0, w_L ← 1                       // initialize weights
/* Stage I: TL with labeled data                                 */
for i in [1, k] do
    f(θ_{T,i} ← fine_tune(f(θ_{S,i}), X_L, Y_L)
/* Stage II: uncertainty-aware self-learning                     */
while cur_val_loss < min_val_loss do
    min_val_loss ← cur_val_loss
    /* create pseudo labels                                      */
    μ_U ← ensemble_vote(f(θ_T), X_U)   // averaging over k models
    Y_U ← post_process(μ_U)            // create binary pseudo labels
    w_U ← entropy(μ_U)                 // update unlabeled weights (Eq. 5.5)
    M ← update_mask(Y_{L∪U})           // update mask (Eq. 5.3)
    for i in [1, k] do
        /* uncertainty-aware self-training                       */
        f(θ_{T,i}) ← train_model(f(θ_{T,i}), X_{L∪U}, Y_{L∪U}, w, M)
    update cur_val_loss                // average val. loss over k models
```

PROSTATE MRI We used two different datasets for the prostate structures segmentation evaluation: the source data was comprised of an internal dataset and the target data was created by using the publicly available Prostate-3T (Litjens et al., 2015) dataset.

Source Data $D_{\mathcal{S}}$: Sixty-two patients who were scheduled for robot-assisted laparoscopic prostatectomy underwent preoperative multiparametric MRI in a 3 T scanner (Signa HDxt 3.0 T; GE Healthcare). As part of the protocol, an axial multi-slice T2w image was acquired with both endorectal and pelvic phased-array coils. The gland, NVB, and EUS were manually segmented by Reader 1, an expert radiologist, using the Editor tool on 3D Slicer (Fedorov et al., 2012). We follow the data splitting described in Section 2.4.2 and split the dataset into 46 cases for training and 16 hold-out test cases. For evaluating the inter-reader variability and the performance of the automatic segmentation, a second reader segmented the test cases for this dataset. This Reader 2 was a research fellow with a medical background and two years of experience in reading prostate MRI. To be clear, for training, only the manual labels of Reader 1 were used as target labels $Y_{\mathcal{S}}$.

	Source	**Target**
Vendor	GE Healthcare, 3T	Siemens, 3T
Coil	endorectal & surface	surface
In-plane res. [mm]	0.27×0.27	$[0.5 - 0.625]$ $\times [0.5 - 0.625]$
Slice thickness [mm]	3.0	$[3.0 - 5.0]$
n_{train}	46	54
n_{test}	16	10

Table 5.1: Dataset details for the *prostate* structures segmentation task.

Target Data $D_{\mathcal{T}}$: For DA, we used the Prostate-3T data (Litjens et al., 2015) as target dataset. The dataset consists of 64 axial T2w scans that were acquired on a 3T Siemens TrioTim using only a pelvic phased-array coil. We selected 25 scans from this dataset for which either segmentations of the prostate zones (PZ and TZ) or the NVBs are available through the NCI-ISBI 2013 challenge (Bloch et al., 2015) and the Cancer Imaging Archive (Clark et al., 2013), respectively. The prostate segmentation for the NCI-ISBI 2013 challenge data is defined as the union of TZ and PZ segmentations. A medical student segmented the structures that were not provided by any of these two ground truth sources, such that for each of these 25 volumes, a three-class segmentation was available in the end. We split the labeled cases of this dataset into 15 training cases and 10 hold-out test cases. The remaining 39 cases remained unlabeled for our semi-supervised DA technique. A comparison between examples from the prostate source and target datasets is visualized in Figure 5.5.

Pre-processing and augmentation: All volumes were resampled to a spacing of $0.5 \times 0.5 \times 3.0$ mm. A bounding box ROI of the prostate was extracted from the center of the volume by cropping the volume to a size of $184 \times 184 \times 32$. Prior to normalization of image intensity to an interval of [0,1], the intensities were cropped to the first and 99th percentile. The training data was augmented by left-right flipping of the volume.

PANCREAS CT For the pancreas CT segmentation, the source and target data consisted of different datasets, too. The source data was aggregated with abdominal datasets from two sites, comprised of scans from patients with healthy pancreas. The target data scans were acquired from patients with pancreas cancer from a third site. Consequently, for the task of pancreas segmentation, the domain shift is not limited to differences in image appearance, but additionally covers the different distributions of healthy pancreas (source domain) and cancerous pancreas (target domain).

	Source	**Target**
Modality	portal venous phase CT	portal venous phase CT
Population	healthy pancreas	pancreatic cancer
In-plane res. [mm]	$[0.59 - 0.98]$ $\times[0.59 - 0.98]$	$[0.61 - 0.98]$ $\times[0.61 - 0.98]$
Slice thickness [mm]	$[0.5 - 5.0]$	$[0.7 - 7.5]$
n_{train}	75	200
n_{test}	14	81

Table 5.2: Dataset details for the *pancreas* segmentation task.

Source Data $D_{\mathcal{S}}$: For the source domain, we used two abdominal datasets: The Cancer Imaging Archive (TCIA) Pancreas-CT dataset (Roth et al., 2015; Roth et al., 2016; Clark et al., 2013) and the Beyond The Cranial Vault (BTCV) abdomen dataset (Xu et al., 2016; Landman et al., 2015). In the TCIA dataset, portal venous phase contrast enhanced 3D CT scans from pre-nephrectomy healthy kidney donors were acquired at the National Institutes of Health Clinical Center (Bethesda, MD, USA). The BTCV dataset was acquired during portal venous contrast phase at the Vanderbilt University Medical Center (Nashville, TN, USA) from metastatic liver patients or post-operative ventral hernia patients. We used the publicly available segmentations (Gibson et al., 2018b) for the TCIA dataset (n=47) and the BTCV abdomen dataset (n=42) as our source data. Lastly, we split this source data into 75 training and 14 hold-out test cases.

Target Data $D_{\mathcal{T}}$: The dataset for the target domain was derived from the Medical Segmentation Decathlon Challenge (Simpson et al., 2019). This dataset consists of portal venous phase CT scans that were acquired from patients undergoing resection of pancreatic masses at Memorial Sloan Kettering Cancer Center (New York, USA). The dataset provides 281 cases with a two-class segmentation with tumor and pancreas outlined individually. For our experiments, we used the union of pancreas and tumor segmentation as foreground structure. For the evaluation on this target domain data, we set the same 81 cases as hold-out test cases as in Xia et al. (2020), the remaining 200 cases were used as training cases for the target domain.

Pre-processing and augmentation: The scans were resampled to a common spacing of $1.0 \times 1.0 \times 3.0$ mm and are cropped to a ROI of $200 \times 128 \times 48$ surrounding the ground truth pancreas segmentation. The intensities (Hounsfield unit) were first clipped to a range of $[-300, 300]$, which represents the intensity of the pancreas and adjacent structures, and subsequently normalized to zero mean and unit variance. We ap-

plied random geometric (translate, scale) and intensity (Gaussian noise, Gaussian blurring) transformations as online augmentations.

5.5.2 *Evaluation Design*

The objectives of our evaluation are three-fold. They include:

1. to assess the feasibility of a fully supervised CNN for automatically segmenting critical structures for PCa therapy,
2. to evaluate the suitability of our proposed semi-supervised DA method to reduce the performance gap of the CNN, which is due to the domain shift, by using only few labeled training images from the new domain,
3. to determine the generalization capability of the methods with respect to another task and type of data (i.e., pancreas CT).

The experimental design for our methods is described in the following paragraphs. For our experiments, we computed the evaluation measures described in Section 2.4.1 and carried out statistical evaluation as outlined in Section 2.4.2.

SUPERVISED MODEL To evaluate our supervised model, we conducted a 5-fold cross validation resulting in a 36/10 train/validation split for training on the source domain data. We compared the supervised method's outcome as a single model (sCNN) (average across the five folds) and as an ensemble of the $k = 5$ models (eCNN) to the segmentations from Reader 1 on the source dataset for the 16 hold-out test cases. To frame the quantitative values of the automatic method, we assessed the inter-reader variance for the manual segmentation (comparing segmentations from Reader 1 and Reader 2). To quantify the domain shift, we ran the sCNN on the target domain test data.

DOMAIN ADAPTATION We evaluated the performance of our DA technique on the prostate target domain dataset with $l = 5$ and $l = 10$ labeled training samples as $D_{\mathcal{T},L}$ plus the $u = 39$ unlabeled images as $D_{\mathcal{T},U}$. We empirically set the lowest number of labeled training samples to $l = 5$, to allow the network to see some variance in the provided labeled dataset (e.g., organ size, relationship of the organ-to-segment and surrounding organs, diseases, imaging contrasts, noise, bias fields etc.). However, it should be possible to run the method even with a smaller number of labeled training samples, but presumably the results' quality will decrease in this scenario. The $k = 5$ models from the k-fold cross validation in the source domain were used to initialize our DA method.

We ran the experiments three times with different train/val splits of the labeled data (resulting from a 3-fold cross-validation data with

10/5 images for train/validation per fold) to compensate for biases introduced by selecting only a small amount of labeled data. We assessed the segmentation quality for our semi-supervised DA technique and compared it to training from scratch (i.e., random network weights initialization) and TL, both with only the l labeled images from the target domain as input.

Additionally, we performed two ablation experiments to determine the impact of the ensembling (ENS) and the uncertainty estimate (H). In the first experiment, we omitted the entropy-based uncertainty-weighting in our DA method (experiment TL+ENS). Instead, we applied $w = 0.5$ for samples with pseudo labels Y_U and $w = 1.0$ for the samples with actual ground truth Y_L. We empirically found these fixed weights to work best on the validation data. In a second experiment, to evaluate the impact of the ensembling of predictions for pseudo labels, we ran the training without the ensemble and only one model for pseudo label prediction (experiment TL+SL).

To compare our approach to another state-of-the art method beyond TL, we evaluated pure uncertainty-aware self-learning. Bian et al. (2020) proposed self-learning with uncertainty-guidance for (unsupervised) domain adaptation. Their method differs clearly from ours, because they propose to use a conditional variational autoencoder for uncertainty estimation and an uncertainty-guided cross-entropy loss on top of uncertainty-aware self-learning. As their method requires data from the source domain to be available, we can not simply apply their method to our source-relaxed problem setting. However, by evaluating the performance of applying only our uncertainty-aware self-learning (experiment ENS+H) for DA, we set out to make a comparison to Bian et al. (2020), and evaluate the impact of TL on our results. To allow for fair comparison with our proposed method, we included the l labeled samples of the target domain in this scenario, too.

5.5.3 *Training*

Following the description of the prostate and pancreas datasets and the experimental setup, we now summarize our method's training details for the prostate structures and pancreas segmentation. For reasons of clarity, we describe the main training procedures and refer to Appendix B for details on the task-specific hyperparameters.

PROSTATE MRI We ran the training for our methods described in Section 5.4 on the prostate datasets on a machine with a 12 GB NVIDIA TitanX Pascal GPU. We trained the networks with Adam optimizer minimizing the loss described in Equation 5.2 until convergence on the validation loss. For the supervised models in the source domain, we set $m = \mathbf{1}$ and $w = 1$ for all cases.

PANCREAS CT For our generalization experiment on the Pancreas CT dataset, we used the anisotropic U-Net described in Chapter 4.3.1, too. Because the pancreas segmentation is a binary segmentation task, we employed a sigmoid function as last layer activation and therefore obtained only a 1-channel output for this experiment. This made the application of the mask in the loss function impractical (as it would be the same as the 1-channel (pseudo) ground truth Y_U), and we simply applied a weighted DSC loss (Equation 5.2), omitting mask m. In contrast to our prostate data experiments, we found the inclusion of dropout regularization with a rate of 0.1 was improving performance on the validation set. Training for this specific dataset was carried out on a 24 GB NVIDIA GeForce RTX 3090 GPU with RMSProp optimizer (Hinton, 2012).

GENERALIZATION CAPABILITY By applying our DA method for the task of pancreas segmentation on CT scans, we aim to investigate its generalization capability regarding other tasks and types of data. Again, we compared our method with $l = 5$ and $l = 10$ labeled training data (and additionally 10 labeled validation cases) fom the source domain against training from scratch and TL. Analogously to the DA experiments on the prostate target dataset, we repeated these experiments three times with different aggregations of labeled train and validation data splits. We used $k = 5$ models that were obtained via 5-fold cross-validation in the source domain.

Additionally, having a large amount of labeled data available in the target domain, we were able to evaluate the upper bound for the target domain. To this end, we trained the supervised model with 200 labeled cases (160/40 training/validation). The methods were evaluated on the 81 hold-out test sets in the target domain. Lastly, because we used the same test dataset as Xia et al. (2020), we were able to make a *relative* comparison to this state-of-the-art technique regarding the DA's performance gain.

5.6 RESULTS

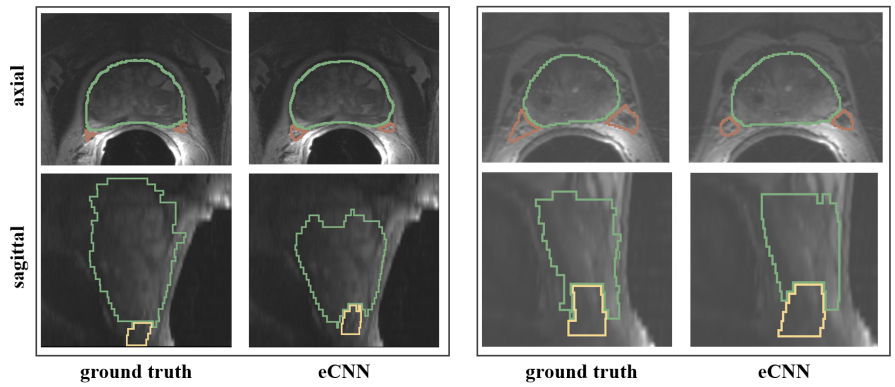
In the following, we report the results that were obtained in the experiments described in Section 5.5. Firstly, the results for the supervised CNN for the prostate structures segmentation in the target domain are summarized in Section 5.6.1. Then, the results for our semi-supervised DA technique for both the prostate and pancreas datasets (Section 5.6.2) are outlined.

5.6.1 *Supervised Learning*

We report the evaluation results for prostate, EUS and NVB for our supervised baseline as well as the inter-reader variance in Table 5.3. The

Method	Prostate			EUS			NVB		
	DSC [%]	ABD [mm]	95-HD [mm]	DSC [%]	ABD [mm]	95-HD [mm]	DSC [%]	ABD [mm]	95-HD [mm]
source data									
sCNN	87.8 ± 2.7	1.17 ± 0.54	4.96 ± 3.30	64.8 ± 12.6	1.54 ± 0.76	5.30 ± 2.70	55.8 ± 7.5	1.44 ± 0.73	7.35 ± 4.91
eCNN	89.3 ± 2.2	0.98 ± 0.46	4.02 ± 2.50	68.3 ± 11.2	1.36 ± 0.64	4.89 ± 2.79	58.3 ± 7.5	1.27 ± 0.66	5.90 ± 3.90
inter-reader-level	86.3 ± 4.9	1.61 ± 1.04	6.94 ± 5.11	46.5 ± 13.9	2.10 ± 1.09	10.28 ± 4.70	54.6 ± 9.1	1.68 ± 1.01	8.12 ± 4.91
target data									
sCNN	62.6 ± 26.1	5.6 ± 7.39	13.24 ± 11.48	34.8 ± 29.7	4.23 ± 5.94	9.06 ± 6.96	17.8 ± 16.4	9.39 ± 10.48	23.77 ± 15.79

Table 5.3: Comparison of segmentation results on source test data for single (sCNN) and ensemble CNN (eCNN) trained on the source data, the inter-reader performance and the performance of sCNN on test data from the target domain.



(a) worst result (b) best result

Figure 5.6: Segmentation visualization of the worst and best result for the supervised eCNN on the test set of the source domain.

average performance of the single networks (sCNN) across the folds are DSCs of 87.8 %, 64.8 % and 55.8 % for prostate, EUS, and NVB.

The ensemble eCNN improved the results to DSCs of 89.3 %, 68.3 %, and 58.3 %. Both approaches obtained better results compared to the inter-reader evaluation, which only achieved DSCs of 86.3 %, 46.5 %, and 54.6 % for the prostate, EUS and NVB, respectively. Visual inspection confirmed these observations except for one subject, where the automatic segmentation did not cover the base of the prostate likely due to a very heterogeneous prostate tissue (see Figure 5.6). Although the DSC values for EUS and NVB may appear quite low, the results' quality is better than expected from these values when inspected visually. As overlap-based metrics generally have lower scores for smaller structures, we refer to the boundary-based evaluation measures for further interpretation. The ABD of the eCNN for the NVB was 1.27 mm and 1.36 mm for the EUS, compared to 0.98 mm for the prostate. The 95-HD was 5.90 mm, 4.89 mm and 4.02 mm for the NVB, EUS and prostate, respectively. This confirms that, for all evaluation measures, the automatic segmentation has a higher rate of agreement with Reader 1 (who also created the training ground truth) than the second human reader (Reader 2) has with Reader 1.

To quantify the effect of the domain shift on our source model's performance in the target domain, we applied the single network (sCNN) to the target test data (see Table 5.4). We can observe a clear performance drop, which highlights the necessity of a DA technique. The DSC for the prostate decreases from 87.7 % on source data to 63.8 % on target data. Similarly, the DSC for EUS decreases from 64.8 % to 29.1 % and the DSC for NVB drops from 55.8 % to 17.7 %.

5.6.2 *Domain Adaptation*

We evaluated our semi-supervised DA method on the prostate dataset in conjunction with an ablation study. Here, we first describe the outcomes for this study. Then, we summarize the results of our method for the task of pancreas segmentation to assess its generalization capability. The quantitative results for our study on the DA methods for the prostate are summarized in Table 5.4 and accompanied with boxplots of their DSC's distribution in Figure 5.7 with corresponding p-values of the Wilcoxon signed-rank test (Section 2.4.2). An example outcome is shown in Figure 5.8.

Applying no DA and simply using the images from the target domain to train a model from scratch, resulted in a rather low DSC of 69.4 % for $l = 5$ and a DSC of 76.0 % for the prostate. Exploiting the external domain knowledge improved the performances significantly. For $l = 5$, we found that the DSC increased to 81.4 %, 84.3 %, 84.9 %, and to 85.5 % with TL, the additional self-learning (TL+SL), the ensemble-based self-learning without uncertainty (TL+ENS), and with uncertainty (TL+ENS+H), respectively. When applying majority voting on the ensemble that resulted from TL+ENS+H, the results could further be improved to a DSC of 86.5 % for the prostate.

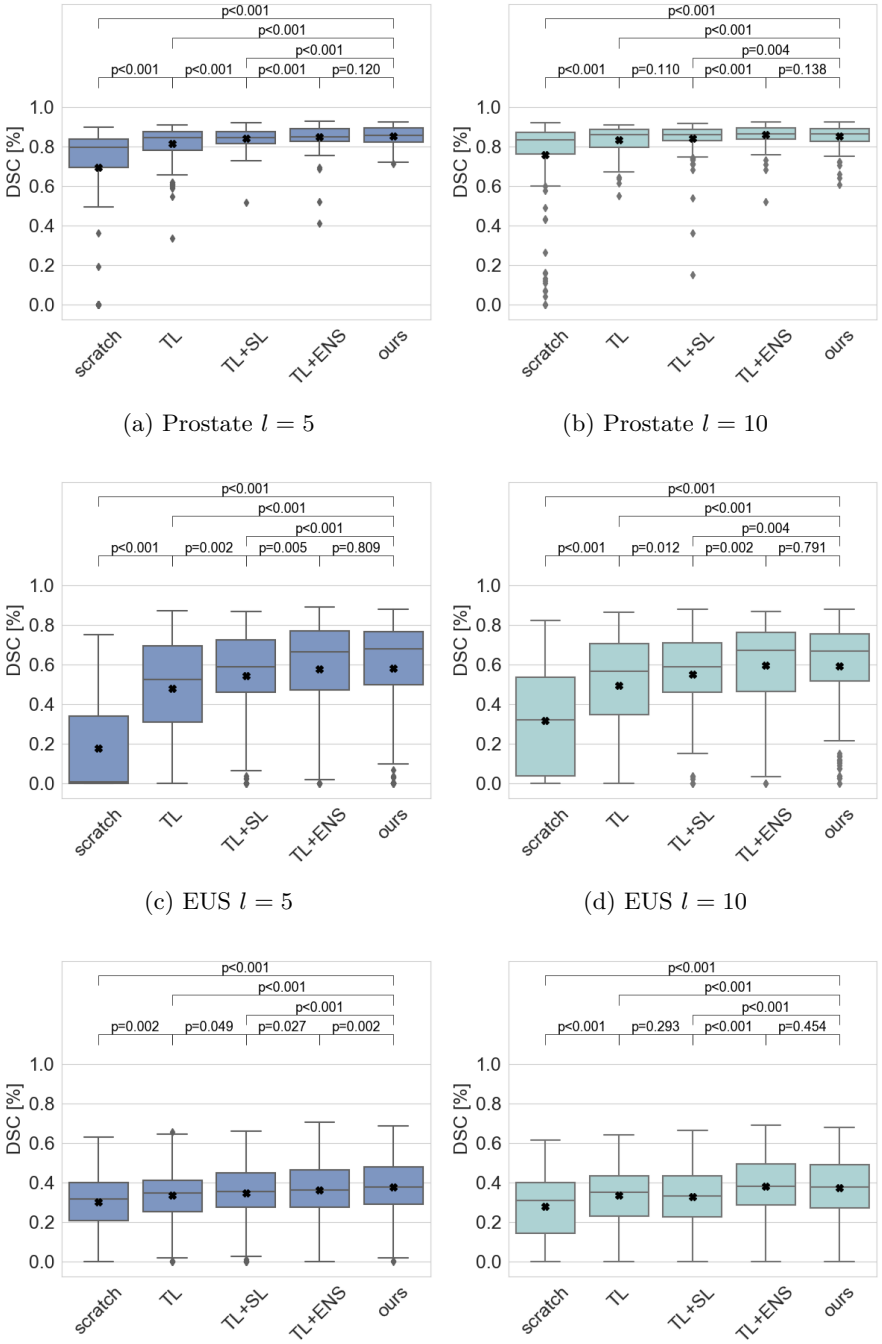
Similar to the prostate, we could also observe improvements for NVB and EUS with each step of our domain adaptation pipeline. Also for $l = 10$, improvements through the self-learning (SL) and ensembling (ENS) components are noted in the results. For this setting, though, the incorporation of entropy (H) as uncertainty information on the subject-level did not contribute to any improvement. We assume that the model predictions together with their post-processing are already of sufficient quality for the self-learning, and do not need to be weighted on a subject-level.

We additionally evaluated the performance of a variant of a state-of-the-art technique: uncertainty-aware self-learning for DA (denoted by ENS+H) in Table 5.4. This technique works substantially better than pure TL, but our method that combines both techniques, works considerably better for $l = 5$ labeled training cases. For $l = 10$ the impact of TL in the DA pipeline diminishes and the results for uncertainty-aware self-learning are in the range of our method's outcome.

Although our method could achieve significant improvement in the target domain for the NVB, the results are rather low (DSCs of 38.7 % for $l = 5$). For the other two structures, however, our DA method achieves outcomes in the range of inter-reader variability, if we compare to the results from the two observers in the source domain.

Method	**Prostate**			**EUS**			**NVB**		
	DSC [%]	ABD [mm]	95-HD [mm]	DSC [%]	ABD [mm]	95-HD [mm]	DSC [%]	ABD [mm]	95-HD [mm]
$l = 5$									
from scratch	69.4 ± 25.3	4.44 ± 7.5	12.41 ± 11.78	17.7 ± 22.4	10.39 ± 11.69	17.16 ± 15.95	30.3 ± 16.7	7.98 ±9.60	24.19 ± 16.50
TL	81.4 ± 8.9	1.98 ± 1.18	8.14 ± 5.47	48.0 ± 26.5	2.88 ± 5.67	6.69 ± 6.42	33.7 ± 14.9	4.98 ± 5.80	18.33 ± 12.16
TL + SL	84.3 ± 4.9	1.57 ± 0.65	5.82 ± 2.98	54.6 ± 23.9	1.73 ± 2.41	5.07 ± 3.24	35.0 ± 15.7	4.21 ±3.16	16.24 ± 9.97
TL + ENS	84.9 ± 4.9	1.51 ± 0.68	5.70 ± 2.93	57.8 ± 25.2	1.43 ± 1.22	4.62 ± 2.17	36.3 ± 16.5	3.83 ± 3.00	12.8 ± 8.42
ENS + H	83.1 ± 4.1	1.86 ± 0.77	7.00 ± 3.83	53.5 ± 27.3	1.87 ± 3.27	4.82 ± 4.01	35.5 ± 17.9	4.46 ± 4.33	16.2 ± 13.39
ours (TL + ENS + H)	85.5 ± 4.7	1.45 ± 0.68	5.57 ± 2.93	58.0 ± 25.7	1.62 ± 2.45	4.68 ± 3.38	37.8 ± 15.9	**3.37 ± 2.22**	12.22 ± 7.97
ours (majority)	**86.5 ± 3.7**	**1.33 ± 0.57**	**5.09 ± 2.27**	**59.2 ± 25.4**	**1.21 ± 0.92**	**3.95 ± 1.99**	**38.7 ± 16.1**	3.48 ± 3.31	**11.36 ± 7.36**
$l = 10$									
from scratch	76.0 ± 21.5	2.55 ± 2.77	8.40 ± 6.60	32.0 ± 25.0	3.51 ± 4.20	7.59 ± 5.19	28.0 ± 17.0	6.25 ± 7.07	19.10 ± 14.51
TL	83.4 ± 7.2	1.61 ± 0.73	6.09 ± 3.37	49.5 ± 26.0	2.00 ± 1.86	5.48 ± 3.15	33.5 ± 17.0	4.11 ± 2.99	15.54 ± 9.66
TL + SL	84.1 ± 8.8	1.53 ± 0.94	5.55 ± 3.16	55.2 ± 21.0	1.55 ± 1.28	4.62 ± 2.48	38.2 ± 17.1	4.63 ± 4.96	15.42 ± 11.96
TL + ENS	86.0 ± 4.9	1.36 ± 0.58	5.15 ± 2.37	59.6 ± 21.4	1.33 ± 1.15	4.29 ± 2.18	38.2 ± 16.4	3.39 ± 3.12	12.03 ± 9.57
ENS + H	85.0 ± 5.8	1.54 ± 0.73	5.15 ± 2.72	**59.8 ± 23.6**	1.51 ± 2.18	4.31 ± 2.05	37.9 ± 16.8	3.61 ± 2.42	13.12 ± 10.64
ours (TL + ENS + H)	85.5 ± 5.4	1.42 ± 0.64	5.15 ± 2.72	59.3 ± 22.7	**1.40 ± 1.10**	4.31 ± 2.05	37.4 ±16.8	3.63 ± 3.28	13.12 ± 10.64
ours (majority)	**86.6 ± 3.90**	**1.29 ± 0.49**	**4.59 ± 1.77**	59.1 ± 23.1	1.41 ± 1.51	**4.25 ± 2.14**	**38.1 ± 17.2**	**3.24 ± 2.51**	**11.34 ± 8.28**

Table 5.4: Evaluation results for training from scratch and the proposed DA method with its ablation experiments on the target (Prostate-3T) test data. 'Ours (Majority)' denotes the approach, where the ensemble of models from our proposed DA method is used to generate a majority vote as outcome. Best results per $l = 5$ and $l = 10$ setting are marked bold.



(a) Prostate $l = 5$

(b) Prostate $l = 10$

(c) EUS $l = 5$

(d) EUS $l = 10$

(e) NVB $l = 5$

(f) NVB $l = 10$

Figure 5.7: Boxplots for evaluation of the methods with $l = 5$ and $l = 10$ labeled images in the target domain. The marker ('x') represents the mean DSC value. P-values for the statistical significant differences between the methods are provided in the top of the plots.

(a) from scratch (b) TL (c) TL+SL (d) Ours (e) GT

Figure 5.8: Example case with segmentation results for the discussed approaches and the ground truth (GT). The quality of segmentation improves with each step/component of our DA pipeline. The DSC improved from 72.6 % for training from scratch to 81.7 % for our proposed DA approach. For the EUS, the DSCs are 0.0 % and 70.6 %, for training from scratch and our method, respectively. Similarly, the DSCs for the NVB improved from 39.2 % (from scratch) to 48.8 % (ours). The training of the CNNs was carried out with $l = 5$ labeled images.

Method	DSC [%]	ABD [mm]	95-HD [mm]
supervised			
source model (source)	69.4 ± 14.6	3.07 ± 2.00	18.99 ± 16.35
source model (target)	63.8 ± 17.4	4.50 ± 6.03	20.29 ± 14.96
target model (target)	77.3 ± 9.3	2.81 ± 1.93	16.00 ± 13.14
DA, $l = 5$			
from scratch	44.4 ± 16.5	9.95 ± 4.65	39.22 ± 12.24
TL	67.8 ± 15.2	4.33 ± 3.30	22.67 ± 14.20
ours	72.6 ± 11.8	3.40 ± 2.54	17.91 ± 13.97
ours (majority)	**73.1 ± 11.6**	**3.25 ± 2.45**	**16.85 ± 13.58**
DA, $l = 10$			
from scratch	53.4 ± 17.5	8.33 ± 4.83	36.86 ± 15.13
TL	69.0 ± 15.1	4.23 ± 3.50	22.14 ± 14.79
ours	72.8 ± 12.1	3.40 ± 2.58	17.81 ± 14.07
ours (majority)	**73.3 ± 12.0**	**3.25 ± 2.48**	**17.09 ± 13.85**

Table 5.5: Results (DSC) for the pancreas CT datasets. The assignments (source) and (target) denote whether the supervised model was evaluated on the source or target test data. Best results for the DA per $l = 5$ and $l = 10$ setting are marked bold.

GENERALIZATION CAPABILITY In this experimental setting, we evaluated our domain adaptation technique to pancreas CT segmentation to demonstrate its generic application.

The results are summarized in Table 5.5 and example segmentation outcomes for different cases are visualized in Figure 5.9. We see that a considerable domain shift exists as the source model's performance drops from a DSC of 69.4 % (source test data) to a DSC of 63.8 % on the target test data. The upper bound for this problem (using the full set of labeled data for supervised training), was found to be a DSC of 77.3%, an ABD of 2.81 mm and a 95-HD of 16 mm. With our DA, the performance could be improved to a DSC of 72.6 % with only five labeled target cases as (labeled) training data. This corresponds to a relative improvement of 13.8 %. Applying the ensembling strategy to our method, the average performance on the test data can be further improved to 73.2 % for the $l = 5$ setting. Considering the ABD and 95-HD of 3.25 mm and 17.1 mm, our method is close to the upper bound.

If we increase our labeled training set size to $l = 10$, we can observe an improvement for the TL results. However, the complete DA pipeline does not lead to much better results than for the $l = 5$ setting. This indicates the high potential that the analysis of the unlabeled data in the target domain can have.

Because we used the same test dataset as Xia et al. (2020), we can make a relative comparison for the performance gain to this state-

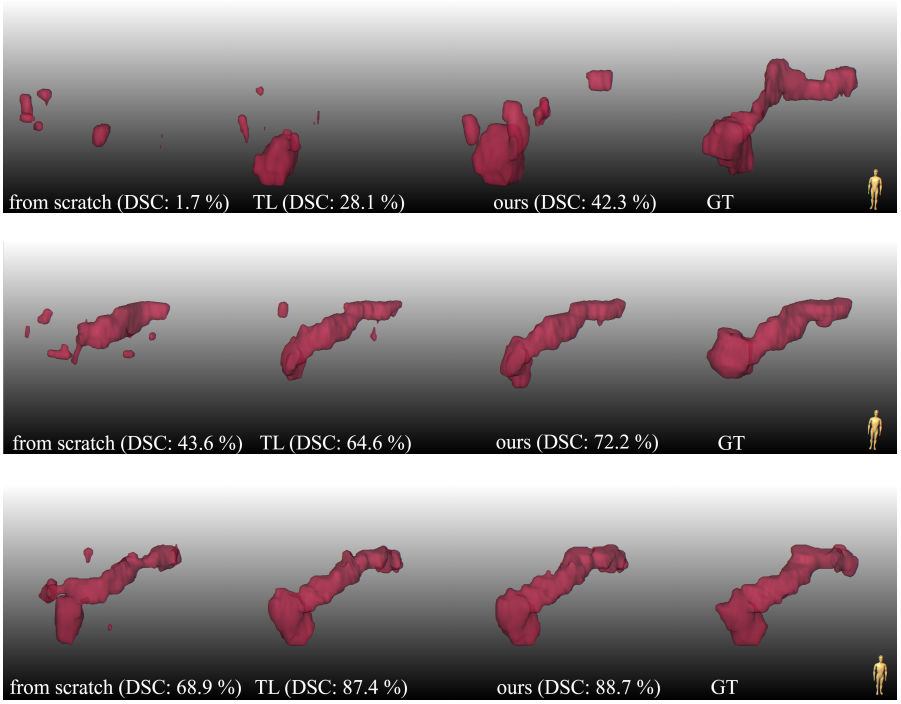


Figure 5.9: Visual segmentation outcomes of training from scratch, TL, our proposed DA method and the manual ground truth (GT) on three different test cases of the target domain. These examples represent different qualities of segmentation outcomes: one of the worst (top), a median (center) and one of the best (bottom) results.

of-the-art technique. The source model from Xia et al. (2020) had a performance DSC of 81.7 % in the source domain which decreased to 70.2 in the target domain. Through their multi-view co-training DA method, they could achieve a DSC of 74.9 % with access to the labeled data in the source domain and a DSC of 74.4 % in the source-relaxed DA setting. Thus, for the source-relaxed setting, they achieved a relative performance gain of 5.9 %. Although there exist some differences in the implementation of their method which make a direct comparison impossible (other backbone architecture, additional segmentation of other organs in the source domain), their lower relative performance gain of 5.9% vs. our's of 13.8 % indicates the effectiveness of our method and motivates using few labeled samples of the target domain.

5.7 DISCUSSION

In our study, we proposed a DA method that exploits the knowledge from an external domain in order to increase the CNN's performance and to reduce the number of labeled samples necessary for training. DA is crucial for the widespread employment of DL models for medical image analysis, given that the characteristics of medical datasets heavily depend on the types of the scanner and protocols used. Without DA,

one would need to create a model for each clinical site involving manual labeling of tens of volumes as the training dataset. In contrast, our study has demonstrated that we only need labeled images as few as $l = 5$ to transfer knowledge to the new domain.

The advantage of our DA method over many others is that it only requires the model that was trained on source data. This is particularly helpful when the entire source dataset cannot be shared with other clinical sites due to the size, or institutional and/or regulatory rules over the protection of data. Lastly, no prior (higher-level) knowledge about the organ-to-segment as in Bateson et al. (2020) is needed, making it easy to apply to other tasks.

Our DA has shown to achieve results in the range of inter-reader variability for EUS and the prostate. For the NVB, though, the quantitative results were rather low. This is likely because the NVB is a thin, tubular structure and is often obscured by the surrounding structures and image artifacts, resulting in inconsistent labeling between the readers.

For both the critical structures and pancreas segmentation, we saw that although increasing the number of labeled samples lead to an improvement in TL, our DA did not lead to substantially different results for different numbers of labeled data. Moreover, in the $l = 10$ setting, the uncertainty-aware self-learning (ENS + H) demonstrated performance en par with our proposed variant (that combines self-learning with TL) for the EUS and NVB. Thus, we can conclude, that the effect of TL diminishes when the number of labeled samples is increased, and it is sufficient to only use the uncertainty-aware self-learning. Furthermore, the ensembling plus its derived uncertainty seem to be very effective as information for the self-learning, which may otherwise degenerate the model's performance due to false pseudo labels.

The major downside of ensembling the models is its higher computational costs. On the other hand, the ensembling can be carried out successively, as the models do not need to be held in the GPU memory during training. Moreover, the ensembling strategy is a very simple and robust method without any requirements for the model architecture such as the inclusion of dropout (Li et al., 2021a) or the use of a conditional variational autoencoder (Wang et al., 2019c). Consequently, this allows for the reuse of any 'off-the-shelf' model to be employed for DA.

LIMITATIONS AND FUTURE WORK Although our evaluation showed that our domain adapted models performed well in the target domain for most structures, our study has limitations.

The trade-off of having such a flexibility concerning source data availability, network architecture, and training, is that the method requires some labeled training and validation data from the new domain. Although few samples suffice to obtain substantial improvement in the new domain, another research question arises from this aspect (similar as in the previous Chapter 4): How do we select the optimal samples

for this task? This question will need to be investigated in the future. The uncertainty measure can again be of help for this question, as the ground truth for samples with high prediction uncertainty is potentially giving higher information in the training.

Moreover, we can assume that our domain-adapted model is likely suffering from a performance drop when applied back to data from the source domain. We have not evaluated this aspect specifically, but as we do not incorporate neither any image data of the source domain nor any regularization, the target model will presumably not capture the distribution of the source domain anymore. Future work needs to investigate solutions for this challenge, for example, training a second encoder with our DA procedure and leaving the decoder and source encoder unchanged.

The ensembling of source models, which aims to provide better pseudo label candidates and uncertainty measures, may not be applicable when only one source model is available. In this case, ensembling could be achieved alternatively for example by Monte-Carlo dropout (Gal and Ghahramani, 2016), different subsets of labeled/unlabeled data from the target domain, or different minima during training from only one network (Huang et al., 2017a). Furthermore, a combination of different training schemes, such as different regularizations, different loss functions, or different learning rates, could be employed to generate models with differentiating minima. We used an ensemble size of $k = 5$, which is relatively small but a compromise between computation time and performance. If enough computation resources are available, the number of models could be increased, and performance may improve further.

Geometric 3D models of the EUS and NVB based on the proposed segmentation technique will allow detailed treatment planning of PCa, for instance, for radical prostatectomy or focal therapy. For these applications, however, the segmentation technique would need to be extended to include other surrounding structures, such as the rectal and bladder walls, which must also be protected from accidental damage. The proposed method can be easily extended to include the structures around the prostate relevant to the therapy planning.

We observed a rather low performance of our method for the NVB structure in the target domain. While the endorectal coil especially affects the shape and appearance of this structure, the low performance is presumably caused to a large extent by disagreement of the different readers involved for the NVB segmentation (an expert radiologist (Reader 1) segmented the source data and a medical student (Reader 3) segmented the target data). Therefore, future work should include a consensus segmentation of the NVB among multiple readers to have a more consistent ground truth for our DA method evaluation.

5.8 SUMMARY

In our work, we proposed a new DA strategy that combines TL and uncertainty-aware self-learning. By exploiting knowledge from an inter-domain level, the proposed strategy outperforms re-training a new model from scratch in the target domain with few labeled data clearly. Our method allows applying a trained network to another domain, for example, another scanner or another acquisition protocol, with only minimum quality loss. This makes automatic segmentation suitable for clinical applications, where the sharing of patient data is often highly restricted. Our supervised method achieves performance comparable to an experienced human reader in the source domain, and the DA gains performance similar to human readers for the prostate and the EUS in the target domain. We demonstrated the generic application of our DA framework by investigating its performance on another challenging task and data, namely pancreas CT segmentation. Moreover, we demonstrated that DL-based automatic segmentation of critical structures for PCa treatment, including the prostate, EUS, and NVB is feasible. The high performance of CNNs allows for a more precise planning of PCa therapy and thus has the potential to reduce the complications in PCa interventions.

6 CONCLUSION

The incorporation of mpMRI is becoming increasingly important in various stages of the clinical workflow for PCa care and research. With its more widespread employment, the effort and number of tasks for processing the mpMRI scans accumulates, and the demand for automating tasks rises. One crucial part of many applications that involve mpMRI is the segmentation of the prostate as well as its interior and adjacent structures. The introduction of deep learning, and in particular CNN-based methods, has led to high improvement gains throughout various biomedical analysis tasks, including the classification and segmentation of prostate MRI scans.

Our work covered in this thesis aims at improving the automatic segmentation of prostate structures in MRI to increase the reproducibility and quality of the analysis of MRI scans. In the following, we describe the contributions in more detail and discuss limitation and future research directions. To conclude, we provide a brief recap of this study's content.

6.1 RESEARCH CONTRIBUTIONS

Although the CNN algorithms regularly set new records on benchmark datasets, the proposed methods for prostate MRI image analysis and other tasks have limitations and challenges, including: (1) neglecting the prostate's detailed anatomy and adjacent structures, (2) relying on MRI data with partly insufficient quality for specific prostate regions (3) demanding large quantities of annotated training data and (4) being sensitive to shifts in the distributions of training and unseen test data. With our work, we set out to tackle these challenges.

In this thesis, we targeted prostate structures which, to our best knowledge, have not yet been considered for automatic segmentation. Specifically, we are the first to investigate and demonstrate the feasibility of automatically segmenting the interior structures DPU, AFS and the adjacent NVB and EUS with outcome quality en par with human inter-reader variability. Within this context, we addressed the other three shortcomings by exploiting supplementary data and knowledge from different levels of clinical data.

- *patient-level data:* Most segmentation errors for the whole gland occur in the apex and base (lower and upper third of the prostate). This is due to the partial volume effect of the axial T2w scans,

which are commonly used for automatic medical image analysis. Since other directions (sagittal and coronal) are acquired as standard of care, we developed a multi-stream CNN-architecture, that can process multi-planar data simultaneously and outputs an isotropic high resolution segmentation. With our method, we achieved improved segmentation accuracy for the extreme parts of the gland when compared to using only the axial scan direction.

- *intra-domain-level data:* We aimed to reduce the CNN method's need for large amounts of labeled training data to obtain sufficient outcome quality. As slice-wise labeling of 3D volumes is a very tedious and time-consuming task requiring medical expert knowledge, various methods have been proposed that leverage unlabeled data to alleviate the need for a large amount of labeled samples (Tajbakhsh et al., 2020). Motivated by the potential of those techniques, we developed a novel SSL method, that combines concepts from temporal ensembling (Laine and Aila, 2017) and self-learning (Scudder, 1965; Agrawala, 1970). For the task of segmenting a detailed interior anatomy of the prostate, our method achieved superior performance over the supervised baseline and other state-of-the-art approaches.

- *inter-domain-level data:* Another prominent problem of CNN methods is their sensitivity to domain shift, leading to a performance drop on data from an unseen domain. There exist several works on DA, that reduce the performance gap in the new (target) domain. However, most of those methods use data from the original (source) domain alongside the data from the target domain. This is a requirement that may often not be met, as medical data is subject to strict sharing policies across organizations or sites. In the context of critical structures segmentation for PCa therapy planning, we developed a simple and effective semi-supervised DA method that relaxes this requirement. Our method has demonstrated to attain performances close to inter-reader variability for the majority of targeted structures. By relying on the concepts of transfer learning and uncertainty-aware self-learning, our method only requires a small amount of manually labeled samples from the new domain. Therefore, it allows for the adaptation of arbitrary off-the-shelf models to new data.

In contrast to several other methods suggested so far for prostate structures segmentation, all of our proposed methods can be applied to the whole 3D volume without requiring multiple or high-end GPUs [1]. Thus, our methods do not involve any slice- or patch-based strategy that would reduce the network's receptive field and therefore the spatial

1 We used either a 11 GB RAM NVIDIA GeForce GTX 1080 Ti or a 12 GB NVIDIA Titan X Pascal GPU for all our prostate structures segmentation experiments.

context. This also spares the user from optimizing aspects such as patch dimensions, how to mine the patches (Bian et al., 2018), or how to handle the patches during inference (Madesta et al., 2020).

Moreover, for our SSL and DA methods proposed in Chapter 4 and Chapter 5, we have shown their effectiveness on other important biomedical segmentation tasks. Our UATS concept improved over its supervised counterpart also for hippocampus segmentation in T1w MRI and skin lesion segmentation in color photographs. Our semi-supervised DA could demonstrate promising improvement for pancreas segmentation in CT volumes, whereas the domain shift resulted not only from other sites, but also from a shift in study population.

Lastly, in the spirit of open science, we released the ground truth segmentations, which were created as part of this thesis for the public PROSTATEx challenge dataset (Litjens et al., 2014a), comprising the high resolution whole gland (Schindele et al., 2020; Clark et al., 2013) and a more detailed zonal anatomy segmentation (Meyer et al., 2020; Clark et al., 2013).

6.2 LIMITATIONS AND FUTURE WORK

We demonstrated that, by exploiting supplementary data and information, (1) CNN-based methods for prostate structure segmentation can be significantly improved and (2) barriers for their development and employment in the clinical workflow can be reduced. However, our methods inherit limitations that we discuss in the following paragraphs. Furthermore, we give an outlook what future research should address.

EVALUATION ASPECTS Considering the differences between manual and automatic predictions, our achieved segmentation performance was as good as human expert performance for all target structures. Nevertheless, our study has some shortcomings on this point.

We based our training data on the manual segmentation of only one medical expert. Since there are no official guidelines on how to segment the prostate structures (Montagne et al., 2021), the task fulfillment depends heavily on the annotator's expertise. In the literature, human expert segmentations showed significant variations due to reader experience differences (radiological vs. non-radiological experts) (Becker et al., 2019), and too little contrast of the tissue boundaries between PZ and TZ for prostates of smaller size (Montagne et al., 2021). This raises the question of how the *actual* ground truth should be determined and created. If only one reader is available, multi-planar data can be leveraged, as we have done for our high-resolution manual reference segmentation for the whole gland in Chapter 3. Furthermore, another interesting evaluation measure to consider in follow-up studies is the *intra-reader* performance, for which the same reader outlines the same structure multiple times. Alternatively, provided that enough medical

experts are available, one can create a consensus segmentation, for example, with majority voting or the STAPLE algorithm (Warfield et al., 2004; Montagne et al., 2021).

Another limitation of our work is that, for the interior anatomy of the prostate (Chapter 5) and the critical structures segmentation in Chapter 4 (EUS and NVB), our test dataset was rather small (20 and 16 cases, respectively). Future research needs to investigate whether the improvement can be confirmed with larger test sets.

Besides consistent label quality and an increased number of test cases, it is also important to evaluate how the methods can handle abnormalities of the gland and its adjacency. This meta-information was not available for the test data in our work, but it will be of high interest to explore whether and how our methods are affected by the presence of, for example, tumors, BPH, prostatitis, cysts, or calcifications.

EXPLOITING OTHER TYPES OF DATA The type of data we used in our studies is from three different levels (patient, intra-domain and inter-domain) within clinical dataset structures and is easily available in the clinical practice. However, on these three and further levels (e.g., organ and other medical fields), there is more data available which can be exploited in future work.

In our multi-stream network architecture in Chapter 3, we evaluated the incorporation of additional T2w scan directions. However, on the level of patient data, it may be further investigated whether other (non-imaging) patient information could be leveraged by employing a multi-task setting (e.g., by including an auxiliary classification task) that may support attaining more robust features.

For our UATS, we made use of the unlabeled data of the intra-domain level that is easier to obtain than fully labeled samples. On this data level, several other techniques exist which have the potential to reduce the workload that is induced by creating fully labeled samples. For example, weakly or scarcely annotated data may be used where objective functions are adapted to work on incomplete ground truth labels (Çiçek et al., 2016). Another field of research is active learning, where the annotator is put into the loop to label only those samples that are supposed to be most rewarding for training (Budd et al., 2021).

In Chapter 5, we exploited knowledge gained in an external domain to improve segmentation performance on a new dataset, thereby reducing the labeling workload. To this end, we have used model weights that represent the knowledge obtained from another domain. If there is additionally the actual image data, ideally labeled, from several domains available, domain generalization techniques can be exploited to apply the trained model on any domain without specific DA (Zhang et al., 2020; Liu et al., 2020b). However, aggregating data from multiple domains limits these technique's feasibility in the medical field. Thus, future work needs to be carried out that relaxes this restriction to, for example,

only one labeled dataset and multiple unlabeled datasets from other domains.

When considering a higher level of clinical data, there is potential in exploiting data from other medical branches, such as MRI acquisitions of the brain or liver. Although these tasks are not directly relevant to prostate segmentation, they can allow for computing more robust features.

Alternatively, one can consider information from a lower level in the data than those proposed thus far. On the level of the organs, prior information about the size of the prostate structures may regularize the training (e.g., similar to Bateson et al. (2020)). Moreover, segmentations of other (proximal) organs can be leveraged. We observed that the qualitative segmentation outcome of the prostate gland was better, when the interior anatomy was targeted by the CNN, in contrast to targeting just one class (the prostate). This indicates that additional segmentation targets, such as lesions or adjacent structures (e.g., seminal vesicles, or pelvic diaphragm) may provide more valuable information than a simple foreground or background label.

METHODOLOGICAL ASPECTS We targeted the automatic segmentation of different substructures and investigated the use of supplementary data for improving their segmentation outcome. Although we assessed these aspects individually with the methods proposed in Chapters 3, 4 and 5, we want to point out that the methods are not exclusive and can be used in conjunction with each other. For instance, the domain adaptation technique (Chapter 5) can be used with the multi-stream architecture Chapter 3, or the semi-supervised techniques introduced in Chapters 4 and 5 can be exchanged. We have not carried out such an experiment because our intention was to investigate the effect of the additional data individually to derive more insights for the design of future studies. We leave it to future work to investigate the effect of combining these methods.

While the incorporation of additional data or information is a rewarding option to improve the segmentation quality, there are also other potential routes to improve the segmentation performance, if no such supplementary input is available. These are outlined in the following.

All our proposed solutions are based on the 3D U-Net architecture that has proven (often in slightly adapted manner) successful for various segmentation tasks in medical images (Isensee et al., 2021). This is in line with our findings that our 3D U-Net variants were sufficient for attaining performances en par with human annotators for prostate structures segmentation if enough data was available. Nonetheless, as our methods are not restricted to the employed U-Net variants, but can be seen complementary for other network architectures, it would be interesting to evaluate whether segmentation performance can be

further improved by evolving the underlying backbone architecture, for example, with deep supervision, residual, or dense connections.

Another aspect to focus future work on is the loss function that drives the learning of the networks. Although the DSC loss, which we incorporated, has demonstrated its effectiveness in several works, there exists a large variety of other training objectives (Ma et al., 2021), that we highly recommend evaluating in future work. For example, training objectives that pay attention to the boundary and foreground edge information have shown promising improvement in segmentation performance for the prostate (Zhu et al., 2020; Qin et al., 2020; Jia et al., 2019).

We have used different data augmentation techniques in our work, but we have not set a particular focus on this aspect and have rather relied on basic geometric transformations. Therefore, more effort needs to be undertaken regarding this limitation, as several works have demonstrated that a more elaborated augmentation strategy leads to increased performance and robustness of the model (Zhang et al., 2020; Sheikh and Schultz, 2020; Hesse et al., 2020).

One other promising route for future work is to concentrate further on learning more meaningful representations of the data (e.g., in the latent space), which can facilitate the learning of the downstream task (i.e., segmentation or classification) (Bengio et al., 2013). Priors of this representation can regularize the training. For instance, in contrastive learning, the objective is to learn embeddings of the data, whose vectors are close to each other, when the data inputs are similar. This similarity can refer to, for example, augmented images (Chen et al., 2020c) or patches from the relatively same position in scans from different patients (Chaitanya et al., 2020). For the task of prostate segmentation, the similarity could be encoded by information about the structure size (e.g., the ratio of zone sizes) that we can obtain from the training labels. Another strategy is to use variational autoencoders (Kingma and Welling, 2013) that regularize the representations by introducing a distribution prior on variables of the latent space. This has the additional benefit, that multiple segmentation candidates can be inferred for an input segmentation (Kohl et al., 2018; Baumgartner et al., 2019).

Lastly, we believe that it will be of high value to explicitly incorporate regularization through shape information into either the learned representation or in the post-processing to improve performance and generalization capacity. There have been efforts to this respect for prostate segmentation (Liu et al., 2020b) and several other tasks (Xie et al., 2021), but it needs to be investigated in future work whether it is beneficial for a more detailed interior anatomy and the adjacent structures segmentation, as these individual shapes are more diverse than the whole gland.

FUTURE RESEARCH DIRECTIONS With the incorporation of additional data and the methodological aspects described above, we aim to learn better features, that can improve the segmentation performance. However, when introducing our segmentation methods into clinical workflows, there remain some open challenges that need to be addressed, even when the segmentation performance and robustness can be further improved. These include the tumor segmentation, out-of-distribution detection and life-long (continuous) learning.

We expanded the segmentation of the prostate into a more detailed subdivision and its adjacent structures. However, employing the methods in PCa therapy planning and monitoring, requires segmenting the tumor, too. To this end, we refer to recent work, such as Dai et al. (2020) and Saha et al. (2021).

Even for the best automatic method, there will be some cases that the automatic methods fail to segment, either due to insufficient image quality or model capacity. This is important to consider in systems where automatic follow-up tasks rely on the segmentation outcome. Therefore, one important future research direction will be to study how the model can recognize its failure and output a heads-up to the human or subsequent algorithm. The research fields of predictive uncertainty and out-of-distribution detection should point to respective solutions for this (e.g., Kohl et al., 2018; Mehrtash et al., 2020).

Finally, future work needs to be carried out on the ability of the methods to adapt to new data or tasks without *forgetting*. For instance, in a case, where we want to extend our prostate zone segmentation algorithm by additionally targeting the tumor, it needs to be avoided to re-train the method on all of the datasets again. We encourage follow-up research that tries to solve this challenge of continuous learning (Karani et al., 2018; Parisi et al., 2019).

6.3 SUMMARY

The research carried out in this thesis addressed the CNN-based segmentation of prostate structures that is needed for various tasks in PCa diagnosis, as well as for therapy planning and monitoring. In this context, our objectives were (1) to provide reliable segmentation results for structures that are relevant for these tasks, and (2) to develop methods that improve the segmentation outcome of CNNs by exploiting supplementary data that is easily available in the clinical routine.

The segmentation methods in our work targeted structures that have not been considered in prior works on automatic segmentation. We extended the common two-zones (TZ and PZ) segmentation by the AFS and the DPU to obtain a more detailed anatomy segmentation of the interior gland and investigated the segmentation of the NVB and EUS, which are critical structures impacting the outcome of PCa therapies.

We achieved performance en par with manual inter-reader variability for all considered structures.

In this context, we contributed novel methods that leverage additional data for their training. Specifically, we (1) included additional scan directions of the T2w acquisitions in a multi-stream 3D CNN. Moreover, we (2) leveraged unlabeled data in a semi-supervised segmentation method and (3) exploited knowledge from another domain within a semi-supervised DA technique. With our studies, we demonstrated the benefit of including supplementary data and knowledge from different levels of the clinical data structure for improving the method's performance and mitigating common problems of CNNs.

One main limitation of our work was the small test dataset size for prostate zones and critical structure segmentation. In the future, our methods should be evaluated on a larger test set and with respect to additional factors such as intra-rater variability and impact of pathological cases. Moreover, elaborating the underlying architecture, loss function and data augmentation techniques as well as regularization of the methods with shape or representation priors are aspects to be investigated in future work. Nonetheless, our results indicate that the methods developed in this thesis have the potential to automate tasks within the diagnosis, research and treatment for PCa and moreover lower the obstacles of CNN implementation in the clinical practice.

APPENDIX A

Additional material for Chapter 4.

A.1 APPENDIX A.1

To train our supervised baselines in Chapter 4, we used a multi-class DSC loss as follows:

$$L_{DSC}(y, \hat{y}) = -\sum_{c \in C} \frac{2 \sum_{i=1}^{N} \hat{y}_{c,i} y_{c,i} + \epsilon}{\sum_{i=1}^{N} y_{c,i} + \sum_{i=1}^{N} \hat{y}_{c,i} + \epsilon}, \quad \text{(A.1)}$$

where y and $\hat{y}$ are the ground truth and prediction for an input sample. The parameter ϵ is a small constant for ensuring numerical stability.

A.2 APPENDIX A.2

The task-specific hyperparameters for the UATS method are listed in Table A.1.

Dataset	Learning rate	Batch size	Confident voxels per class
Prostate	5e-5	2	PZ: 50%, TZ: 50%, DPU: 10%, AFS: 10%, Background: 50%
Skin	1e-5	8	Lesion: 50%, Background: 50%
Hippocampus	4e-5	4	C1: 50%, C2: 50%, Background: 50%

Table A.1: Task-specific hyperparameter settings for our experiments in Chapter 4.

A.3 APPENDIX A.3

The bias correction for the temporal ensemble was introduced in the work from Laine and Aila Laine and Aila, 2017. In their algorithm, the ensemble $\hat{E}_i$ is initialized as zero vector and accumulated over the epochs as: $\hat{E}_i \leftarrow \alpha \hat{E}_i + (1 - \alpha)\hat{Y}_i$, where $\hat{Y}_i$ is the current epoch's prediction. To correct for the startup-bias, the training targets E_i are created by dividing $\hat{E}_i$ by $(1 - \alpha)^t$, with t being the current epoch.

We found that enabling the bias correction achieved better performance. However, during the preparation of this manuscript, we also found that t did not get updated in our sourcecode after every epoch,

Method	PZ	TZ	DPU	AFS	∅
supervised ANISO	78.2 ± 6.0	87.1 ± 6.3	74.3 ± 7.0	39.0 ± 15.9	69.7
$t = epoch, \alpha = 0.6$	79.0 ± 6.1	87.2 ± 5.2	74.6 ± 6.3	43.4 ± 10.9	71.0
$t = 1, \alpha = 0$	77.7 ± 5.6	86.6 ± 4.7	73.5 ± 6.1	42.8 ± 9.7	70.2
$t = 1, \alpha = 0.2$	78.5 ± 5.8	87.0 ± 5.5	75.9 ± 7.0	38.8 ± 9.7	70.0
$t = 1, \alpha = 0.4$	79.9 ± 6.0	87.7 ± 5.8	76.1 ± 7.1	43.2 ± 11.4	71.7
$t = 1, \alpha = 0.6$	79.5 ± 5.8	87.8 ± 5.6	76.0 ± 7.0	46.2 ± 10.8	72.4
$t = 1, \alpha = 0.8$	78.4 ± 6.1	87.3 ± 5.5	75.7 ± 7.7	51.5 ± 11.5	73.2

Table A.2: DSC [%] for different prostate zones in evaluation of the bias correction effect. Fold1. $\alpha = 0$ corresponds to no bias correction.

but always kept the value 1. As $\alpha = 0.6$ in our setting, this resulted in dividing E_i by 0.4 for the loss calculation. Figure A.1 plots the effect that the division of the training target by a constant $(1 - \alpha)$ has on the negative cDSC output, that is incorporated in our losses. The higher α is, the lower the output for the negative cDSC is. This effect is magnified when the predictions $\hat{Y}_i$ are getting closer to the ensemble target E_i. As the ensemble prediction is not only resued for L_{Cons}, but also as pseudo label for the unlabeled samples in L_{Task}, these unlabeled samples are weighted higher in L_{Task} when the predictions are of good quality.

We ran experiments to investigate what effect altering E_i with different values for α and with the actual bias correction have. We conducted the training on the first fold for the prostate zone segmentation and evaluated the method on the test dataset. The results are summarized in Table A.2. As we can see, performing the actual bias correction and simply dividing the ensemble by a constant has a positive effect on the average performance across classes. The higher α is, the higher the overall performance is. Not altering the ensemble performs similar to the supervised baseline.

A.4 APPENDIX A.4

An example for a confidence mask for our UATS method of a prostate case with the settings from Table A.1 is depicted in Figure A.2.

A.5 APPENDIX A.5

To give an example of improvement through our UATS method, we included Figure A.3.

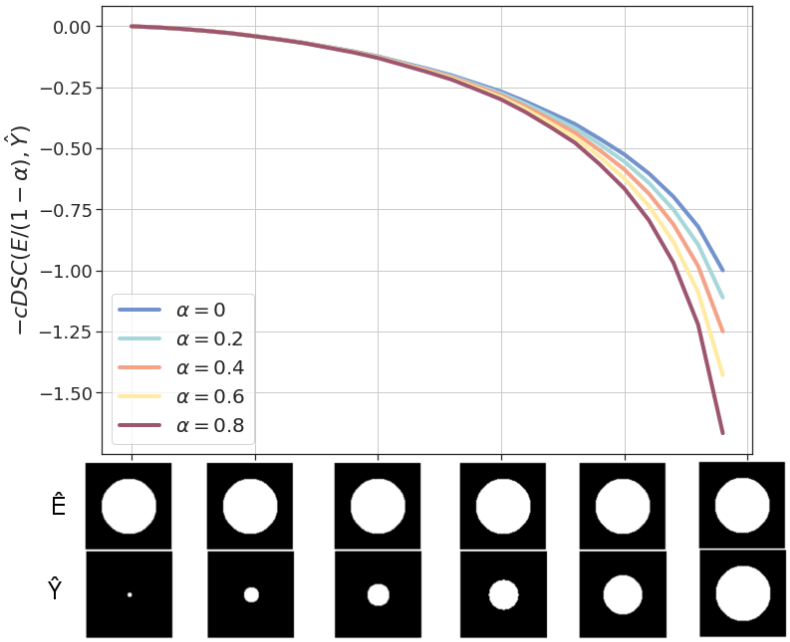


Figure A.1: Effect of dividing the ensemble ground truth by $(1-\alpha)$ on the cDSC score. For simplicity, we assumed the ensembe $\hat{E}$ to be a binary segmentation.

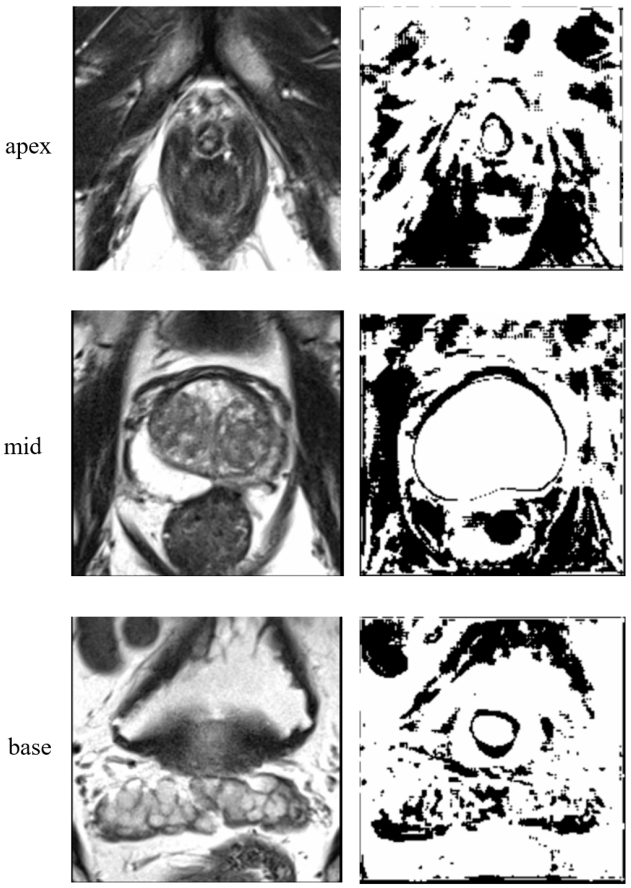
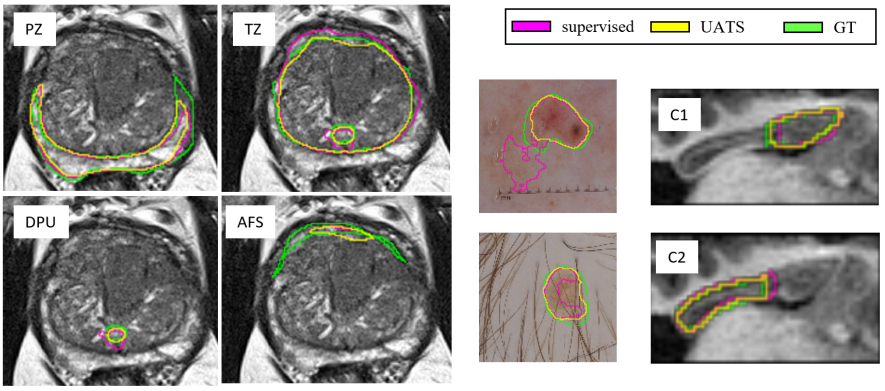


Figure A.2: Example confidence masks for our UATS method for one prostate case on apex, mid-gland and base-level of the prostate. Masks are calculated based on the softmax confidence.



(a) Prostate (b) Skin (c) Hippocampus

Figure A.3: Example cases that visualize the prediction of the supervised and UATS method, as well as the manual ground truth.

B

APPENDIX B

We summarize the hyperparameter details for our experiments on domain adpatation in Chapter 5 in this appendix.

Parameters were found empirically on the validation set. For prostate structure segmentation, the Adam optimizer was used. LR was reduced by a factor of 0.8 when the validation loss did not decrease by value of 0.00001. To obtain the pseudo labels, we applied thresholds of 0.5, 0.5, 0.5 and 0.999 for the prostate, NVB, EUS and background, respectively. More task-specific hyprparameters are provided in Table B.1.

For pancreas segmentation, the RMSprop optimizer was used. The learning rate was reduced by a factor of 0.9 when the validation loss did not decrease by value of 0.005. We applied a threshold of 0.25 to the foreground (pancreas) class. Table B.2.

	supervised + DA Stage II	**DA Stage I (TL)**
max epochs	300	300
batch size	2	2
learning rate (LR)	$1e^{03}$	$1e^{04}$
early stop patience	40	30
LR decay patience	10	10

Table B.1: Overview of method-specific hyperparameters for the **prostate structures** segmentation.

	supervised	**DA Stage I (TL)**	**DA Stage II**
max epochs	1500	500	500
batch size	4	1	4
learning rate (LR)	$6e^{05}$	$1e^{05}$	$1e^{05}$
early stop patience	150	50	75
LR decay patience	50	20	25

Table B.2: Overview of method-specific hyperparameters for the **pancreas** structures segmentation.

BIBLIOGRAPHY

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi (2021). "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." In: *Information Fusion* 76, pp. 243–297. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2021.05.008 (cit. on p. 20).

Agrawala, A. (1970). "Learning with a probabilistic teacher." In: *IEEE Transactions on Information Theory* 16.4, pp. 373–379. DOI: 10.1109/TIT.1970.1054472 (cit. on pp. 23, 130).

Aldoj, N., F. Biavati, F. Michallek, S. Stober, and M. Dewey (2020). "Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-Net." In: *Scientific reports* 10.1, pp. 1–17. DOI: 10.1038/s41598-020-71080-0 (cit. on pp. 63, 64).

Bai, W., O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert (2017). "Semi-supervised Learning for Network-Based Cardiac MR Image Segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 10434, pp. 253–260. DOI: 10.1007/978-3-319-66185-8_29 (cit. on pp. 65, 78, 81, 88).

Bashkanov, O., A. Meyer, D. Schindele, M. Schostak, K.-D. Tönnies, C. Hansen, and M. Rak (2021). "Learning Multi-Modal Volumetric Prostate Registration With Weak Inter-Subject Spatial Correspondence." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1817–1821. DOI: 10.1109/ISBI48211.2021.9433848 (cit. on p. 30).

Bateson, M., J. Dolz, H. Kervadec, H. Lombaert, and I. B. Ayed (2021). "Constrained Domain Adaptation for Image Segmentation." In: *IEEE Transactions on Medical Imaging* 40.7, pp. 1875–1887. DOI: 10.1109/TMI.2021.3067688 (cit. on p. 103).

Bateson, M., H. Kervadec, J. Dolz, H. Lombaert, and I. Ben Ayed (2020). "Source-Relaxed Domain Adaptation for Image Segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 490–499. DOI: 10.1007/978-3-030-59710-8_48 (cit. on pp. 101, 104, 106, 125, 133).

Baumgartner, C. F., K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu (2019). "Phiseg: Capturing uncertainty in medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 119–127. DOI: 10.1007/978-3-030-32245-8_14 (cit. on pp. 20, 134).

Becker, A. S., K. Chaitanya, K. Schawkat, U. J. Muehlematter, A. M. Hötker, E. Konukoglu, and O. F. Donati (2019). "Variability of manual segmentation of the prostate in axial T2-weighted MRI: A multi-reader study." In: *European Journal of Radiology* 121, p. 108716. DOI: 10.1016/j.ejrad.2019.108716 (cit. on p. 131).

Bengio, Y., A. Courville, and P. Vincent (2013). "Representation learning: A review and new perspectives." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50 (cit. on p. 134).

Bermúdez-Chacón, R., P. Márquez-Neila, M. Salzmann, and P. Fua (2018). "A domain-adaptive two-stream U-Net for electron microscopy image segmentation." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 400–404 (cit. on p. 102).

Bian, C., X. Yang, J. Ma, S. Zheng, Y.-A. Liu, R. Nezafat, P.-A. Heng, and Y. Zheng (2018). "Pyramid network with online hard example mining for accurate left atrium segmentation." In: *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 237–245. DOI: 10.1007/978-3-030-12029-0_26 (cit. on p. 131).

Bian, C., C. Yuan, J. Wang, M. Li, X. Yang, S. Yu, K. Ma, J. Yuan, and Y. Zheng (2020). "Uncertainty-aware domain alignment for anatomical structure segmentation." In: *Medical Image Analysis* 64, p. 101732. DOI: 10.1016/j.media.2020.101732 (cit. on pp. 20, 103, 115).

Bloch, N., A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani (2015). "NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures." In: *The Cancer Imaging Archive*. DOI: 10.7937/K9/TCIA.2015.zF0vlOPv (cit. on p. 112).

Blum, A. and T. Mitchell (1998). "Combining labeled and unlabeled data with co-training." In: *Proceedings of the Annual Conference on Computational Learning Theory*, pp. 92–100. DOI: 10.1145/279943.2799620 (cit. on p. 24).

Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra (2015). "Weight uncertainty in neural network." In: *International Conference on Machine Learning (ICML)*, pp. 1613–1622 (cit. on p. 19).

Bortsova, G., F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne (2019). "Semi-supervised medical image segmentation via learning consistency under transformations." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 810–818. DOI: 10.1007/978-3-030-32226-7_90 (cit. on p. 66).

Brosch, T., J. Peters, A. Groth, T. Stehle, and J. Weese (2018). "Deep Learning-Based Boundary Detection for Model-Based Segmentation with Application to MR Prostate Segmentation." In: (cit. on pp. 33, 34).

Budd, S., E. Robinson, and B. Kainz (2021). "A survey on active learning and human-in-the-loop deep learning for medical image analysis." In: *Medical Image Analysis*, p. 102062. DOI: j.media.2021.102062 (cit. on p. 132).

Cao, X., H. Chen, Y. Li, Y. Peng, S. Wang, and L. Cheng (2020). "Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation." In: *IEEE Transactions on Medical Imaging* 40.1, pp. 431–443. DOI: 10.1109/TMI.2020.3029161 (cit. on p. 69).

Cha, S.-H. (2007). "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions." In: *International Journal of Mathematical Models and Methods in Applied Sciences* 1.4, pp. 300–307 (cit. on p. 73).

Chaitanya, K., E. Erdil, N. Karani, and E. Konukoglu (2020). *Contrastive learning of global and local features for medical image segmentation with limited annotations* (cit. on pp. 65, 134).

Chapelle, O., B. Schölkopf, and A. Zien, eds. (2006). *Semi-Supervised Learning*. MIT Press. ISBN: 9780262033589 (cit. on p. 22).

Chen, C., Q. Dou, H. Chen, and P.-A. Heng (2018). "Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-ray Segmentation." In: *International Workshop on Machine Learning in Medical Imaging*, pp. 143–151. DOI: 10.1007/978-3-030-00919-9_17 (cit. on p. 103).

Chen, C., Q. Dou, H. Chen, J. Qin, and P. A. Heng (2020a). "Unsupervised Bidirectional Cross-Modality Adaptation via Deeply Synergistic Image and Feature Alignment for Medical Image Segmentation." In: *IEEE Transactions on Medical Imaging* 39.7, pp. 2494–2505. DOI: 10.1109/TMI.2020.2972701 (cit. on p. 103).

Chen, L., W. Zhang, Y. Wu, M. Strauch, and D. Merhof (2020b). "Semi-supervised Instance Segmentation with a Learned Shape Prior." In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing (MICCAI Workshop)*, pp. 94–102. DOI: 10.1007/978-3-030-61166-8_10 (cit. on p. 65).

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (2020c). "A simple framework for contrastive learning of visual representations." In: *International Conference on Machine Learning (ICML)*, pp. 1597–1607 (cit. on p. 134).

Chen, W., Y. Zhang, J. He, Y. Qiao, Y. Chen, H. Shi, E. X. Wu, and X. Tang (2019). "Prostate Segmentation using 2D Bridged U-net." In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. DOI: 10.1109/IJCNN.2019.8851908 (cit. on p. 33).

Chen, Y., L. Xing, L. Yu, W. Liu, B. Pooya Fahimian, T. Niedermayr, H. P. Bagshaw, M. Buyyounouski, and B. Han (2021). "MR to ultrasound image registration with segmentation-based learning for HDR prostate brachytherapy." In: *Medical Physics* 48.6, pp. 3074–3083. DOI: 10.1002/mp.14901 (cit. on p. 30).

Cheng, R., N. Lay, F. Mertan, B. Turkbey, H. R. Roth, L. Lu, W. Gandler, E. S. McCreedy, T. Pohida, P. Choyke, M. J. McAuliffe, and R. M. Summers (2017). "Deep learning with orthogonal volumetric HED segmentation and 3D surface reconstruction model of prostate MRI." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 749–753. DOI: 10.1109/ISBI.2017.7950627 (cit. on pp. 33, 35, 36, 53).

Cheplygina, V., M. de Bruijne, and J. P. Pluim (2019). "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis." In: *Medical Image Analysis* 54, pp. 280–296. DOI: 10.1016/j.media.2019.03.009 (cit. on p. 24).

Chilali, O., P. Puech, S. Lakroum, M. Diaf, S. Mordon, and N. Betrouni (Dec. 2016). "Gland and Zonal Segmentation of Prostate on T2w MR Images." In: *Journal of Digital Imaging* 29.6, pp. 730–736. DOI: 10.1007/s10278-016-9890-0 (cit. on pp. 63, 64).

Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 424–432. DOI: 10.1007/978-3-319-46723-8_49 (cit. on pp. 17, 37, 62, 70, 132).

Clark, K., B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior (2013). "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository." In: *Journal of Digital Imaging* 26.6, pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7 (cit. on pp. 40, 75, 112, 113, 131).

Clark, T., J. Zhang, S. Baig, A. Wong, M. A. Haider, and F. Khalvati (2017). "Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted MRI using convolutional neural networks." In: *Journal of Medical Imaging* 4.4, p. 041307 (cit. on pp. 63, 64).

Codella, N. C. F., D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern (2018). "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 168–172. DOI: 10.1109/ISBI.2018.8363547 (cit. on p. 76).

Cui, W., Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye (2019). "Semi-supervised Brain Lesion Segmentation with an Adapted Mean Teacher Model." In: *International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 554–565. DOI: 10.1007/978-3-030-20351-1_43 (cit. on pp. 66, 68).

Dai, Z., E. Carver, C. Liu, J. Lee, A. Feldman, W. Zong, M. Pantelic, M. Elshaikh, and N. Wen (2020). "Segmentation of the Prostatic Gland and the Intraprostatic Lesions on Multiparametic Magnetic Resonance Imaging Using Mask Region-Based Convolutional Neural Networks." In: *Advances in Radiation Oncology* 5.3, pp. 473–481. DOI: 10.1016/j.adro.2020.01.005 (cit. on p. 135).

de Vente, C., P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta (2020). "Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-parametric MRI." In: *IEEE Transactions on Biomedical Engineering*, pp. 374–383. DOI: 10.1109/TBME.2020.2993528 (cit. on pp. 60, 106).

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255 (cit. on pp. 13, 102).

Dice, L. R. (1945). "Measures of the amount of ecologic association between species." In: *Ecology* 26.3, pp. 297–302. DOI: 10.2307/1932409 (cit. on p. 26).

Dou, Q., C. Ouyang, C. Chen, H. Chen, and P.-A. Heng (2018). "Unsupervised Cross-Modality Domain Adaptation of Convnets for Biomedical Image Segmentations with Adversarial Loss." In: *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 691–697. DOI: 10.24963/ijcai.2018/96 (cit. on p. 103).

Drake, R. L., A. Wayne Vogl, and A. W. M. Mitchell, eds. (2010). *Gray's anatomy for students*. 2. edition. Churchill Livingstone, Elsevier. ISBN: 978-0443069529 (cit. on p. 7).

Engelen, J. E. van and H. H. Hoos (2020). "A survey on semi-supervised learning." In: *Machine Learning* 109.2, pp. 373–440. DOI: 10.1007/s10994-019-05855-6 (cit. on p. 24).

Engelhardt, S., R. De Simone, P. M. Full, M. Karck, and I. Wolf (2018). "Improving surgical training phantoms by hyperrealism: deep unpaired image-to-image translation from real surgeries." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 747–755. DOI: 10.1007/978-3-030-00928-1_84 (cit. on p. 1).

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun (2017). "Dermatologist-level classification of skin cancer with deep neural networks." In: *Nature* 542.7639, pp. 115–118. DOI: 10.1038/nature21056 (cit. on p. 1).

Falkner, S., A. Klein, and F. Hutter (2018). "BOHB: Robust and efficient hyperparameter optimization at scale." In: pp. 1437–1446 (cit. on p. 39).

Fang, K. and W.-J. Li (2020). "DMNet: Difference minimization network for semi-supervised segmentation in medical images." In: *International Conference on Medical Image Computing and Computer-Assisted*

Intervention (MICCAI), pp. 532–541. DOI: 10.1007/978-3-030-59710-8_52 (cit. on p. 66).

Fedorov, A., R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, et al. (2012). "3D Slicer as an image computing platform for the Quantitative Imaging Network." In: *Journal of Magnetic Resonance Imaging* 30.9, pp. 1323–1341. DOI: 10.1016/j.mri.2012.05.001 (cit. on pp. 41, 75, 111).

Fedorov, A., S. Khallaghi, A. C. Sánchez, A. Lasso, S. Fels, K. Tuncali, E. S. Neubauer, T. Kapur, C. Zhang, W. Wells, P. L. Nguyen, P. Abolmaesumi, and C. Tempany (2015). "Open-source image registration for MRI-TRUS fusion-guided prostate interventions." In: *International Journal of Computer Assisted Radiology* 10.6, pp. 925–934. DOI: 10.1007/s11548-015-1180-7 (cit. on p. 30).

Fotedar, G., N. Tajbakhsh, S. Ananth, and X. Ding (2020). "Extreme Consistency: Overcoming Annotation Scarcity and Domain Shifts." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 699–709. DOI: 10.1007/978-3-030-59710-8_68 (cit. on pp. 67, 103).

Gal, Y. and Z. Ghahramani (2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In: *International Conference on Machine Learning (ICML)*, pp. 1050–1059 (cit. on pp. 18–20, 65, 73, 92, 126).

Gasser, T. (2015). *Basiswissen Urologie*. 6th ed. Springer. DOI: 10.1007/978-3-662-45131-1 (cit. on pp. 9, 10).

Ghafoorian, M., A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. G. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken, A. Fedorov, P. Abolmaesumi, B. Platel, and W. M. Wells (2017). "Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 10435, pp. 516–524. DOI: 10.1007/978-3-319-66179-7_59 (cit. on pp. 102, 105).

Ghavami, N., Y. Hu, E. Gibson, E. Bonmati, M. Emberton, C. M. Moore, and D. C. Barratt (2019). "Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration." In: *Medical Image Analysis* 58, p. 101558. DOI: 10.1016/j.media.2019.101558 (cit. on pp. 27, 28).

Ghose, S., A. Oliver, R. Martí, X. Lladó, J. C. Vilanova, J. Freixenet, J. Mitra, D. Sidibé, and F. Meriaudeau (2012). "A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images." In: *Computer Methods and Programs in Biomedicine* 108.1, pp. 262–287. DOI: 10.1016/j.cmpb.2012.04.006 (cit. on p. 32).

Ghosh, S. (2019). "Automated segmentation of prostate zones using deep semi-supervised learning." MA thesis. Otto-von-Guericke University Magdeburg (cit. on p. 59).

Gibson, E., F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt (2018a). "Automatic multi-organ segmentation on abdominal CT with dense v-networks." In: *IEEE Transactions on Medical Imaging* 37.8, pp. 1822–1834. DOI: 10.1109/TMI.2018.2806309 (cit. on pp. 98, 100).

Gibson, E., Y. Hu, N. Ghavami, H. U. Ahmed, C. Moore, M. Emberton, H. J. Huisman, and D. C. Barratt (2018b). "Inter-site Variability in Prostate Segmentation Accuracy Using Deep Learning." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 506–514. DOI: 10.1007/978-3-030-00937-3_58 (cit. on pp. 101, 113).

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). "Generative adversarial nets." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680 (cit. on p. 102).

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio (2016). *Deep learning*. http://www.deeplearningbook.org. MIT Press. ISBN: 0262035618 (cit. on pp. 13, 37, 99, 107).

Grall, A., A. Hamidinekoo, P. Malcolm, and R. Zwiggelaar (2019). "Using a conditional generative adversarial network (cGAN) for prostate segmentation." In: *Conference on Medical Image Understanding and Analysis*, pp. 15–25. DOI: 10.1007/978-3-030-39343-4_2 (cit. on p. 33).

Greer, M. D., J. H. Shih, T. Barrett, S. Bednarova, I. Kabakus, Y. M. Law, H. Shebel, M. J. Merino, B. J. Wood, P. A. Pinto, P. L. Choyke, and B. Turkbey (2018). "All over the map: An interobserver agreement study of tumor location based on the PI-RADSv2 sector map." In: *Journal of Magnetic Resonance Imaging* 48.2, pp. 482–490. DOI: 10.3389/fonc.2019.00826 (cit. on p. 60).

Greer, P., J. Martin, M. Sidhom, P. Hunter, P. Pichler, J. H. Choi, L. Best, J. Smart, T. Young, M. Jameson, et al. (2019). "A multi-center prospective study for implementation of an MRI-only prostate treatment planning workflow." In: *Frontiers in Oncology* 9, p. 826 (cit. on p. 30).

Guan, H. and M. Liu (2021). "Domain Adaptation for Medical Image Analysis: A Survey." In: *CoRR*. URL: https://arxiv.org/abs/2102.09508 (cit. on pp. 100–102).

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger (2017). "On calibration of modern neural networks." In: *International Conference on Machine Learning (ICML)*, pp. 1321–1330 (cit. on p. 18).

Guo, Y., Y. Gao, and D. Shen (2016). "Deformable MR Prostate Segmentation via Deep Feature Learning and Sparse Patch Matching."

In: *IEEE Transactions on Medical Imaging* 35.4, pp. 1077–1089. DOI: 10.1109/TMI.2015.2508280 (cit. on p. 32).

Hambarde, P., S. N. Talbar, N. Sable, A. Mahajan, S. S. Chavan, and M. Thakur (2019). "Radiomics for peripheral zone and intra-prostatic urethra segmentation in MR imaging." In: *Biomedical Signal Processing and Control* 51, pp. 19–29. DOI: 10.1016/j.bspc.2019.01.024 (cit. on p. 105).

Hang, W., W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin (2020). "Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 562–571. DOI: 10.1007/978-3-030-59710-8_55 (cit. on p. 67).

Haskins, G., U. Kruger, and P. Yan (2020). "Deep learning in medical image registration: a survey." In: *Machine Vision and Applications* 31.1, pp. 1–18. DOI: 10.1007/s00138-020-01060-x (cit. on p. 56).

Hassanzadeh, T., L. G. C. Hamey, and K. Ho-Shon (2019). "Convolutional Neural Networks for Prostate Magnetic Resonance Image Segmentation." In: *IEEE Access* 7, pp. 36748–36760. DOI: 10.1109/ACCESS.2019.2903284 (cit. on pp. 33, 34).

Hautmann, R. and J. E. Gschwend (2014). *Urologie*. 5th ed. Springer. DOI: 10.1007/978-3-642-34319-3 (cit. on pp. 5, 6, 9, 10, 96).

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep Residual Learning for Image Recognition." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on pp. 17, 18, 34).

Herman, G. T., J. Zheng, and C. A. Bucholtz (1992). "Shape-based interpolation." In: *IEEE Computer Graphics and Applications* 12.3, pp. 69–79. DOI: 10.1109/38.135915 (cit. on pp. 41, 76).

Hesse, L. S., G. Kuling, M. Veta, and A. L. Martel (2020). "Intensity augmentation to improve generalizability of breast segmentation across different MRI scan protocols." In: *IEEE Transactions on Biomedical Engineering* 68.3, pp. 759–770. DOI: 10.1109/TBME.2020.3016602 (cit. on pp. 103, 134).

Hinton, G. (2012). *Lecture 6.5 - rmsprop: Divide the gradient by a running average of its recent magnitude*. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (cit. on p. 116).

Hosny, A., M. Schwier, C. Berger, E. P. Örnek, M. Turan, P. Tran, L. Weninger, F. Isensee, K. Maier-Hein, R. McKinley, M. T. Lu, U. Hoffmann, B. Menze, S. Bakas, A. Fedorov, and H. Aerts (2019). "ModelHub.AI: Dissemination Platform for Deep Learning Models." In: *CoRR*. URL: http://arxiv.org/abs/1911.13218 (cit. on p. 98).

Hu, X., D. Zeng, X. Xu, and Y. Shi (2021). "Semi-supervised Contrastive Learning for Label-Efficient Medical Image Segmentation." In: *International Conference on Medical Image Computing and Computer-*

Assisted Intervention (MICCAI), pp. 481–490. DOI: 10.1007/978-3-030-87196-3_45 (cit. on p. 65).

Huang, G., Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger (2017a). "Snapshot Ensembles: Train 1, get M for free." In: *CoRR*. URL: http://arxiv.org/abs/1704.00109 (cit. on p. 126).

Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger (2017b). "Densely connected convolutional networks." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708. DOI: https://doi.org/10.1109/CVPR.2017.243 (cit. on p. 17).

Huang, H., N. Zhou, L. Lin, H. Hu, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong (2021). "3D Graph-S 2 Net: Shape-Aware Self-ensembling Network for Semi-supervised Segmentation with Bilateral Graph Convolution." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 416–427. DOI: 10.1007/978-3-030-87196-3_39 (cit. on p. 65).

Hüllermeier, E. and W. Waegeman (2021). "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." In: *Machine Learning* 110.3, pp. 457–506. DOI: 10.1007/s10994-021-05946-3 (cit. on pp. 19, 21).

Huo, Y., Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman (2018). "Synseg-net: Synthetic segmentation without target modality ground truth." In: *IEEE Transactions on Medical Imaging* 38.4, pp. 1016–1025. DOI: 10.1109/TMI.2018.2876633 (cit. on p. 103).

Hutchinson, R. and Y. Lotan (2017). "Cost consideration in utilization of multiparametric magnetic resonance imaging in prostate cancer." In: *Translational Andrology and Urology* 6.3, p. 345. DOI: 10.21037/tau.2017.01.13 (cit. on p. 12).

Inoue, S., H. Shiina, T. Hiraoka, Y. Mitsui, M. Sumura, S. Urakami, and M. Igawa (2009). "Retrospective analysis of the distance between the neurovascular bundle and prostate cancer foci in radical prostatectomy specimens: its clinical implication in nerve-sparing surgery." In: *British Journal of Urology International* 104.8, pp. 1085–1090. DOI: 10.1111/j.1464-410X.2009.08592.x (cit. on p. 97).

Ioffe, S. and C. Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In: *International Conference on Machine Learning (ICML)*, pp. 448–456 (cit. on p. 39).

Isensee, F., P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein (2021). "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." In: *Nature methods* 18.2, pp. 203–211. DOI: 10.1038/s41592-020-01008-z (cit. on pp. 18, 27, 33, 34, 51, 56, 133).

Jacob, S. (2008). "Chapter 4 - Abdomen." In: *Human Anatomy*. Churchill Livingstone, pp. 71–123. DOI: 10.1016/B978-0-443-10373-5.50007-5 (cit. on p. 6).

Jia, H., Y. Xia, W. Cai, M. Fulham, and D. D. Feng (2017). "Prostate segmentation in MR images using ensemble deep convolutional neural networks." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 762–765. DOI: 10.1109/ISBI.2017.7950630 (cit. on pp. 32, 33).

Jia, H., Y. Song, H. Huang, W. Cai, and Y. Xia (2019). "HD-Net: Hybrid Discriminative Network for Prostate Segmentation in MR Images." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 110–118. DOI: 10.1007/978-3-030-32245-8_13 (cit. on pp. 33, 35, 134).

Jia, H., Y. Xia, Y. Song, D. Zhang, H. Huang, Y. Zhang, and W. Cai (2020). "3D APA-Net: 3D Adversarial Pyramid Anisotropic Convolutional Network for Prostate Segmentation in MR Images." In: *IEEE Transactions on Medical Imaging* 39.2, pp. 447–457. DOI: 10.1109/TMI.2019.2928056 (cit. on pp. 18, 33–35).

Kamnitsas, K., C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker (2017). "Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks." In: *Information Processing in Medical Imaging (IPMI)*. Vol. 10265, pp. 597–609. DOI: 10.1007/978-3-319-59050-9_47 (cit. on p. 103).

Karani, N., K. Chaitanya, C. Baumgartner, and E. Konukoglu (2018). "A lifelong learning approach to brain MR segmentation across scanners and protocols." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 476–484. DOI: 10.1007/978-3-030-00928-1_54 (cit. on pp. 102, 105, 135).

Karani, N., E. Erdil, K. Chaitanya, and E. Konukoglu (2021). "Test-time adaptable neural networks for robust medical image segmentation." In: *Medical Image Analysis* 68, p. 101907. DOI: 10.1016/j.media.2020.101907 (cit. on pp. 104, 106).

Karimi, D., G. Samei, C. Kesch, G. Nir, and S. E. Salcudean (2018). "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models." In: *International Journal of Computer Assisted Radiology and Surgery* 13.8, pp. 1211–1219. DOI: https://doi.org/10.1007/s11548-018-1785-8 (cit. on pp. 33, 34).

Karimi, D., S. K. Warfield, and A. Gholipour (2021). "Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations." In: *Artificial Intelligence in Medicine* 116, p. 102078. DOI: 10.1016/j.artmed.2021.102078 (cit. on pp. 102, 105).

Kaur, B., P. Lemaître, R. Mehta, N. M. Sepahvand, D. Precup, D. Arnold, and T. Arbel (2019). "Improving pathological structure segmentation via transfer learning across diseases." In: *Domain Adaptation and Representation Transfer and Medical Image Learning with*

Less Labels and Imperfect Data (MICCAI Workshop), pp. 90–98. DOI: 10.1007/978-3-030-33391-1_11 (cit. on pp. 102, 105).

Kendall, A., V. Badrinarayanan, and R. Cipolla (2017). "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding." In: *British Machine Vision Conference (BMVC)*. DOI: 10.5244/C.31.57 (cit. on pp. 21, 104).

Kendall, A. and Y. Gal (2017). "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *Advances in Neural Information Processing Systems (NeurIPS)* 30, pp. 5574–5584 (cit. on p. 19).

Kingma, D. and J. Ba (2015). "Adam: A method for stochastic optimization." In: *International Conference on Learning Representations (ICLR)* (cit. on pp. 39, 77).

Kingma, D. P. and M. Welling (2013). "Auto-encoding variational bayes." In: *CoRR*. URL: https://arxiv.org/abs/1312.6114 (cit. on p. 134).

Kohl, S. A. A., B. Romera-Paredes, C. Meyer, J. D. Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. A. Eslami, D. J. Rezende, and O. Ronneberger (2018). "A Probabilistic U-Net for Segmentation of Ambiguous Images." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6965–6975 (cit. on pp. 20, 103, 134, 135).

Kovashka, A., O. Russakovsky, L. Fei-Fei, and K. Grauman (2016). *Crowdsourcing in computer vision*. Now Foundations and Trends. ISBN: 978-1-68083-212-9. DOI: 10.1561/0600000071 (cit. on p. 2).

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks." In: *Advances in Neural Information Processing Systems (NeurIPS)* 25, pp. 1097–1105. DOI: 10.1109/CVPR.2009.5206848 (cit. on p. 13).

Kwak, J. T., S. Sankineni, S. Xu, B. Turkbey, P. L. Choyke, P. A. Pinto, M. Merino, and B. J. Wood (2016). "Correlation of magnetic resonance imaging with digital histopathology in prostate." In: *International Journal of Computer Assisted Radiology and Surgery* 11.4, pp. 657–666. DOI: 10.1007/s11548-015-1287-x (cit. on pp. 2, 30, 62).

Laine, S. and T. Aila (2017). "Temporal ensembling for semi-supervised learning." In: *International Conference on Learning Representation (ICLR)* (cit. on pp. 24, 25, 72, 73, 78, 87, 130, 137).

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6402–6413 (cit. on pp. 19–21).

Landman, B., Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein (2015). "Multi-atlas Labeling beyond the Cranial Vault–Workshop and Challenge." In: DOI: 10.7303/syn3193805 (cit. on p. 113).

Laukamp, K. R., F. Thiele, G. Shakirin, D. Zopfs, A. Faymonville, M. Timmer, D. Maintz, M. Perkuhn, and J. Borggrefe (2019). "Fully automated detection and segmentation of meningiomas using deep

learning on routine multiparametric MRI." In: *European Radiology* 29.1, pp. 124–132. DOI: 10.1007/s00330-018-5595-8 (cit. on p. 1).

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989). "Backpropagation Applied to Handwritten Zip Code Recognition." In: *Neural Computation* 1.4, pp. 541–551. DOI: 10.1162/neco.1989.1.4.541 (cit. on p. 13).

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791 (cit. on p. 13).

Lee, S. E., S.-S. Byun, H. J. Lee, S. H. Song, I. H. Chang, Y. J. Kim, M. C. Gill, and S. K. Hong (2006). "Impact of variations in prostatic apex shape on early recovery of urinary continence after radical retropubic prostatectomy." In: *Urology* 68.1, pp. 137–141. DOI: 10.1016/j.urology.2006.01.021 (cit. on pp. 6, 55).

Leest, M. van der, E. Cornel, B. Israël, R. Hendriks, A. R. Padhani, M. Hoogenboom, P. Zamecnik, D. Bakker, A. Y. Setiasti, J. Veltman, et al. (2019). "Head-to-head comparison of transrectal ultrasound-guided prostate biopsy versus multiparametric prostate resonance imaging with subsequent magnetic resonance-guided biopsy in biopsy-naive men with elevated prostate-specific antigen: a large prospective multicenter clinical study." In: *European Urology* 75.4, pp. 570–578. DOI: 10.1016/j.eururo.2018.11.023 (cit. on p. 1).

Li, K., S. Wang, L. Yu, and P. A. Heng (2021a). "Dual-Teacher++: Exploiting Intra-Domain and Inter-Domain Knowledge With Reliable Transfer for Cardiac Segmentation." In: *IEEE Transactions on Medical Imaging* 40.10, pp. 2771–2782. DOI: 10.1109/TMI.2020.3038828 (cit. on pp. 104, 125).

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar (2017). "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization." In: *Journal of Machine Learning Research* 18.1, pp. 6765–6816 (cit. on p. 39).

Li, S., Z. Zhao, K. Xu, Z. Zeng, and C. Guan (2021b). "Hierarchical Consistency Regularized Mean Teacher for Semi-supervised 3D Left Atrium Segmentation." In: *CoRR*. URL: https://arxiv.org/abs/2105.10369 (cit. on pp. 67, 91).

Li, X., L. Yu, H. Chen, C. Fu, and P. Heng (2020a). "Transformation Consistent Self-ensembling Model for Semi-supervised Medical Image Segmentation." In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2, pp. 523–534. DOI: 10.1109/TNNLS.2020.2995319 (cit. on p. 67).

Li, Y., J. Chen, X. Xie, K. Ma, and Y. Zheng (2020b). "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 614–623. DOI: 10.1007/978-3-030-59710-8_60 (cit. on pp. 19, 66).

Liao, S., Y. Gao, A. Oto, and D. Shen (2013). "Representation learning: a unified deep learning framework for automatic prostate MR segmentation." In: pp. 254–261. DOI: 10.1007/978-3-642-40763-5_32 (cit. on p. 32).

Litjens, G., O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman (2014a). "Computer-Aided Detection of Prostate Cancer in MRI." In: *IEEE Transactions on Medical Imaging* 33.5, pp. 1083–1092. DOI: 10.1109/TMI.2014.2303821 (cit. on pp. 40, 75, 131).

Litjens, G., O. Debats, W. van de Ven, N. Karssemeijer, and H. Huisman (2012). "A pattern recognition approach to zonal segmentation of the prostate on MRI." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 413–420. DOI: 10.1007/978-3-642-33418-4_51 (cit. on pp. 63, 64).

Litjens, G., J. Futterer, and H. Huisman (2015). "Data From Prostate-3T [Data set]." In: *The Cancer Imaging Archive*. DOI: 10.7937/K9/TCIA.2015.QJTV5IL5 (cit. on pp. 111, 112).

Litjens, G., R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, et al. (2014b). "Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge." In: *Medical Image Analysis* 18.2, pp. 359–373. DOI: 10.1016/j.media.2013.12.0029 (cit. on pp. 26, 27, 40, 53).

Litjens, G., R. Toth, and W. van de Ven *et al.* (2014c). "Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge." In: *Medical Image Analysis* 18.2, pp. 359–373. DOI: 10.1016/j.media.2013.12.002 (cit. on pp. 31, 37, 43, 97).

Litjens, G., O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman (2017a). "ProstateX Challenge data [Dataset]." In: *The Cancer Imaging Archive*. DOI: 10.7937/K9TCIA.2017.MURS5CL (cit. on pp. 40, 75).

Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez (2017b). "A survey on deep learning in medical image analysis." In: *Medical Image Analysis* 42, pp. 60–88. DOI: 10.1016/j.media.2017.07.005 (cit. on p. 1).

Litwin, M. and H.-J. Tan (2017). "The diagnosis and treatment of prostate cancer: a review." In: *Jama* 317.24, pp. 2532–2542. DOI: 10.1001/jama.2017.7248 (cit. on pp. 1, 9, 10).

Liu, Q., M. Fu, H. Jiang, and X. Gong (2020a). "Densely Dilated Spatial Pooling Convolutional Network using benign loss functions for imbalanced volumetric prostate segmentation." In: *Current Bioinformatics* 15.7, pp. 788–799. DOI: 10.2174/1574893615666200127124145 (cit. on p. 34).

Liu, Q., Q. Dou, and P.-A. Heng (2020b). "Shape-Aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains." In: *International Conference on Medical Image Computing and Computer-*

Assisted Intervention (MICCAI), pp. 475–485. DOI: 10.1007/978-3-030-59713-9_46 (cit. on pp. 56, 104, 132, 134).

Liu, Q., Q. Dou, L. Yu, and P. A. Heng (2020c). "MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data." In: *IEEE Transactions on Medical Imaging* 39.9, pp. 2713–2724. DOI: 10.1109/TMI.2020.2974574 (cit. on p. 100).

Liu, Y., G. Yang, M. Hosseiny, A. Azadikhah, S. A. Mirak, Q. Miao, S. S. Raman, and K. Sung (2020d). "Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation." In: *IEEE Access* 8, pp. 151817–151828. DOI: 10.1109/ACCESS.2020.3017168 (cit. on pp. 64, 65).

Liu, Y., G. Yang, S. A. Mirak, M. Hosseiny, A. Azadikhah, X. Zhong, R. E. Reiter, Y. Lee, S. S. Raman, and K. Sung (2019). "Automatic Prostate Zonal Segmentation Using Fully Convolutional Network With Feature Pyramid Attention." In: *IEEE Access* 7, pp. 163626–163632. DOI: 10.1109/ACCESS.2019.2952534 (cit. on pp. 63, 64).

Loeb, S., M. A. Bjurlin, J. Nicholson, T. L. Tammela, D. F. Penson, H. B. Carter, P. Carroll, and R. Etzioni (2014). "Overdiagnosis and overtreatment of prostate cancer." In: *European Urology* 65.6, pp. 1046–1055. DOI: 10.1016/j.eururo.2013.12.062 (cit. on p. 1).

Long, J., E. Shelhamer, and T. Darrell (2015). "Fully convolutional networks for semantic segmentation." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965 (cit. on pp. 14, 15).

Lozoya, R. C., A. Iannessi, J. Brag, S. Patriti, and E. Oubel (2018). "Assessing the relevance of multi-planar MRI acquisitions for prostate segmentation using deep learning techniques." In: *Proceedings of SPIE - The International Society for Optical Engineering*, p. 45. DOI: 10.1117/12.2293514 (cit. on pp. 33, 35, 36, 53).

Lu, Y., K. Zheng, W. Li, Y. Wang, A. P. Harrison, C. Lin, S. Wang, J. Xiao, L. Lu, C.-F. Kuo, et al. (2021). "Contour transformer network for one-shot segmentation of anatomical structures." In: *IEEE Transactions on Medical Imaging* 40.10, pp. 2672–2684. DOI: 10.1109/tmi.2020.3043375 (cit. on p. 65).

Luo, X., W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang (2021). "Efficient Semi-Supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 318–329. DOI: 10.1007/978-3-030-87196-3_30 (cit. on p. 67).

Ma, J., J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel (2021). "Loss odyssey in medical image segmentation." In: *Medical Image Analysis* 71, p. 102035. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102035 (cit. on p. 134).

Maaten, L. van der and G. Hinton (2008). "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9.11, pp. 2579–2605 (cit. on p. 97).

Madesta, F., R. Schmitz, T. Rösch, and R. Werner (2020). "Widening the Focus: Biomedical Image Segmentation Challenges and the Underestimated Role of Patch Sampling and Inference Strategies." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 289–298. DOI: 10.1007/978-3-030-59719-1_29 (cit. on pp. 68, 131).

Makni, N., A. Iancu, O. Colot, P. Puech, S. Mordon, and N. Betrouni (2011). "Zonal segmentation of prostate using multispectral magnetic resonance images." In: *Medical Physics* 38.11, pp. 6093–6105. DOI: 10.1118/1.3651610 (cit. on pp. 63, 64).

Mårtensson, G. et al. (2020). "The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study." In: *Medical Image Analysis* 66, p. 101714. DOI: https://doi.org/10.1016/j.media.2020.101714 (cit. on pp. 100, 101).

Massey Jr, F. J. (1951). "The Kolmogorov-Smirnov test for goodness of fit." In: *Journal of the American statistical Association* 46.253, pp. 68–78. DOI: 10.2307/2280095 (cit. on p. 28).

McNeal, J. E. (1981). "The zonal anatomy of the prostate." In: *Prostate* 2.1, pp. 35–49. DOI: 10.1002/pros.2990020105 (cit. on p. 6).

Mehrtash, A., A. Sedghi, M. Ghafoorian, M. Taghipour, C. M. Tempany, W. M. Wells, T. Kapur, P. Mousavi, P. Abolmaesumi, and A. Fedorov (2017a). "Classification of Clinical Significance of MRI Prostate Findings Using 3D Convolutional Neural Networks." In: *Proceedings of SPIE - The International Society for Optical Engineering* 10134 (cit. on p. 60).

Mehrtash, A., W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur (2020). "Confidence calibration and predictive Uncertainty Estimation for Deep Medical Image Segmentation." In: *IEEE Transactions on Medical Imaging* 39.12, pp. 3868–3878 (cit. on pp. 2, 20, 21, 92, 106, 109, 135).

Mehrtash, A., M. Pesteie, J. Hetherington, P. A. Behringer, T. Kapur, W. M. Wells III, R. Rohling, A. Fedorov, and P. Abolmaesumi (2017b). "DeepInfer: Open-Source Deep Learning Deployment Toolkit for Image-Guided Therapy." In: *Proceedings of SPIE - The International Society for Optical Engineering*. Vol. 10135. DOI: 10.1117/12.2256011 (cit. on p. 98).

Meyer, A., A. Mehrtash, M. Rak, D. Schindele, M. Schostak, C. Tempany, T. Kapur, P. Abolmaesumi, A. Fedorov, and C. Hansen (2018). "Automatic High Resolution Segmentation of the Prostate from Multi-Planar MRI." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 177–181 (cit. on p. 36).

Meyer, A., G. Chlebus, M. Rak, D. Schindele, M. Schostak, B. van Ginneken, A. Schenk, H. Meine, H. K. Hahn, A. Schreiber, and

C. Hansen (2021a). "Anisotropic 3D Multi-Stream CNN for Accurate Prostate Segmentation from Multi-Planar MRI." In: *Computer Methods and Programs in Biomedicine* 200, p. 105821. DOI: https://doi.org/10.1016/j.cmpb.2020.105821 (cit. on pp. 33, 36).

Meyer, A., S. Ghosh, D. Schindele, M. Schostak, S. Stober, C. Hansen, and M. Rak (2021b). "Uncertainty-aware temporal self-learning (UATS): Semi-supervised learning for segmentation of prostate zones and beyond." In: *Artificial Intelligence in Medicine* 116, p. 102073. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2021.102073 (cit. on pp. 64, 105).

Meyer, A., M. Rak, D. Schindele, S. Blaschke, M. Schostak, A. Fedorov, and C. Hansen (2019). "Towards Patient-Individual PI-Rads v2 Sector Map: CNN for Automatic Segmentation of Prostatic Zones From T2-Weighted MRI." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 696–700. DOI: 10.1109/ISBI.2019.8759572 (cit. on pp. 82, 105).

Meyer, A., D. Schindele, D. F. von Reibnitz, M. Rak, M. Schostak, and C. Hansen (2020). "PROSTATEx zone segmentations [Dataset]." In: *The Cancer Imaging Archive*. DOI: 10.7937/TCIA.2019.DEG7ZG1U (cit. on p. 131).

Milletari, F., N. Navab, and S.-A. Ahmadi (2016). "V-Net: Fully convolutional neural networks for volumetric medical image segmentation." In: *International Conference on 3D Vision (3DV)*, pp. 565–571. DOI: 10.1109/3DV.2016.79 (cit. on pp. 17, 33, 34, 39).

Miyato, T., S.-i. Maeda, M. Koyama, and S. Ishii (2018). "Virtual adversarial training: a regularization method for supervised and semi-supervised learning." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8, pp. 1979–1993. DOI: 10.1109/TPAMI.2018.2858821 (cit. on pp. 26, 66).

Mohareri, O., G. Nir, J. Lobo, R. Savdie, P. Black, and S. Salcudean (2015). "A system for MR-ultrasound guidance during robot-assisted laparoscopic radical prostatectomy." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, pp. 497–504. DOI: 10.1007/978-3-319-24553-9_61 (cit. on p. 30).

Montagne, S., D. Hamzaoui, A. Allera, M. Ezziane, A. Luzurier, R. Quint, M. Kalai, N. Ayache, H. Delingette, and R. Renard-Penna (2021). "Challenge of prostate MRI segmentation on T2-weighted images: inter-observer variability and impact of prostate morphology." In: *Insights into imaging* 12.1, pp. 1–12. DOI: 10.1186/s13244-021-01010-9 (cit. on pp. 131, 132).

Mooij, G., I. Bagulho, and H. Huisman (2018). "Automatic segmentation of prostate zones." In: *CoRR*. URL: https://arxiv.org/abs/1806.07146 (cit. on pp. 63, 64).

Mungovan, S. F., J. S. Sandhu, O. Akin, N. A. Smart, P. L. Graham, and M. I. Patel (Mar. 2017). "Preoperative Membranous Urethral Length

Measurement and Continence Recovery Following Radical Prostatectomy: A Systematic Review and Meta-analysis." In: *European Urology* 71.3, pp. 368–378. DOI: 10.1016/j.eururo.2016.06.023 (cit. on pp. 2, 96).

Neal, R. M. (2008). "Chapter 5: MCMC Using Hamiltonian Dynamics." In: *Handbook of Markov Chain Monte Carlo*. Chapman and Hall. ISBN: 9780429138508. DOI: 10.1201/b10905 (cit. on p. 19).

Nguyen, L. N., L. Head, K. Witiuk, N. Punjani, R. Mallick, S. Cnossen, D. A. Fergusson, I. Cagiannos, L. T. Lavallée, C. Morash, and R. H. Breau (Oct. 2017). "The Risks and Benefits of Cavernous Neurovascular Bundle Sparing during Radical Prostatectomy: A Systematic Review and Meta-Analysis." In: *Journal of Urology* 198.4, pp. 760–769. DOI: 10.1016/j.juro.2017.02.3344 (cit. on pp. 2, 96).

Nie, D., Y. Gao, L. Wang, and D. Shen (2018). "ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 11073, pp. 370–378. DOI: 10.1007/978-3-030-00937-3_43 (cit. on pp. 19, 20, 65, 68, 87).

Nielsen, M. A. (2015). *Neural networks and deep learning*. Vol. 25. Determination Press (cit. on p. 17).

Oliver, A., A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow (2018). "Realistic evaluation of deep semi-supervised learning algorithms." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3235–3246 (cit. on p. 92).

Ouali, Y., C. Hudelot, and M. Tami (2020). "An Overview of Deep Semi-Supervised Learning." In: *CoRR*. URL: https://arxiv.org/abs/2006.05278 (cit. on pp. 22–24).

Pan, H., Y. Feng, Q. Chen, C. Meyer, and X. Feng (2019). "Prostate segmentation from 3D MRI using a two-stage model and variable-input based uncertainty measure." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 468–471. DOI: 10.1109/ISBI.2019.8759300 (cit. on pp. 18, 33, 34).

Pan, S. J. and Q. Yang (2009). "A survey on transfer learning." In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191 (cit. on p. 99).

Parisi, G. I., R. Kemker, J. L. Part, C. Kanan, and S. Wermter (2019). "Continual lifelong learning with neural networks: A review." In: *Neural Networks* 113, pp. 54–71. DOI: 10.1016/j.neunet.2019.01.012 (cit. on p. 135).

Peng, J. and Y. Wang (2021). "Medical image segmentation with limited supervision: A review of deep network models." In: *IEEE Access* 9, pp. 36827–36851. DOI: 10.1109/ACCESS.2021.3062380 (cit. on p. 65).

Peng, J., G. Estrada, M. Pedersoli, and C. Desrosiers (2020). "Deep co-training for semi-supervised image segmentation." In: *Pattern*

Recognition 107, p. 107269. DOI: 10.1016/j.patcog.2020.107269 (cit. on p. 66).

Perone, C. S., P. Ballester, R. C. Barros, and J. Cohen-Adad (2019). "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling." In: *NeuroImage* 194, pp. 1–11. DOI: 10.1016/j.neuroimage.2019.03.026 (cit. on p. 103).

Porpiglia, F., R. Bertolo, E. Checcucci, D. Amparore, R. Autorino, P. Dasgupta, P. Wiklund, A. Tewari, E. Liatsikos, and C. Fiori (2018). "Development and validation of 3D printed virtual models for robot-assisted radical prostatectomy and partial nephrectomy: urologists' and patients' perception." In: *World Journal of Urology* 36.2, pp. 201–207. DOI: 10.1007/s00345-017-2126-1 (cit. on p. 30).

Qin, X., Y. Zhu, W. Wang, S. Gui, B. Zheng, and P. Wang (2020). "3D multi-scale discriminative network with multi-directional edge loss for prostate zonal segmentation in bi-parametric MR images." In: *Neurocomputing* 418, pp. 148–161. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.07.116 (cit. on pp. 27, 64, 65, 134).

Qiu, W., J. Yuan, E. Ukwatta, Y. Sun, M. Rajchl, and A. Fenster (May 2014). "Dual optimization based prostate zonal segmentation in 3D MR images." In: *Medical Image Analysis* 18.4, pp. 660–673. DOI: 10.1016/j.media.2014.02.009 (cit. on pp. 63, 64).

Quiñonero-Candela, J., M. Sugiyama, N. D. Lawrence, and A. Schwaighofer (2009). *Dataset shift in machine learning*. MIT Press. ISBN: 9780262170055 (cit. on p. 98).

Rawla, P. (2019). "Epidemiology of prostate cancer." In: *World journal of oncology* 10.2, pp. 63–89. DOI: 10.14740/wjon1191 (cit. on p. 9).

Rieke, N., J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, and et al. (2020). "The future of digital health with federated learning." In: *NPJ Digital Medicine* 3.1, pp. 1–7. DOI: 10.1038/s41746-020-00323-1 (cit. on p. 98).

Riepe, T., M. Hossainzadeh, P. Brand, and H. Huisman (2020). "Anisotropic multi-planar automatic prostate segmentation." In: *Abstract at International Society for Magnetic Resonance in Medicine (ISMRM)*. URL: http://indexsmart.mirasmart.com/ISMRM2020/PDFfiles/3518.html (cit. on pp. 33, 36).

Robinson, R., O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, et al. (2018). "Real-Time Prediction of Segmentation Quality." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 11073, pp. 578–585. DOI: 10.1007/978-3-030-00937-3_66 (cit. on p. 92).

Roels, J., J. Hennies, Y. Saeys, W. Philips, and A. Kreshuk (2019). "Domain adaptive segmentation in volume electron microscopy imaging." In: *Proceedings of the IEEE International Symposium on Biomedical*

Imaging (ISBI), pp. 1519–1522. DOI: 10.1109/ISBI.2018.8363602 (cit. on p. 104).

Ronneberger, O., P. Fischer, and T. Brox (2015). "U-Net: convolutional networks for biomedical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28 (cit. on pp. 15–17).

Roth, H. R., A. Farag, E. B. Turkbey, L. Lu, J. Liu, and R. M. Summers (2016). "Data from pancreas-CT." In: *The Cancer Imaging Archive*. DOI: 10.7937/K9/TCIA.2016.tNB1kqBU (cit. on p. 113).

Roth, H. R., L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers (2015). "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 556–564. DOI: 10.1007/978-3-319-24553-9_68 (cit. on p. 113).

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). "Learning representations by back-propagating errors." In: *nature* 323.6088, pp. 533–536. DOI: 10.1038/323533a0 (cit. on p. 14).

Rundo, L., C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi, M. C. Gilardi, S. Vitabile, G. Mauri, H. Nakayama, and P. Cazzaniga (2019). "USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets." In: *Neurocomputing* 365, pp. 31–43. DOI: 10.1016/j.neucom.2019.07.006 (cit. on pp. 63, 64).

Saar, M., J. Linxweiler, A. Borkowetz, S. Fussek, K. Urbanova, L. Bellut, G. Kristiansen, and B. Wullich (2020). "Current Role of Multiparametric MRI and MRI Targeted Biopsies for Prostate Cancer Diagnosis in Germany: A Nationwide Survey." In: *Urologia Internationalis* 104.9-10, pp. 731–740. DOI: 10.1159/000508755 (cit. on p. 12).

Saha, A., M. Hosseinzadeh, and H. Huisman (2021). "End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction." In: *Medical Image Analysis* 73, p. 102155. DOI: 10.1016/j.media.2021.102155 (cit. on p. 135).

Schindele, D., A. Meyer, D. F. von Reibnitz, V. Kiesswetter, M. Schostak, M. Rak, and C. Hansen (2020). "High Resolution Prostate Segmentations for the ProstateX-Challenge [Dataset]." In: *The Cancer Imaging Archive*. DOI: 10.7937/TCIA.2019.DEG7ZG1U (cit. on p. 131).

Scudder, H. (1965). "Probability of error of some adaptive pattern-recognition machines." In: *IEEE Transactions on Information Theory* 11.3, pp. 363–371. DOI: 10.1109/TIT.1965.1053799 (cit. on pp. 23, 130).

Sedai, S., B. Antony, R. Rai, K. Jones, H. Ishikawa, J. Schuman, W. Gadi, and R. Garnavi (2019). "Uncertainty guided semi-supervised

segmentation of retinal layers in OCT images." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 282–290. DOI: 10.1007/978-3-030-32239-7_32 (cit. on pp. 19, 65).

Seok, J., S. Yoon, C. H. Ryu, S.-k. Kim, J. Ryu, and Y.-S. Jung (2021). "A Personalized 3D-Printed Model for Obtaining Informed Consent Process for Thyroid Surgery: A Randomized Clinical Study Using a Deep Learning Approach with Mesh-Type 3D Modeling." In: *Journal of Personalized Medicine* 11.6, p. 574. DOI: 10.3390/jpm11060574 (cit. on p. 1).

Shah, V., T. Pohida, B. Turkbey, H. Mani, M. Merino, P. A. Pinto, P. Choyke, and M. Bernardo (2009). "A method for correlating in vivo prostate magnetic resonance imaging and histopathology using individualized magnetic resonance-based molds." In: *Review of Scientific Instruments* 80.10, p. 104301. DOI: 10.1063/1.3242697 (cit. on p. 30).

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas (2015). "Taking the human out of the loop: A review of Bayesian optimization." In: *Proceedings of the IEEE* 104.1, pp. 148–175. DOI: 10.1109/JPROC.2015.2494218 (cit. on p. 39).

Shamir, R. R., Y. Duchin, J. Kim, G. Sapiro, and N. Harel (2019). "Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations." In: *CoRR*. URL: https://arxiv.org/abs/1906.11031 (cit. on pp. 71–73).

Sheikh, R. and T. Schultz (2020). "Feature Preserving Smoothing Provides Simple and Effective Data Augmentation for Medical Image Segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 116–126. DOI: 10.1007/978-3-030-59710-8_12 (cit. on pp. 103, 134).

Shin, H.-C., H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers (2016). "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." In: *IEEE Transactions on Medical Imaging* 35.5, pp. 1285–1298. DOI: 10.1109/TMI.2016.2528162 (cit. on p. 102).

Siegel, R. L., K. D. Miller, and A. Jemal (Jan. 2020). "Cancer statistics, 2020." In: *CA: A Cancer Journal for Clinicians* 70.1, pp. 7–30. DOI: 10.3322/caac.21590 (cit. on pp. 1, 7).

Simonyan, K. and A. Zisserman (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *International Conference on Learning Representations (ICLR)* (cit. on p. 35).

Simpson, A. L., M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. (2019). "A large annotated medical image dataset for the development and evaluation of segmentation algorithms." In: *CoRR*. URL: https://arxiv.org/abs/1902.09063 (cit. on pp. 76, 113).

Singh, P., V. K. Verma, P. Mazumder, L. Carin, and P. Rai (2020). "Calibrating CNNs for Lifelong Learning." In: *Advances in Neural Information Processing Systems (NeurIPS)* 33, pp. 15579–15590 (cit. on p. 2).

Siversson, C., F. Nordström, T. Nilsson, T. Nyholm, J. Jonsson, A. Gunnlaugsson, and L. E. Olsson (2015). "MRI only prostate radiotherapy planning using the statistical decomposition algorithm." In: *Medical Physics* 42.10, pp. 6090–6097. DOI: 10.1118/1.4931417 (cit. on p. 30).

Sohn, K., H. Lee, and X. Yan (2015). "Learning structured output representation using deep conditional generative models." In: *Advances in neural information processing systems (NeurIPS)* 28, pp. 3483–3491 (cit. on p. 20).

Somford, D. M., J. J. Fütterer, T. Hambrock, and J. O. Barentsz (2008). "Diffusion and Perfusion MR Imaging of the Prostate." In: *Magnetic Resonance Imaging Clinics of North America* 16.4, pp. 685–695. DOI: 10.1016/j.mric.2008.07.002 (cit. on p. 11).

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting." In: *Journal of Machine Learning Research* 15.1, pp. 1929–1958 (cit. on pp. 20, 37).

Stabile, A., F. Giganti, A. B. Rosenkrantz, S. S. Taneja, G. Villeirs, I. S. Gill, C. Allen, M. Emberton, C. M. Moore, and V. Kasivisvanathan (2020). "Multiparametric MRI for prostate cancer diagnosis: current status and future directions." In: *Nature Reviews Urology* 17.1, pp. 41–61. DOI: 10.1038/s41585-019-0212-4 (cit. on p. 12).

Standring, S., N. Ananad, and H. Gray, eds. (2016). *Gray's anatomy: The anatomical basis of clinical practice*. 41. edition. Elsevier. ISBN: 978-0702052309 (cit. on pp. 5–8, 92).

Sun, J., Y. Zhang, J. Zhu, J. Wu, and Y. Kong (2021). "Semi-Supervised Medical Image Semantic Segmentation with Multi-scale Graph Cut Loss." In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 624–628 (cit. on p. 65).

Sun, Y., H. M. Reynolds, B. Parameswaran, D. Wraith, M. E. Finnegan, S. Williams, and A. Haworth (2019). "Multiparametric MRI and radiomics in prostate cancer: a review." In: *Australasian Physical & Engineering Sciences in Medicine* 42.1, pp. 3–25. DOI: 10.1007/s13246-019-00730-z (cit. on p. 30).

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2015). "Intriguing properties of neural networks." In: *International Conference on Learning Representations (ICLR)* (cit. on pp. 21, 97).

Tajbakhsh, N., J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and L. Jianming (2016). "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?"

In: *IEEE Transactions on Medical Imaging* 35.5, pp. 1299–1312. DOI: 10.1109/TMI.2016.2535302 (cit. on pp. 99, 102).

Tajbakhsh, N., Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding (2019). "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1251–1255 (cit. on p. 65).

Tajbakhsh, N., L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding (2020). "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation." In: *Medical Image Analysis* 63, p. 101693. DOI: https://doi.org/10.1016/j.media.2020.101693 (cit. on pp. 62, 65, 130).

Tarvainen, A. and H. Valpola (2017). "Mean Teachers are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1195–1204 (cit. on pp. 25, 66, 91).

To, M. N. N., D. Q. Vu, B. Turkbey, P. L. Choyke, and J. T. Kwak (2018). "Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging." In: *International Journal of Computer Assisted Radiology* 13.11, pp. 1687–1696. DOI: 10.1007/s11548-018-1841-4 (cit. on pp. 33, 34).

Toth, R., J. Ribault, J. Gentile, D. Sperling, and A. Madabhushi (2013). "Simultaneous Segmentation of Prostatic Zones Using active appearance models with Multiple Coupled levelsets." In: *Computer Vision and Image Understanding* 117.9, pp. 1051–1060. DOI: 10.1016/j.cviu.2012.11.013 (cit. on pp. 63, 64).

Tsai, Y.-H., W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker (2018). "Learning to Adapt Structured Output Space for Semantic Segmentation." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7472–7481. DOI: 10.1109/CVPR.2018.00780 (cit. on p. 103).

Tschandl, P., C. Rosendahl, and H. Kittler (2018). "The HAM10000 Dataset, a Large Collection of Multi-source Dermatoscopic Images of Common Pigmented Skin Lesions." In: *Scientific Data* 5.180161. DOI: 10.1038/sdata.2018.161 (cit. on p. 76).

Turkbey, B., A. B. Rosenkrantz, M. A. Haider, A. R. Padhani, G. Villeirs, K. J. Macura, C. M. Tempany, P. L. Choyke, F. Cornud, D. J. Margolis, et al. (2019). "Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2." In: *European Urology* 76.3, pp. 340–351. DOI: 10.1016/j.eururo.2019.02.033 (cit. on pp. 2, 3, 5, 7, 9–11, 30, 31, 60, 61, 97).

Umapathy, L., W. Unger, F. Shareef, H. Arif, D. R. Martín, M. I. Altbach, and A. Bilgin (2020). "A Cascaded Residual UNET for Fully Automated Segmentation of Prostate and Peripheral Zone in

T2-weighted 3D Fast Spin Echo Images." In: *CoRR*. URL: https://arxiv.org/abs/2012.13501 (cit. on pp. 33, 34).

Valindria, V. V., I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker (2018). "Domain adaptation for MRI organ segmentation using reverse classification accuracy." In: *International Conference on Medical Imaging with Deep Learning (MIDL)* (cit. on p. 102).

Valverde, S., M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, and X. Lladó (2019). "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks." In: *NeuroImage: Clinical* 21, p. 101638. DOI: 10.1016/j.nicl.2018.101638 (cit. on pp. 102, 105).

Van der Kwast, T. H. and M. J. Roobol (2013). "Defining the threshold for significant versus insignificant prostate cancer." In: *Nature Reviews Urology* 10.8, pp. 473–482. DOI: 10.1038/nrurol.2013.112 (cit. on p. 9).

Venkataramani, R., H. Ravishankar, and S. Anamandra (2019). "Towards continuous domain adaptation for medical imaging." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 443–446. DOI: 10.1109/ISBI.2019.8759268 (cit. on pp. 104, 105).

Venturini, L., A. T. Papageorghiou, J. A. Noble, and A. I. L. Namburete (2020). "Uncertainty estimates as data selection criteria to boost omni-supervised learning." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 689–698. DOI: 10.1007/978-3-030-59710-8_67 (cit. on p. 20).

Verma, S., P. L. Choyke, S. C. Eberhardt, A. Oto, C. M. Tempany, B. Turkbey, and A. B. Rosenkrantz (2017). "The current state of MR imaging–targeted biopsy techniques for detection of prostate cancer." In: *Radiology* 285.2, pp. 343–356. DOI: 10.1148/radiol.2017161684 (cit. on p. 1).

Wake, N., J. E. Nussbaum, M. I. Elias, C. V. Nikas, and M. A. Bjurlin (2020). "3D printing, augmented reality, and virtual reality for the assessment and management of kidney and prostate cancer: a systematic review." In: *Urology* 143, pp. 20–32. DOI: 10.1016/j.urology.2020.03.066 (cit. on pp. 2, 30, 96).

Wang, B., Y. Lei, J. J. Jeong, T. Wang, Y. Liu, S. Tian, P. Patel, X. Jiang, A. B. Jani, H. Mao, W. J. Curran, T. Liu, and X. Yang (2019a). "Automatic MRI prostate segmentation using 3D deeply supervised FCN with concatenated atrous convolution." In: *Proceedings of SPIE - The International Society for Optical Engineering*, p. 141. ISBN: 9781510625471. DOI: 10.1117/12.2512551 (cit. on p. 33).

Wang, B., Y. Lei, S. Tian, T. Wang, Y. Liu, P. Patel, A. B. Jani, H. Mao, W. J. Curran, T. Liu, and X. Yang (2019b). "Deeply supervised 3D fully convolutional networks with group dilated convolution for

automatic MRI prostate segmentation." In: *Medical Physics* 46.4, pp. 1707–1718. DOI: 10.1002/mp.13416 (cit. on p. 34).

Wang, J., C. Bian, M. Li, X. Yang, K. Ma, W. Ma, J. Yuan, X. Ding, and Y. Zheng (2019c). "Uncertainty-guided domain alignment for layer segmentation in OCT images." In: *CoRR*. URL: http://arxiv.org/abs/1908.08242 (cit. on p. 125).

Wang, M., Q. Zhang, S. Lam, J. Cai, and R. Yang (2020a). "A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning." In: *Frontiers in Oncology* 10, p. 2177. DOI: 10.3389/fonc.2020.580919 (cit. on p. 1).

Wang, P., J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers (2021a). "Self-paced and self-consistent co-training for semi-supervised image segmentation." In: *Medical Image Analysis*, p. 102146. DOI: 10.1016/j.media.2021.102146 (cit. on p. 67).

Wang, S., J. Frisbie, Z. Keepers, Z. Bolten, A. Hevaganinge, E. Boctor, S. Leonard, J. Tokuda, A. Krieger, and M. M. Siddiqui (2020b). "The Use of Three-dimensional Visualization Techniques for Prostate Procedures: A Systematic Review." In: *European Urology Focus*. online ahead of print. DOI: 10.1016/j.euf.2020.08.002 (cit. on p. 96).

Wang, W., Q. Xia, Z. Hu, Z. Yan, Z. Li, Y. Wu, N. Huang, Y. Gao, D. Metaxas, and S. Zhang (2021b). "Few-shot Learning by a Cascaded Framework with Shape-constrained Pseudo Label Assessment for Whole Heart Segmentation." In: *IEEE Transactions on Medical Imaging*. online ahead of print. DOI: 10.1109/TMI.2021.3053008 (cit. on p. 67).

Wang, Y., Y. Zhang, J. Tian, C. Zhong, Z. Shi, Y. Zhang, and Z. He (2020c). "Double-uncertainty weighted method for semi-supervised learning." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 542–551. DOI: 10.1007/978-3-030-59710-8_53 (cit. on pp. 67, 91).

Warfield, S. K., K. H. Zou, and W. M. Wells (2004). "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." In: *IEEE Transactions on Medical Imaging* 23.7, pp. 903–921. DOI: 10.1109/TMI.2004.828354 (cit. on p. 132).

Wein, A. J., L. R. Kavoussi, A. W. Partin, A. C. Novick, and C. A. Peters, eds. (2012). *Campbell-Walsh Urology*. 10. edition. Elsevier Saunders. ISBN: 978-1416069119 (cit. on p. 8).

Wilcoxon, F. (1992). "Individual comparisons by ranking methods." In: *Breakthroughs in Statistics*. Springer, pp. 196–202. DOI: 10.2307/3001968 (cit. on p. 28).

Xia, Y., D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth (2020). "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation." In: *Medical Image Analysis* 65, p. 101766. ISSN: 1361-8415. DOI: 10.

1016/j.media.2020.101766 (cit. on pp. 66, 68, 104, 105, 113, 116, 123, 124).

Xie, X., J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu (2021). "A survey on incorporating domain knowledge into deep learning for medical image analysis." In: *Medical Image Analysis*, p. 101985. DOI: 10.1016/j.media.2021.101985 (cit. on p. 134).

Xu, Z., C. P. Lee, M. P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. G. Abramson, and B. A. Landman (2016). "Evaluation of six registration methods for the human abdomen on clinically acquired CT." In: *IEEE Transactions on Biomedical Engineering* 63.8, pp. 1563–1572. DOI: 10.1109/TBME.2016.2574816 (cit. on p. 113).

Yan, W., Y. Wang, M. Xia, and Q. Tao (2019). "Edge-Guided Output Adaptor: Highly Efficient Adaptation Module for Cross-Vendor Medical Image Segmentation." In: *IEEE Signal Processing Letters* 26.11, pp. 1593–1597. ISSN: 1070-9908. DOI: 10.1109/LSP.2019.2940926 (cit. on p. 103).

Yang, H., C. Shan, A. F. Kolen, et al. (2020). "Deep Q-Network-Driven Catheter Segmentation in 3D US by Hybrid Constrained Semi-Supervised Learning and Dual-UNet." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 646–655. DOI: 10.1007/978-3-030-59710-8_63 (cit. on p. 67).

Yang, J., N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, and J. S. Duncan (2019). "Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 255–263. DOI: 10.1007/978-3-030-32245-8_29 (cit. on p. 103).

Yang, X., P. Rossi, A. B. Jani, H. Mao, T. Ogunleye, W. J. Curran, and T. Liu (2015). "A 3D neurovascular bundles segmentation method based on MR-TRUS deformable registration." In: *SPIE Medical Imaging: Image Processing*, p. 941319. DOI: 10.1117/12.2077828 (cit. on p. 105).

Yu, F. and V. Koltun (2016). "Multi-Scale Context Aggregation by Dilated Convolutions." In: *International Conference on Learning Representations (ICLR)* (cit. on p. 18).

Yu, L., X. Yang, H. Chen, J. Qin, and P.-A. Heng (2017). "Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images." In: *AAAI Conference on Artificial Intelligence*, pp. 66–72 (cit. on pp. 33, 34).

Yu, L., S. Wang, X. Li, C.-W. Fu, and P.-A. Heng (2019). "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 605–613. DOI: 10.1007/978-3-030-32245-8_67 (cit. on pp. 67, 68, 91).

Yuan, Y., W. Qin, X. Guo, M. Buyyounouski, S. Hancock, B. Han, and L. Xing (2019). "Prostate Segmentation with Encoder-Decoder Densely Connected Convolutional Network (Ed-Densenet)." In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 434–437. ISBN: 978-1-5386-3641-1. DOI: 10.1109/ISBI.2019.8759498 (cit. on pp. 33, 34).

Yusim, I., M. Krenawi, E. Mazor, V. Novack, and N. J. Mabjeesh (2020). "The use of prostate specific antigen density to predict clinically significant prostate cancer." In: *Scientific Reports* 10.1, pp. 1–6. DOI: 10.1038/s41598-020-76786-9 (cit. on p. 30).

Zabihollahy, F., N. Schieda, S. Krishna Jeyaraj, and E. Ukwatta (2019). "Automated Segmentation of Prostate Zonal Anatomy on T2-weighted (T2W) and Apparent Diffusion Coefficient (ADC) Map MR Images using U-Nets." In: *Medical Physics* 46.7, pp. 3078–3090. DOI: 10.1002/mp.13550 (cit. on pp. 63, 64).

Zhang, L., X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, et al. (2020). "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation." In: *IEEE Transactions on Medical Imaging* 39.7, pp. 2531–2540. DOI: 10.1109/TMI.2020.2973595 (cit. on pp. 103, 132, 134).

Zhao, Y.-X., Y.-M. Zhang, M. Song, and C.-L. Liu (2019). "Multi-view semi-supervised 3D whole brain segmentation with a self-ensemble network." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 256–265. DOI: 10.1007/978-3-030-32248-9_29 (cit. on p. 66).

Zhou, H.-Y., A. Oliver, J. Wu, and Y. Zheng (2018). "When semi-supervised learning meets transfer learning: Training strategies, models and datasets." In: *CoRR*. URL: http://arxiv.org/abs/1812.05313 (cit. on p. 98).

Zhou, Y., Y. Wang, P. Tang, S. Bai, W. Shen, E. Fishman, and A. Yuille (2019). "Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training." In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 121–140. DOI: 10.1109/WACV.2019.00020 (cit. on p. 66).

Zhou, Z.-H. and M. Li (2005). "Tri-training: Exploiting unlabeled data using three classifiers." In: *IEEE Transactions on Knowledge and Data Engineering* 17.11, pp. 1529–1541. DOI: 10.1109/TKDE.2005.186 (cit. on p. 24).

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros (2017a). "Unpaired image-to-image translation using cycle-consistent adversarial networks." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2223–2232. DOI: 10.1109/ICCV.2017.244 (cit. on pp. 103, 104).

Zhu, Q., B. Du, B. Turkbey, P. L. Choyke, and P. Yan (2017b). "Deeply-supervised CNN for prostate segmentation." In: *International Joint*

Conference on Neural Networks (IJCNN), pp. 178–184. ISBN: 978-1-5090-6182-2. DOI: 10.1109/IJCNN.2017.7965852 (cit. on pp. 33, 34).

Zhu, Q., B. Du, J. Wu, and P. Yan (2018). "A Deep Learning Health Data Analysis Approach: Automatic 3D Prostate MR Segmentation with Densely-Connected Volumetric ConvNets." In: *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. DOI: 10.1109/IJCNN.2018.8489136 (cit. on pp. 33, 34).

Zhu, Q., B. Du, and P. Yan (2020). "Boundary-Weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation." In: *IEEE Transactions on Medical Imaging* 39.3, pp. 753–763. DOI: 10.1109/TMI.2019.2935018 (cit. on pp. 33, 34, 97, 102, 134).