Domain Adaptation for Novel Imaging Modalities with Application to Prostate MRI

Eleni Chiou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

September 8, 2022

I, Eleni Chiou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The need for training data can impede the adoption of novel imaging modalities for deep learning-based medical image analysis. Domain adaptation can mitigate this problem by exploiting training samples from an existing, *densely-annotated source domain* within a novel, *sparsely-annotated target domain*, by bridging the differences between the two domains. In this thesis we present methods for *adapting* between diffusion-weighed (DW)-MRI data from multiparametric (mp)-MRI acquisitions and VERDICT (Vascular, Extracellular and Restricted Diffusion for Cytometry in Tumors) MRI, a richer DW-MRI technique involving an optimized acquisition protocol for cancer characterization. We also show that the proposed methods are general and their applicability extends beyond medical imaging.

First, we propose a semi-supervised domain adaptation method for prostate lesion segmentation on VERDICT MRI. Our approach relies on stochastic generative modelling to translate across two heterogeneous domains at pixel-space and exploits the inherent uncertainty in the cross-domain mapping to generate multiple outputs conditioned on a single input. We further extend this approach to the unsupervised scenario where there is no labeled data for the target domain. We rely on stochastic generative modelling to translate across the two domains at pixel space and introduce two loss functions that promote semantic consistency.

Finally we demonstrate that the proposed approaches extend beyond medical image analysis and focus on unsupervised domain adaptation for semantic segmentation of urban scenes. We show that relying on stochastic generative modelling allows us to train more accurate target networks and achieve state-of-the-art performance on two challenging semantic segmentation benchmarks.

Impact statement

The work presented in this thesis is part of ongoing research efforts to develop generalisable and adaptive learning paradigms. We propose domain adaptation methods that allow training deep learning models with less supervision via knowledge transfer from auxiliary datasets. The proposed methods can alleviate the burden of collecting large-scale labeled data in many applications where large related datasets are available. In medical imaging applications it is very critical since it allows to: i) reduce the cost of annotating large medical datasets - manual annotation can be very costly and time consuming when it is carried out for every new target domain, ii) facilitate the quick adoption of novel and more informative imaging modalities for learning-based image analysis, thereby allowing for improved diagnosis and clinical decision making. Finally, even though our work was primarily motivated by particular medical imaging applications, the proposed solutions are general and can address data scarcity problems beyond medical image analysis.

Acknowledgements

First and foremost, I would like to thank my PhD advisors, Laura Panagiotaki and Iasonas Kokkinos. Laura, for bringing me to UCL and giving me substantial room to work on topics that I was excited about. Iasonas, for his constant support and guidance along the way. His breadth of knowledge and dedication to research have served as an inspiration to me.

I would like to thank all the CMIC colleagues and friends (An, Giulia, Elena, Maura, Rica, Ioanna, Kiko, Pere, Beppe, Paddy) for making the lab such a better place and for all the fun moments we shared, either within the walls of the lab or somewhere in London.

I would also like to thank my oldest friends Nikolas and Katerina for cheering me up and supporting me throughout this journey. Special thanks go to Nikos for his patience, support and encouragement and for always being there for me.

Last but not least, I want to thank my family for their unconditional support and love: my parents, Antonia and Theodoros; my brother, Giannis; and my aunt and uncle, Aristea and Dennis. Thank you for always loving me, believing in me and being my greatest supporters.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Deep learning-based medical image analysis has the potential to transform healthcare by achieving high accuracy and efficiency on various diagnostic and treatment processes [8, 9, 10]. In certain cases, where large, homogeneous datasets with high-quality annotations are available, deep learning models have been shown to surpass the performance of clinical experts [11, 12, 13, 14].

However, medical imaging is an evolving field and advanced imaging techniques are constantly developed to replace or supplement existing techniques for improved diagnosis. At the same time, annotating large-scale datasets for every newly developed imaging technique is not always a feasible solution due to human labor and expertise required. Thus, the successful adoption of novel imaging techniques for learning-based medical image analysis can be hindered by the need of manual annotation of a large corpus of training data. This can result in inertia, favoring earlier imaging techniques that come with larger training sets.

A potential solution to this problem is the development of domain adaptation methods that leverage training samples from an existing, *densely-annotated* domain within a novel, *sparsely-annotated* domain, by bridging the differences between the two domains. This facilitates training powerful deep-learning models for novel medical imaging modalities or acquisition protocols, effectively compensating for the limited amount of training data.

Recent, pixel-level domain adaptation methods establish a transformation between the two domains that bridges the difference in their statistics while preserving the semantics of the translated samples [15, 16, 17, 18, 19, 20]. However, these approaches rely on the strong assumption that the translation is a deterministic function mapping a single source to a single target image. We address the challenge of adapting to a more informative target domain where multiple target samples can emerge from a single source sample.

We initially focus on prostate lesion characterization on an advanced diffusion-weighted imaging (DWI) technique called VERDICT (Vascular, Extracellular and Restricted Diffusion for Cytometry in Tumors) magnetic resonance imaging (MRI). VERDICT MRI is a non-invasive imaging technique combining an advanced DWI acquisition protocol and a mathematical model to estimate microstructural feature of tumour in-vivo [21, 6, 22]. Compared to the naive DWI from multiparametric (mp)-MRI acquisitions, VERDICT MRI has a richer acquisition protocol to probe the underlying microstructure and reveal changes in tissue features similar to histology. However, the limited availability of labeled training data prevents the training of robust deep neural networks that could directly exploit the information in the raw VERDICT MRI. On the other hand, large-scale, clinical mp-MRI datasets exist. We propose domain adaptation methods that exploit labeled mp-MRI data to improve the generalization capabilities of prostate lesion characterization on VERDICT MRI.

We also demonstrate that the proposed methods are general and applicable to visual scene understanding. Specifically we focus on semantic segmentation of urban scenes and exploit synthetically generated datasets (GTA5 [23], SYNTHIA [24]) that come with rich ground-truth to train models that can perform well in real images (Cityscapes [25]) with different appearance properties.

1.1 Thesis contributions and outline

The main goal of the thesis is to address data scarcity, i.e, the limited availability or even complete lack of carefully annotated data required to build accurate predictive models. Specifically, we focus on the development of domain adaptation methods that allow leveraging labeled data coming from a related, densely-labeled source domain to train models that perform well on data coming from a sparsely-labeled or unlabeled target domain. We propose a semi-supervised and an unsupervised domain adaptation method for lesion segmentation on VERDICT MRI and we also extend our methods to unsupervised domain adaptation for semantic segmentation of urban scenes.

In Chapter 2 we provide a brief overview of the MRI sequences involved in prostate cancer characterization. We discuss the different MRI imaging techniques developed for prostate cancer characterization, their strengths as well as their limitations. We also provide a brief overview of machine learning techniques for automatic assessment of prostate mp-MRI.

In Chapter 3 we perform some preliminary analysis on VERDICT MRI. In particular we investigate the potential of model-free prostate lesion classification on the raw VERDICT MRI data using fully convolutional networks (FCNs). We also examine whether the raw VERDICT MRI allows for better classification of prostate lesions compared to the raw diffusion-weighed (DW) data and the apparent diffusion coefficient (ADC) map from the mp-MRI acquisition. Our results indicate that: i) FCNs trained on VERDICT MRI achieve good performance in differentiating between malignant and benign lesions and ii) FCNs trained and evaluated on VERDICT MRI perform better than FCNs trained and evaluated on the naive DW data and the ADC map from mp-MRI acquisitions. This chapter contains material from [26, 27].

In Chapter 4 we propose a semi-supervised domain adaptation approach for lesion segmentation. We rely on stochastic generative modelling to translate DW-MRI from mp-MRI to VERDICT MRI and exploit the inherent uncertainty in the cross-domain mapping to generate multiple outputs conditioned on a single input. In addition, we enforce semantic consistency between the real and synthetic images by exploiting both source-domain and target-domain lesion segmentation supervision to train target-domain networks operating on the synthetic images. This results in training networks that can generate diverse outputs while at the same time preserving critical structures. We further accommodate the statistical discrepancies between real and synthetic data by introducing residual adapters in the segmentation network. These capture domain-specific properties and allow the segmentation network to generalize better across the two domains. When compared to its deterministic counter- parts, our approach yields substantial improvements across a broad range of dataset sizes, increasingly strong baselines, and evaluation metrics. This chapter contains material from [28, 29].

In Chapter 5, we propose an unsupervised domain adaptation approach for prostate lesion segmentation. We rely on stochastic generative modelling to translate across the source and the target domain at pixel space and introduce two new loss functions that promote semantic consistency. Firstly, we introduce a semantic cycle-consistency loss in the source domain to ensure that the translation preserves the semantics. Secondly, we introduce a pseudo-labelling loss, where we translate target data to source, label them using a source-domain network, and use the generated pseudo-labels to supervise the target-domain network. When compared to several unsupervised domain adaptation approaches, our approach yields substantial improvements, that consistently carry over to the semi-supervised and supervised learning settings. This chapter contains material from [30, 31].

In Chapter 6, we extend our unsupervised domain adaptation approach for semantic segmentation of urban scenes. We rely on stochastic generative modelling to capture inherent translation ambiguities. This allows us to (i) train more accurate target networks by generating multiple outputs conditioned on the same source image, (ii) impute robust pseudo-labels for the target data by averaging the predictions of a source network on multiple translated versions of a single target image and (iii) train and ensemble diverse networks in the target domain by modulating the degree of stochasticity in the translations. We report improvements over strong recent baselines, leading to state-of-the-art unsupervised domain adaptation results on two challenging semantic segmentation benchmarks. This chapter contains material from [32].

Finally, in Chapter 7 we present conclusions and suggest future research directions.

1.2 List of publications

Publications and submissions under review that are included as part or whole in this thesis are listed below.

- **E. Chiou**, E. Panagiotaki and I. Kokkinos. “Beyond deterministic translation for unsupervised domain adaptation”, BMVC, 2022. [32]

- **E. Chiou**, F. Giganti, Punwani, I. Kokkinos, and E. Panagiotaki. “Unsupervised domain adaptation with semantic consistency across heterogeneous modalities for MRI prostate lesion segmentation”, DART@MICCAI, 2021. [30]

- **E. Chiou**, F. Giganti, S. Punwani, I. Kokkinos, and E. Panagiotaki. “Prostate lesion segmentation on VERDICT-MRI driven by unsupervised domain adaptation”, ISMRM, 20201. [31]

- **E. Chiou**, F. Giganti, Punwani, I. Kokkinos, and E. Panagiotaki. “Harnessing uncertainty in domain adaptation for MRI prostate lesion segmentation”, MICCAI 2020. [28]

- **E. Chiou**, F. Giganti, S. Punwani, I. Kokkinos, and E. Panagiotaki. “Domain adaptation for prostate lesion segmentation on VERDICT MRI”, ISMRM, 2020. [29]

- **E. Chiou**, F. Giganti, E. Bonet-Carne, S. Punwani, I. Kokkinos, and E. Panagiotaki. “Automatic classification of benign and malignant prostate lesions: a comparison using VERDICT-MRI and ADC maps”, ISMRM, 2019. [27]

- **E. Chiou**, F. Giganti, E. Bonet-Carne, S. Punwani, I. Kokkinos, and E. Panagiotaki. “Prostate cancer classification on VERDICT DW-MRI using convolutional neural networks”, MLMI@MICCAI, 2018. [26]

Other publications that I have co-authored during my PhD but are not included in this thesis are listed below.

- **E. Chiou**, V. Valindria, F. Giganti, S. Punwani, I. Kokkinos, and E. Panagiotaki. “Synthesizing VERDICT maps from standard DWI data using GANs”, CDMRI@MICCAI, 2021. [33]

- V. Valindria, M. Palombo, **E. Chiou**, S. Singh, S. Punwani, and E. Panagiotaki. “Synthetic Q-space learning with deep regression networks for prostate cancer characterisation with VERDICT”, ISBI, 2021. [34]

- V. Valindria, S. Singh, **E. Chiou**, T. Mertzanidou, B. Kanber, S. Punwani, M. Palombo and E. Panagiotaki. “Non-invasive Gleason score classification with VERDICT-MRI”, ISMRM, 2021. [35]

- M. Palombo, V. Valindria, S. Singh, **E. Chiou**, F. Giganti, H. Pye, H.C. Whitaker, D. Atkinson, S. Punwani, D.C Alexander and E. Panagiotaki. “Joint estimation of relaxation and diffusion tissue parameters for prostate cancer grading with relaxation-VERDICT MRI”, medRxiv 2021. [36]

Chapter 2

Background

In this Chapter we provide a brief overview of the MRI sequences involved in prostate cancer characterization. In Sec. 2.2, Sec. 2.3 we discuss the different MRI imaging techniques developed for prostate cancer characterization, their strengths as well as their limitations. We also provide a brief overview of machine learning techniques for automatic assessment of prostate mp-MRI.

2.1 Prostate cancer diagnosis

Prostate cancer is the second most common cancer among men worldwide [37]. Early diagnosis and treatment are important to reduce the mortality rate. The standard procedure to provide a diagnosis of the disease is to carry out a systematic transrectal ultrasound-guided (TRUS) biopsy when elevated levels of prostate specific antigen (PSA) are reported in the blood. Usually, 10-12 biopsy cores (tissue samples) are sampled randomly from the prostate [38, 39]. These samples are further evaluated based on the Gleason grading system [1, 40]. As illustrated in Fig. 2.1, Gleason grading system defines five histological patterns or grades ranging from 1 (well differentiate glands) to 5 (no glandular differentiation) based on the degree of differentiation of the cells. Initially, biopsy cores are assigned a grade and then the grades of the two most prevalent patterns are combined to produce the final Gleason Score ranging from 2 to 10 with higher score associated with worse prognosis; carcinoma of Gleason Score 2-4 is considered as well-differentiated, 5-7 as moderately differentiated and 8-10 as poorly differentiated.

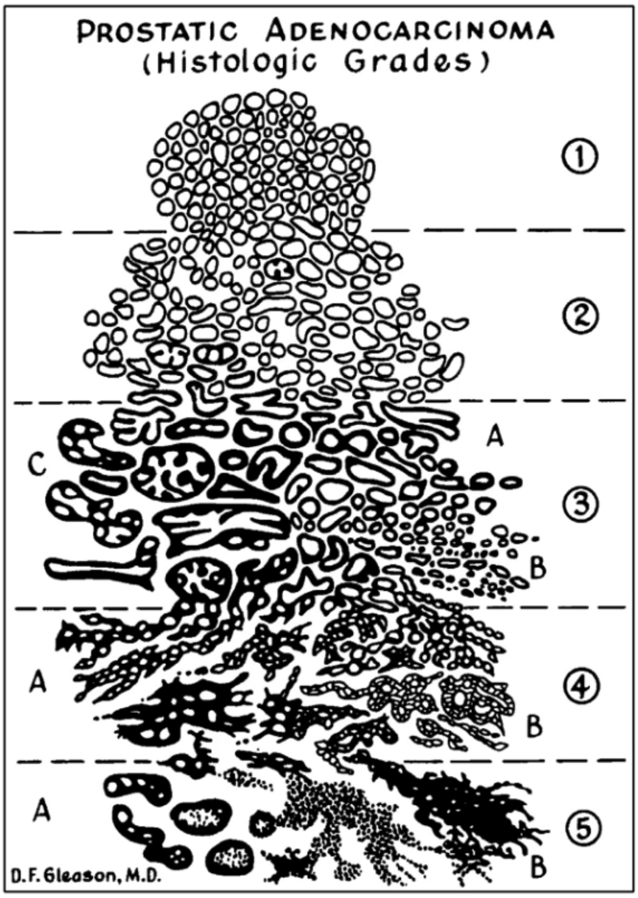


Figure 2.1: Histological patterns or grades of prostate adenocarcinoma based on Gleason grading system; it defines five patterns ranging from 1 (well differentiate glands) to 5 (no glandular differentiation) based on the degree of differentiation of the cells [1].

However, despite the importance of TRUS biopsy for prostate cancer diagnosis, it is a suboptimal diagnostic process [41, 42]. The biopsies are sampled systematically but randomly from the prostate meaning that there is a high chance of missing significant cancers or detecting insignificant cancers [43]. In particular, overdiagnosis of insignificant disease occurs in up to 50% of the cases while underdiagnosis occurs in 18% of cases, especially in the anterior apical regions of the prostate [44]. As a consequence, repeated biopsies with associated patient discomfort and additional risks and costs are often necessitated [45].

Transperineal template-guided mapping (TTM) biopsy offers a diagnostic alternative to TRUS biopsy providing better diagnostic accuracy [46, 47]. Transperineal biopsies are obtained by sampling the entire prostate at 5 mm intervals [48].

Thus, it provides additional information allowing for improved diagnostic accuracy. Unfortunately though, it comes with additional risks for patients and increased cost since it requires histological examination of a larger number of cores [49] and general anesthesia since the biopsy is taken through the perineum. To address these limitations, mp-MRI of the prostate is recommended before biopsy [50]. Pre-biopsy mp-MRI reduces the number of biopsies and overdiagnosis of insignificant disease and improves the detection of clinically significant prostate cancer [51, 41, 52]; clinically significant cancer is defined as Gleason score of 7 or greater. Clinically significant prostate cancer has the potential to metastasize while insignificant not metastasize and mostly results in indolent or slowly growing low-grade tumors. Therefore accurate discrimination between clinically significant and non-clinically significant prostate cancer is critical for risk stratification and clinical decision making.

2.2 Multiparametric magnetic resonance imaging of the prostate

MRI is a popular imaging technique providing useful insights about the human body anatomy and pathology. mp-MRI, which consists of T2 weighed (T2W) imaging, dynamic contrast enhanced (DCE) imaging, diffusion weighed (DW) imaging and the corresponding ADC maps, has become a useful tool for prostate cancer detection.

Radiological interpretation, and reporting of prostate mp-MRI examination relies on the Prostate Imaging Reporting and Data System (PIRADS) [53]. PI-RADS assesses the likelihood of clinically significant prostate cancer on a 5-point scale for each lesion. PI-RADS 1 and 2 lesions have been classified as *clinically significant cancer is highly unlikely to be present* and *clinically significant cancer is unlikely to be present* respectively. PI-RADS 3 lesions has been classified as *the presence of clinically significant cancer is equivocal* and *clinically significant cancer is unlikely to be present*. Finally, PI-RADS 4 and 5 lesions have been classified as *clinically significant cancer is likely to be present* and *clinically significant cancer is highly* *likely* respectively.

Below we present the basic concepts of MRI and then describe in detail the different MRI sequences used in prostate imaging.

2.2.1 Basics on magnetic resonance imaging

MRI relies on the interaction between an applied magnetic field and the nucleus of hydrogen atoms that are abundant in the human body. The hydrogen nucleus has a positive charge and possesses spin - rotates around its own axis and creates magnetic field. In normal environment, the spin magnetic moments of the nuclei are randomly oriented and therefore produce no overall magnetic field. When a strong external magnetic field B_0 is applied, the magnetic moments of the nuclei align with the direction of this field and start precessing around it with a precessional frequency or Lamor frequency ω_0 which is proportional to the strength of the magnetic field (Fig. 2.2).

$$\omega_0 = \gamma B_0 \tag{2.1}$$

The net magnetization vector M_0 of spinning nuclei can be decomposed into two components that are perpendicular to each other: a longitudinal component M_z and a transversal component M_{xy} with respect to the main magnetic field B_0 such that $M_0 = M_z + M_{xy}$. When a radiofrequency (RF) pulse of the same frequency as the Larmor frequency is applied perpendicular to the magnetic field B_0, the nuclei gain energy and the net magnetization moves away from B_0 (longitudinal axis) lying at an angle to it (towards the transverse axis). As a consequence, the longitudinal magnetization M_z decreases while the transverse magnetization M_{xy} increases.

When the RF pulse is removed the nuclei return to equilibrium and the net magnetization vector realigns with B_0; the return to equilibrium is called relaxation. During relaxation two independent processes occur: the longitudinal and transverse relaxation. The longitudinal relaxation or T1 recovery is caused by the nuclei releasing the absorbed energy to the surrounding environment allowing the longitudinal magnetization to recover. The return of magnetization follows an exponential pro-

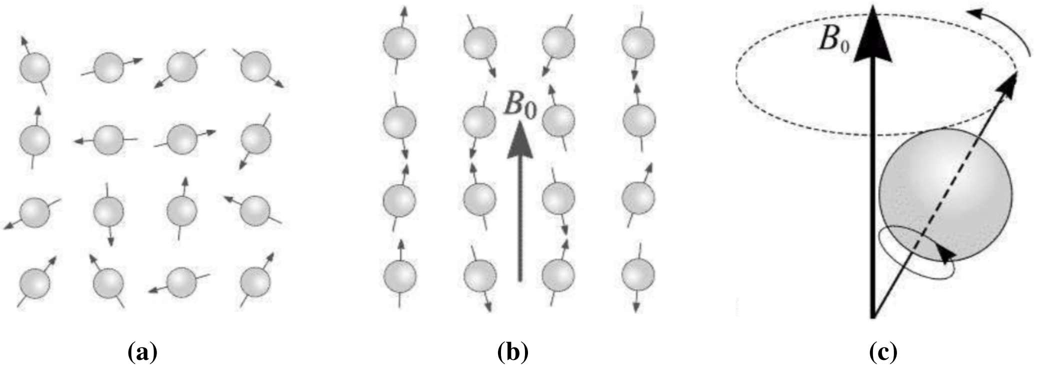


Figure 2.2: Impact of magnetic field B_0 on the magnetic moments of the nuclei. a) random orientation of the spin magnetic moments of the nuclei, b) alignment of the magnetic moments of the nuclei with the direction of the external magnetic field B_0, c) precession of the magnetic moments of the nuclei around the magnetic field B_0 [2].

cess characterized by a tissue specific time constant T1. Transverse or T2 decay occurs because of the interaction of neighboring nuclei (spin-to-spin interaction) causing loss of coherence (dephasing) of the transverse magnetization. The decay of the transverse magnetization follows also an exponential process characterized by a tissue specific time constant T2.

2.2.2 Common MRI sequences for prostate imaging

MRI sequences are mainly characterized by two intrinsic parameters: the repetition time (TR) and the echo time (TE). TR is the time between two successive RF pulses applied to the same slice and determines the amount of T1 weighing on the contrast of the image while TE is the time between the excitation and the collection of the signal and determines the contribution of T2 weighting on the contrast of the image. As we mentioned previously and also shown in Fig. 2.3, mp-MRI consists of T2W imaging, DCE imaging and DW imaging [4, 54, 3], which are discussed below in more detail.

2.2.2.1 T2W Imaging

T2W imaging is the first MRI sequence used to provide insights about prostate anatomy and pathology [55, 56]. It provides great information regarding the prostate's zonal anatomy and it is also used to detect and evaluate abnormalities

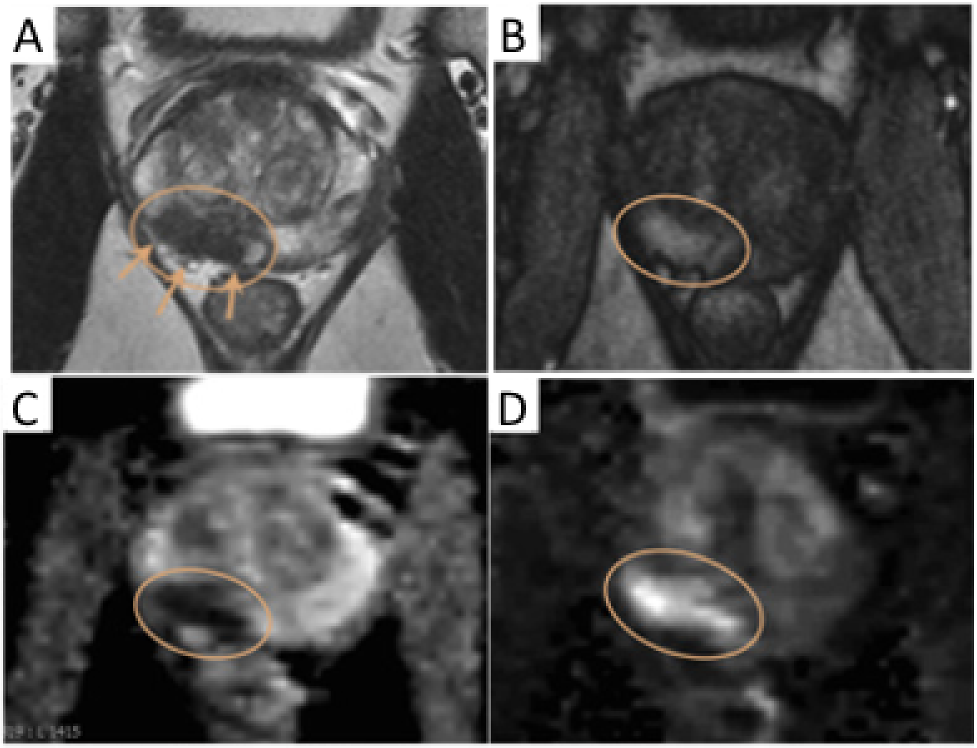


Figure 2.3: Multi-parametric magnetic resonance imaging (mp-MRI) of a patient with suspicion of clinically significant prostate cancer. It consists of T2-weighted (T2W) imaging (A), dynamic contrast enhanced (DCE) imaging (B), diffusion weighed (DW) imaging (D) and the corresponding ADC maps (C). The lesion (orange circle) is appeared as a low-signal intensity structure in T2W image (A), a high-signal intensity structure on the DCE image, a focal "white" area on the DW image (b1400) (D) and a focal "black" area with a low ADC value (C) [3].

of the prostate.

Prostate contains three histological zones: the peripheral zone, transition zone and the central zone. The peripheral zone is the largest part of the prostate comprising almost the 70 % of the prostate gland while the central and transition zone comprise the 25% and 5% of the prostate gland respectively [57]. The peripheral zone is characterized by higher signal intensity on T2W sequences compared to the central and the transition zone which are visualized as a low intensity structures. The central and peripheral zones have similar intensities and can be distinguished based on their anatomical location [55, 56].

Prostate cancer is visualized as a low intensity mass in the peripheral zone and it is relatively easy to be detected given that the peripheral zone is characterised by high signal intensity. Cancer detection in the transition zone can be challenging since it is characterized by heterogeneous signal intensity which overlaps with the signal intensity characteristics of cancer. However, there are some studies indicating that cancer detection in the transition zone is still possible since it is usually appeared as homogeneous low intensity structure with ill-defined margins and lack of capsule [58].

Despite the importance of T2W imaging in providing anatomical and pathological information, it has some important limitations [59, 54, 60, 61]. As we mentioned earlier, detection of cancer in transition zone is challenging leading to reduced sensitivity. In addition, benign abnormalities such as prostatitis and benign prostatic hyperplasia mimic cancer in transition zone leading to low specificity.

2.2.2.2 DCE Imaging

Functional sequences such as DCE imaging is used to improve the diagnostic performance. DCE-MRI provides information about the vascularity of prostate cancer tissue by assessing the enhancement pattern of tissue over time after the administration of a contrast agent material such as gadolinium [62, 63]. The contrast agent passes from the plasma to the extravascular-extracellular space (EES) with a rate that depends on the vascular permeability, the vascular surface area and the blood flow. In tumours, genetic mutations promote the development of new blood vessels that are highly disorganised, abnormal and are characterized by increased permeability [64] leading to a different enhancement pattern compared to the one observed in normal tissues.

DCE imaging usually relies on a T1-weighed sequence to measure the enhancement pattern of the tissues since the presence of contrast agent in the extracellular space shortens the relaxation time increasing contrast in T1-weighed images. Qualitative analysis of DCE images involves examination of the enhancement pattern in different locations. Malignant tumours are usually characterized by a signal having an early, rapid and high enhancement followed by a fast washout. On the other hand, normal tissue is characterized by slower enhancement for a few minutes after the injection. However, qualitative analysis is inherently subjective and thus less reliable. Semi-quantitative analysis aim at quantifying the measured signal by extracting parameters such as the time to peak, maximum slope, peak enhancement. In addition, several quantitative methods relying on pharmacokinetic modelling have been proposed [65, 66]. These methods model the rate of transfer of the contrast agent between plasma and EES; k_{trans} corresponds to the rate of transfer from plasma to EES while k_{ep} corresponds to the rate of transfer from EES to plasma. These two constants are characterized by large values in cancer.

Combined DCE imaging and T2W imaging yields to improved diagnostic performance compared to single T2W imaging. Nevertheless, DCE-MRI requires the administration of a contrast agent and is characterized by low specificity since it has some limitations in discriminating cancer between prostatitis in the peripheral zone and highly vascularized benign prostatic hyperplasia in the transition zone.

2.2.2.3 DW Imaging

DW imaging has been demonstrated to be the most important component of mp-MRI compared to T2W-MRI and DCE-MRI since it allows noninvasive assessment of the complex tissue microstructure [67, 68, 54, 69]. DW imaging explores the random displacement (also called Brownian motion) of water molecules caused by the thermal energy carried by these molecules [70, 71]. As the displacement of water molecules is influenced by tissue microstructure, by measuring this displacement pattern, DW imaging is able to distinguish different microstructural environments. In general, the movement of water molecules depends on the cellularity of the tissue and the integrity of the cell membrane; higher restriction in the motion of water molecules is observed in tissues with high cellular density such as tumours.

Typically, a T2-weighed sequence and diffusion-sensitizing gradients are used for DW imaging. In particular, the intensity of the DW-MR signals becomes sensitive to diffusion by applying a pair of dephasing and rephasing gradients; the movement of water molecules between the application of the two gradients leads to imperfect rephasing and corresponding signal loss. In general, the resulting MR signal is inversely proportional to the movement of the water molecules. Tumours, which are usually more cellular than normal tissues and do not permit great move-

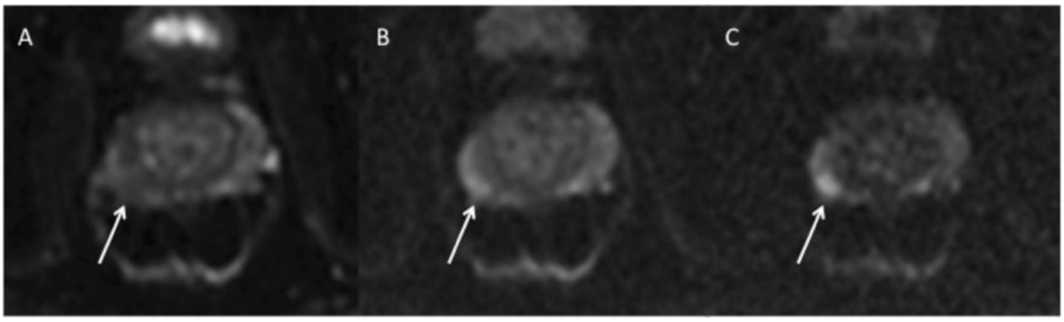


Figure 2.4: Diffusion-weighed images acquired using different b-values; 50 s/mm^2 (A), 800 s/mm^2 (B) and 1500 s/mm^2 (C). High b-values suppress prostatic tissue allowing for better contrast between normal prostate and tumors (arrow) [4]

ment of water molecules, are characterized by high-intensity signals. Thus, the DW-MR signal can provide information regarding the microstructural organization of biological tissues.

A critical parameter in DW imaging is the b-value that describes the magnitude and duration of the diffusion-sensitising gradients as well as the diffusion time. The sensitivity of the DW-MR signal to the water diffusion can be changed by varying the b-value; the higher the b-value characterizing the diffusion-sensitizing gradients, the greater the degree of signal attenuation from water molecules. Images acquired with zero b-values are equivalent to T2-weighed images.

In prostate, DW imaging is usually performed using b-values ranging between 0 and 1500 s/mm^2 [72, 5] (Fig. 2.4). Higher b-values (e.g. b=2000 s/mm^2) are useful in discriminating normal prostatic tissue mimicking cancer in lower b-values [61]. In general, high b-values suppress prostatic tissue allowing for better visualization of the tumour. However, images acquired with very high b-values tend to have decreased signal-to-noise (SNR) and artifacts.

One of the disadvantages of qualitative assessment of DW images is that the intensity of the signal depends on both water diffusion and T2 relaxation time. Thus, areas with a very long T2 relaxation time may have high intensity on DW images, a phenomenon called T2 shine-through [72, 5]. This issue can be addressed by calculating the apparent diffusion coefficient (ADC) as follows:

$$ADC = -\frac{1}{b} \ln \frac{S}{S_0}, \tag{2.2}$$

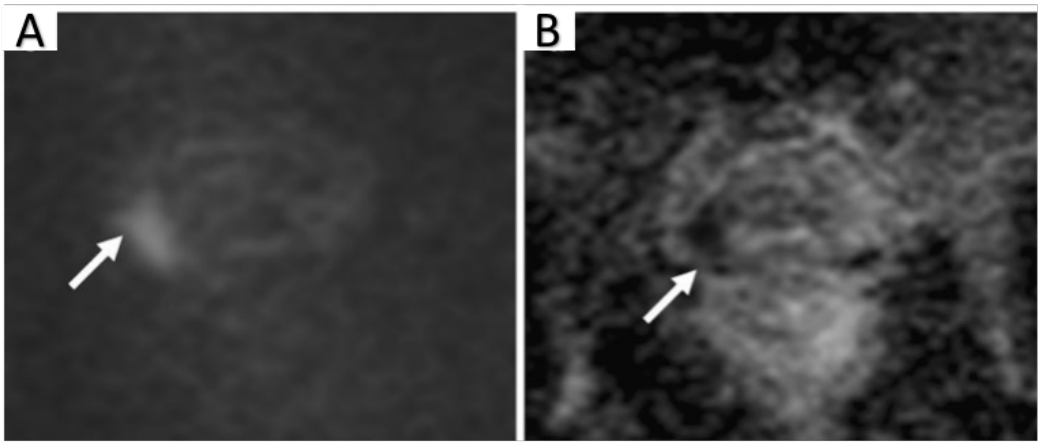


Figure 2.5: Diffusion-weighed (DW) image (A) and the corresponding ADC map (B). The lesion (arrow) appears as a focal hyperintense mass on the DW image and as a focal hypointense mass on the apparent diffusion coefficient map (ADC) map [5].

where S_0 and S are the images acquired with b0 (b $=$ 0s/mm^2) and b respectively. A large value in the DW image corresponds to a low value in the ADC map and vice versa. Thus, lesions appear as focal “black” areas in the ADC maps (Fig. 2.5).

2.2.3 Machine learning for cancer characterization

mp-MRI is very important for non-invasive localization, scoring and staging of abnormalities that may correspond to clinically significant prostate cancer . However, radiological interpretation of mp-MRI often leads to over-diagnosis of low-grade or non-clinically significant tumours and subsequent over-treatment [41].

Aiming to address this limitation and to speed up the radiological interpretation of MRI sequences by assisting radiologists, several studies focus on the development of machine learning techniques for automatic assessment of prostate mp-MRI.

In this section we provide an overview of computer-aided diagnosis (CAD) systems for prostate cancer diagnosis on mp-MRI sequences. CAD systems for prostate cancer are composed of several subsystems, i.e., prostate segmentation, image registration, and lesion classification or segmentation [73]. Below, we review the methods used in each subsystem.

2.2.3.1 Segmentation

Prostate segmentation is usually performed to limit further analysis on the organ of interest. In this section we present the different segmentation approaches used in CAD systems for prostate cancer.

- Atlas-based segmentation. Litjens et al. [74] used an atlas-based segmentation approach similar to the one proposed in [75] to segment the prostate. The segmentation is performed in two steps : 1) image registration and 2) atlas label image fusion. In the image registration step, new subject images are registered to the atlas images using a non-rigid registration algorithm. The registered images are applied to the label images. In the label image fusion step, the labelled images for each subject are combined to one labelled image. In [76], Litjens et al. extended the atlas-based segmentation approach used in [74] applying an atlas selection mechanism presented in [77] to improve label image fusion.

- Model-based segmentation. Viswanath et al. [78] used a novel active shape model (ASM) called MANTRA (Multi-Attribute, Non-Initializing, Texture Reconstruction Based Active Shape Model) proposed in [79] to segment the prostate. Reba et al. [80, 81, 82] performed prostate segmentation using a level set method. In their work, the speed function controlling the evolution of the surface is estimated by fusing image intensity and prostate shape features using a non-negative matrix factorization (NMF) approach.

Despite the success of these methods, they require careful feature engineering to achieve good performance. The multi-atlas based methods require good features for identifying correspondences between a new image and each atlas image, while the deformable model relies on discriminative features, i.e., intensity to segment the prostate.

2.2.3.2 Registration

Multi-modal image registration is an important component of CAD systems since it enables the integration of information obtained from different MRI sequences.

In this section we present some of the image registration methods used in CAD systems for prostate cancer.

Viswanath et al. [78] performed an affine registration via maximization of mutual information (MI) to correct the misalignment between different MRI modalities. Giannini et al. [83] applied affine registration between T2W and DW images using bladder contours to focus registration. Kiraly et al. [84] used a 3D non-rigid registration proposed in [85] to align the different MRI modalities. Registration is performed by maximizing mutual information using a stochastic analog of gradient descent. Yang et al. [86] used also non-rigid registration based on mutual information to register T2W and ADC images.

2.2.3.3 Prostate lesion detection and classification

In this section we give an overview of the methods used for prostate lesion detection, classification and segmentation.

- Support Vector Machines (SVM). Artan et al. [87] developed a framework that combines SVM and conditional random fields (CRFs) for prostate cancer segmentation. They trained and evaluated their method on a dataset that included DCE-MRI in addition to DW-MRI and T2W-MRI and was obtained from 21 biopsy-confirmed prostate cancer patients. The tumour regions were contoured by a radiologist using as guidance the histological slides. Litjens et al. [74] used SVM to classify candidate ROIs obtained after several steps. Initially, likelihood maps representing the probability of each voxel being malignant are obtained using a k-nearest neighbour (k-nn) classifier. Then, local maxima obtained from the likelihood maps are used to segment the candidate ROIs using a region growing and morphology based method. They trained and evaluated their method on dataset obtained from 288 patients who underwent MR-guided biopsy. The dataset included DW-MRI, T2W-MRI, DCE-MRI. Niaf et al. [88] used SVM to classify malignant and benign lesions. They trained and evaluated their method on a dataset obtained from 30 patients who underwent T2W-MRI, DW and DCE-MRI imaging prior to radical prostatectomy.

- Linear Discriminant Analysis (LDA). Litjens et al. [76] applied LDA to perform voxel-wise classification and obtain likelihood maps indicating the probability of malignancy in each voxel. Then, they performed candidate selection and extracted hand crafted statistical features for each candidate. After the extraction of features, candidate classification was performed using LDA. They evaluated their approach on a large consecutive cohort of 347 patients with MR-guided biopsy as the reference standard. This set contained 165 patients with cancer and 182 patients without prostate cancer. All patients underwent T2W-MRI, DCE-MRI and DW-MRI.

- Probabilistic Models. Niaf et al. [88] used a Naive Bayes (NB) classifier to classify malignant and benign prostatic lesions. They proposed method was trained and evaluated on a dataset obtained from 30 patients who underwent T2W-MRI, DW and DCE-MRI imaging prior to radical prostatectomy. In addition, Giannini [83] et al. used a NB classifier to estimate voxel malignancy probability. The dataset used in this study includes included T2W-MRI and DW imaging data obtained from 10 patients. A radiologist contoured the lesion of the T2W images using the histopathologic sections as guidance.

- Neural Networks. One major limitation of the aforementioned is that they employ ad-hoc and handcrafted which are empirically designed and have limited generalization power to different domains. To address this limitation, several deep learning methods have been recently proposed. Compared to previous methods, which use handcrafted features, deep learning methods are able to effectively learn feature hierarchies from the data. Reda et al. [80] used neural networks and performed prostate cancer classification using an approach consisting of two main stages. In the first stage they used different auto-encoders for DW images acquired for different b-values and obtained initial probability maps. Then, they used a stacked non-negativity constraint auto-encoder to estimate the final classification based on the initial probability maps. They trained and evaluated their approach on DW-MRI data obtained from 53 patients. Malignant and benign lesions were contoured on the DW-MRI data by a radiologist. Kiraly et al. [84] used deep encoder-decoder networks to discriminate between benign and malignant lesions. They use simple point locations as ground truth and train the network to output Gaussian kernels around those points. This approach facilitates simultaneous localization and classification within a single run. Mehrtash et al. [89] proposed a 3D CNN to perform image based classification for prostate cancer. In [90], Tsehay et al. adopted a network architecture from an edge detector proposed in [91] to generate image probability map and detect lesions after applying thresholding. Wang et al. [92] proposed a framework for joint multimodal registration and prostate cancer detection. The proposed architecture consists of two sub-networks; a tissue deformation network that performs multimodal registration and a dual convolutional neural network that performs image classification and generates class probability maps. They also performed a post processing step to detect prostate cancer on the class probability maps. [84, 89, 90, 91, 92] trained and evaluated the proposed methods on PROSTATEx Challenge dataset [93]. PROSTATEx consists of T2W-MRI, DCE-MRI, and DW-MRI data obtained from 347 patients. MR-guided biopsy is used as the reference standard. Cao et al. [94] proposed a CNN for simultaneous detection and Gleason score prediction of prostate lesions. They trained their model on a large prostate mp-MRI dataset of 417 patients who underwent 3T mp-MRI exams prior to robotic-assisted laparoscopic prostatectomy. Finally, Mehta et al. [95] proposed a deep learning framework for automatic assessment of prostate cancer on mp-MRI, consisting of three sub-modules: a CNN that performs zone segmentation, a CNN that performs segmentation of clinically significant cancer, and a report-generator that generates an automatic web-based report. The proposed method was trained on PROSTATEx dataset [93] and externally validated using the Prostate Imaging Compared to Transperineal Ultrasound guided biopsy for significant prostate cancer Risk Evaluation (PICTURE) study dataset [96].

Several deep learning methods have been also proposed for prostate lesion segmentation. Hambarde et al. [97] relied on U-Net to achieve segmentation of prostate gland and prostate lesions. The proposed method was trained and validated on 1174 and 2071 T2W-MR images of 40 patients and tested on 250 and 415 T2W-MR images of 10 patients for prostate capsule segmentation and prostate lesion segmentation, respectively. A radiologist marked prostate gland and prostate lesion on the images which served as the groundtruth. Chen [98] proposed a multiple branch UNet for the segmentation of prostate lesions on mp-MRI images. They used mp-MRI data from 136 patient. Each patient underwent T2W-MRI and DW-MRI. A radiologist contoured malignant prostate lesions based on the radiology report. Liu [99] et al. proposed a multi-scale segmentation network with a cascading pyramid convolution module and a double-input channel attention module for prostate lesion segmentation. They used a dataset obtained from 171 patients to train the proposed method and a dataset obtained 17 patients to test the model. The two datasets were acquired from different centers but both include ADC, T2W-MRI, and DW-MRI. A radiologist annotated the images according to the pathology report. Duran et al. [100] propose an attention-based CNN for joint multi-class segmentation of prostate and prostate lesions. The dataset used for training and evaluation was obtained from 219 patients. All patients underwent radical prostatectomy, meaning that the majority of patients had clinically significant cancer and a high number of lesions in the dataset. A radiologist outlined the prostate lesions based on T2W-MRI, ADC maps and DCE-MRI. The prostatectomy specimens were analyzed and used as groundtruth for the outlined lesions.

2.3 Advanced imaging of the prostate

As we mentioned in Sec. 2.2.3, several studies focus on the development of machine learning to address some of the limitations of mp-MRI. Another dominant research direction for improving non-invasive prostate cancer characterization is the development of advanced imaging techniques aiming at improving the

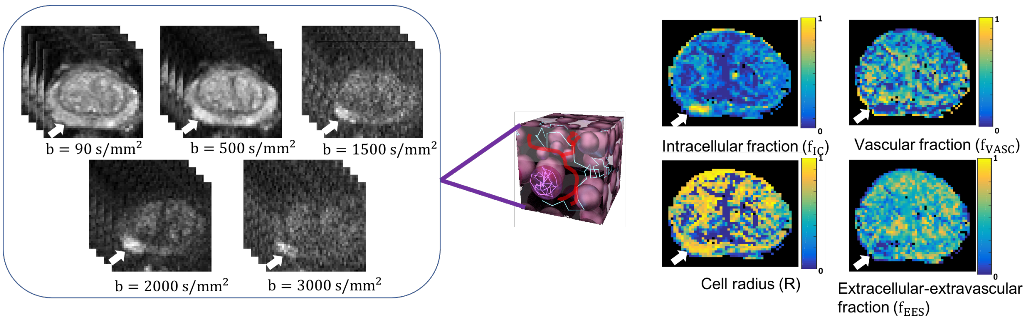

b = 90 s/mm²
b = 500 s/mm²
b = 1500 s/mm²
b = 2000 s/mm²
b = 3000 s/mm²
Intracellular fraction (f_{IC})
Vascular fraction (f_{VASC})
Cell radius (R)
Extracellular-extravascular fraction (f_{EES})
1
0

Figure 2.6: VERDICT MRI. It combines a mathematical model and an optimized diffusion-weighted (DW) acquisition protocol to access microstructural features such as cell size, density, and vascular volume fraction, all of which change in cancer. The intracellular volume fraction (f_{IC}, f_{VASC}) is significant higher for malignant tumours (arrow) than for benign or normal prostate tissue since increased cellularity is a common characteristic in cancer. The extracellular-extravascular volume fraction f_{EES} is significantly lower in tumours compared to benign or normal tissue while there is not significant change in the estimate of the cell radius (R).

quality of the acquired data and providing information similar to histology. Recent advanced imaging techniques include VERDICT MRI, luminal water imaging (LWI) [101, 102, 103], hybrid multi-dimensional MRI (HM-MRI) [104, 105, 103] and restriction spectrum imaging (RSI) [106, 107, 103]. In this thesis we focus on VERDICT MRI and we describe it in more detail in section Sec. 2.3.1 and in Sec. 2.3.2 we give a brief overview of LWI and HM-MRI.

2.3.1 VERDICT MRI

VERDICT MRI is a non-invasive microstructural imaging technique for cancer characterization [21, 6]. It combines a mathematical model and an optimized DW acquisition protocol to access microstructural features such as cell size, density, and vascular volume fraction, all of which change in cancer (Fig. 2.6).

DW-MRI is an integral component of mp-MRI since it provides information about cancer aggressiveness and improves specificity [67, 68]. However, mp-MRI studies use DW-MRI in its simplest form by deriving the ADC map. This simplified model of water diffusion lacks biological specificity as it fails to discriminate the variety of histological changes that occur in cancer [69]. VERDICT MRI improves on ADC maps by modelling directly the underlying microstructure.

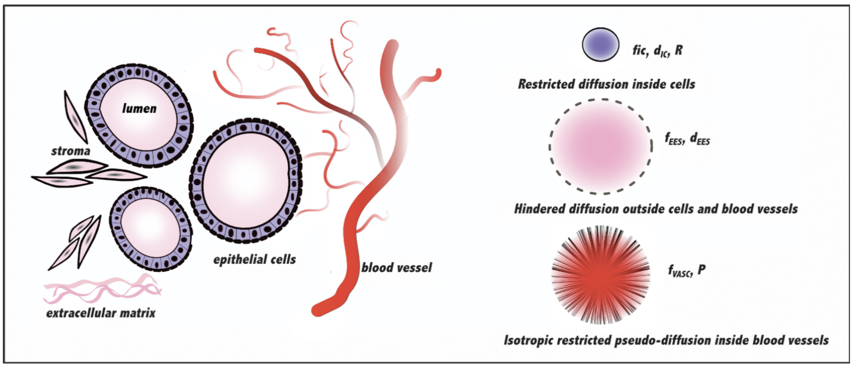

lumen
stroma
epithelial cells
extracellular matrix
blood vessel
fic, d_{IC}, R
Restricted diffusion inside cells
f_{EES}, d_{EES}
Hindered diffusion outside cells and blood vessels
f_{VASC}, P
Isotropic restricted pseudo-diffusion inside blood vessels

Figure 2.7: Schematic representation of the prostate tissue and the corresponding components of the VERDICT model. The color indicates the assignment of the tissue compartments to the model components [6].

VERDICT model is a three compartment model characterizing water diffusion in three primary compartments: 1) vascular, 2) extracellular-extravascular, and 3) intracellular space allowing the estimation of intracellular, extracellular-extravascular and vascular volume fractions, as well as cell radius. Fig. 2.7 shows a schematic representation of the VERDICT model for the prostate tissue. The intracellular compartment (IC) has three parameters: intracellular volume fraction (f_{IC}), diffusivity (d_{IC}) and cell radius (R). The extracellular-extravascular space (EES) compartment has EES volume fraction (f_{EES}) and EES diffusivity (d_{EES}) as parameters. The vascular compartment has vascular volume fraction (f_{VASC}) and pseudo-diffusivity (P) as parameters. The estimated parameters provide information about the cellular and vascular structure of the tissue which change with disease.

The intracellular volume fraction is significant higher for malignant tumours than for benign or normal prostate tissue since increased cellularity is a common characteristic of malignant tumours. The extracellular-extravascular volume fraction is significantly lower in tumours compared to benign or normal tissue while there is not significant change in the estimate of the cell radius.

An initial pre-clinical study demonstrated that the application of VERDICT in colorectal tumour xenographs can reveal differences in the microstructural features of different tissue types [21]. An subsequent in-vivo study in patients with prostate cancer demonstrated the potential of VERDICT in discriminating cancer and benign tissue [6]. In addition, a recent study indicated that the intracellular volume fraction

(FIC) map provides better differentiation of Gleason 4 cancer from benign and/or Gleason 3+3 compared to the ADC map [22].

2.3.2 Other advanced prostate imaging techniques

Other advanced imaging techniques include LWI [101, 102, 103] and HM-MRI [104, 105, 103] and are described briefly below.

LWI [101, 102, 103] is an MRI method that employs multicomponent modelling of T2 mapping data. It has been developed based on the fact that the composition and lumen percentage of the prostatic tissue changes significantly with the presence of cancer and the Gleason grade on cancer. It introduces a new parameter called luminal water fraction (LWF), which is proportional to the fractional volume of luminal space in prostatic tissue. LWF can reveals important information for diagnostic purposes because of the difference in composition and lumen percentage between normal and cancerous tissues.

HM-MRI [104, 105, 103] is an MRI method that relies on the fact that fractional volumes of the prostate gland components stroma, epithelium, and lumen correlate strongly with cancer presence, Gleason grade. It uses two-dimensional MRI sampling to measure the change in ADC and T2 on TE and b-value, respectively. Thus, HM-MRI could provide quantitative estimates of tissue composition by exploiting the coupled T2 and ADC values associated with each tissue component and use these as a source of information about the underlying tissue microstructure.

Chapter 3

Model-free prostate cancer characterization on VERDICT-MRI using deep learning

In this chapter we investigate the potential of model-free prostate lesion characterization on VERDICT MRI and examine whether raw VERDICT MRI allow for better classification of prostate lesions compared to the raw DW data and the ADC map from the mp-MRI acquisition. This chapter contains material from [26, 27], which were published at MLMI@MICCAI 2018 and ISMRM 2019.

3.1 Introduction

As we discussed in Sec. 2.3.1, VERDICT MRI is a microstructural imaging technique aiming to decode the information contained in DW images acquired with different diffusion weightings (b-values) and to derive microstructural features that allow prostate cancer characterization in-vivo [21, 6]. Currently, the standard procedure to perform prostate cancer characterization using advanced diffusion models, such as VERDICT MRI, is to fit biophysical models to the DW-MR signal to quantify and map microstructural tissue parameters that change with cancer. Biophysical models assume a simplified tissue structure and rely on numerical estimation of the DW-MR signal in such an environment. However, this approach has some limitations [108]: i) the models have to be simple enough for the fitting to work stably, ii) the models are handcrafted meaning that they may discard information in a sub-optimal way. Thus biophysical model, may not allow to fully exploit the rich information encoded in the DW-MR signal.

Several studies aim to address these limitations by relying on data-driven approaches and in particular neural networks. Golkov et al. trained a simple multi-layer perceptron to discriminate between several tissue types in the brain by using directly the DW images as inputs rather that using scalar tissue parameters obtained from model fitting [109]. This allowed them to fully exploit the unique information provided by the DW-MR signal without potential information loss due to model simplicity. The results of their work show that model-free diffusion MRI can be used to estimate arbitrary tissue properties in various settings where ground truth training datasets are available. In a follow-up work [110], they demonstrated that abnormality detection is also feasible without the requirement for labels. Their work uses raw DW images from a healthy population as reference and any deviation in the patient dataset from the healthy reference dataset can be detected using novelty detection methods. Despite the fact it eliminates the requirement for labeled data, this approach does not provide information regarding the underlying pathology causing changes to DW-MR signal.

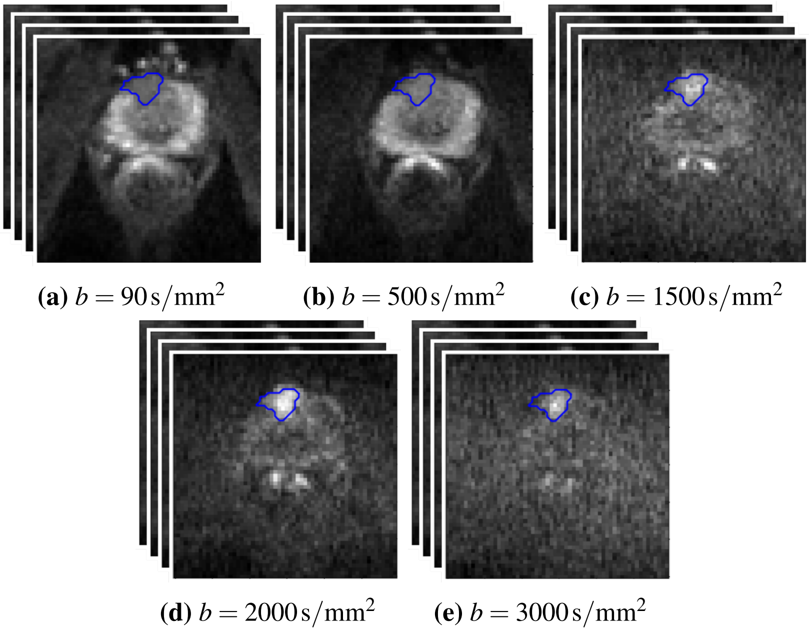
In addition, several recent studies rely on machine learning, and in particular deep learning, to estimate microstructural tissue parameters from the DW-MR signal [111, 112, 113, 114, 115]. The results indicate that deep learning can be used to avoid instabilities accompanying model fitting and reduce scan time since only a subset of the DW images contain relevant information.

In this chapter, we first aim to investigate the potential of model-free prostate lesion characterization using the raw DW-MR data from VERDICT MRI acquisitions. We rely on fully convolutional networks (FCNs) trained end-to-end using as input the raw DW images. Second, we examine whether raw VERDICT MRI allows for better classification of prostate lesions compared to the raw DW data and the ADC map from the mp-MRI acquisition.

3.2 Datasets

VERDICT MRI data: In this study we use VERDICT MRI data from 103 patients (median age, 62.2 years; range, 49.5–82.0 years)acquired as part of the INNOVATE clinical trial [116]. DW images (Fig. 3.1) were acquired with pulsed-gradient spin-echo sequence (PGSE) using an optimised imaging protocol for VERDICT prostate characterization with 5 b-values (90, 500, 1500, 2000, 3000 $\mathrm{s/mm^2}$) in 3 orthogonal directions, on a 3T scanner (Achieva, Philips Healthcare, NL) [117]. Also, images with $\mathrm{b} = 0\,\mathrm{s/mm^2}$ were acquired before each b-value acquisition. Compared to the naive DW-MRI from mp-MRI acquisitions, VERDICT-MRI has a richer acquisition protocol to probe the underlying microstructure and reveal changes in tissue features similar to histology. The DW-MRI sequence was acquired with a voxel size of $1.25\times1.25\times5\,\mathrm{mm^3}$, 5 mm slice thickness, 14 slices, a field of view of $220\times220\,\mathrm{mm^2}$ and the images were reconstructed to a 176×176 matrix size. The data was registered using rigid registration [118]. A dedicated radiologist highly experienced in prostate mp-MRI reporting (reporting more than 1000 scans per year) contoured malignant and benign lesions on the registered VERDICT MRI using mp-MRI for guidance. Lesions in Prostate Imaging Reporting and Data System (PI-RADS) category 3 are considered benign while lesions in PI-RADS category 4, 5 are considered malignant. In total, there are 134 lesions; 61 benign and 73 malignant. We note here that only index lesions (the largest lesion with the highest score) were annotated.

Standard DW data from mp-MRI: The DW-MRI data from the mp-MRI acquisition was acquired with diffusion-weighted echo-planar imaging sequence with 4 b-values (0, 150, 500, 1000, 2000 $\mathrm{s/mm^2}$). The DW data was acquired with the following imaging parameters: a repetition time msec/echo time msec, $2753/80$; field of view, 220×220 mm; section thickness, 5 mm; no intersection gap; acquisition matrix, 168×169mm. The ADC map was calculated by the scanner software. A dedicated radiologist highly experienced in prostate mp-MRI reporting (reporting more than 1000 scans per year) contoured malignant and benign lesions on DW images. Lesions in PI-RADS category 3 are considered benign while lesions in PI-RADS category 4, 5 are considered malignant. We note here that only index lesions were annotated.



(a) $b = 90\,\mathrm{s/mm^2}$ **(b)** $b = 500\,\mathrm{s/mm^2}$ **(c)** $b = 1500\,\mathrm{s/mm^2}$

(d) $b = 2000\,\mathrm{s/mm^2}$ **(e)** $b = 3000\,\mathrm{s/mm^2}$

Figure 3.1: VERDICT MRI data acquired with 5 b-values in 3 orthogonal directions. Malignant regions (noted in blue) are seen as a focus of high signal intensity on DW-MRI of $b = 2000, 3000\,\mathrm{s/mm^2}$ and as a focus of low signal intensity on the corresponding $b = 90\,\mathrm{s/mm^2}$ image.

3.3 Methods

We consider a dataset with paired image-label data: $\mathcal{D} = \{(x_i, y_i)\}, i \in [1, D]$, where $x_i \in \mathbb{R}^{H\times W\times 20}$ is a 20-channel DW image and $y_i \in \mathcal{L}^{H\times W}$ the corresponding labeling. We consider two classes (malignant, benign/normal) and consider a label set $\mathcal{L} = \{0, 1\}$ where 0 corresponds to benign/normal/background and 1 to malignant. We note here that only index lesions are annotated meaning that there might be areas that are mistakenly considered as normal. Nevertheless, evaluating the performance of a model in discriminating malignant lesions can still provide meaningful information. Our task is to train a model F that performs pixel-wise classification. We train the model using pixel-wise cross-entropy loss resulting in a training objective of the following form:

$$\mathcal{L}_{CE} = \sum_{(x_i,y_i)\in\mathcal{D}} \sum_{w,h} y_i \log(p_i^{(w,h,1)}) + (1-y_i)\log(p_i^{(w,h,0)}), \tag{3.1}$$

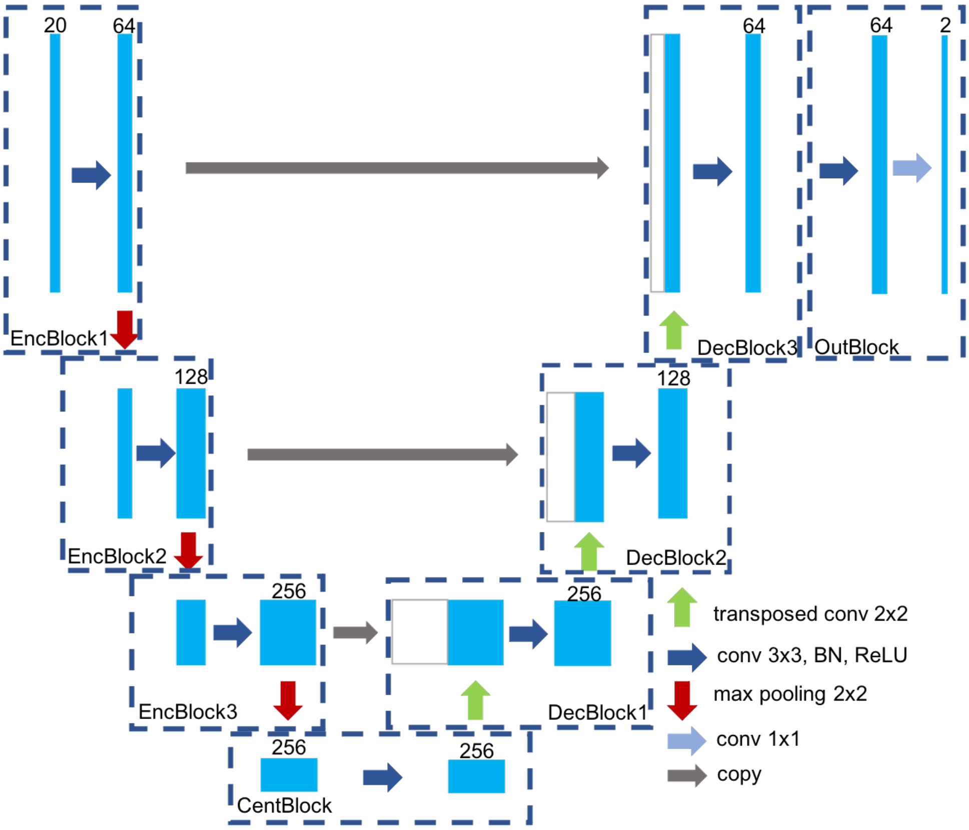
where $p_i = F(x_i)$ the softmax output of model F given the input image x_i.

We consider two different models F parameterized by FCNs trained end-to-end. FCNs have shown great success on pixel-wise classification tasks on both natural and medical images [119, 120, 121, 122]. In this study we consider U-Net [120] and ResNet [123] with an effective decoder module proposed in [122]; both architectures use an encoder-decoder structure. We modify both architectures and make several design choices to avoid overfitting that may arise due to the small dataset we have in our disposal. Specifically, we opt for shallower networks composed of a smaller number of layers and channels compared to the original architectures. The original models have a large number of trainable parameters that are sufficient to overfit a small training set. Reducing the number on parameters, we reduce the complexity of the model and thus avoid overfitting [124].

Below we provide a detailed description of the modified architectures we used in our experiments.

Network Architectures.

MRI-U-Net: The first model (MRI-U-Net) shown in Fig. 3.2 is based on the U-Net architecture proposed in [120]. U-Net consists of two main modules; an encoder module and a symmetric decoder module. MRI-U-Net has fewer convolutional layers to avoid overfitting. The encoder module is composed by 3 encoder blocks (EncBlock$_k$, $k = 1, \ldots 3$). Each encoder block consists of a convolution layer followed by batch normalization (BN) [125], a rectified-linear unit (ReLU) [126] and a 2x2 max pooling operation with stride 2. Each encoder block doubles the number of feature maps by applying 3x3 convolutions and halves the spatial dimension of the feature maps by applying maxpooling. The central module consists of a convolution layer followed by BN and a ReLU. The decoder module is composed by 3 decoder blocks (EncBlock$_k$, $k = 1, \ldots 3$). Each decoder block consists of a 2x2 transposed convolution with stride 2 to upsample the low resolution feature maps and a convolutional layer followed by BN and a ReLU. Concatenation of the upsampled feature maps with the corresponding encoder feature maps is performed before the convolutions. Each convolutional layer performs 2D convolutions of the



(a) Visual representation of MRI-U-Net architecture.

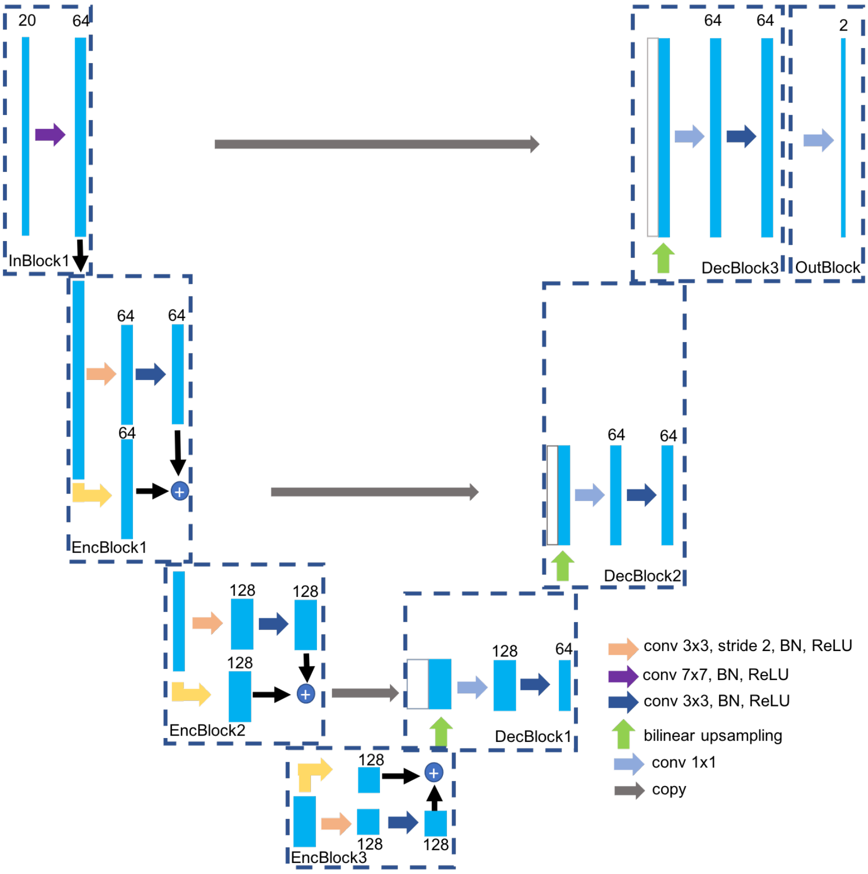
Block	layer	kernel size	# filters	stride	BN	ReLU	dropout
EncBlock1	conv1 pool1	3x3 2x2	64 n/a	1 2	yes no	yes no	no no
EncBlock2	conv2 pool2	3x3 2x2	128 n/a	1 2	yes no	yes no	no no
EncBlock3	conv3 pool3	3x3 2x2	256 n/a	1 2	yes no	yes no	no no
CentBlock	conv4	3x3	256	1	yes	yes	yes
DecBlock1	transpConv1 conv5	2x2 3x3	256 256	2 1	no yes	no yes	no yes
DecBlock2	transpConv2 conv6	2x2 3x3	128 128	2 1	no yes	no yes	no no
DecBlock3	convTransp3 conv7	2x2 3x3	64 64	2 1	no yes	no yes	no no
OutBlock	conv8	1x1	2	1	no	no	no

(b) Detailed description of each layer in MRI-U-Net.

Figure 3.2: MRI-UNet. MRI-U-Net consists of two main modules; an encoder module and a symmetric decoder module. The encoder module is composed by 3 encoder blocks. Each encoder block consists of a convolution layer followed by a 2x2 maxpooling operation with stride 2. The central module consists of a convolution layer. The decoder module is composed by 3 decoder blocks. Each decoder block consists of a 2x2 transposed convolution with stride 2 to upsample the low resolution feature maps and a convolutional layer to refine the feature maps. Concatenation of the upsampled feature maps with the corresponding encoder feature maps is performed before the convolutions. The last block (OutBlock) consists of a convolutional layer that performs 1x1 convolutions to map the input to the desired number of classes. To prevent overfitting, we apply dropout to layers conv4, conv5.

input with 3x3 kernels to halve the number of features maps. The last block (OutBlock) consists of a convolutional layer that performs 1x1 convolutions to map the input to the desired number of classes. As in previous studies [127, 128], we apply dropout [129] on layers conv4, conv5 that have a large number of parameters to prevent feature coadaptation and overfitting.

MRI-ResNet: The second network (MRI-ResNet) has also an encoder-decoder structure as shown in Fig. 3.3. The encoder module is a modified version of ResNet-18 proposed in [123]. We remove the maxpooling layer in the beginning of the network and the global average pooling layer at the end of the network. We also replace the last fully-connected layer with a convolutional layer and decrease the number of convolutional layers. The encoder module is composed by an input block (InBlock) and 3 encoder blocks (EncBlock$_k$, $k = 1, \ldots 3$). The input block consists of a 7x7 convolution allowing for a larger receptive field. Each encoder block consists of 2 convolutional layers. The first convolutional layer of each encoder block halves the spatial dimension of the feature maps by applying 3x3 convolutions with stride 2. Shortcut connections, which has been proposed to address the degradation problem in deep convolutional neural networks [123], are added to each pair of convolutional layers in each encoder block. Specifically, given x, the input of an encoder block, the output y is given by $y = \mathcal{F}(x) + G(x)$, where $\mathcal{F}$ represents multiple convolutional layers (3x3 convolutions) and G represents the projection shortcut used to match dimensions (done by 1x1 convolutions with stride 2). All the convolutional layers are followed by BN and a ReLU apart from cases where residual connections are considered; in those cases ReLU is applied after the addition. As in previous studies [127, 128], we apply dropout [129] on layers conv6, conv7 that have a large number of parameters to prevent feature coadaptation and overfitting. The decoder module is similar to the one proposed in [122] and has 3 decoder blocks (DecBlock$_k$, $k = 1, \ldots 3$). Each decoder block consists of a bilinear upsampling and two convolutional layers followed by BN and a ReLU. Bilinear upsampling upsamples the low resolution feature maps by a factor of 2. The upsampled features maps are then concatenated with the corresponding encoder feature maps. Both convolutional layers



(a) Visual representation of MRI-U-Net architecture.

Block	layer	kernel size	# filters	stride	BN	ReLU	dropout	upsampling factor
InBlock	conv1	7x7	64	1	yes	yes	no	-
EncBlock1	conv2	3x3	64	2	yes	yes	no	-
	conv3	3x3	64	1	yes	no	no	-
	conv_proj1	1x1	64	2	yes	no	no	-
EncBlock2	conv4	3x3	128	2	yes	yes	no	-
	conv5	3x3	128	1	yes	no	no	-
	conv_proj2	1x1	128	2	yes	no	no	-
EncBlock3	conv6	3x3	128	2	yes	yes	yes	-
	conv7	3x3	128	1	yes	no	yes	-
	conv_proj3	1x1	128	2	yes	no	no	-
DecBlock1	bilUp	-	-	-	-	-	-	2
	conv8	1x1	128	1	yes	yes	no	-
	conv9	3x3	64	1	yes	yes	no	-
DecBlock2	bilUp	-	-	-	-	-	-	2
	conv10	1x1	64	1	yes	yes	no	-
	conv11	3x3	64	1	yes	yes	no	-
DecBlock3	bilUp	-	-	-	-	-	-	2
	conv12	1x1	64	1	yes	yes	no	-
	conv13	3x3	64	1	yes	yes	no	-
InBlock	conv14	7x7	64	1	yes	yes	no	-

(b) Detailed description of each layer in MRI-U-Net.

Figure 3.3: MRI-UNet.The encoder module is composed by an input block (InBlock) and 3 encoder blocks. The input block consists of a 7x7 convolution allowing for a larger receptive field. Each encoder block consists of 2 convolutional layers. Shortcut connections are added to each pair of convolutional layers in each encoder block; 1x1 convolutions with stride 2 are used to match dimensions. The decoder module has 3 decoder blocks. Each decoder block consists of a bilinear upsapmling and two convolutional layers. The upsampled features maps are concatenated with the corresponding encoder feature maps. The convolutional layers halve and refine the number of features maps by performing 1x1 and 3x3 convolutions respectively. We apply dropout to layers conv6, conv7 to prevent overfitting.

halve the number of features maps by performing 1x1 and 3x3 convolutions respectively. We first apply 1x1 convolutions to reduce the number of parameters of our model since the concatenated features usually contain a large number of channels.

Training settings.

We implement both networks using Pytorch [130]. We employ a 10-fold cross validation approach to train and test the networks. We repeat each 10-fold cross validation 5 times. We train the networks for 200 epochs and select the parameters (i.e., number of layers, batch size, learning rate, weight decay, dropout rate) that has the smallest loss on a validation set (20% of the training set). We use stochastic gradient descent (SGD) with a mini-batch size of 32, a constant learning rate of 1e-5, a momentum of 0.9 and a weight decay of 1e-3. We employ dropout as a regularization strategy with dropout rate 0.5.

Evaluation metrics.

We evaluate the binary pixel-wise classification using average sensitivity, specificity, area under the receiver operating characteristic (ROC) curve (AUC) and precision. Sensitivity, specificity and precision are defined as

- sensitivity $= \frac{\text{TP}}{\text{P}}$, where TP is the number of true positive pixels and P is the number of positive pixels.
- specificity $= \frac{\text{TN}}{\text{N}}$, where TN is the number of true negative pixels and N is the number of negative pixels.
- precision $= \frac{\text{TP}}{\text{FP}+\text{TP}}$, where FP is the number of false positive pixels.

3.4 Results

3.4.1 Model-free prostate lesion characterization

We perform three different experiments and report the results. In each experiment we train the models considering different parts of the image during training. In the 1st experiment we train the models using predefined malignant and benign/normal ROIs. This allows us to avoid using areas of the image that belong to malignant class and have not been annotated; as we mentioned earlier only index lesions were

Networks	Regions	AUC		sensitivity		specificity		precision	
		Slice-level	Subject-level	Slice-level	Subject-level	Slice-level	Subject-level	Slice-level	Subject-level
MRI-UNet	malignant vs benign ROIs	85.7 (±9.2)	83.4(±10.1)	75.7(±9.1)	73.2(±9.9)	75.4(±6.1)	73.2(±6.9)	86.2(±10.3)	83.1(11.8)
	malignant vs all	74.1 (±10.5)	72.9(±10.2)	75.6(±10.4)	74.8(±11.0)	47.6(±7.4)	42.1(±8.0)	2.0(±11.1)	1.5(12.2)
MRI-ResNet	malignant vs benign ROIs	86.7 (±10.8)	84.3(±11.9)	71.8(±9.6)	69.7(±10.1)	83.3(±6.6)	81.9(±7.1)	90.0(±11.2)	88.4(12.4)
	malignant vs all	71.9 (±11.1)	69.8(±12.0)	71.6(±9.9)	70.8(±10.8)	57.1(±6.9)	55.5(±7.8)	4.7(±11.9)	2.9(12.7)

Table 3.1: 1st experiment: Average AUC, sensitivity, specificity, precision of MRI-UNet and MRI-ResNet when evaluation is performed on i) malignant and benign regions of interest (ROIs) (malignant vs benign ROIs) and ii) the entire image (malignant vs all). Using MRI-ResNet results in slightly improved performance (AUC of 86.7%) when we compare the performance on predefined ROIs. When we evaluate the performance on the entire image, we observe that it drops significantly especially in terms of specificity and precision. This is normal since the models have been trained only on a subset of the entire image and therefore they encounter unseen samples at inference time. In addition, we observe that the two networks behave differently for different metrics and evaluation protocols.

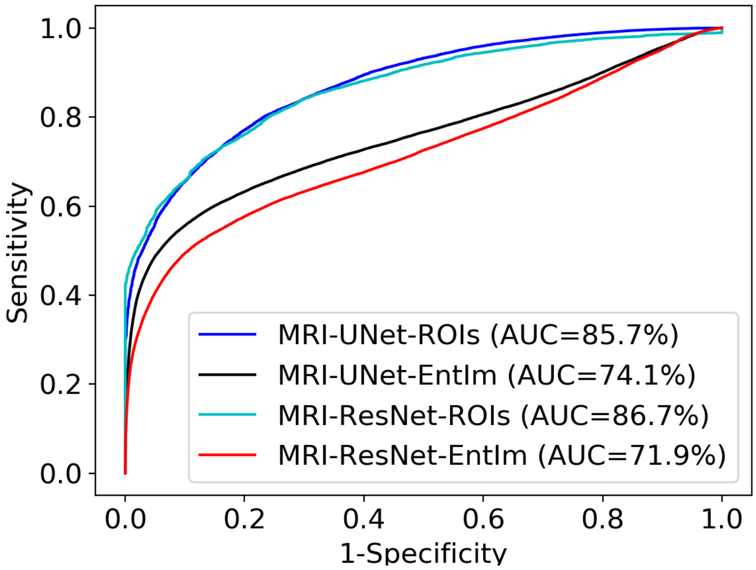


Figure 3.4: 1st experiment: Receiver operating characteristic (ROC) curves of MRI-U-Net and MRI-ResNet when evaluation is performed on i) predefined malignant and benign regions of interest (ROIs) and ii) the entire image (EntIm).Using MRI-ResNet results in slightly improved performance (AUC of 86.7%) when we compare the performance on predefined ROIs.

annotated. In the 2nd experiment we randomly sample and use normal/background ROIs during training; this allows us increase the number of negative labelled ROIs and to improve performance. In the 3rd experiment we train the models using the entire image as input.

1st experiment. In the first experiment we train the networks on predefined malignant and benign/normal ROIs and ignore the rest of the pixels. The main objective of this experiment is to evaluate the ability of the models to discriminate between malignant and benign lesions. Table 3.1 shows the performance of the two networks on i) predefined malignant and benign ROIs (malignant vs benign ROIs) and ii) the entire image (malignant vs all) at slice-level and subject-level. We observe that slice-level and subject-level results are similar - the subject-level performance is slightly lower. In the second case regions which are not labelled as malignant are considered as benign/normal. Fig. 3.4 shows the receiver operating characteristic (ROC) curves of MRI-U-Net and MRI-ResNet when evaluation is performed on i) predefined malignant and benign ROIs and ii) the entire image. Using MRI-ResNet results in slightly improved performance (AUC of 86.7%) when we compare the performance on predefined ROIs. When we evaluate the performance of the two networks in the entire image we observe that it drops significantly especially in terms of specificity and precision. This is normal since the models have been trained only on a subset of the entire image and therefore they encounter unseen samples at inference time. Finally, we observe that the two networks behave differently for different metrics and evaluation protocols. For instance, MRI-ResNet results in slightly improved performance (AUC of 86.7%) when we compare the performance on predefined ROIs while MRI-U-Net yield to better performance when we evaluate the models on the entire image. In addition, MRI-U-Net has better sensitivity while MRI-ResNet has better specificity and precision.

2nd experiment. In the previous experiment we consider only malignant and benign ROIs during training. However, at inference time we need to evaluate the performance on the entire image. This means that the network is asked to classify normal prostate tissue and background voxels that it has not seen during train-

Networks	Regions	AUC		sensitivity		specificity		precision	
		Slice-level	Subject-level	Slice-level	Subject-level	Slice-level	Subject-level	Slice-level	Subject-level
MRI-UNet	malignant vs benign ROIs	89.0 (±8.9)	85.3(±9.2)	82.9(±8.8)	79.9(±9.1)	77.9(±5.4)	75.1(±6.2)	88.8(±9.5)	85.4(10.3)
	malignant vs all	94.2 (±9.8)	92.1(±9.9)	82.7(±8.9)	80.2(±10.2)	91.2(±6.2)	89.8(±6.9)	13.4(±10.3)	10.2(11.8)
MRI-ResNet	malignant vs benign ROIs	87.6 (±10.3)	84.8(±10.8)	86.4(±8.2)	83.5(±9.8)	72.6(±5.9)	70.4(±7.0)	86.7(±10.0)	84.8 (10.1)
	malignant vs all	94.0 (±10.9)	91.4(±11.4)	86.4(±9.4)	84.3(±10.1)	88.8(±6.0)	85.2(±7.4)	11.2(±10.9)	5.8(10.3)

Table 3.2: 2nd experiment: Average AUC, sensitivity, specificity, precision of MRI-UNet and MRI-ResNet when evaluation is performed on i) malignant and benign/normal/background ROIs (malignant vs benign/normal/background ROIs) and ii) the entire image (malignant vs all) when we use additional negative labelled ROIs. Using MRI-U-Net results in slightly improved performance(AUC of 89.7%) when we compare the performance on predefined ROIs. Using additional negative labelled ROIs improves performance when classification is performed on the entire image. However, precision remains still low when we evaluate on the entire image. Regarding the performance of the two models we observe that MRI-U-Net performs better in terms of specificity and precision while MRI-ResNet performs better in terms of sensitivity.

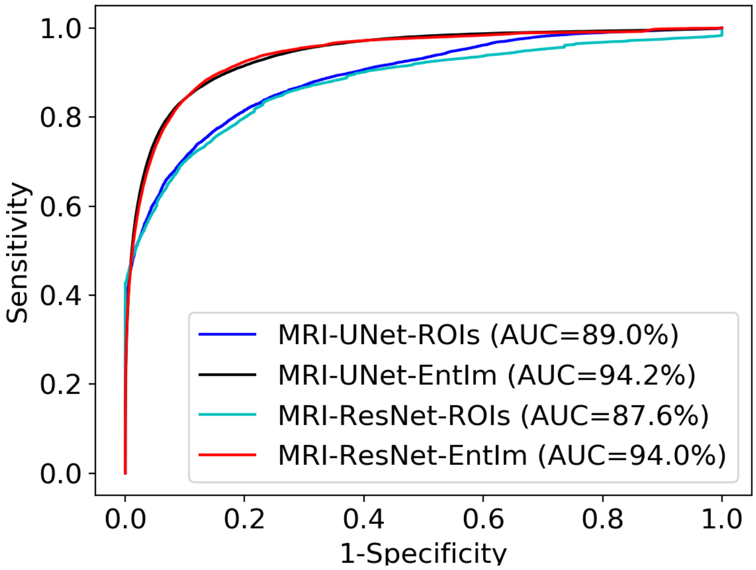


Figure 3.5: 2nd experiment. Receiver operating characteristic (ROC) curves of MRI-UNet and MRI-ResNet when evaluation is performed on i) predefined malignant and benign/normal/background ROIs and ii) the entire image (EntIm). Using MRI-U-Net results in slightly improved performance (AUC of 89.0%) when we compare the performance on predefined ROIs. We observe no difference when we evaluate the performance on the entire image.

ing. To address this issue, we increase the number of negative labelled ROIs by randomly selecting and using normal/background ROIs during training. As in the previous experiment we consider two classes; malignant (positive class) and benign/normal/background (negative class). Table 3.2 shows the classification performance of the networks on i) predefined malignant and benign/normal/background

ROIs (malignant vs benign/normal/background ROIs) and ii) the entire image (malignant vs all) at slice-level and subject-level. We observe that slice-level and subject-level results are similar - the subject-level performance is slightly lower. Fig. 3.5 shows the ROC curves of MRI-UNet and MRI-ResNet when evaluation is performed on i) predefined malignant and benign/normal/background ROIs and ii) the entire image. Using additional negative labelled ROIs improves performance when classification is performed on the predefined regions or the entire image. However, precision remains still low when we evaluate on the entire image meaning that a large proportion of pixels in negative class are classified as positive. Regarding the performance of the two models we observe that MRI-U-Net performs better in terms of specificity and precision while MRI-ResNet performs better in terms of sensitivity.

Networks	Regions	AUC	sensitivity	specificity	precision
MRI-U-Net	malignant vs all	85.0%	52.4%	97.8%	58.2%
MRI-ResNet	malignant vs all	89.2%	55.1%	97.2%	55.4%

Table 3.3: 3rd experiment: Average AUC, sensitivity, specificity, precision of MRI-UNet and MRI-ResNet when training and evaluation is performed only on the entire image. We observe that MRI-U-Net behaves better in terms of precision while MRI-ResNet behaves better in terms of AUC and sensitivity.

3rd experiment. Finally, we train the two networks using the entire image instead of using predefined ROIs. The main objective of this experiment is to evaluate the ability of the models to discriminate between malignant and benign lesions/background. Accurate automatic classification of malignant lesions could assist and accelerate the accurate radiological interpretation of VERDICT-MRI data. As in the previous experiments, we consider two different classes; malignant (positive class) and benign/normal/background (negative class). The only difference with the previous experiment (2nd experiment) is that during training we consider all the voxel corresponding to normal prostate tissue and background instead of random sampling ROIs. Table 3.3 shows the classification performance of the networks when they are trained and evaluated on the entire image. We observe that MRI-U-Net behaves better in terms of precision while MRI-ResNet behaves better in terms

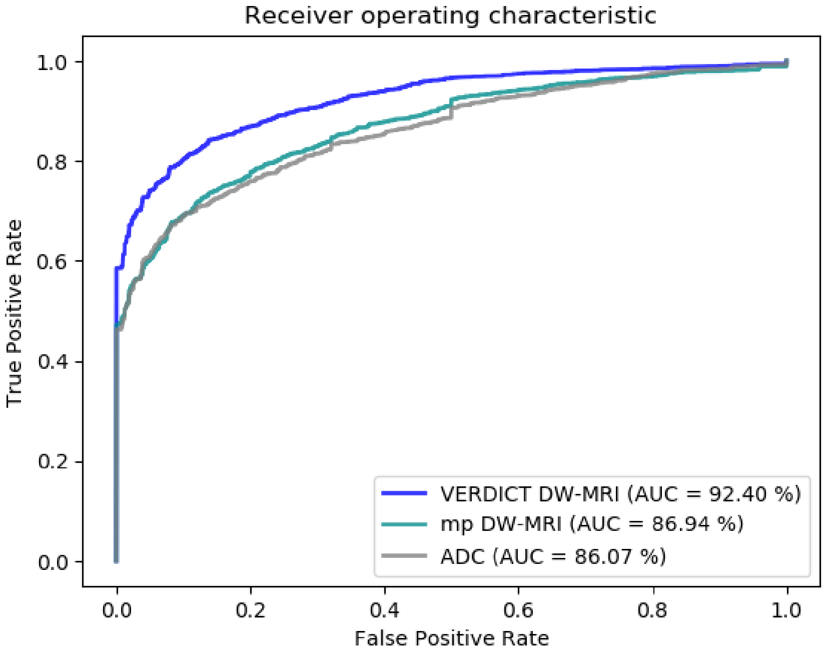


Figure 3.6: Receiver operating characteristic (ROC) curves of MRI-UNet when training and evaluation is performed on the raw VERDICT MRI data, the ADC map and the raw DW-MRI data from the mp-MRI acquisition. MRI-U-Net achieves better performance (area under the curve (AUC) of 92.40%) on VERDICT MRI data. When the ADC map and the raw DW-MRI data from the mp-MRI acquisition are used, it achieves an AUC of 86.07% and 86.94% respectively.

of AUC and sensitivity. In addition, compared to the previous experiment (2nd experiment), we observe that sensitivity drops significantly while precision increases. The drop in sensitivity might be related to the fact that only the index lesions are annotated meaning that may exist malignant ROIs that are mistakenly considered as normal/background during training and evaluation.

3.4.2 VERDICT MRI vs standard DW-MRI imaging from mp-MRI

In this section we examine whether raw VERDICT MRI allows for better classification of prostate lesions compared to the naive DW imaging and the ADC map from the mp-MRI acquisition. We note here that we use only a subset of the patients, in particular 18, for whom we have paired VERDICT MRI and mp-MRI data as well as annotations. In this experiment we use only MRI-U-Net. Since the number of patients we have in our disposal is smaller compared to the previous experiments

(Sec. 3.4.1), we reduce the number of layers of the network to avoid overfitting. We train and evaluate the network on labelled malignant and benign ROIs. Fig. 3.6 shows the receiver operating characteristic (ROC) curves of MRI-U-Net when training and test is performed on the raw VERDICT MRI data, the ADC map and the raw DW-MRI data from the mp-MRI acquisition. The results show that MRI-U-Net achieves better performance (AUC of 92.40%) on VERDICT MRI data. When the ADC map and the raw DW-MRI data from the mp-MRI acquisition are used, it achieves an AUC of 86.07% and 86.94% respectively. This indicates that VERDICT MRI may encode richer information allowing for better discrimination of the different types of lesions. However, training and evaluation is performed on a very small number of subjects and thus further analysis is required.

3.5 Conclusion

In this chapter we investigate the potential of model-free prostate lesion classification on the raw VERDICT MRI data using FCNs and we examine whether raw VERDICT MRI allows for better classification of prostate lesions compared to the raw DW data and the ADC map from the mp-MRI acquisition. To this end, we adapt and evaluate two FCNs (MRI-U-Net, MRI-ResNet) architectures. Previous studies that are based on mp-MRI data to provide an automated solution for prostate lesion classification use DW data from mp-MRI acquisitions. In this work, we use DW data from VERDICT MRI that has a richer acquisition protocol compared to mp-MRI; VERDICT MRI data is acquired for 5 b-values in 3 orthogonal directions.

Our preliminary results indicate that i) FCNs trained on VERDICT MRI achieve good performance in differentiating between malignant and benign lesions and (Sec. 3.4.1) ii) FCNs trained and evaluated on VERDICT MRI perform better than FCNs trained and evaluated on the naive DW data and the ADC from mp-MRI acquisitions (Sec. 3.4.2). However, the current work has some important limitations that have to be addressed in the future.

Firstly, our current work relies on annotations of index lesions only, not allowing us to train accurate models. Incorrectly labeled data hinder the generalisation of the discriminate models – labeling errors may be memorised leading to undesired biases [131, 132] – and have detrimental effects on the validity of model evaluation, potentially leading to incorrect conclusions. Thus, it is important that we obtain annotations for the all the lesions in the entire volume. Future work will focus on collecting a dataset where all lesions are annotated. We also note here that annotating large-scale is a time consuming task and requires expertise; only radiologists with experience in prostate mp-MRI reporting are able to accurately annotate the lesions. Thus, obtaining large labeled datasets, especially for new imaging techniques, is not always feasible. Future work will also focus on domain adaptation that allows to mitigate this problem by exploiting training samples from an existing, densely-annotated domain within a novel, sparsely-annotated domain.

Second, as we mentioned in Sec. 3.2, our current work relies on labels corresponding to PI-RADS scores. However, the ultimate goal is to accurate differentiate between clinically significant and non-clinically significant prostate cancer; clinically significant cancer is defined as Gleason score of 7 or greater. Recent studies [92, 95, 133, 134, 135, 136, 94] focus on the development of deep learning methods for discriminating between clinically significant and non-clinically significant cancer on mp-MRI where biopsy-based [92, 95, 133, 134, 135] or prostatectomy-based [136, 94] annotations serve as ground truth. The results of these studies indicate that deep learning models achieve high diagnostic accuracy. Future work will focus on obtaining biopsy-based ground for pairs of VERDICT MRI and DW-MRI data from mp-MRI acquisitions; this will allow to examine more thoroughly whether FCNs that rely on VERDICT MRI can better discriminate between clinically significant and non-clinically significant cancer than models that rely on the naive DW data from mp-MRI acquisitions.

Chapter 4

Semi-supervised domain adaptation for lesion segmentation

As we discussed in Sec. 3.5, accurately annotating large training datasets is a challenging task impeding the adoption of novel imaging modalities for learning-based medical image analysis. In this chapter we propose a semi-supervised domain adaptation method to address this problem. We demonstrate the effectiveness of our approach on VERDICT MRI; however, it is quite general can be applied in other applications where the amount of labeled training data for the domain of interest is limited. Our code is publicly available at `https://github.com/elchiou/DA`. This chapter contains material from [28, 29], which were published at MICCAI 2020 and at ISMRM 2020.

4.1 Introduction

Domain adaptation can be used to exploit training samples from an existing, densely-annotated domain within a novel, sparsely-annotated domain, by bridging the differences between the two domains. This can facilitate the training of powerful convolutional neural networks (CNNs) for novel medical imaging modalities or acquisition protocols, effectively compensating for the limited amount of training data available to train CNNs in the new domain.

The assumption underlying most domain adaptation methods is that one can align the two domains either by extracting domain-invariant representations (features), or by establishing a ‘translation’ between the two domains at the signal level, where in any domain the ‘resident’ and the translated signals are statistically indistinguishable.

In particular for medical imaging, [137] and [138] rely on adversarial training to align the feature distributions between the source and the target domain for medical image classification and segmentation respectively. Pixel-level distribution alignment is performed by [17, 18, 19, 20], who use CycleGAN [139] to map source domain images to the style of the target domain; they further combine the translation network with a task-specific loss to penalize semantic inconsistency between the source and the synthetic images. The synthetic images are used to train models for image segmentation in the target domain. Ouyang et al. [140] perform adversarial training to learn a shared, domain-invariant latent space which is exploited during segmentation. They show that their approach is effective in cases where target-domain data is scarce. Similarly, [141] embed the input images from both domains onto a domain-specific style space and a shared content space. Then, they use the content-only images to train a segmentation model that operates well in both domains. However, their approach does not necessarily preserve crucial semantic information in the content-only images.

These methods rely on the strong assumption that the two domains can be aligned. However, recent works on the closely related problem of unsupervised image translation [142, 143, 144] have highlighted that this is a strong assumption and is frequently violated in practice. As a natural image example, an image taken at night can have many day-time counterparts, revealed by light; similarly in medical imaging, a different imaging protocol can reveal structures that had previously passed unnoticed. In technical terms, the translation can be one-to-many, or, stated in probabilistic terms, multi-modal [142, 143, 144]. In particular, we consider the source domain $\mathcal{X}_S$ which is to be mapped to fthe target domain $\mathcal{X}_T$). We are given samples that are drawn from the two marginal distributions $P(\mathcal{X}_S)$ and $P(\mathcal{X}_T)$ and the translation generates a diverse set of output $\mathcal{X}_{S\to T}$, corresponding to different modes in the distribution $P(\mathcal{X}_T|\mathcal{X}_S)$. Using a one-to-one translation network in such

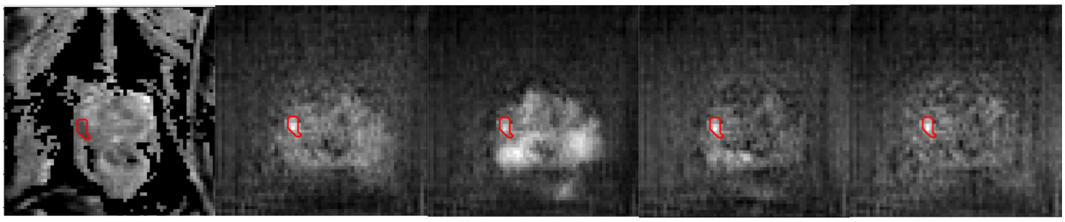


Figure 4.1: One-to-many mapping from one mp-MRI image (left) to four VERDICT MRI translations: our network can generate samples with both local and global structure variation, while at the same time preserving the critical structure corresponding to the prostate lesion, shown as a red circle. We note that the lesion area is annotated by a physician on the leftmost image, but is not used as input to the translation network - instead the translation network learns to preserve lesion structures thanks to the end-to-end discriminative training (details in text).

a setting can harm performance, since the translation may predict the mean of the underlying multi-modal distribution, rather than provide diverse, realistic samples from it.

In our work we accommodate the inherent uncertainty in the cross-domain mapping and, as shown in Figure 4.1, generate multiple outputs conditioned on a single input, thereby allowing for better generalization of the segmentation network in the target domain. As in recent studies [17, 18, 19, 20], we use GANs [145] to align the source and target domains, but go beyond their one-to-one, deterministic mapping approaches. In addition, inspired by [17, 18, 19, 15], we enforce semantic consistency between the real and synthetic images by exploiting source-domain lesion segmentation supervision to train target-domain networks operating on the synthetic images. This results in training networks that can generate diverse outputs while at the same time preserving critical structures - such as the lesion area in Figure 4.1. We further accommodate the statistical discrepancies between real and synthetic data by introducing residual adapters (RAs) [146, 29] in the segmentation network. These capture domain-specific properties and allow the segmentation network to generalize better across the two domains.

We demonstrate the effectiveness of our approach in prostate lesion segmentation on VERDICT MRI. As we discussed in Sec. 3.4.1 and Sec. 3.5, annotating large amounts of training data for the adoption of VERDICT MRI for learning-based medical image analysis is challenging due to labor and expertise required.

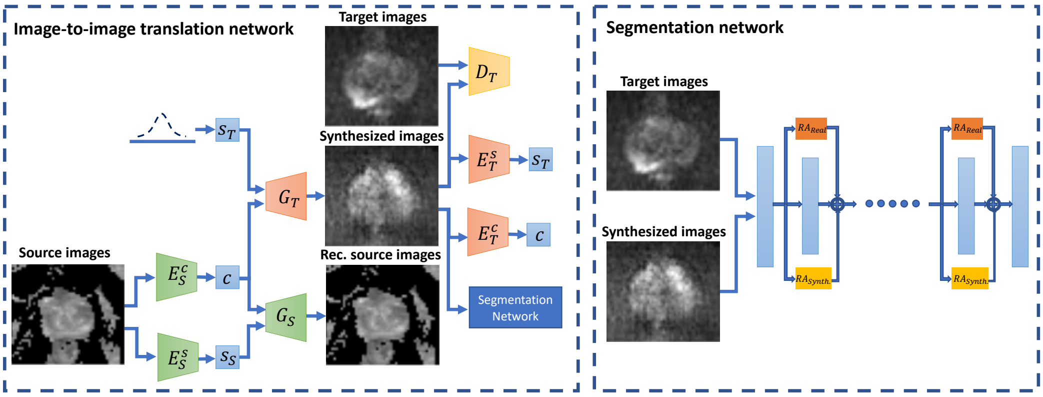


Figure 4.2: Overview of our domain adaptation framework: we train a noise-driven domain translation network in tandem with a discriminatively supervised segmentation network in the target domain; GAN-type losses align the translated samples with the target distribution, while residual adapters allow the segmentation network to compensate for remaining discrepancies.

On the other hand, large scale, labeled mp-MRI datasets exist [93, 41]. As shown experimentally, our approach largely improves the generalization capabilities of a lesion segmentation model on VERDICT MRI by exploiting label DW-MRI data from mp-MRI acquisitions.

4.2 Method

Our approach relies on a unified network for cross-modal image synthesis and segmentation, that is trained end-to-end with a combination of objective functions. As shown in Fig. 4.2, at the core of this network is an image-to-image translation network that maps images from the source ('S') to the target ('T') domain. The translation network is trained in tandem with a segmentation network that operates in the target domain, and is trained with both the synthetic and the few real annotated target-domain images. Beyond these standard components, our approach relies on three additional components: firstly, we sample a latent variable from a Gaussian distribution when translating to the target domain; this represents structures that cannot be accounted by a deterministic mapping, and can result in one-to-many translation when needed. Secondly, we introduce residual adapters (RAs) to a common backbone network for semantic segmentation, allowing the discriminative training to accommodate any remaining discrepancies between the real and

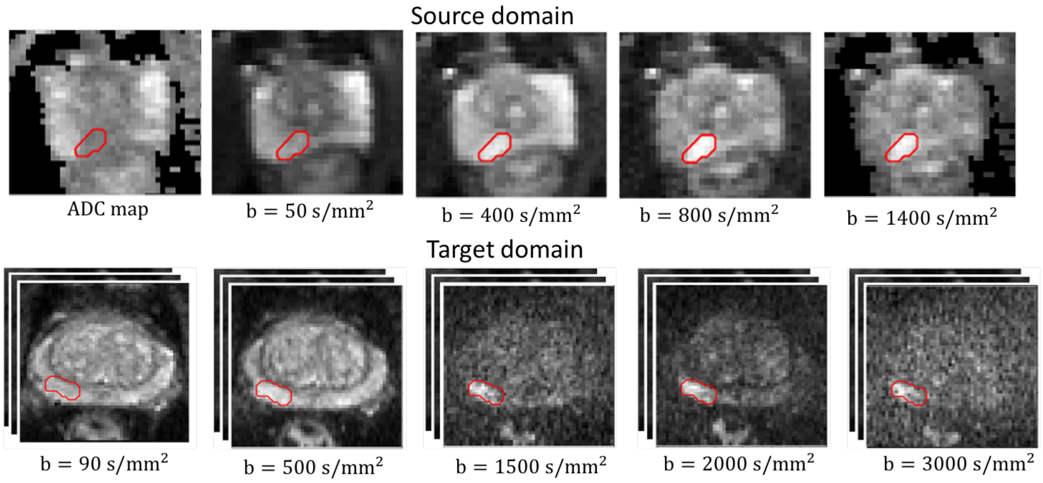


Figure 4.3: Example of the data coming from the source and the target domains: standard DW-MRI (source domain) consists of 5 input channels (4 b-values and the ADC map) while VERDICT MRI consists of 15 input channels (5 b-values in 3 orthogonal directions)

synthetic target domain images. Finally, we use a dual translation network from the target to the source domain, allowing us to use cycle-consistency in domain adaptation [147, 139, 143]; the cycle constraint allows us to disentangle the deterministic, transferable part from the stochastic, non-transferable part, which is filled in by Gaussian sampling, as mentioned earlier.

Problem formulation. Having provided a broad outline of our method, we now turn to a more detailed technical description. We consider the problem of domain adaptation in prostate lesion segmentation. We assume that the source domain, $\mathcal{X}_S$, contains N_S images, $x_S \in \mathcal{X}_S$, with associated segmentation masks, $y_S \in \mathcal{Y}_S$. Similarly, the sparsely labeled target domain, $\mathcal{X}_T$, consists of N_T images, $x_T \in \mathcal{X}_T$. A subset $\tilde{\mathcal{X}}_T$ of $\mathcal{X}_T$ comes with associated segmentation masks, $y_T \in \mathcal{Y}_T$. Our goal is to train a model that provides accurate predictions in the target domain. As shown in Fig. 4.3, there is a substantial domain gap between the two domains precluding the naive approach of training a network on the source domain and then deploying it in the target domain. The proposed framework consists of two main components, i.e. an image-to-image translation network and a segmentation network described below.

Segmentation Network

The segmentation network (Fig. 4.2), *Seg*, operates on image-label pairs of both real, $\mathcal{X}_T$, and synthetic data, $\mathcal{X}_{S\to T}$, translated from source to target. An encoder-decoder network [122, 120] is the main backbone which serves both domains. To compensate further for differences in the feature statistics of real and synthetic data we install residual adapter modules [146] in parallel to each of the convolutional layer of the backbone. Introducing residual adapters ensures that most of the parameters stay the same with the network, but also that the new unit introduces a small, but effective modification that accommodates the remaining statistical discrepancies of the two domains.

More formally, let l be a convolutional layer in the segmentation network and $F^l \in \mathbb{R}^{k\times k\times C_i\times C_o}$ be a set of filters for that layer, where $k\times k$ is the kernel size and C_i, C_o are the number of input and output feature channels respectively. Let also $Z^l_j \in \mathbb{R}^{1\times 1\times C_i\times C_o}$ be a set of domain-specific residual adapter filters of domain j, where $j \in \{1,2\}$, installed in parallel with the existing set of filters F_l. Given an input tensor $x_l \in \mathbb{R}^{H\times W\times C_i}$, the output $y_l \in \mathbb{R}^{H\times W\times C_o}$ of layer l is given by

$$y_l = F^l * x_l + Z^l_j * x_l. \tag{4.1}$$

We train the segmentation network by optimizing the following objective

$$\begin{aligned}&\mathcal{L}_{Seg}(Seg, \tilde{\mathcal{X}_T}, \mathcal{Y}_T, \mathcal{X}_{S\to T}, \mathcal{Y}_S) = \\ &\qquad \mathcal{L}_{DSC}(Seg, \tilde{\mathcal{X}_T}, \mathcal{Y}_T) + \mathcal{L}_{DSC}(Seg, \mathcal{X}_{S\to T}, \mathcal{Y}_S).\end{aligned} \tag{4.2}$$

The dice loss, $\mathcal{L}_{DSC}$, based on dice coefficient, is given by

$$\mathcal{L}_{DSC}(Seg, \mathcal{X}, \mathcal{Y}) = -\frac{2\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})}\sum_{k=1}^{K} p_k y_k}{\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})}\sum_{k=1}^{K}(p_k^2 + y_k^2)}, \tag{4.3}$$

where K the number of voxels in the input images and $p = Seg(x)$, the softmax output of the segmentation network. We adopt this objective function since it is a differentiable approximation of a criterion that is well-adapted to our task [148, 149].

Stochastic Translation Network

Recently, several studies [142, 143] have pointed out that cross-domain mapping is inherently multi-modal and proposed approaches to produce multiple outputs conditioned on a single input. Here we use MUNIT [143] to illustrate the key idea. As it is illustrated in Figure 4.2 the image-to-image translation network consists of content encoders E^c_S, E^c_T, style encoders E^s_S, E^s_T, generators G_S, G_T and domain discriminators D_S, D_T for both domains. The content encoders E^c_S, E^c_T map images from the two domains onto a domain-invariant content space $\mathcal{C}$ ($E^c_S : \mathcal{X}_S \rightarrow \mathcal{C}$, $E^c_T : \mathcal{X}_T \rightarrow \mathcal{C}$) and the style encoders E^s_S, E^s_T map the images onto domain-specific style spaces $\mathcal{S}_S$ ($E^s_S : \mathcal{X}_S \rightarrow \mathcal{S}_S$) and $\mathcal{S}_T$ ($E^s_T : \mathcal{X}_T \rightarrow \mathcal{S}_T$). The content code can be understood as the underlying anatomy which is the information that we want transfer during the translation while the style codes capture information related to the imaging modalities. Image-to-image translation is performed by combining the content code extracted from a given input and a random style code sampled from the target-style space. For instance, to translate an image $x_S \in \mathcal{X}_S$ to $\mathcal{X}_T$ we first extract its content code $c = E^c_S(x_S)$. The generator G_T uses the extracted content code c and a randomly drawn style code $s_T \in \mathcal{S}_T$ to produce the final output $x_{S\rightarrow T} = G_T(c, s_T)$. By sampling random style codes from the style spaces $\mathcal{S}_S$ and $\mathcal{S}_T$ the generators G_S and G_T are able to produce diverse outputs. We train the networks with a loss function that consists of domain adversarial, self-reconstruction, latent reconstruction, cycle-consistency and segmentation losses.

Domain adversarial loss. We utilize GANs to match the distribution between the synthetic and the real images of the two domains. The adversarial discriminators D_T, D_S aim at discriminating between real and synthetic images, while the generators G_T, G_S aim at generating realistic images that fool the discriminators. For G_T and D_T the GAN loss is defined as

$$\begin{split}\mathcal{L}^T_{GAN}(E^c_S, G_T, D_T, \mathcal{S}_T, \mathcal{X}_S) = \\ \mathbb{E}_{x_S\sim\mathcal{X}_S, s_T\sim\mathcal{S}_T}[\log(1 - D_T(G_T(E^c_S(x_S), s_T)))] + \mathbb{E}_{x_T\sim\mathcal{X}_T}[\log(D_T(x_T))].\end{split} \tag{4.4}$$

Self-reconstruction loss. Given the encoded content and style codes of a source-domain image the generator G_S should be able to decode them back to the original one.

$$\mathcal{L}^{S}_{recon}(G_S, E^s_S, E^c_S, \mathcal{X}_S) = \mathbb{E}_{x_S \sim \mathcal{X}_S}[\|G_S(E^c_S(x_S), E^s_S(x_S)) - x_S\|_1]. \tag{4.5}$$

Latent reconstruction loss. To encourage the translated image to preserve the content of the source image, we require that a latent code c sampled from the latent distribution can be reconstructed after decoding and encoding.

$$\begin{aligned}\mathcal{L}^{c_S}_{recon}(E^c_S, G_T, E^c_T, \mathcal{X}_S, \mathcal{S}_T) = \\ \mathbb{E}_{x_S \sim \mathcal{X}_S, s_T \sim \mathcal{S}_T}[\|E^c_T(G_T(E^c_S(x_s), s_T)) - E^c_S(x_s)\|_1].\end{aligned} \tag{4.6}$$

Similarly, to align the style representation with a Gaussian prior distribution, we enforce the same constrain for the latent style code.

$$\begin{aligned}\mathcal{L}^{s_T}_{recon}(E^c_S, G_T, E^s_T, \mathcal{X}_S, \mathcal{S}_T) = \\ \mathbb{E}_{x_S \sim \mathcal{X}_S, s_T \sim \mathcal{S}_T}[\|E^s_T(G_T(E^c_S(x_s), s_T)) - s_T)\|_1].\end{aligned} \tag{4.7}$$

Cycle-consistency loss. To facilitate training we enforce cross-cycle consistency which implies that if we translate an image to the target domain and then translate it back to the source domain using the extracted source-domain style code, we should be able to obtain the original image.

$$\begin{aligned}\mathcal{L}^{S}_{cyc}(E^c_S, E^s_S, G_T, E^c_T, G_S, \mathcal{X}_S, \mathcal{S}_T) = \\ \mathbb{E}_{x_S \sim \mathcal{X}_S, s_T \sim \mathcal{S}_T}[\|G_S(E^c_T(G_T(E^c_S(x_S), s_T)), E^s_S(x_S)) - x_S\|_1].\end{aligned} \tag{4.8}$$

$\mathcal{L}^S_{GAN}$, $\mathcal{L}^T_{recon}$, $\mathcal{L}^{c_T}_{recon}$, $\mathcal{L}^{s_S}_{recon}$, $\mathcal{L}^T_{cyc}$ are defined in a similar way.

Segmentation loss. To enforce the generator to preserve the lesions, we enrich the network with segmentation supervision on the synthetic images. The segmentation loss on the synthetic images is given by

$$\mathcal{L}^{Synth}_{Seg}(Seg, G_T, E^c_S, \mathcal{X}_S, \mathcal{Y}_S, \mathcal{S}_T) = \mathcal{L}_{DSC}(Seg, G_T(E^c_S(\mathcal{X}_S), \mathcal{S}_T), \mathcal{Y}_S). \tag{4.9}$$

The full objective is given by

$$\begin{aligned}\min_{E_S^c,E_S^s,E_T^c,E_T^s,G_S,G_T}\max_{D_S,D_T} & \lambda_{GAN}(\mathcal{L}_{GAN}^S+\mathcal{L}_{GAN}^T)+\lambda_x(\mathcal{L}_{recon}^S+\mathcal{L}_{recon}^T)\\ & +\lambda_c(\mathcal{L}_{recon}^{c_S}+\mathcal{L}_{recon}^{c_T})+\lambda_s(\mathcal{L}_{recon}^{s_S}+\mathcal{L}_{recon}^{s_T})\\ & +\lambda_{cyc}(\mathcal{L}_{cyc}^S+\mathcal{L}_{cyc}^T)+\mathcal{L}_{Seg}^{Synth},\end{aligned} \tag{4.10}$$

where λ_{GAN}, λ_x, λ_c, λ_s, λ_{cyc} are weights that control the importance of each term.

Implementation details

We implement our model using Pytorch [130]. The content encoders consist of several convolutional layers and residual blocks followed by instance normalization [150]. The style encoders consist of convolutional layers followed by fully connected layers. The decoders include residual blocks followed by upsampling and convolutional layers. The residual blocks are followed by adaptive instance normalization (AdaIN) [151] layers to adjust the style of the output image. In particular, AdaIN aligns the mean and variance of the content code to the style of the target domain since the affine parameters of AdaIN are generated by a multilayer perceptron given a random style code sampled from the style space of the target domain. Sampling different style code allows us to obtain different affine parameters and generate diverse translations that adopt the appearance properties of the target domain. The encoder of the segmentation network is a standard ResNet [123] consisting of several convolutional layers while the decoder consists of several upsampling and convolutional layers. For training we use Adam optimizer, a batch size of 32 and a learning rate of 0.0001. We make our code available at `https://github.com/elchiou/DA`.

4.3 Datasets

VERDICT MRI: We use VERDICT MRI data collected from 60 men with a suspicion of cancer. We have provided the acquisition details for VERDICT MRI in Sec. 3.2. A dedicated radiologist, highly experienced in prostate mp-MRI, contoured the lesions on VERDICT MRI using mp-MRI for guidance.

DW-MRI from mp-MRI acquisition: We use DW-MRI data from the ProstateX challenge dataset [93] which consists of training mp-MRI data acquired from 204 patients. The DW-MRI data were acquired with a single-shot echo planar imaging sequence with a voxel size of $2 \times 2 \times 3.6$ mm^3, 3.6 mm slice thickness. Three b-values were acquired ($50, 400, 800$ s/mm^2), and subsequently, the ADC map and a b-value image at b $= 1400$ s/mm^2 were calculated by the scanner software. In this study, we use DW-MRI data from 80 patients. Since the ProstateX dataset provides only the position of the lesion, a dedicated radiologist manually annotated the lesions on the ADC map using as reference the provided position of the lesion.

4.4 Experiments

4.4.1 Quantitative results

In this section we evaluate the performance of our approach and the impact of the ratio of synthetic to real data on the performance. We also provide qualitative results and quantitative results related to the effect of sampling random style codes on the performance.

Performance evaluation. We first compare our approach to several baselines. i)VERDICT MRI only: we train the segmentation network only on VERDICT MRI. ii) Finetuning: we pre-train on mp-MRI and then perform finetuning using the VERDICT MRI data. iii) RAs: we pre-train on mp-MRI, then we install RAs in parallel to each of the convolutional layers of the pre-trained network and update them using VERDICT MRI. iv) MUNIT: we use MUNIT to map from source to target without segmentation supervision. v) CycleGAN + $\mathcal{L}_{Seg}^{Synth}$: we use CycleGAN and segmentation supervision to perform the translation, an approach similar to the one proposed in [20]. vi) CycleGAN + $\mathcal{L}_{Seg}^{Synth}$ + RAs: we use (v) for the translation and introduce RAs to the segmentation network. We evaluate the performance based on the average recall, precision, dice similarity coefficient (DSC), and average precision (AP). We report the results in Table 4.1. The proposed approach yields substantial improvements and outperforms all baselines including CycleGAN, which indicates that accommodating the uncertainty in the cross-domain mapping allows

Model	Recall	Precision	DSC	AP
VERDICT MRI only	67.1 (±14.2)	59.6 (±11.5)	62.4 (±13.4)	63.5 (±13.1)
Finetuning	68.4 (±12.4)	62.5 (±13.5)	64.7 (±11.2)	65.8 (±14.7)
RAs	66.6 (±11.6)	67.0 (±8.8)	65.7 (±10.2)	66.6 (±12.6)
MUNIT	65.2 (±10.2)	64.2 (±13.7)	64.4 (±11.3)	68.2 (±12.0)
CycleGAN + $\mathcal{L}_{Seg}^{Synth}$	64.5 (±10.4)	66.1 (±10.1)	64.8 (±8.7)	70.1 (±9.8)
CycleGAN + $\mathcal{L}_{Seg}^{Synth}$ + RAs	60.9 (±10.7)	**74.0** (±11.8)	66.6 (±13.6)	71.6 (±11.3)
MUNIT + $\mathcal{L}_{Seg}^{Synth}$ (Ours)	**71.8** (±7.8)	68.0 (±6.8)	69.8 (±7.9)	73.5 (±8.1)
MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + RAs (Ours)	69.2 (±8.6)	71.2 (±9.7)	**69.9** (±9.0)	**75.4** (±9.7)

Table 4.1: Average recall, precision, dice similarity coefficient (DSC), and average precision (AP) across 5 folds. The results are given in mean (±std) format.

us to learn better representations for the target domain. Compared to the naive MUNIT without segmentation supervision, $\mathcal{L}_{Seg}^{Synth}$, our approach performs better since it successfully preserves the lesions during the translation. Finally, introducing RAs in the segmentation networks further improves the performance of both CyclgeGAN + $\mathcal{L}_{Seg}^{Synth}$ and MUNIT + $\mathcal{L}_{Seg}^{Synth}$.

Impact of sampling on the performance. To experimentally validate that sampling different style codes enhances the performance, we perform two experiments: i) we keep the style code of the target domain fixed during the translation and ii) we use the encoded style code of the source domain. We evaluate the performance based on the mean recall, precision, dice similarity coefficient (DSC), and average precision (AP) across 5 folds. The results (Table 4.2) show that indeed sampling different style codes improves the performance.

Model	Recall	Precision	DSC	AP
MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + fixed s_T	68.4 (±9.0)	65.3 (±7.9)	67.0 (±8.8)	70.0 (±8.9)
MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + encoded s_S	61.5 (±12.4)	68.5 (±9.2)	64.5 (±10.6)	67.4 (±10.6)
MUNIT + $\mathcal{L}_{Seg}^{Synth}$ (Ours)	**71.8** (±7.8)	**68.0** (±6.8)	**69.8** (±7.9)	**73.5** (±8.1)

Table 4.2: Impact of sampling on the performance. Average recall, precision, dice similarity coefficient (DSC), and average precision (AP) across 5 folds. The results are given in mean (±std) format.

Impact of the ratio of synthetic to real data on the performance. Using synthetic data is motivated by the fact that annotating large datasets can be challenging in medical applications. We therefore evaluate the impact of the ratio of synthetic to real data. To this end, we first vary the percentage of real data while keeping fixed the amount of synthetic data (Fig. 4.4 (top, left)). We compare our approach to a segmentation network trained only on real data and to [20] where CycleGAN is used for the generation of synthetic data. Our approach outperforms both baselines. Figure 4.4 (top, right) shows the performance when we vary the percentage of synthetic samples while fixing the percentage of real ones. The AP of our approach increases as we increase the amount of synthetic data. The baseline also improves but we systematically outperform it. Figure 4.4 (bottom) shows the performance of our approach when we vary the percentage of real data while keeping fixed the percentage of synthetic. Here, we also vary the ratio of real to synthetic data in a mini-batch during training. Note that when the percentage of real data is small, a large ratio of synthetic to real data in the mini-batch delivers better results.

Impact of residual adapters in the performance. Figure 4.5 shows the impact of residual adapters in the performance for different dataset sizes. We vary the percentage of real data while keeping fixed the amount of synthetic data. Introducing residual adapters in the segmentation network while using MUNIT and segmentation supervision (MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + RAs) during the translation systematically improves performance for different dataset sizes.

4.4.2 Qualitative results

Figure 4.6 shows the mapping from mp-MRI to VERDICT. Our approach is able to generate multiple outputs while preserving the critical structure corresponding to the prostate lesion. In this work, we do not provide quantitative results for the image-to-image translation. Instead, we evaluate the quality of the translated images visually and based on the results we obtain for the task at hand.

In Figure 4.7 we present lesion segmentation results produced by the different models for two patients. i) MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + RAs (Ours): we use stochastic translation and segmentation supervision for the translation and introduce RAs in the segmentation network. ii) CycleGan + $\mathcal{L}_{Seg}^{Synth}$ + RAs: we use deterministic translation and segmentation supervision to perform the translation, and introduce residual adapters (RAs) in the segmentation network. iii) VERDICT MRI only: we train the segmentation network only on real VERDICT MRI.

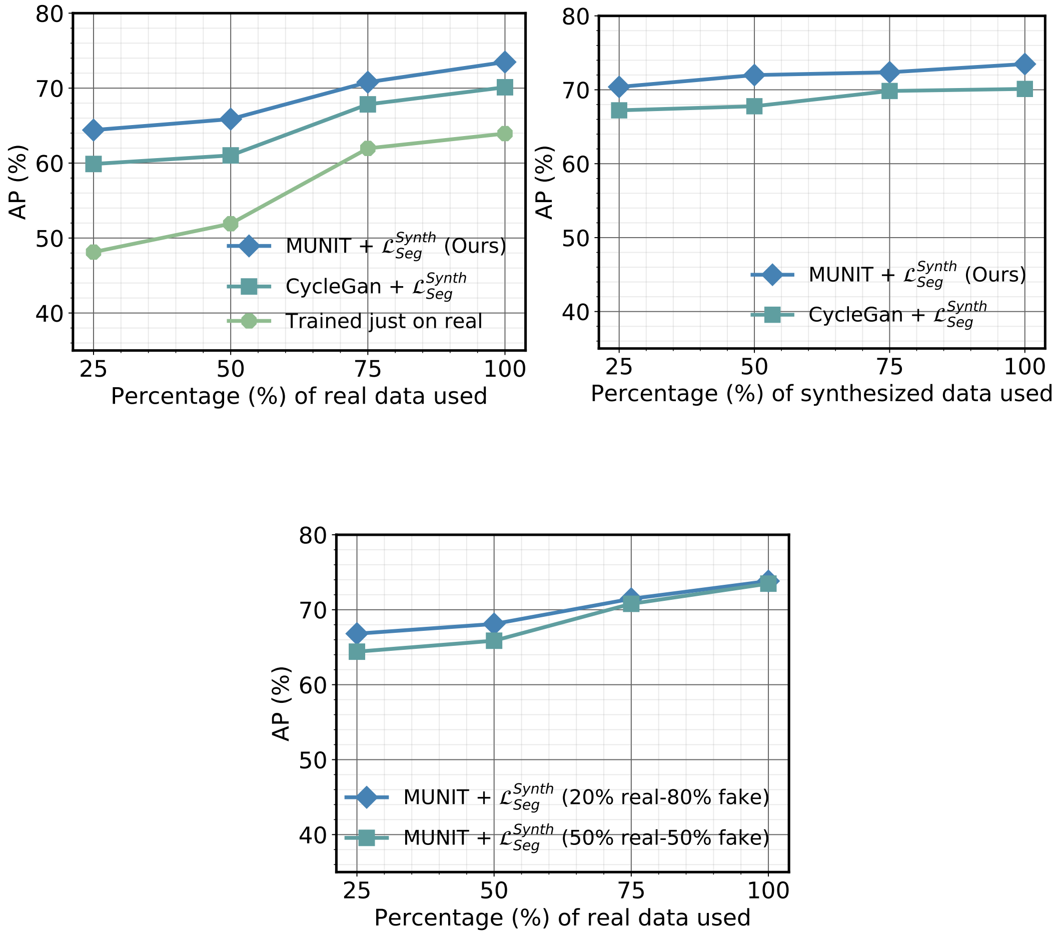


Figure 4.4: Impact of the ratio of synthetic to real data on the performance. (Top, left) Average precision (AP) as a function of the percentage of real samples used given a constant number of synthetic ones. (Top, right) AP as a function of the number of synthetic examples used given a constant number of real ones. (Bottom) AP as a function of the percentage of real data used given a constant number of synthetic ones. Here, the ratio of real to synthetic data in a mini-batch also varies during training.

4.5 Conclusion

In this chapter we propose a domain adaptation approach for lesion segmentation on VERDICT-MRI. Our approach relies on stochastic generative modelling to generate multiple outputs conditioned on a single input allowing the extraction of richer representations for the task of interest in the target domain. Compared to its deterministic counterparts, our approach yields substantial improvements across a broad

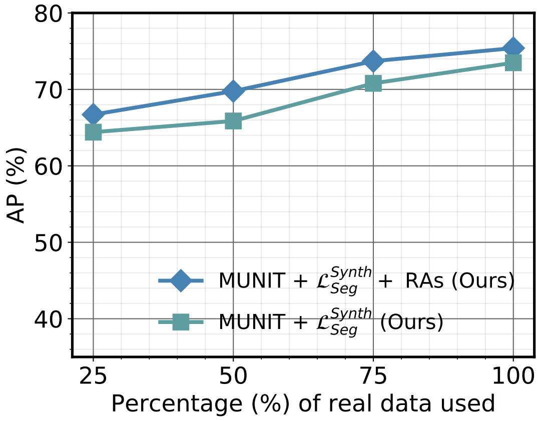


Figure 4.5: Impact of residual adapters (RAs) in the average precision (AP) for different dataset sizes. We vary the percentage of real data while keeping fixed the amount of synthetic data. Introducing residual adapters in the segmentation network while using MUNIT and segmentation supervision (MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + RAs) during the translation systematically improves performance for different dataset sizes.

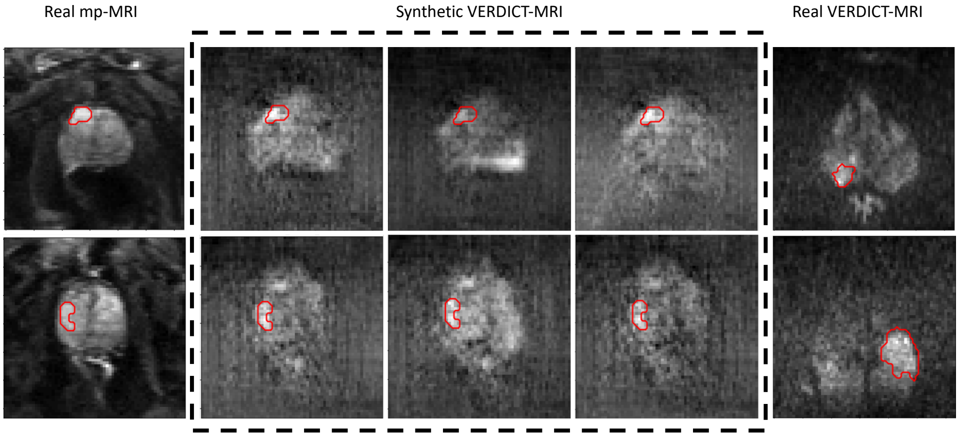


Figure 4.6: One-to-many mapping from one mp-MRI (left) to three VERDICT MRI translations (middle) for two different patients (rows): Our network can generate samples with both local and global structure variation, while at the same time preserving the critical structure corresponding to the prostate lesion. The right column shows two real VERDICT MRI samples as an example of data from the target domain.

range of dataset sizes, increasingly strong baselines, and evaluation measures.

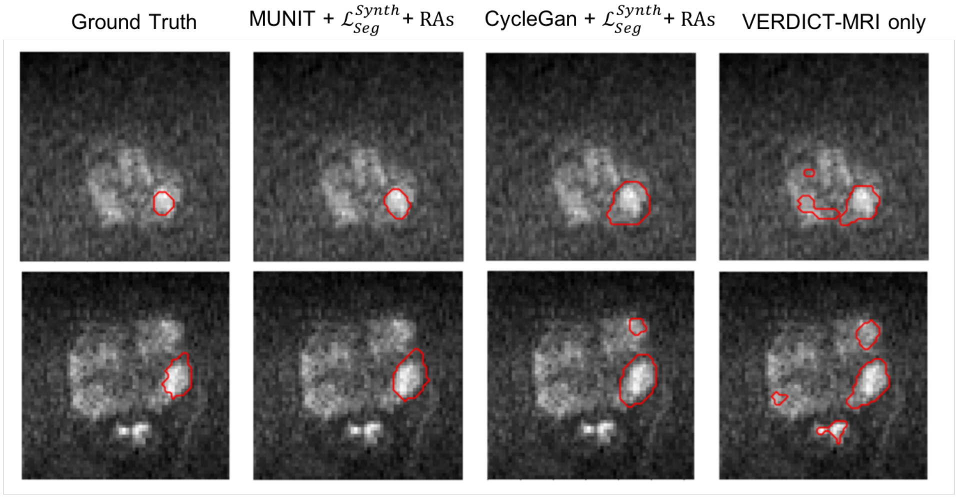

Ground Truth
MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + RAs
CycleGan + $\mathcal{L}_{Seg}^{Synth}$ + RAs
VERDICT-MRI only

Figure 4.7: Lesion segmentation results for two patients. MUNIT + $\mathcal{L}_{Seg}^{Synth}$ + RAs (Ours): we use stochastic translation and segmentation supervision for the translation and introduce RAs in the segmentation network. Cyclegan + $\mathcal{L}_{Seg}^{Synth}$ + RAs: we use deterministic translation and segmentation supervision for the translation and introduce RAs in the segmentation network. VERDICT MRI only: Trained only on real data.

Chapter 5

Unsupervised domain adaptation for lesion segmentation

In Chapter 4 we proposed a semi-supervised domain adaptation that relies on stochastic translation. In this chapter we turn our attention on the unsupervised domain adaptation where facing new challenges due to the complete lack of labeled data in the target domain. In particular, as in Chapter 4, we rely on stochastic generative modelling to translate across the source and the target domain at pixel space and we introduce two new loss functions that promote semantic consistency. We demonstrate the effectiveness of our approach on VERDICT MRI; however, it is quite general can be applied in other application where the amount of labeled training data for the domain of interest is limited. This chapter contains material from [30, 31], which were published at DART@MICCAI 2020 and at ISMRM 2021.

5.1 Introduction

In this work we address the challenge of adapting across two heterogeneous domains where both the distribution and the dimensionality of the input features are different (Fig. 4.3) in cases where there is a complete lack of labeled data for the target domain. As in Chapter 4, we rely on stochastic translation [143] to align the two domains at pixel-level; in Chapter 4 we showed that stochastic translation yields clear improvements in heterogeneous domain adaptation tasks compared to deterministic, CycleGAN-based [139] translation approaches. However, these improvements have been obtained with semi-supervised learning, where a few labeled target-domain images are available, whereas our goal is unsupervised domain adaptation. To this end we introduce a *semantic cycle-consistency* loss on the cycle-reconstructed source images; if a source image is translated to the target domain and then back to the source domain, we require that critical structures are preserved. We also introduce a *pseudo-labeling* loss that allows us to use the unlabeled target data to supervise the target-domain segmentation network. In particular we translate the target data to the source domain, predict their labels according to a pre-trained source-domain segmentation network and use the generated pseudo-labels to supervise the target-domain segmentation network. This allows us to use exclusively target-domain statistics and train highly discriminative models. As in Chapter 4 we demonstrate the effectiveness of our approach in prostate lesion segmentation on VERDICT MRI. As shown experimentally, our approach largely improves the generalization capabilities of a lesion segmentation model on VERDICT MRI by leveraging labeled DW data from mp-MRI acquisitions.

Most domain adaptation methods align the two domains either by extracting domain-invariant features or by aligning the two domains at the raw pixel space. Ren et al. [137] and Kamnitsas et al. [138], rely on adversarial training to align the feature distributions between the source and the target domain for medical image classification and segmentation respectively. Pixel-level approaches [17, 18, 19, 20], use GAN-based methods [139, 143] to align the source and the target domains at pixel level. Chen et al. [152] align simultaneously the two domains at pixel- and feature-level by utilizing adversarial training. Ouyang et al. [140] combine a variational autoencoder (VAE)-based feature prior matching and pixel-level adversarial training to learn a domain-invariant latent space which is exploited during segmentation. Similarly, [141] perform pixel-level adversarial training to extract content-only images and use them to train a segmentation model that operates well in both domains. Other studies exploit unlabeled target domain data during the discriminative training. Bateson et al. [153] and Guodong et al.

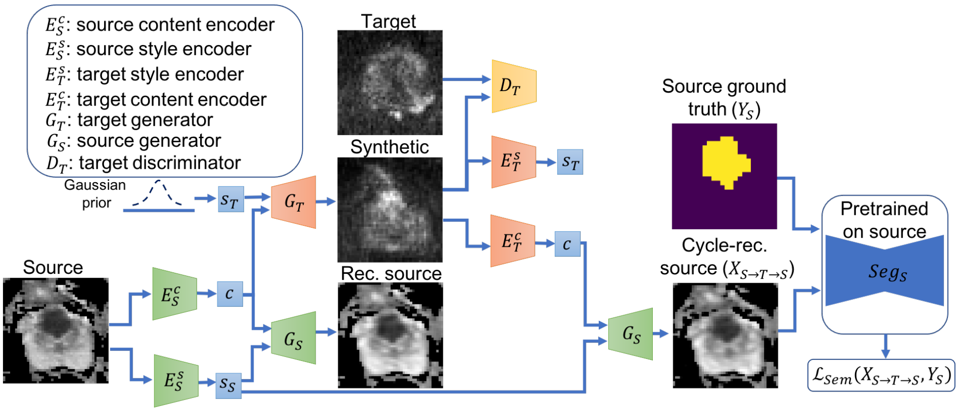

E_S^c: source content encoder
E_S^s: source style encoder
E_T^s: target style encoder
E_T^c: target content encoder
G_T: target generator
G_S: source generator
D_T: target discriminator
Gaussian prior
s_T
G_T
Target
Synthetic
Rec. source
Source
E_S^c
c
E_S^s
s_S
G_S
D_T
E_T^s
s_T
E_T^c
c
G_S
Source ground truth (Y_S)
Cycle-rec. source ($X_{S\to T\to S}$)
Pretrained on source
Seg_S
$\mathcal{L}_{Sem}(X_{S\to T\to S}, Y_S)$

Figure 5.1: We force a network for stochastic translation across domains to preserve semantics through a semantic segmentation-based loss. The image-to-image translation network translates source-domain images to the style of the target domain by combining a domain-invariant content code c with a random code s_T. We introduce a semantic cycle-consistency loss, $\mathcal{L}_{Sem}$, on the cycle-reconstructed images that ensures that the prostate lesions are successfully preserved.

[154] use entropy minimization on the prediction of target data as an extra regularization while [155] propose a teacher-student framework to train a model using labeled and unlabeled target data as well as labeled source data.

5.2 Method

Problem formulation

We consider the problem of domain adaptation in prostate lesion segmentation. Let $\mathcal{X}_S \subset \mathbb{R}^{H\times W\times C_S}$ be a set of N_S source images and $\mathcal{Y}_S \subset \{0,1\}^{H,W}$ their segmentation masks. The sample $x_S \in \mathcal{X}_S$ is a $H \times W \times C_S$ image and the entry $y_S^{(h,w)}$ of the mask y_S provides the label of voxel (h,w) as a one-hot vector. Let also $\mathcal{X}_T \subset \mathbb{R}^{H\times W\times C_T}$ be a set of N_T unlabeled target images. Sample $x_T \in \mathcal{X}_T$ is an $H \times W \times C_T$ image.

Stochastic translation with semantic cycle-consistency

We rely on stochastic translation [143] to learn the mapping between the two domains and introduce a semantic cycle-consistency loss to enforce the cross-domain mapping to preserve critical structures.

The image-to-image translation network (Fig. 5.1) consists of content encoders E_S^c, E_T^c, style encoders E_S^s, E_T^s, generators G_S, G_T and domain discriminators D_S, D_T for both domains. The content encoders E_S^c, E_T^c extract a domain-invariant content code $c \in \mathcal{C}$ ($E_S^c : \mathcal{X}_S \to \mathcal{C}$, $E_T^c : \mathcal{X}_T \to \mathcal{C}$) while the style encoders E_S^s, E_T^s extract domain-specific style codes $s_S \in \mathcal{S}_S$ ($E_S^s : \mathcal{X}_S \to \mathcal{S}_S$) and $s_T \in \mathcal{S}_T$ ($E_T^s : \mathcal{X}_T \to \mathcal{S}_T$). Image-to-image translation is performed by combining the content code ($c = E_S^c(X_S)$) extracted from a given input ($x_S \in \mathcal{X}_S$) and a random style code s_T sampled from the target-style space. We note that the random style-code sampled from a Gaussian distribution represents structures that cannot be accounted by a deterministic mapping and results in one-to-many translation. We train the networks with a loss function consisting of domain adversarial, self-reconstruction, latent reconstruction and semantic cycle-consistency losses.

Domain adversarial loss.

$$\mathcal{L}_{GAN}^T = \mathbb{E}_{c_S \sim \mathcal{C}, s_T \sim \mathcal{S}_T}[\log(1 - D_T(G_T(c_S, s_T)))] + \mathbb{E}_{x_T \sim \mathcal{X}_T}[\log(D_T(x_T))].$$

Self-reconstruction loss.

$$\mathcal{L}_{recon}^S = \mathbb{E}_{x_S \sim \mathcal{X}_S}[||G_S(E_S^c(x_S), E_S^s(x_S)) - x_S||_1].$$

Latent reconstruction loss.

$$\mathcal{L}_{recon}^{c_S} = \mathbb{E}_{x_S \sim \mathcal{X}_S, s_T \sim \mathcal{S}_T}[||E_T^c(G_T(E_S^c(x_S), s_T)) - E_S^c(x_S)||_1].$$

$$\mathcal{L}_{recon}^{s_T} = \mathbb{E}_{x_S \sim \mathcal{X}_S, s_T \sim \mathcal{S}_T}[||E_T^s(G_T(E_S^c(x_S), s_T)) - s_T)||_1].$$

Cycle-consistency loss.

$$\mathcal{L}_{cyc}^S = \mathbb{E}_{x_S \sim \mathcal{X}_S, s_T \sim \mathcal{S}_T}[||G_S(E_T^c(G_T(E_S^c(x_S), s_T)), E_S^s(x_S)) - x_S||_1].$$

$\mathcal{L}_{GAN}^S$, $\mathcal{L}_{recon}^T$, $\mathcal{L}_{recon}^{c_T}$, $\mathcal{L}_{recon}^{s_S}$, $\mathcal{L}_{cyc}^T$ are defined in a similar way.

Semantic cycle-consistency loss. Recent studies [17, 19, 20] enforce semantic consistency between the real source and the synthetic target images by exploiting a target-domain segmentation network trained on a few available labeled target-domain images. However, in the unsupervised scenario, where there is no supervision available for the target domain, such approach is not feasible. To this end we introduce a semantic cycle-consistency loss or lesion segmentation loss on the cycle-reconstructed source images $x_{S \to T \to S}$; if a source image is translated to the target domain and then back to the source domain, we require that critical structures, corresponding to lesions, are preserved. The naive cycle-consistency loss, introduced in [139], penalizes inconsistencies in the entire image and may fail to

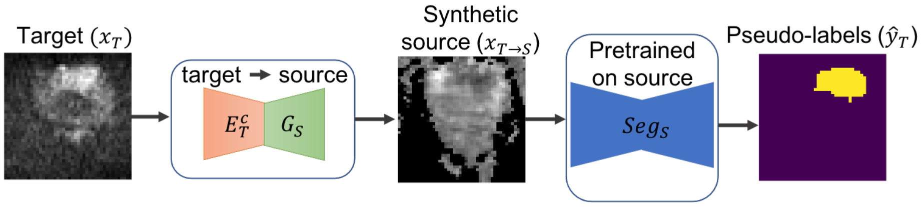


Figure 5.2: Pseudo-labeling through translation to the source domain: We translate the target data to the source domain and predict their pseudo-labels according to a pre-trained source-domain segmentation network Seg_S.

preserve small structures corresponding to lesions. In contrast our semantic cycle-consistency loss penalizes inconsistencies in the label space enforcing the translation network to preserve the lesions. The semantic cycle-consistency loss is a soft generalization of the dice score given by

$$\mathcal{L}_{Sem} = -\frac{2\sum_{h,w} p^{(h,w)} y_S^{(h,w)}}{\sum_{h,w}(p^{(h,w)^2} + y_S^{(h,w)^2})}, \tag{5.1}$$

where $p^{(h,w)}$ is the predictive probability of class 1 for voxel (h,w) provided by the pre-trained source network Seg_S. The full objective is given by

$$\begin{aligned} \min_{E^c_S,E^s_S,E^c_T,E^s_T,G_S,G_T} \max_{D_S,D_T} \ & \lambda_{GAN}(\mathcal{L}^S_{GAN} + \mathcal{L}^T_{GAN}) + \lambda_x(\mathcal{L}^S_{recon} + \mathcal{L}^T_{recon}) \\ & + \lambda_c(\mathcal{L}^{cS}_{recon} + \mathcal{L}^{cT}_{recon}) + \lambda_s(\mathcal{L}^{sS}_{recon} + \mathcal{L}^{sT}_{recon}) + \lambda_{cyc}(\mathcal{L}^S_{cyc} + \mathcal{L}^T_{cyc}) \\ & + \lambda_{sem}\mathcal{L}_{sem}, \end{aligned} \tag{5.2}$$

where λ_{GAN}, λ_x, λ_c, λ_s, λ_{cyc}, λ_{sem} control the importance of each term.

Pseudo-labeling through translation to the source

We generate pseudo-labels for the target images by translating them to the source domain and predicting their labels according to the pre-trained source-domain segmentation network Seg_S, trained on the labeled source data (Fig. 5.2).

Given a synthetic source image $x_{T\to S}$ and the segmentation network Seg_S we obtain a soft-segmentation map, $p_{x_{T\to S}} = Seg_S(x_{T\to S})$, where each vector $p^{(h,w)}_{x_{T\to S}}$ corresponds to a probability distribution over classes. Assuming that high-scoring pixel-wise predictions on synthetic source samples are correct, we obtain a segmen-

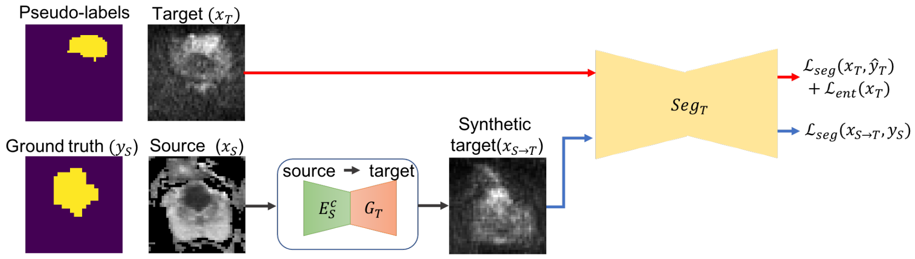

Pseudo-labels
Target (x_T)
Ground truth (y_S)
Source (x_S)
source → target
E_S^c
G_T
Synthetic target($x_{S\to T}$)
Seg_T
$\mathcal{L}_{seg}(x_T, \hat{y}_T)$
$+ \mathcal{L}_{ent}(x_T)$
$\mathcal{L}_{seg}(x_{S\to T}, y_S)$

Figure 5.3: We use data that have exclusively target-domain statistics to train the target segmentation network (Seg_T). We translate the source data to the target domain and supervise Seg_T using the ground-truth segmentation masks. We also use target pseudo-labels to supervise Seg_T.

tation mask $\hat{y}_T$ by selecting high-scoring pixels with a fixed threshold. Each entry $\hat{y}_T^{(h,w)}$ can be either a discrete one-hot vector for high-scoring pixels or a zero-vector for low-scoring pixels. The pseudo-labeling configuration is defined as follows

$$\hat{y}_T^{(h,w,c)} = \begin{cases} 1, & \text{if } c = \arg\max_c p_{x_{T\to S}}^{(h,w)} \text{ and } p_{x_{T\to S}}^{(h,w,c)} > \text{threshold} \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

Segmentation Network

The target-domain segmentation network, Seg_T, is an encoder-decoder network [122, 120]. We supervise Seg_T using both the synthetic target images and the corresponding source labels and the real target images and their pseudo-labels (Fig. 5.3). Given an image x and its segmentation mask y, the segmentation loss is defined as

$$\mathcal{L}_{Seg}(x,y) = -\frac{2\sum_{h,w} p^{(h,w,1)} y^{(h,w,1)}}{\sum_{h,w}(p^{(h,w,1)2} + y^{(h,w,1)2})}, \tag{5.4}$$

where $p^{(h,w,1)}$ is the predictive probability of class 1 for voxel (h,w).

As in recent studies [156, 153], to further regularize the network on the target-domain data for which we have not obtained pseudo-labels, we apply entropy-based regularization. The loss $\mathcal{L}_{ent}$ is defined as follows

$$\mathcal{L}_{Ent}(x_T) = \sum_{h,w} \frac{-1}{\log C} \sum_{c=1}^{C} p_{x_T}^{(h,w,c)} \log p_{x_T}^{(h,w,c)}, \tag{5.5}$$

where $p^{(h,w,c)}$ is the predictive probability of class c, $c = \{0,1\}$, for voxel (h,w). The full objective is given by $\min_{Seg_T} L_{Seg} + \mathcal{L}_{Ent}$.

Implementation details

We implement our framework using Pytorch [130].

Image-to-image translation network: The content encoders consist of convolutional layers and residual blocks followed by instance normalization [150]. The style encoders consist of convolutional layers followed by fully connected layers. The decoders include residual blocks followed by upsampling and convolutional layers. The residual blocks are followed by adaptive instance normalization (AdaIN) [151] layers to adjust the style of the output image. The discriminators consist of convolutional layers. For training we use Adam optimizer, a batch size of 32, a learning rate of 0.0001 and set losses weights to $\lambda_{GAN} = 1$, $\lambda_x = 10$, $\lambda_c = 1$, $\lambda_s = 1$, $\lambda_{sem} = 10$. We train the translation network for 50000 iterations.

Segmentation network: The encoder of the segmentation network is a standard ResNet [123] consisting of convolutional layers while the decoder consists of upsampling and convolutional layers. For training we use stochastic gradient decent and a batch size of 32. We split the training set into 80% training and 20% validation to select the learning rate, the number of iterations and the threshold to perform pseudo-labeling.

Datasets

We use VERDICT MRI data from 90 patients since annotations for 30 more patients become available and we use DW-MRI data from 80 patients from the ProstateX challenge dataset [93]. We have provided the acquisition details for both VERDICT MRI and DW-MRI from mp-MRI acquisitions in Sec. 4.3.

5.3 Results

We evaluate the performance based on the average recall, precision, dice similarity coefficient (DSC), and average precision (AP) across 5-folds.

We compare our approach to several baselines. i)VERDICT-MRI: train using VERDICT-MRI only. ii) VERDICT-MRI + Synth (MUNIT + $\mathcal{L}_{Sem}$) : train us-

Model	Recall	Precision	DSC	AP
VERDICT-MRI (Oracle)	66.2 (8.1)	70.5 (9.9)	68.9 (9.2)	72.1 (10.4)
VERDICT-MRI + Synth (MUNIT + $\mathcal{L}_{Sem}$)	71.1 (8.9)	72.5 (10.4)	72.1 (8.7)	76.7 (9.6)
mp-MRI + EntMin (ADVENT [156])	50.8 (12.3)	48.0 (11.4)	49.8 (13.0)	51.4 (13.9)
Synth (MUNIT)	51.5 (13.3)	60.6 (11.9)	53.6 (12.7)	60.2 (13.0)
Synth (MUNIT + $\mathcal{L}_{Sem}$)	55.1 (13.9)	62.4 (12.8)	55.3 (10.9)	62.0 (13.4)
Synth (MUNIT + $\mathcal{L}_{Sem}$) + EntMin	54.7 (11.5)	**69.2** (10.3)	57.1 (10.8)	63.4 (12.8)
Synth (MUNIT + $\mathcal{L}_{Sem}$) + PsLab	59.8 (10.1)	64.8 (11.1)	61.5 (10.3)	64.9 (10.1)
Synth (MUNIT + $\mathcal{L}_{Sem}$) + EntMin + PsLab (Proposed)	**61.4** (9.9)	66.9 (10.7)	**62.1** (9.8)	**65.6** (10.9)

Table 5.1: Average recall, precision, dice similarity coefficient (DSC), and average precision (AP) across 5 folds. The results are given in mean (±std) format.

ing real VERDICT-MRI and the synthetic VERDICT-MRI obtained from MUNIT with semantic cycle-consistency loss. iii) mp-MRI + EntMin (ADVENT [156]): train the model by minimizing the segmentation loss, $\mathcal{L}_{Seg}(x_S, y_S)$, on the raw mp-MRI and the entropy loss, $\mathcal{L}_{Ent}(x_T)$, on VERDICT-MRI, an approach proposed in [156, 153, 154]. iv) Synth (MUNIT): use the naive MUNIT to map from source to target and train only on the synthetic data. v) Synth (MUNIT + $\mathcal{L}_{Sem}$): use MUNIT with semantic cycle-consistency loss to translate from source to target. vi) Synth (MUNIT + $\mathcal{L}_{Sem}$) + EntMin: use (v) and entropy-based regularization on VERDICT-MRI data. vii) Synth (MUNIT + $\mathcal{L}_{Sem}$) + PsLab: use (v) and pseudo-labels to train the segmentation network on real VERDICT-MRI. viii) Synth (MUNIT + $\mathcal{L}_{Sem}$) + EntMin + PsLab: use (vi) and pseudo-labels to train the segmentation network on real VERDICT-MRI.

We report the results in Table 5.1. We observe that the performance is poor when the segmentation network is trained on the mp-MRI and VERDICT-MRI data (mp-MRI + EntMin (ADVENT [156])). However, we observe that when we train the network with synthetic VERDICT-MRI and real VERDICT-MRI (Synth (MUNIT + $\mathcal{L}_{Sem}$) + EntMin) the performance improves. This justifies our assumption that pixel-level alignment is beneficial in cases where there is a large distribution shift. The performance further improves when we use pseudo-labels obtained from confident predictions (Synth (MUNIT + $\mathcal{L}_{Sem}$) + EntMin + PsLab). We also observe that compared to the naive MUNIT without the semantic cycle-consistency loss (Synth (MUNIT)) our approach (Synth (MUNIT + $\mathcal{L}_{Sem}$)) performs better since it successfully preserves the lesions. When combining real and synthetic

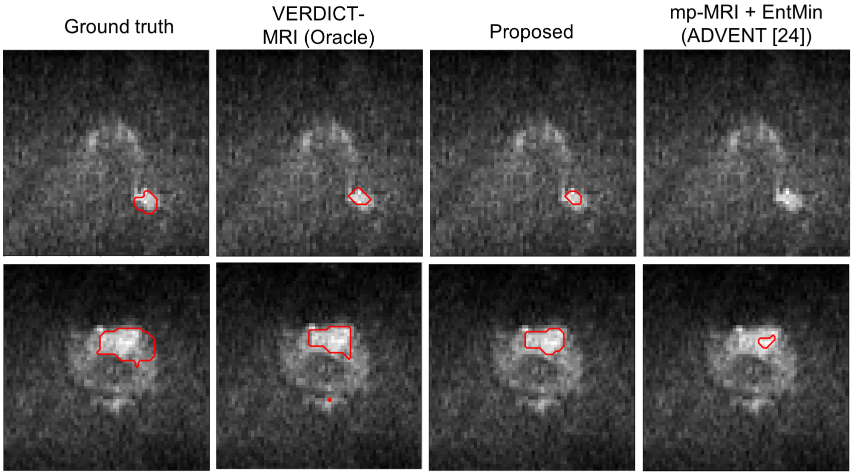


Figure 5.4: Lesion segmentation results for two patients - the proposed approach performs well on the target domain despite the fact that it does not utilize any manual target annotations during training.

data (VERDICT-MRI + Synth (MUNIT + $\mathcal{L}_{Sem}$)) to train the network in a fully-supervised manner we get better results compared to the oracle, where we use only the real VERDICT-MRI. In Figure 5.4 we present lesion segmentation results produced by the different models for two patients. The results indicate that the proposed approach performs well despite the fact that it does not use any manual annotations during training.

So far we have considered only the unsupervised case. However, our approach can also be used in a semi-supervised setting. To evaluate the performance of our method when labeled target data is available, we perform additional experiments varying the percentage of labeled data; we use the pseudo-labels (PsLab) and entropy minimization (EntMin) for the unlabeled data. Figure 5.5 shows that the performance of our method improves as the percentage of real data increases and always outperforms the baseline that is trained only on the target domain.

5.4 Conclusion

In this chapter we proposed a domain adaptation approach for lesion segmentation. Our approach relies on appearance alignment along with pseudo-labeling to train a target domain classifier using exclusively target domain statistics. We demonstrate the effectiveness of our approach for lesion segmentation on VERDICT-MRI which

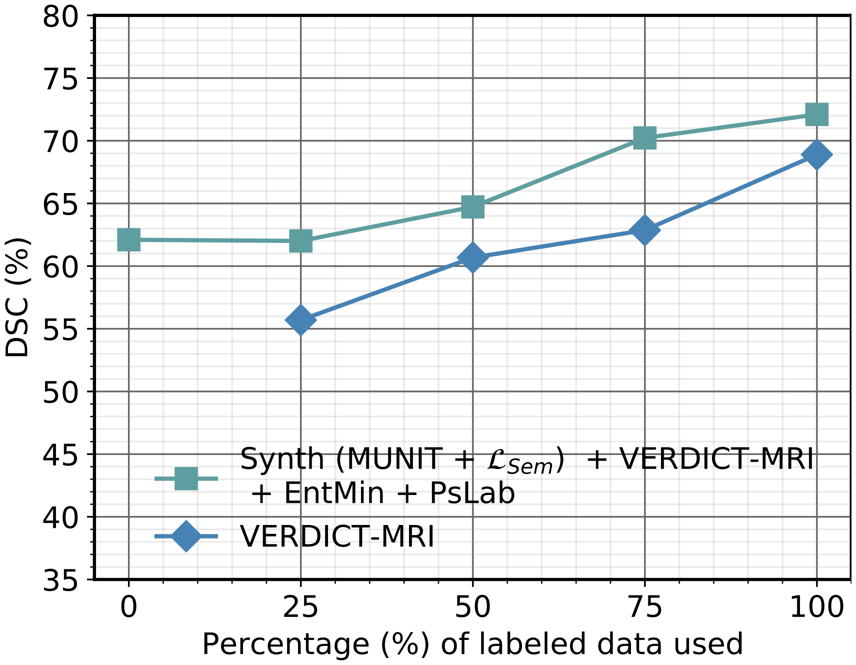


Figure 5.5: Performance as we vary the percentage of labeled target data used for training. We observe that our method improves with more supervision and the improvements introduced by our method over the baseline of target-only training carry over all the way to the fully-supervised regime.

is an advanced imaging technique for cancer characterization. When compared to several unsupervised domain adaptation approaches, our approach yields substantial improvements, that consistently carry over to the semi-supervised and supervised learning settings.

Chapter 6

Unsupervised domain adaptation for semantic segmentation of urban scenes

In this chapter, we further extend the approach proposed in Chapter 5 to unsupervised domain adaptation for semantic segmentation of urban scenes. We focus on semantic segmentation of urban scenes since there are two benchmarks available allowing us to compare our approach with recent state-of-the-art methods. As in Chapter 4, 5, we rely on stochastic generative modelling to capture inherent translation ambiguities. This allows us to (i) train more accurate target networks by generating multiple outputs conditioned on the same source image, leveraging both accurate translation and data augmentation for appearance variability, (ii) impute robust pseudo-labels for the target data by averaging the predictions of a source network on multiple translated versions of a single target image and (iii) train and ensemble diverse networks in the target domain by modulating the degree of stochasticity in the translations. We report improvements over strong recent baselines, leading to state-of-the-art unsupervised domain adaptation (UDA) results on two challenging semantic segmentation benchmarks. Our code is publicly available at `https://github.com/elchiou/Beyond-deterministic-translation-for-UDA`. This chapter contains material from [32] which was published at BMVC 2022.

6.1 Introduction

Unsupervised domain adaptation aims at accommodating the differing statistics between a 'source' and a 'target' domain, where the source domain comes with input-label pairs for a task, while the target domain only contains input samples. Successfully solving this problem can allow us for instance to exploit synthetically generated datasets that come with rich ground-truth to train models that can perform well in real images with different appearance properties. Translation-based approaches [15, 16, 157, 7, 158] rely on establishing a transformation between the two domains (often referred to as 'pixel space alignment') that bridges the difference in their statistics while preserving the semantics of the translated samples. This translation can then be used as a mechanism for generating supervision in the 'target' domain based on ground-truth originally available in a 'source' domain.

In this work we address a major shortcoming of this approach - namely the assumption that this translation is a deterministic function, mapping a single source to a single target image. Recent works on the closely related problem of unsupervised image translation [143, 159, 144, 142] have highlighted that this is a strong assumption and is frequently violated in practice. For instance a nighttime scene can have multiple daytime counterparts where originally invisible structures are revealed by the sun and also illuminated from different directions during the day. To mitigate this problem these techniques introduce methods for multimodal, or stochastic translation, that allows an image from one domain to be associated with a whole distribution of images in another.

Firstly, as in Chapter 4, 5, we use stochastic translation [143, 159, 144, 142] across the source and target domains. We show that allowing for stochastic translations yields clear improvements over the deterministic CycleGAN-based baseline, as well as all published pixel space alignment-based techniques. We attribute this to the ability of the multimodal translation to generate more diverse and sharper samples, that provide better training signals to the target-domain network.

Secondly, we exploit the ability to sample multiple translations for a given image in order to obtain better pseudo-labels for the unlabelled target images: we generate multiple translations of every target image into the source domain, label each according to a source-domain CNN, and average the resulting predictions to form a reliable estimate of the class probability. This is used as supervision for target-domain networks, and is shown to be increasingly useful as the number of averaged samples per image grows.

Thirdly, we modify the variance of the latent style code in order to train and ensemble complementary target-domain networks, each of which is adapted to handle a different degree of appearance variability. The results of ensembling these networks on the target data are then used to train a single target-domain network that outperforms all methods that also rely on ensembling-based supervision in the target domain.

We show that each of our proposed contributions yields additional improvements over strong recent baselines, leading to state-of-the-art UDA results on two challenging semantic segmentation benchmarks.

6.2 Related work

UDA approaches aim at learning domain invariant representations by aligning the distributions of the two domains at feature/output level [160, 161, 162, 163, 164, 165, 166] or at image level [15, 16, 7]. Based on the observation that the source and the target domain share a similar semantic layout, [160, 156] rely on adversarial training to align the raw output and entropy distributions respectively. However, such a global alignment does not guarantee that individual target samples are correctly classified. Category-based feature alignment methods [167, 168, 169, 170, 161, 162] attempt to address this problem by mapping target-domain features closer to the corresponding source-domain features.

Image-level UDA methods aim at aligning the two domain at the raw pixel space. [15, 16, 157, 7] rely on CycleGAN [139] to translate source domain images to the style of the target domain. Two recent works [171, 163] bypass the need for training an image translation network by relying on simple Fourier transform and global photometric alignment respectively.

Complementary to the idea of translation is the use of self-training [172, 173, 174, 175] which has been originally used in semi-supervised learning. Self-training iteratively generates pseudo-labels for the target domain based on confident predictions and uses those to supervise the model, implicitly encouraging category-based feature alignment between the source and the target domain. Another direction pursued in [176, 177] is to leverage the unlabeled target data by using consistency regularization to make the model predictions invariant to perturbations imposed in the input images.

Two recent works [16, 7] that rely on both image-level alignment and self-training are more closely related to our work. [16] relies on CycleGan to translate source images to the style of the target domain. They train the image translation network and the segmentation network alternatively and introduce a perceptual supervision based on the segmentation network to enforce semantic consistency during translation. They also generate pseudo-labels for the target data based on high confident predictions of the target network and use those to supervise the target network. [7] improves upon [16] by replacing the single-domain perceptual supervision with a cross-domain perceptual supervision using two segmentation networks operating in the source and the target domain respectively. In addition, they rely on both the source and the target networks to generate pseudo labels for the target data. Similar to these works we rely on image-to-image translation to translate source images to the style of the target domain, but we go beyond their one-to-one mapping approach which allows us to leverage both accurate translation and data augmentation for appearance variability. In addition, as in [7] we use source and target networks to generate pseudo-labels, but we exploit stochasticity in the translation to generate more robust pseudo-labels that allow us to train accurate target-domain networks.

6.3 Methods

We start in Sec. 6.3.1 by introducing the background of using translation in UDA, and then introduce our technical contributions from Sec. 6.3.2 onwards. Our presentation gradually introduces different components, loss terms, and processes used in UDA, and we summarize how everything is pieced together in Sec. 6.3.5.

6.3.1 Domain translation and UDA

In UDA we consider a source dataset with paired image-label data: $\mathcal{S} = \{(x_s^i, y_s^i)\}, i \in [1, S]$ and a target dataset comprising only image data $\mathcal{T} = \{x_t^i\}, i \in [1, T]$. Our task is to learn a segmentation system that provides accurate predictions in the target domain; we assume a substantial domain gap, precluding the naive approach of training a network on $\mathcal{S}$ and then deploying it in the target domain.

Output-space alignment UDA approaches [160] train a single segmentation network, F on both the source and the target images, using a cross-entropy loss in the source domain and an adversarial loss in the target domain to statistically align the predictions on target images to the distribution of source predictions. This results in a training objective of the following form:

$$\mathcal{L}(F) = \sum_{(x,y)\in\mathcal{S}} \mathcal{L}_{ce}(F(x), y) + \sum_{x\in\mathcal{T}} \mathcal{L}_{adv}(F(x)), \tag{6.1}$$

where $F(x)$ the softmax output.

In [156] entropy-based adversarial training is used to align the target entropy distribution to the source entropy distribution instead of aligning the raw predictions, resulting in the following objective:

$$\mathcal{L}(F) = \sum_{(x,y)\in\mathcal{S}} \mathcal{L}_{ce}(F(x), y) + \sum_{x\in\mathcal{T}} \mathcal{L}_{adv}(E(F(x))), \tag{6.2}$$

where $E(F(x)) = -F(x)\log(F(x))$ is the weighed self-information.

Given that the network provides low-entropy predictions on source images, adversarial entropy minimization promotes low-entropy predictions in the target domain. The entropy-based objective forces the target points to be classified confidently, and aims at reducing misclassifications by aligning the decision boundaries of F with low-density areas of the target domain - reflecting a desired property under the cluster assumption [178]. Still, having a single network F that successfully operates in both domains can be challenging due to the broader intra-class variability caused by the domain gap.

Pixel-space alignment approaches try to mitigate this problem by establishing a relation between the distributions of the source and target domain images and using that to supervise a network that only operates with target-domain images. In its simplest form, adopted also in [179, 15, 180, 16, 157] this relation is a deterministic translation function $\mathbf{T}$ that maps source images to the target domain, resulting in the following objective:

$$\mathcal{L}(F_t) = \sum_{(x,y)\in\mathcal{S}} \mathcal{L}_{ce}(F_t(\mathbf{T}[x]),y) + \sum_{x\in\mathcal{T}} \mathcal{L}_{adv}(E(F_t(x))), \tag{6.3}$$

where the difference with respect to Eq. 6.2 is that the translated version of x, $\mathbf{T}[x]$ is passed to the target-domain segmentation network, F_t. A straightforward way of obtaining such a translation function is through unsupervised translation between the two domains [139]; more sophisticated approaches [15, 16, 157] train the translation network in tandem with the UDA task, using for instance semantic losses to ensure the semantics of the source domains are preserved during cyclic translation. Other methods that implicitly use translation include [171], where a Fourier domain-based approach is used to align the two domains, effectively bypassing the need for a pixel-level translation network.

This approach creates a target-adapted variant of the source-domain dataset, allowing us to train a single network that is tuned exclusively to the statistics of the target domain. This reduces the intra-class variance and puts less strain on the segmentation network, but relies on the strong assumption that such a deterministic translation function exists. In this work we relax this assumption and work with a *distribution on translated images*. This better reflects most UDA scenarios and provides us with novel and simple tools to improve UDA performance, as described below.

6.3.2 Stochastic translation and UDA

We propose to replace the determinstic translation function $\mathbf{T}[x]$, with a distribution over images given by $\mathbf{T}[x,\mathbf{v}], \mathbf{v} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, where $\mathbf{v}$ is a random vector sampled

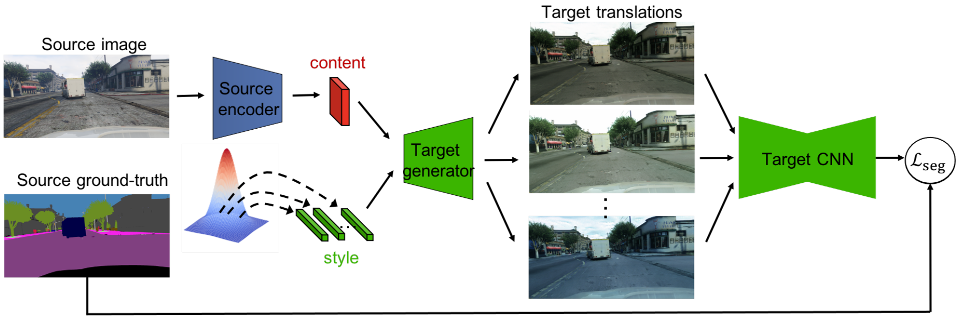


Figure 6.1: Unsupervised Domain Adaptation (UDA) with stochastic translation: we rely on a content-style separation network to associate a synthetic image from the GTA5 dataset (source) with a distribution of image translations to the target domain. These translations preserve the content signal and adopt the appearance properties of the Cityscapes dataset (target) through randomly sampled style codes. We use the resulting images to train a target-domain network tasked with predicting the labels of the respective source-domain image, irrespective of the style variation. Stochasticity in UDA allows the translation networks to generate multiple, sharp outputs that better capture the diversity of the scenes in the target domain, and train the target-domain network with a more representative set of images.

from a normal distribution with zero mean and unit covariance [143]. For instance when translating a nighttime scene into its daytime scene, the random argument can reflect the (unpredictable) position of the sun, clouds or obscured objects. For the synthetic-to-real case that we handle in our experiments we can see from Fig. 6.1 that the translation network can indeed generate scenes illuminated differently as well as different cloud patterns, allowing us to capture more faithfully the range of scenes encountered in the target domain. We note that **T** remains deterministic and can be expressed by a neural network, but has a random argument which results in a distribution on translated images.

This change is reflected in the UDA training objective by replacing the loss of the translated image with the *expected loss* of the translated image:

$$\mathcal{L}(F_t) = \sum_{(x,y)\in\mathcal{S}} \mathbf{E}_{\mathbf{v}}\left[\mathcal{L}_{ce}(F_t(\mathbf{T}[x,\mathbf{v}]),y)\right] + \sum_{x\in\mathcal{T}} \mathcal{L}_{adv}(E(F_t(x))), \tag{6.4}$$

where the expectation is taken with respect to the random vector $\mathbf{v} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, driving the stochastic translation. We note that during training we create minibatches by first sampling images from $\mathcal{S}$ and then sampling $\mathbf{v}$ once per image, effectively replacing the integration in the expectation with a Monte Carlo approximation.

Our stochastic translation network relies on MUNIT [143]: we start from reconstructing images in each domain through content and style encodings, where content is fed to the first layer of a generator whose subsequent layers are modulated by style-driven Adaptive Instance Normalization [151] - this amounts to minimizing the following domain-specific autoencoding objectives:

$$L_s = \sum_{x \in \mathcal{S}} \|x - G_s(C_s(x), S_s(x))\|,$$
$$L_t = \sum_{x \in \mathcal{T}} \|x - G_t(C_t(x), S_t(x))\|,$$

where C_s, S_s, G_s are the content-encoder, style-encoder and generator networks for the source domain s respectively, while C_t, S_t, G_t are those of the target domain t.

The basic assumption is that the commonalities between two domains are captured by the shared content space, allowing us to pass content from the source image to its target counterparts, as also shown in Fig. 6.1. The uncertainty in the translation is captured by a domain-specific style encoding that is inherently uncertain given the source image.

This results in the following stochastic translation function from source to target:

$$\mathbf{T}[x, \mathbf{v}] \doteq G_t(C_s(x), \mathbf{v}), \ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \ x \in \mathcal{S},$$

where we encode the content of the source image through $C_s(x)$ and then pass it to the target-domain generator G_t that is driven by a random style code $\mathbf{v}$. A similar translation is established between the target and source domains, and adversarial losses on both domains ensure that the resulting translations appear as realistic samples of the respective domains.

The alignment of the shared latent space for content is enforced by a cycle translation objective:

$$L^c_{cycle} = \|C_t(G_t(C_s(x), \mathbf{v})) - C_s(x)\|_2, \ x \in \mathcal{S}, \ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

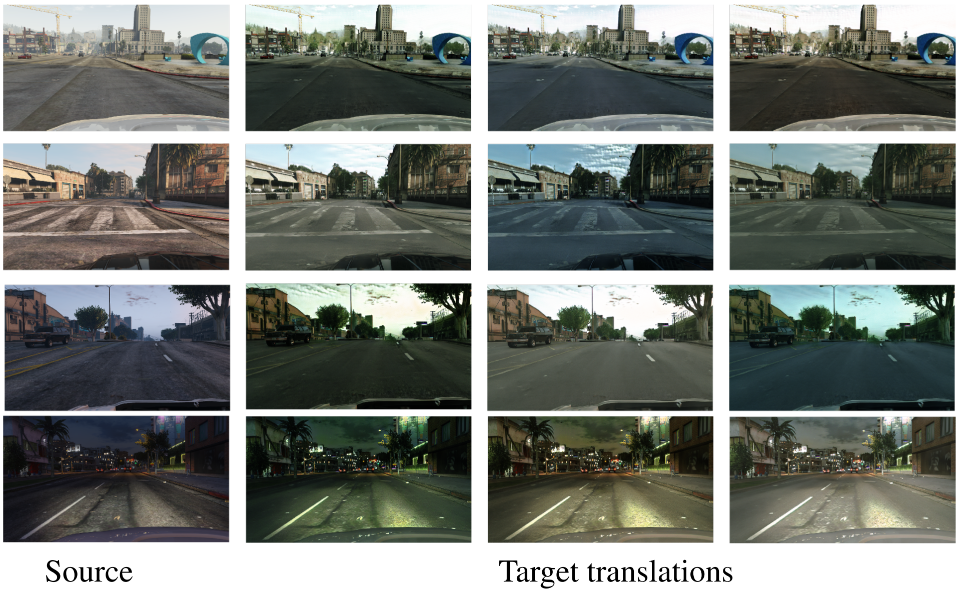


Figure 6.2: Diverse translations of images from the GTA source dataset to the Cityscapes target dataset: we observe that even though the content and pixel semantics stay intact, we generate diverse variants of the same scene, effectively capturing more faithfully the data distribution in the target domain.

ensuring that regardless of the random style code, we can recover the original content $C_s(x)$ by encoding the translated image through the respective content encoder. A similar loss is used for the style code:

$$L^s_{cycle} = \|S_t(G_t(C_s(x), \mathbf{v})) - \mathbf{v}\|_2,\ x \in \mathcal{S},\ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

We preserve semantic information during translation by imposing a semantic consistency constraint to our stochastic translation network using a fixed segmentation network F pretrained on source and target data using Eq. 6.2. Given an image x we obtain the predicted labels before translation as $p = \text{argmax}(F(x))$ and enforce semantic consistency during translation using an objective of the following form:

$$L_{sem} = \mathcal{L}_{ce}(F(\mathbf{T}[x, \mathbf{v}]), p). \tag{6.5}$$

The losses are applied to translations to both domains since unlike UDA, there is no special ‘source’ and ‘target’ domain.

We argue that stochastic translation provides us with a natural mechanism to handle UDA problems with large domain gaps where things may unavoidably get 'lost in translation'; the content cycle constraint can help preserve semantics during translation, while the random style allows the translated image appearance to vary freely, avoiding a deterministic and blunt translation.

This is demonstrated in Fig. 6.2, where we show some of the samples obtained by our method: we observe that our method generates sharp samples of high variability and noticeable diversity. As we show in the experimental results section, this results in substantially improved UDA accuracy. We also note that our approach includes deterministic translation as a special case, since the network can always learn to ignore the source of stochasticity if that is not useful - hence deterministic translation-based results provide effectively a lower bound on what our method can deliver.

6.3.3 Stochastic translation and pseudo-labelling

Having shown how stochastic translation from the source to the target domain can be integrated in the basic formulation of UDA, we now turn to exploiting stochastic translation from the target to the source domain, which is freely provided by the cycle-consistent formulation of [143].

In particular we consider a complementary segmentation network, F_s, that operates in the source domain and can be directly supervised from the labeled source dataset based on a cross-entropy objective:

$$\mathcal{L}(F_s) = \sum_{(x,y)\in\mathcal{S}} \mathcal{L}_{ce}(F_s(x), y). \tag{6.6}$$

This network can provide labels for the target-domain images, once these are translated from the target to the source domain; these pseudo-labels of the target data can in turn be used to supervise the target-domain network through a cross-entropy loss.

In the case of deterministic translation pseudo-labels would be obtained by the following expression:

$$\hat{y}(x) = F_s(\mathbf{I}[x]), \quad x \in \mathcal{T}, \tag{6.7}$$

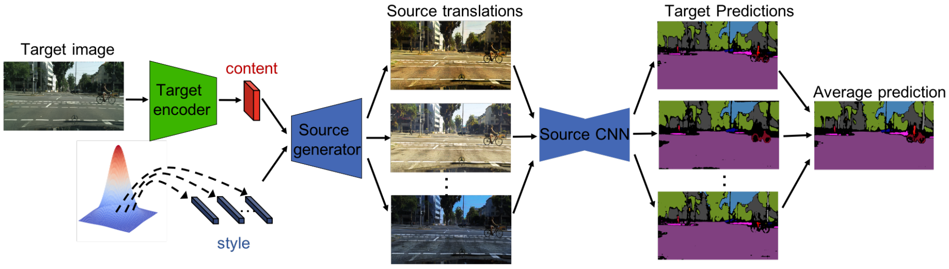


Figure 6.3: Stochastic translation for pseudo-labeling: the target image (left) results in multiple target-domain translations (middle) which are processed by the source-domain network, F_s and averaged to produce pseudo-labels for the target image; the latter are used to supervise the target-domain network F_t through a cross-entropy loss.

where $\mathbf{I}$ is the inverse transform from the target to the source domain, and $\hat{y}$ indicates the pixel-level posterior distribution on labels.

In our case however we have a whole distribution on translations for every image in $\mathcal{T}$. We realise that we can exploit multiple samples to obtain a better estimate of the pseudo-labels. In particular we form the following Monte Carlo estimate of pseudo-labels:

$$\begin{aligned} \hat{y}(x) &= E_{\mathbf{v}}\left[F_s(\mathbf{I}[x,\mathbf{v}])\right], \quad x \in \mathcal{T}, \mathbf{v} \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \\ &\simeq \frac{1}{K}\sum_{k=1}^{K} F_s(\mathbf{I}[x,\mathbf{v}_k]), \end{aligned}$$

where $\mathbf{v}_k$ are independently sampled from the normal distribution. As shown in Fig. 6.3 the label maps obtained through this process tend to have fewer errors and be more confident, since averaging the results obtained by different translations can be expected to cancel out the fluctuation of the predictions around their ground-truth value.

Our experimental results indicate that using $K = 10$ yields substantially better results than using a single sample. We also note that pseudo-label generation is a one-off process done prior to training the target-domain network, and consequently the number of samples, K, does not affect training time.

6.3.4 Stochasticity-driven training of diverse network ensembles

An experimental approach that has been recently adopted by several recent works [171, 7] consists in ensembling different networks trained for UDA, and using their predictions as an enhanced pseudo-labeling mechanism. For instance in [171] this was accomplished by modifying one of the main design parameters of their phase-driven translation algorithm. A main recipe for successful network ensembling is to generate complementary networks, so that they make uncorrelated errors, which hopefully cancel out.

Based on the understanding that the stochasticity driving our translation mechanism can be seen as implementing appearance-level dataset augmentation in the target domain, we introduce a simple twist to the translation mechanism that allows us to train networks that operate in different regimes. For this we scale by a constant the variance of the normal distribution used to sample the random style code - this amounts to generating more diverse translations than those suggested by the image statistics of the target domain. On one hand this trains a target network that can handle a broader range of inputs, but on the other hand it may waste capacity to handle unrepresentative samples.

We train two such networks, one with the variance left intact and the other with the variance scaled by 10, and average their predictions with those of the source-domain network described in the previous subsection as shown in Fig. 6.4. Our results show that this triplet of networks yields a clear boost over the baseline operating with a single network.

Further following common practice in UDA we use the resulting ensembling results as pseudo-labels in the next round of training - this yields further improvements, as documented in detail in the experimental results section.

6.3.5 Training objectives

Having described the components of our method, we now summarize the losses used for training our networks.

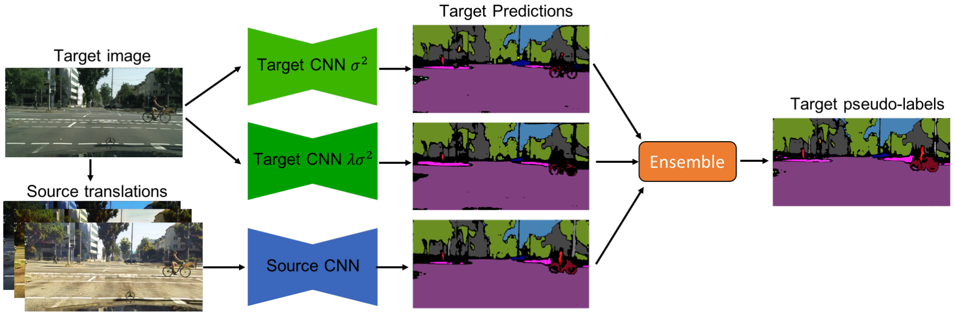


Figure 6.4: Ensembling of a triplet of networks — two target networks trained with different degrees of stochasticity in the translation (σ^2) and a source network — for robust pseudo-labeling.

Firstly, we train our stochastic translation network using the process of [143] and introduce a semantic consistency loss as in [15] to preserve semantics during translation.

For the target-domain network the basic objective has already been provided in Eq. 6.4, where $\mathcal{L}_{ce}$ is the standard cross-entropy loss and $\mathcal{L}_{adv}$ is the adversarial entropy minimization objective [156]. A more sophisticated objective can train this network with pseudo-labels, obtained either from a source-domain network as described in Sec. 6.3.3 or from the ensembling of multiple networks, as described in Sec. 6.3.4. In that case the objective becomes:

$$\begin{split}\mathcal{L}(F_t) = \sum_{(x,y)\in\mathcal{S}} \mathbf{E}_{\mathbf{v}}\left[\mathcal{L}_{ce}(F_t(\mathbf{T}[x,\mathbf{v}]),y)\right] + \\ \sum_{x\in\mathcal{T}} \mathcal{L}_{adv}(E(F_t(x))) + \sum_{x\in\mathcal{T}} \mathcal{L}^{\theta}_{ce}(F_t(x), \text{argmax}(\hat{y})),\end{split} \tag{6.8}$$

where the cross entropy loss $\mathcal{L}^{\theta}_{ce}(F_t(x))$ is only applied on pseudo-labels where the dominant class has a score above a certain threshold θ. Similar to [172] we use class-wise confidence thresholds to address the inter-class imbalance and avoid ignoring hard classes. Specifically, for each class c the threshold θ_c equals to the probability ranked at $r * N_c$, where N_c is the number of pixels predicted to belong in class c and r is the proportion of pseudo-labels we want retain. We provide more details in the Appendix (Sec. 6.6).

Finally, for the source-domain network, we observed experimentally that we obtain better results by adding an entropy-based regularization to the output of F_s when it is driven by translated target images - this ensures that the source network will correctly classify the source images, while placing its boundaries far from areas populated by synthetic source-domain images. The objective function for the source network becomes:

$$\mathcal{L}(F_s) = \sum_{(x,y)\in\mathcal{S}} \mathcal{L}_{ce}(F_s(x),y) + \sum_{x\in\mathcal{T}} \mathbf{E}_{\mathbf{v}}\left[\mathcal{L}_{adv}(F_s(\mathbf{I}[x,\mathbf{v}]))\right], \tag{6.9}$$

forming the source-domain counterpart to the objective encountered in Eq. 6.4. When pseudolabels are available for the target domain the objective function becomes:

$$\begin{aligned}\mathcal{L}(F_s) = \sum_{(x,y)\in\mathcal{S}} \mathcal{L}_{ce}(F_s(x),y) + \sum_{x\in\mathcal{T}} \mathbf{E}_{\mathbf{v}}\left[\mathcal{L}_{adv}(F_s(\mathbf{I}[x,\mathbf{v}]))\right] \\ + \sum_{x\in\mathcal{T}} \mathcal{L}^{\theta}_{ce}(F_s(\mathbf{I}[x,\mathbf{v}]), \operatorname{argmax}(\hat{y})),\end{aligned} \tag{6.10}$$

forming the source-domain counterpart of the objective in Eq. 6.8.

6.4 Experiments

We evaluate the proposed approach on two common UDA benchmarks for semantic segmentation. In particular we use the synthetic dataset GTA5 [23] or SYNTHIA [24] with ground-truth annotations as the source domain and the Cityscapes [25] dataset as the target domain with no available annotations during training. We evaluate the performance using the mean intersection over union score (mIoU) across semantic classes on the Cityscapes validation set.

6.4.1 Datasets

Cityscapes [25] is a real-world dataset of diverse urban street scenes collected from different cities. We use 2975 training images and 500 validation images with resolution 2048 × 1024. We resize the images to 1024 × 512. We train the image translation network and the segmentation network using the training set and report the results on the validation set.

GTA5 [23] consists of 24966 synthesized images captured from a video game. The original images have resolution 1914 × 1052 and they are resized to 1024 × 512 for training. GTA5 provides pixel-level semantic annotations of 33 classes. Similar to other studies, we use the 19 common classes between GTA5 and Cityscapes.

SYNTHIA [24] consists of synthesized images rendered from a virtual city. We use SYNTHIA-RAND-CITYSCAPES subset which has 9400 annotated images with resolution 1280 × 760. We use the 16 common classes between SYNTHIA and Cityscapes for training and we evaluate the performance on 16 classes and a subset of 13 classes following previous studies [16, 156, 171, 157].

6.4.2 Implementation details

Stochastic translation network: We rely on MUNIT [143] to establish a stochastic translation across the source and target domain. Images from the source and the target domain are resized to 1024 × 512 and cropped to 400 × 400. We train the network for 600000 iterations with batch size 1 and a learning rate starting 0.0001 and decreasing by half every 100000 iterations.

Semantic segmentation network We train two different architectures, i.e., DeepLabV2 [181] with ResNet101 [182] backbone, and FCN-8s [119] with VGG-16 [128] backbone. We train DeepLabV2 with ResNet101 using Stochastic Gradient Descent optimizer with initial learning rate 2.5×10^{-4}, momentum 0.9 and weight decay 1×10^{-4}. The learning rate is adjusted according to the poly learning rate scheduler with a power of 0.9. We train FCN-8s with VGG-16 using ADAM with initial learning rate 1×10^{-5} and momentum 0.9 and 0.99. The learning is decreased by a factor $\gamma = 0.1$ every 50000 iterations. We use the same discriminator for both the DeepLabV2 and FCN-8s. The discriminator used to adapt the entropy maps is similar to [183]. It has 4 convolutional layers, each followed by a leaky-ReLU layer with negative slope of 0.2. The last layer is a binary classification layer classifying the inputs either as source or target.

6.4.3 Results

Stochastic translation: We start by examining in how stochastic translation improves performance compared to deterministic translation. In all cases the segmen-

Method	Output space	Pixel space	mIoU
ADVENT [156]	✓		43.8
ADVENT *	✓		42.9
ADVENT *+ CycleGAN*	✓	✓	45.1
Ours	✓	✓	46.2
Ours w/ L_{sem}	✓	✓	**46.6**

Table 6.1: GTA to Cityscapes UDA using stochastic translation: We train ADVENT using synthetic images obtained from deterministic translation (CycleGAN) and stochastic translation (Ours). We observe a clear improvement thanks to pixel-space alignment based on stochastic translation. * denotes our retrained models.

tation model is DeepLabV2 [181] and the source and target datasets are GTA5 [23] and Cityscapes [25] respectively.

In Table 6.1 we start with an apples-to-apples comparison that builds on directly on the ADVENT baseline [156]; the first two rows compare the originally published and our reproduced numbers respectively. The third row shows the substantial improvement attained by training the system of ADVENT using translated images - which amount to training with Eq. 6.3. The forth row reports our stochastic translation-based result, amounting to training with Eq. 6.4. We observe a substantial improvement, that can be attributed solely to the stochasticity of the translation. The last row shows that imposing a semantic consistency constraint as described in Eq. 6.5 further improves the performance.

Pseudo labeling: As discussed in Section 6.3.3 we translate from the target to the source domain and generate pseudo labels for the target data. The first three rows in Table 6.2 show the impact of the number of samples K, on performance. Averaging the predictions of multiple translations for a given target image improves the performance and allows to obtain better pseudo labels for the target domain. Our results show that using 10 samples yields better performance. In rows 4, 5 of the same table we report the performance obtained from the two target networks trained with different degrees of stochasticity in the translation as described in Sec. 6.3.4. Averaging the prediction of the three networks gives the best results, indicating the

F_s, K=1	F_s, K=5	F_s, K=10	F_t, $\sigma^2 = 1$	F_t, $\sigma^2 = 10$	mIoU
✓					43.3
	✓				44.0
		✓			44.4
			✓		46.6
				✓	46.1
		✓	✓		47.7
		✓		✓	47.6
			✓	✓	47.7
		✓	✓	✓	**48.2**

Table 6.2: Performance of different models and their combinations. The first 3 rows show the performance of the source network F_s when averaging the predictions of multiple translations K, of a target image while rows 4, 5 show the performance of the target networks F_t, trained with different degrees of stochasticity (σ^2) in the translation. Averaging the predictions of multiple translations and combining the three models allows us to obtain better pseudo-labels for the target domain.

complementary of the model predictions. We also provide qualitative results in the Appendix (Sec. 6.6).

Network ensembling: Table 6.3 shows the results obtained in three rounds of pseudo-labeling and training, following the approach of [171, 16, 7]. In the first round ($R = 0$) we train the target and source networks with Eq. 6.4 and Eq. 6.9 respectively using the synthetic and real data and average the predictions of the three models to generate pseudo-labels for the target data. In the second round (R=1) we use the generated pseudo-labels as ground-truth labels to train the target and source networks with Eq. 6.8 and Eq. 6.10 respectively. We observe that the pseudo-labels obtained by ensembling improve the performance of each individual network, as well as the ensemble obtained in the last round (R=2).

Benchmark results We use DeepLabV2 [181] with ResNet101 [182] backbone, and FCN-8s [119] with VGG-16 [128] for the segmentation and compare with [160, 156, 16, 184, 171, 157, 174, 185, 186, 177, 7] which use exactly the same experimental settings. We report both the results obtained using a single target network and the results obtained by ensembling. We provide qualitative results in the

Model	mIoU
F_s (R=0)	44.4
F_t, $\sigma^2 = 1$ (R=0)	46.6
F_t, $\sigma^2 = 10$ (R=0)	46.1
Ens (R=0)	48.2
F_s (R=1)	49.1
F_t, $\sigma^2 = 1$ (R=1)	50.1
F_t, $\sigma^2 = 10$ (R=1)	50.9
Ens (R=1)	52.0
F_s (R=2)	51.3
F_t, $\sigma^2 = 1$ (R=2)	53.0
F_t, $\sigma^2 = 10$ (R=2)	52.9
Ens (R=2)	54.3

Table 6.3: Ablation study on GTA to Cityscapes. Averaging the predictions (Ens) of a source network F_s, and two target networks F_t trained with different degrees of stochasticity (σ^2) in the translation allows to obtain robust pseudo-labels, while using multiple rounds R of pseudo-labeling and training improves the overall performance.

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
VGG backbone																				
AdaptSegNet[160]	87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0
AdvEnt[156]	86.9	28.7	78.7	28.5	25.2	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	81.5	26.0	17.2	18.9	11.7	1.6	36.1
BDL [16]	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	23.3	82.7	25.8	28.0	2.3	25.7	19.9	41.3
LTIR [184]	**92.5**	**54.5**	83.9	34.5	25.5	31.0	30.4	18.0	84.1	39.6	**83.9**	53.6	19.3	81.7	21.1	13.6	17.7	12.3	6.5	42.3
FDA-MBT [171]	86.1	35.1	80.6	30.8	20.4	27.5	30.0	26.0	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	**29.9**	14.6	24.0	42.2
PCEDA [157]	90.7	49.8	81.9	23.4	18.5	**37.3**	35.5	34.3	82.9	36.5	75.8	**61.8**	12.4	83.2	19.2	26.1	4.0	14.3	21.8	42.6
DPL-Dual (Ensemble) [7]	89.2	44.0	83.5	**35.0**	24.7	27.8	**38.3**	25.3	84.2	39.5	81.6	54.7	**25.8**	83.3	29.3	49.0	5.2	30.2	32.6	46.5
Ours	91.1	43.2	84.1	34.6	25.5	25.8	33.7	31.3	84.7	**44.9**	83.1	55.3	23.5	81.6	23.1	34.3	6.3	**32.7**	34.8	46.0
Ours (Ensemble)	91.0	40.7	**84.7**	33.8	**27.1**	30.9	33.1	**35.1**	**85.3**	44.7	82.9	56.8	23.4	**86.2**	**36.5**	**50.3**	2.8	27.8	**36.6**	**47.9**
ResNet101 backbone																				
AdvEnt[156]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
BDL [16]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
LTIR [184]	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
FDA-MBT [171]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
PCEDA [157]	91.0	49.2	85.6	37.2	29.7	33.7	38.1	39.2	85.4	35.4	85.1	61.1	32.8	84.1	**45.6**	46.9	0.0	34.2	44.5	50.5
TPLD [174]	**94.2**	**60.5**	82.8	36.6	16.6	39.3	29.0	25.5	85.6	44.9	84.4	60.6	27.4	84.1	37.0	47.0	**31.2**	36.1	**50.3**	51.2
Wang et al. [186]	90.5	38.7	**86.5**	41.1	32.9	**40.5**	48.2	42.1	86.5	36.8	84.2	**64.5**	**38.1**	87.2	34.8	50.4	0.2	41.8	54.6	52.6
PixMatch [177]	91.6	51.2	84.7	37.3	29.1	24.6	31.3	37.2	86.5	44.3	85.3	62.8	22.6	87.6	38.9	52.3	0.65	37.2	50.0	50.3
DPL-Dual (Ensemble) [7]	92.8	54.4	86.2	41.6	32.7	36.4	**49.0**	34.0	85.8	41.3	**86.0**	63.2	34.2	87.2	39.3	44.5	18.7	**42.6**	43.1	53.3
Ours	93.3	56.5	85.9	41.0	33.1	34.8	43.8	**43.8**	86.6	46.5	82.5	61.1	30.4	87.0	39.7	50.7	8.8	34.9	46.8	53.0
Ours (Ensemble)	93.4	55.8	86.4	**44.4**	**36.1**	34.6	45.0	39.8	**86.9**	**48.0**	84.4	61.7	30.9	**87.7**	44.9	**55.9**	11.1	38.4	45.4	**54.3**

Table 6.4: Quantitative comparison on GTA5→Cityscapes. We present per-class IoU and mean IoU (mIoU) obtained using VGG and ResNet101 backbones.

The results for the **GTA-to-Cityscapes** benchmark are summarized in Table 6.4. Our results show that our method achieves state-of-the-art performance and outperforms previous methods. When compared with other approaches relying on both deterministic translation and multiple rounds of pseudo-labeling and training [16, 7, 171], our approach performs better while at the same time is simpler. In particular, [16] and [7] train both the image translation and segmentation networks multiple times and use complex warm-up stages [7]. On the other hand we train the image translation network only once and use the same image translation network in all rounds of pseudo-labeling and training.

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*sky*	*person*	*rider*	*car*	*bus*	*motocycle*	*bicycle*	mIoU	mIoU*
VGG backbone																		
AdvEnt[156]	67.9	29.4	71.9	6.3	0.3	19.9	0.6	2.6	74.9	74.9	35.4	9.6	67.8	21.4	4.1	15.5	31.4	36.6
BDL [16]	72.0	30.3	74.5	0.1	0.3	24.6	10.2	25.2	80.5	80.0	54.7	23.2	72.7	24.0	7.5	44.9	39.0	46.1
FDA-MBT [171]	84.2	35.1	78.0	**6.1**	0.4	27.0	8.5	22.1	77.2	79.6	55.5	19.9	74.8	24.9	14.3	40.7	40.5	47.3
PCEDA [157]	79.7	35.2	78.7	1.4	0.6	23.1	10.0	28.9	79.6	81.2	51.2	**25.1**	72.2	24.1	16.7	**50.4**	41.1	48.7
DPL-Dual (Ensemble) [7]	83.5	38.2	**80.4**	1.3	**1.1**	**29.1**	**20.2**	32.7	81.8	83.6	55.9	20.3	79.4	26.6	7.4	46.2	43.0	50.5
Ours	83.3	40.9	80.3	1.4	0.6	24.8	16.9	31.1	**82.4**	**84.1**	**57.4**	20.1	83.2	30.3	16.0	44.5	43.6	51.5
Ours (Ensemble)	**88.7**	**41.6**	80.3	1.0	0.7	23.6	14.3	**33.1**	81.9	81.1	57.2	21.1	**84.1**	**33.4**	**19.1**	44.3	**44.1**	**52.3**
ResNet101 backbone																		
AdvEnt[156]	85.6	42.2	79.7	-	-	-	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	-	48.0
LTIR [184]	92.6	53.2	79.2	-	-	-	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	-	49.3
BDL [16]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
FDA-MBT [171]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	**31.1**	83.9	40.8	**38.4**	51.1	-	52.5
PCEDA [157]	85.9	44.6	80.8	-	-	-	24.8	23.1	79.5	83.1	57.2	29.3	73.5	34.8	32.4	48.2	-	53.6
TPLD [174]	80.9	44.3	82.2	19.9	0.3	**40.6**	20.5	30.1	77.2	80.9	60.6	25.5	84.8	41.1	24.7	43.7	47.3	53.5
Wang et al. [186]	79.4	34.6	**83.5**	**19.3**	**2.8**	35.3	**32.1**	**26.9**	78.8	79.6	**66.6**	30.3	86.1	36.6	19.5	**56.9**	48.0	54.6
PixMatch [177]	**92.5**	**54.6**	79.8	4.7	0.08	24.1	22.8	17.8	79.4	76.5	60.8	24.7	85.7	33.5	26.4	54.4	46.1	54.5
DPL-Dual (Ensemble) [7]	87.5	45.7	82.8	13.3	0.6	33.2	22.0	20.1	83.1	86.0	56.6	21.9	83.1	40.3	29.8	45.7	47.0	54.2
Ours	85.8	41.7	82.4	7.6	1.9	33.2	26.5	18.4	83.3	86.5	62.0	29.7	83.9	52.1	34.6	51.4	48.8	56.8
Ours (Ensemble)	87.2	44.1	82.1	6.5	1.4	33.1	24.7	17.9	**83.4**	**86.6**	62.4	30.4	**86.1**	**58.5**	36.8	52.8	**49.6**	**57.9**

Table 6.5: Quantitative comparison on SYNTHIA→Cityscapes. We present per-class IoU and mean IoU (mIoU) obtained using VGG and ResNet101 backbones. mIoU and mIoU* are the mean IoU computed on the 16 classes and the 13 subclasses respectively.

The results for the **SYNTHIA-to-Cityscapes** benchmark are reported in Table 6.5. Following the evaluation protocol of previous studies [16, 157, 171, 156, 7] we report the mIoU of our method on 13 and 16 classes. We observe that our method outperforms previous state-of-the art methods by a large margin (+3.7 compared to DPL[7]). We note here that the domain gap between SYNTHIA and Cityscapes is much larger compared to the domain gap between GTA and Cityscapes. We attribute the substantial improvements obtained by our method to the stochasticity in the translation which allows us to better capture the range of scenes encountered in the two domains and to generate sharp samples even in cases where there is a large domain gap between the two domains.

6.5 Conclusion

In this work we have introduced stochastic translation in the context of UDA for semantic segmentation of urban scenes and showed that we can reap multiple benefits by acknowledging that certain structures are 'lost in translation' across two domains. The networks trained directly through stochastic translation clearly outperforms all comparable counterparts, while we have also shown that we retain our edge when combining our approach with more involved UDA approaches such as pseudo-labeling and ensembling.

6.6 Appendix

We report results from additional ablation studies and class-wise IoU for ablation studies already presented in Sec. 6.4.3. We report results on GTA-to-Cityscapes using DeepLab-V2 with ResNet-101. We also report qualitative results.

6.6.1 Quantitative results

Training objective of the source domain network.

As we mentioned in Sec. 6.3.5, for the source domain network we observed experimentally that we obtained better results by adding an entropy-based adversarial loss $\mathcal{L}_{adv}$, to the output of source-domain network F_s when it is driven by translated target images. In Table 6.6 we report results obtained with and without the entropy-based adversarial loss (Eq. 6.9). Adding the entropy-based regularization improves performance for most classes.

Loss	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
$\mathcal{L}_{CE}$	90.2	38.0	81.2	**29.1**	16.2	**24.4**	23.7	15.5	**84.0**	**38.8**	78.5	56.9	24.0	85.0	**36.4**	**47.0**	0.3	**31.8**	26.8	43.6
$\mathcal{L}_{CE}+\mathcal{L}_{adv}$	**90.5**	**39.4**	**82.0**	29.0	**21.4**	23.6	**28.6**	**17.8**	83.9	38.2	**79.8**	**56.9**	**26.0**	**85.1**	32.2	44.1	**3.8**	31.5	**30.1**	**44.4**

Table 6.6: Better performance is achieved by adding an entropy-based regularization $\mathcal{L}_{adv}$ to the output of source-domain network F_s when it is driven by translated target images.

Selection of r for pseudolabel generation.

In Table 6.7 and Table 6.8 we provide the per-class IoU and mIoU obtained on the validation set for different values of r in the first (R=0) and second (R=1) round of pseudo-labeling respectively. In both rounds the best performance is achieved for $r = 0.6$. In the second round of pseudolabeling the networks provide more confident predictions since the performance remains the same for almost all classes when $r \leq 0.5$.

r	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
1.0	92.1	47.8	84.3	36.5	27.9	31.5	36.6	24.5	85.4	41.2	81.6	61.4	30.1	86.3	37.6	47.3	1.3	28.7	32.7	48.2
0.8	97.4	69.8	93.4.	52.0	42.5	47.1	55.2	37.7	94.2	53.5	91.1	78.2	40.2	94.4	47.4	55.7	2.2	40.9	55.4	60.4
0.7	98.2	**72.1**	95.9.	63.4.	51.5	**55.2**	**64.0**	**42.5**	96.5	65.0	93.7	86.9	51.7	96.4	57.0	61.2	**2.7**	53.1	65.8	67.0
0.6	98.5	70.2	96.3	73.4	**57.5**	54.1	58.7	33.3	97.2	77.4	**93.8**	90.8	60.3	97.5	65.5	69.6	2.5	65.8	**71.9**	**70.2**
0.5	**98.6**	59.4	96.5	**79.0**	55.2	44.6	54.3	20.7	97.6	81.4	93.8	92.2	61.0	**97.9**	**70.6**	**74.1**	1.3	**73.3**	70.3	69.6
0.4	98.6	50.1	**96.6**	76.5	42.2	44.6	54.6	20.6	**97.7**	**82.3**	93.8	**92.3**	**61.1**	97.9	70.6	74.1	0.5	73.4	70.5	68.3
0.3	98.6	50.1	96.6	76.6	38.3	44.7	54.6	20.6	97.7	82.3	93.8	92.3	61.1	97.9	70.6	74.1	0.2	73.4	70.5	68.1

Table 6.7: Per-class IoU and mean (mIoU) obtained using different values of r for class-wise confidence threshold selection in the first round (R=0) of pseudolabeling. We observe that $r = 0.6$ gives the best results.

r	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
1.0	93.0	53.3	85.8	41.2	33.1	33.4	39.1	29.7	86.4	45.4	84.5	60.0	29.3	86.9	45.8	57.7	2.7	34.6	45.8	52.0
0.8	96.8	**67.4**	93.6	57.3	45.2	46.7	**54.2**	**41.9**	94.2	59.2	92.4	74.8	37.1	94.4	57.1	71.3	4.6	49.2	60.8	63.1
0.7	97.2	67.3	94.2	64.5	49.4	**49.0**	51.9	37.1	95.3	67.6	**92.5**	80.8	46.3	95.9	68.6	77.9	**5.4**	59.7	70.5	66.9
0.6	**97.3**	65.4	**94.5**	67.8	**51.3**	44.0	48.5	36.0	**95.7**	71.9	92.5	84.4	52.5	96.9	85.3	81.2	4.9	67.9	74.7	69.1
0.5	97.3	65.4	94.5	**68.1**	49.3	42.8	48.6	36.3	95.7	**72.2**	92.5	**84.7**	**52.7**	**97.0**	**87.5**	**81.6**	3.0	**68.6**	**75.1**	**69.1**
0.4	97.3	65.4	94.5	68.1	49.3	42.8	48.6	36.3	95.7	72.2	92.5	84.7	52.7	97.0	87.5	81.6	1.3	68.6	75.1	69.0
0.3	97.3	65.4	94.5	68.1	49.3	42.8	48.6	36.3	95.7	72.2	92.5	84.7	52.7	97.0	87.5	81.6	0.4	68.6	75.1	69.0

Table 6.8: Per-class IoU and mean (mIoU) obtained using different values of r for class-wise confidence threshold selection in the second round (R=1) of pseudolabeling. We observe that $r = 0.6$ gives the best results.

Class-wise IoU for ablation studies reported in Sec. 6.4.3.

In Table 6.9 we report the per-class IoU obtained from deterministic and stochastic translation (mIoU results only are reported in Table 6.1). In Table 6.10 we report the per-class IoU obtained from multiple rounds R of pseudo-labeling and training (mIoU only results are provided in Table 6.3). Multiple rounds of pseudolabeling and training yield improved performance.

6.6.2 Qualitative results

Diverse translation obtained using stochastic translation.

Fig. 6.5 shows diverse translations of images from the SYNTHIA source dataset to

Model	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bicycle	mIoU
ADVENT	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
ADVENT *	87.2	38.5	78.2	25.9	24.6	30.4	36.3	21.7	84.0	28.7	76.7	60.1	28.8	80.0	28.0	45.2	0.7	19.7	19.9	42.9
ADVENT *+ CycleGAN*	**91.9**	**51.5**	83.1	30.8	23.6	**32.0**	32.1	24.3	83.8	**38.5**	**82.3**	58.7	28.5	84.1	33.3	35.9	0.6	21.7	20.0	45.1
Ours	90.2	37.6	**84.1**	**33.0**	**25.1**	30.1	36.8	**28.4**	83.8	36.1	82.2	58.1	**29.6**	84.6	**34.4**	45.4	1.0	**26.2**	**30.8**	46.2
Ours w/ L_{sem}	92.1	49.9	83.5	29.1	24.7	30.3	**38.3**	27.2	**84.8**	34.4	81.1	**60.4**	28.1	**85.2**	33.0	**45.7**	**2.5**	23.8	30.4	**46.6**

Table 6.9: GTA to Cityscapes UDA using stochastic translation: We train ADVENT using synthetic images obtained from deterministic translation (CycleGAN) and stochastic translation (Ours). We observe a clear improvement thanks to pixel-space alignment based on stochastic translation. * denotes our retrained models.

Model	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bicycle	mIoU
source (R=0)	90.5	39.4	82.0	29.0	21.4	23.6	28.6	17.8	83.9	38.2	79.8	56.9	26.0	85.1	32.2	44.1	3.8	31.5	30.1	44.4
target, $\sigma^2 = 1$ (R=0)	92.1	49.9	83.5	29.1	24.7	30.3	38.3	27.2	84.8	34.4	81.1	60.4	28.1	85.2	33.0	45.7	2.5	23.8	30.4	46.6
target, $\sigma^2 = 10$ (R=0)	90.9	43.0	83.4	30.6	29.3	30.6	34.1	27.1	84.4	36.2	79.9	60.6	29.5	84.5	32.5	40.3	3.1	29.2	26.4	46.1
Ens (R=0)	92.1	47.8	84.3	36.5	27.9	31.5	36.6	24.5	85.4	41.2	81.6	61.4	30.1	86.3	37.6	47.3	1.3	28.7	32.7	48.2
source (R=1)	92.1	48.4	84.3	36.4	29.5	30.5	35.9	26.5	85.4	42.9	82.1	59.8	29.6	85.5	38.2	52.9	3.4	32.7	37.3	49.1
target, $\sigma^2 = 1$ (R=1)	92.1	47.5	85.1	38.3	29.4	32.9	35.4	32.1	85.9	46.8	81.7	60.5	30.4	86.6	35.7	51.1	4.4	34.9	41.0	50.1
target, $\sigma^2 = 10$ (R=1)	92.9	55.2	85.1	38.1	30.6	32.8	39.8	34.8	85.9	42.2	84.0	59.0	26.1	85.4	47.9	46.3	10.1	28.4	42.8	50.9
Ens (R=1)	93.0	53.3	85.8	41.2	33.1	33.4	39.1	29.7	86.4	45.4	84.5	60.0	29.3	86.9	45.8	57.7	2.7	34.6	45.8	52.0
source (R=2)	92.3	48.2	85.1	40.7	34.3	29.8	38.5	28.2	86.5	46.7	83.3	60.9	30.2	86.9	41.3	53.1	10.4	38.4	40.5	51.3
target, $\sigma^2 = 1$ (R=2)	93.3	56.5	85.9	41.0	33.1	34.8	43.8	43.8	86.6	46.5	82.5	61.1	30.4	87.0	39.7	50.7	8.8	34.9	46.8	53.0
target, $\sigma^2 = 10$ (R=2)	93.4	56.3	85.6	40.6	33.5	35.9	43.5	41.1	85.7	43.8	84.1	60.6	29.2	87.2	44.2	53.7	13.7	33.8	39.2	52.8
Ens (R=2)	93.4	55.8	86.4	44.4	36.1	34.6	45.0	39.8	86.9	48.0	84.4	61.7	30.9	87.7	44.9	55.9	11.1	38.4	45.4	54.3

Table 6.10: Ablation study on GTA→Cityscapes. Averaging the predictions (Ens) of a source network F_s, and two target networks F_t trained with different degrees of stochasticity (σ^2) in the translation allows to obtain robust pseudo-labels, while using multiple rounds R of pseudo-labeling and training improves the overall performance.

the Cityscapes target dataset. Fig. 6.6 and Fig. 6.7 show diverse translations of images from the Cityscapes target dataset to the SYNTHIA and GTA source datasets respectively. We observe that stochastic translation generates diverse samples that capture more faithfully the data distribution of the source domain and preserve the content of the original image allowing us to obtain more robust pseudolabels for the target data.

Stochastic versus deterministic translation.

Fig. 6.8 shows stochastic and deterministic translation of images from the GTA source dataset to the Cityscapes target dataset while Fig. 6.9 shows stochastic and deterministic translation of images from the SYNTHIA source dataset to the Cityscapes target dataset. Fig. 6.10 shows stochastic and deterministic translation of images from the Cityscapes target dataset to the GTA source dataset while Fig. 6.11

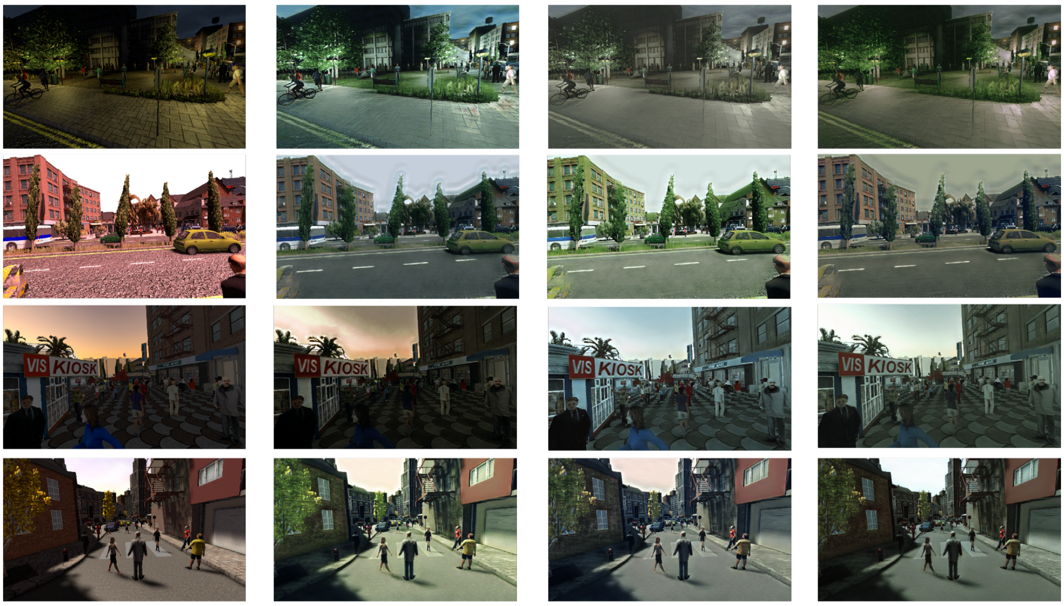


Figure 6.5: Diverse translations of images from the SYNTHIA source dataset to the Cityscapes target dataset: we observe that even though the content and pixel semantics stay intact, we generate diverse variants of the same scene, effectively capturing more faithfully the data distribution in the target domain.

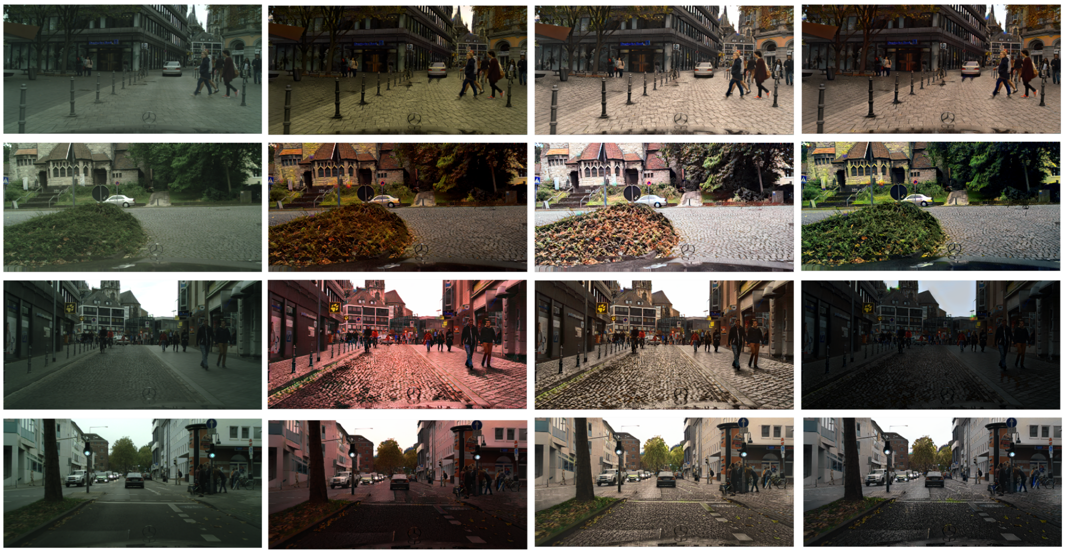


Figure 6.6: Diverse translations of images from the Cityscapes target dataset to the SYNTHIA source dataset: we observe that even though the content and pixel semantics stay intact, we generate diverse variants of the same scene, effectively capturing more faithfully the data distribution of the source domain. This allows us to generate more robust pseudolabels.

shows stochastic and deterministic translation of images from the Cityscapes target dataset to the SYNTHIA source dataset. We observe that stochastic translation

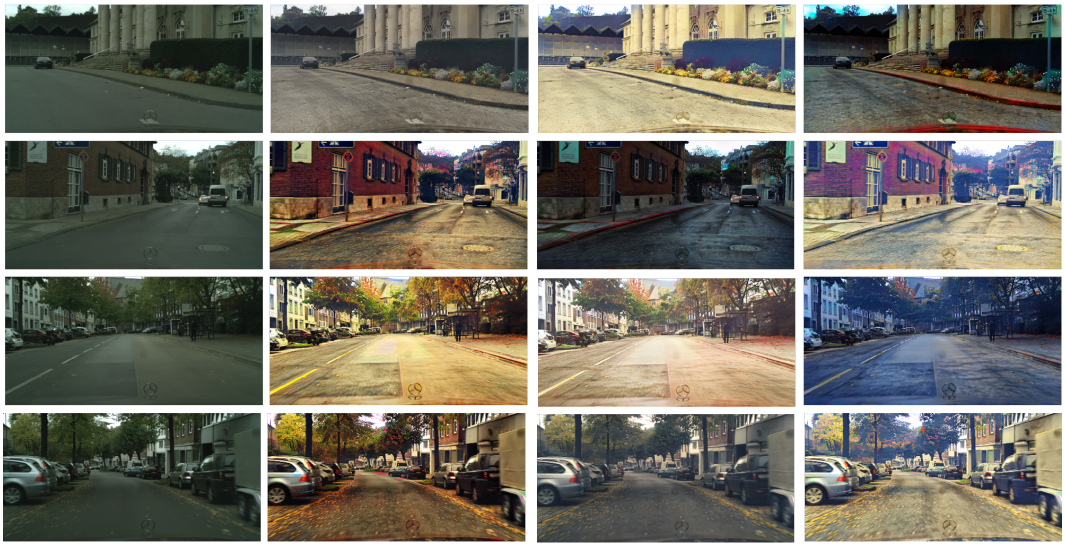


Figure 6.7: Diverse translations of images from the Cityscapes target dataset to the GTA source dataset: we observe that even though the content and pixel semantics stay intact, we generate diverse variants of the same scene, effectively capturing more faithfully the data distribution of the source domain. This allows us to generate more robust pseudolabels.

generates sharp samples of noticeable diversity compared to the deterministic translation that generates a single output.

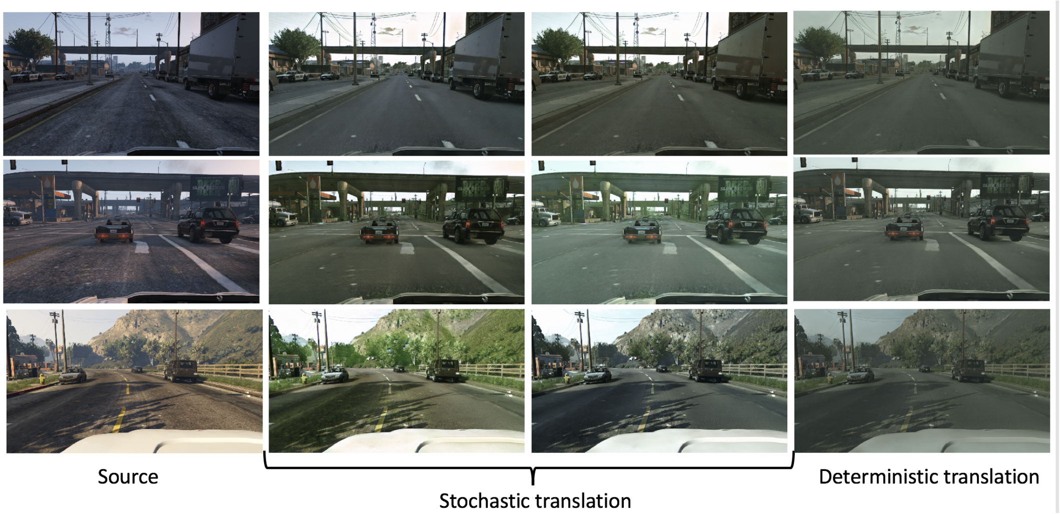


Figure 6.8: Stochastic and deterministic translation of images from the GTA source dataset to the Cityscapes target dataset.

Multiple rounds of pseudolabeling.

Fig. 6.12 shows the pseudo-labels obtained from the first (R=0) and second (R=1) round of pseudolabeling. We observe that the pseudolabels we obtained in the sec-

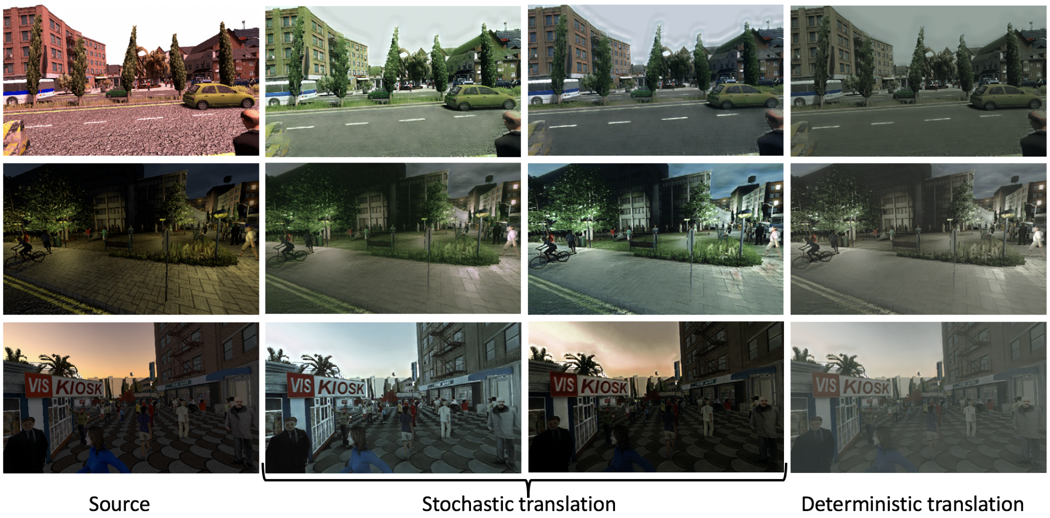


Figure 6.9: Stochastic and deterministic translation of images from the SYNTHIA source dataset to the Cityscapes target dataset.

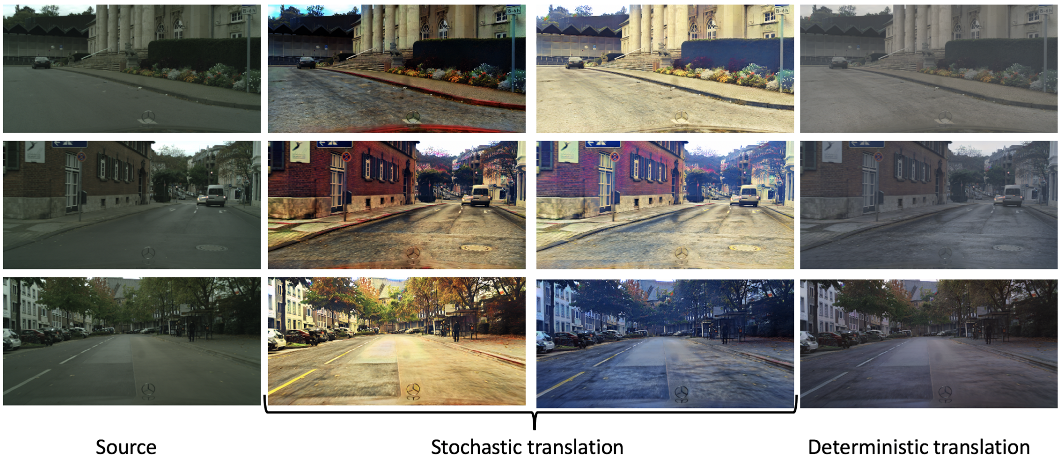


Figure 6.10: Stochastic and deterministic translation of images from the Cityscapes target dataset to the GTA source dataset.

ond round are more accurate allowing us to train more accurate models in the last round of training.

Robust pseudolabeling through ensembling.

Fig. 6.13 shows the pseudo-labels obtained by averaging the predictions of two target networks $F_{t,\sigma^2=1}$, $F_{t,\sigma^2=10}$ and a one source network F_s. Averaging the predictions allows us to generate more accurate pseudolabels.

Ensembling for improved segmentation performance.

Fig. 6.14 shows the predictions obtained by averaging the predictions of two target

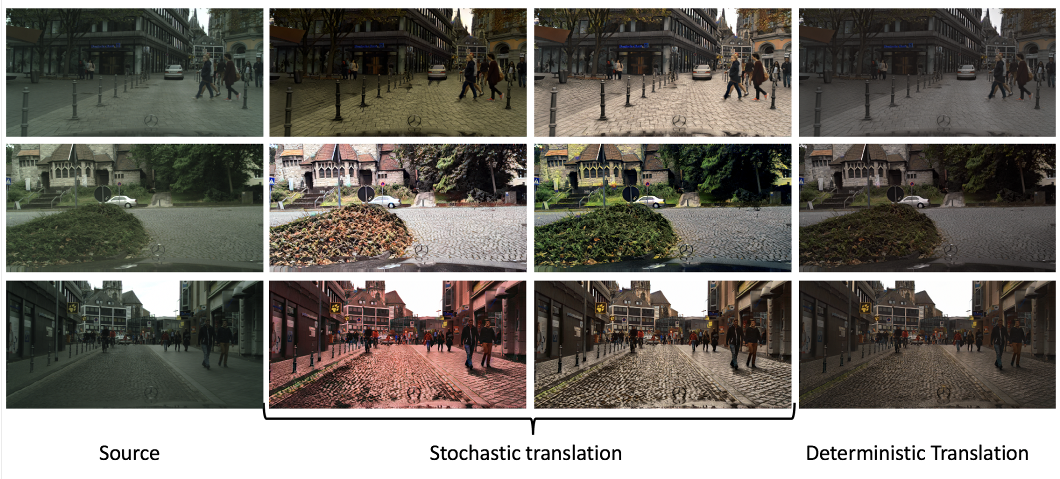


Figure 6.11: Stochastic and deterministic translation of images from the Cityscapes target dataset to the SYNTHIA source dataset.

networks $F_{t,\sigma^2=1}$, $F_{t,\sigma^2=10}$ and a one source network F_s. Averaging the predictions allows us to further improve performance by better distinguishing similar structures (e.g., road, sidewalk) and identifying small objects.

Qualitative comparison of the segmentation results.

Fig. 6.15 shows results segmentation results obtained by our method and DPL[7]. Our method generates better predictions that are closer to the ground-truth.

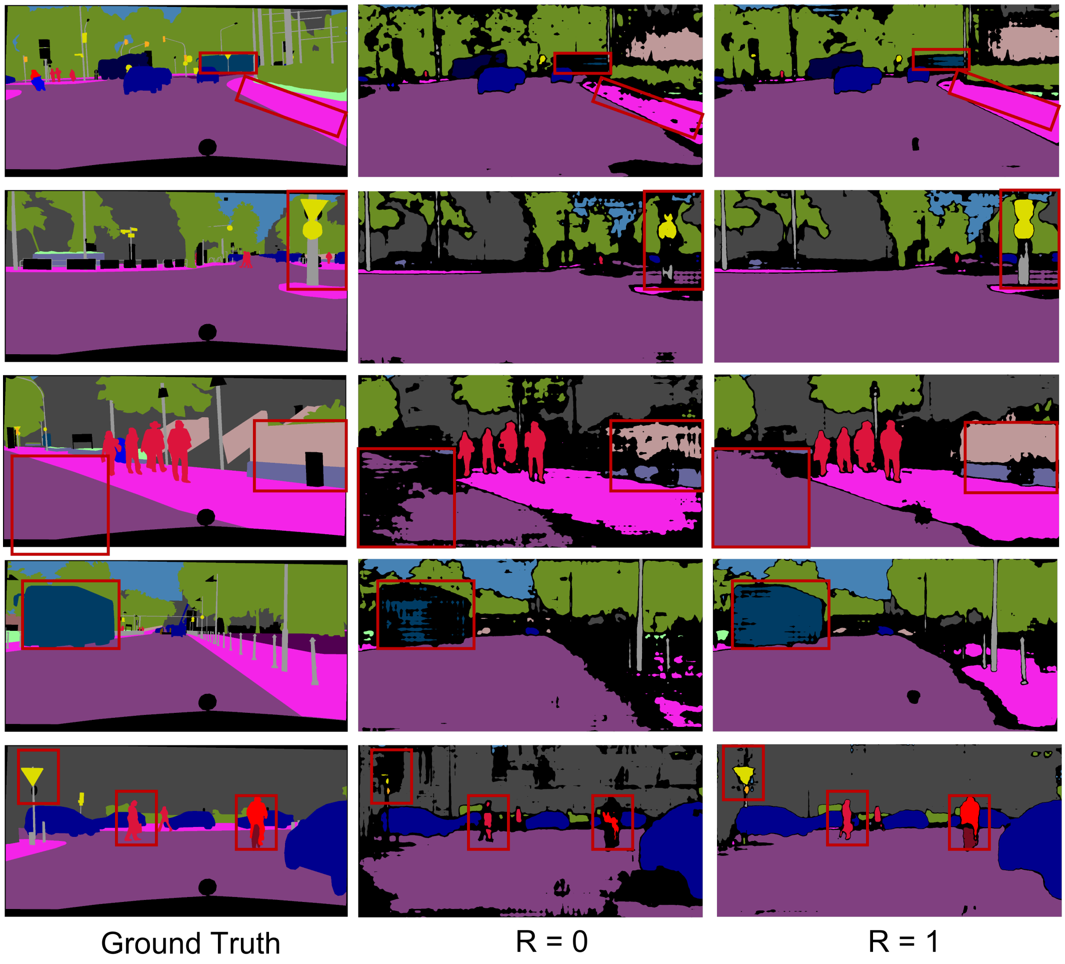


Figure 6.12: Visualization of pseudolabels obtained from the first (R=0) and second (R=1) round. Pseudolabels obtained in the second round are more accurate allowing us to train more robust models in the last round of training.

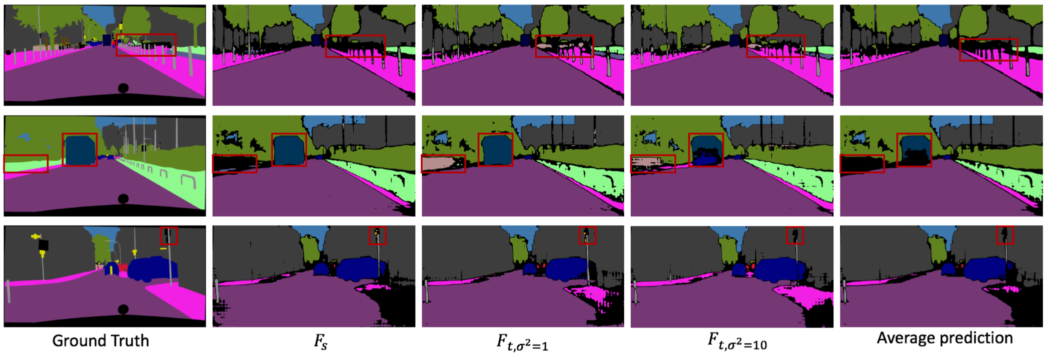


Figure 6.13: Visualization of pseudolabels obtained by averaging the predictions of two target networks $F_{t,\sigma^2=1}$, $F_{t,\sigma^2=10}$ and a one source network F_s. Averaging the predictions allows us to generate more accurate pseudolabels.

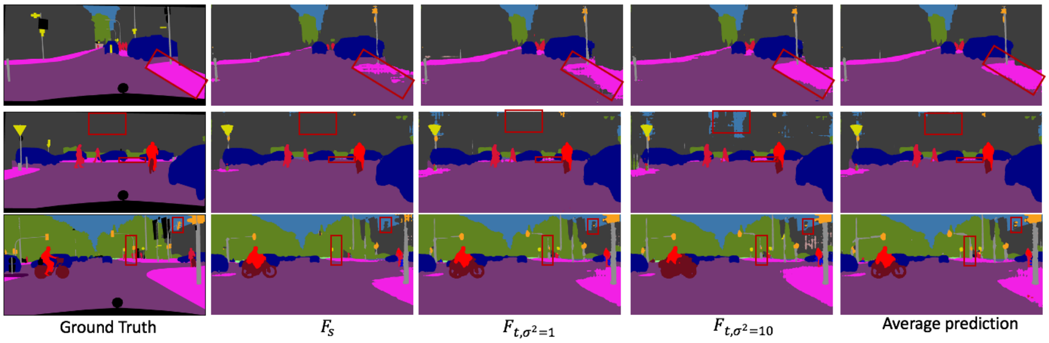


Figure 6.14: Visualization of results obtained by averaging the predictions of two target networks $F_{t,\sigma^2=1}$, $F_{t,\sigma^2=10}$ and a one source network F_s. Averaging the predictions allows us to further improve performance by better distinguishing similar structures (e.g., road, sidewalk) and identifying small objects.

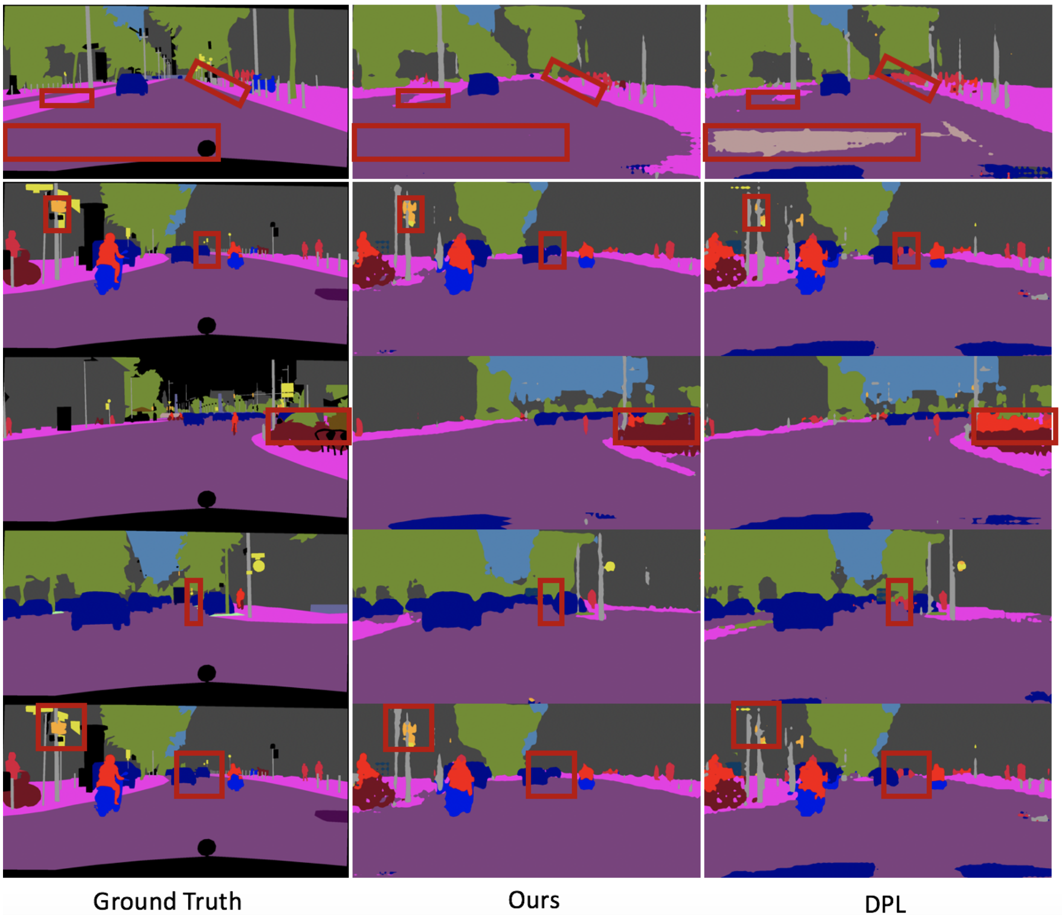


Figure 6.15: Qualitative comparison of our method with DPL[7]. Our method generates better predictions that are closer to the ground-truth.

Chapter 7

Conclusions

The main objective of the thesis was the development of methods that address the scarcity of carefully annotated data required to train accurate deep learning models. We focused on VERDICT MRI, an advanced imaging modality and we exploited labeled DW-MRI data from mp-MRI to train deep learning models that generalize well on VERDICT MRI. To this end, we proposed a semi-supervised and an unsupervised domain adaptation approach for prostate lesion segmentation. We further extended our approach for unsupervised domain adaptation in semantic segmentation of natural images. In Sec. 7.1 we provide a summary of our contributions and in Sec. 7.2 we discuss future directions.

7.1 Summary of contributions

Model-free prostate lesion characterization on VERDICT MRI

In Chapter 3, we investigated the potential of model-free prostate lesion classification on the raw VERDICT MRI data using FCNs. We also examined whether the raw VERDICT MRI data allows for better classification of prostate lesions compared to the raw DW data and the ADC map from the mp-MRI acquisition. Our results indicate that i) FCNs trained on VERDICT MRI achieve good performance in differentiating between malignant and benign lesions and ii) FCNs trained and evaluated on VERDICT MRI perform better than FCNs trained and evaluated on the raw DW data and the ADC from mp-MRI acquisitions.

Semi-supervised domain adaptation for lesion segmentation

In Chapter 4, we proposed a semi-supervised domain adaptation approach for lesion segmentation. Our approach relies on stochastic generative modelling to translate across two heterogeneous domains at pixel-space and exploits the inherent uncertainty in the cross-domain mapping to generate multiple outputs conditioned on a single input. In addition, we enforce semantic consistency between the real and synthetic images by exploiting both source-domain and target-domain lesion segmentation supervision to train target-domain networks operating on the synthetic images. This results in training networks that can generate diverse outputs while at the same time preserving critical structures. We further accommodate the statistical discrepancies between real and synthetic data by introducing residual adapters in the segmentation network. These capture domain-specific properties and allow the segmentation network to generalize better across the two domains. When compared to its deterministic counterparts, our approach yields substantial improvements across a broad range of dataset sizes, increasingly strong baselines, and evaluation metrics without increasing computational complexity. Specifically, our method requires the same computational resources as its deterministic counterparts during training. During test time only the segmentation network is used to provide the predictions for a given test image.

Unsupervised domain adaptation for lesion segmentation

In Chapter 5, we proposed an unsupervised domain adaptation approach for lesion segmentation. As in Chapter 4, we rely on stochastic generative modelling to translate across two heterogeneous domains at pixel-space and introduce two new loss functions that promote semantic consistency. Firstly, we introduce a semantic cycle-consistency loss in the source domain to ensure that the translation preserves the semantics. Secondly, we introduce a pseudo-labelling loss, where we translate target data to source, label them using a source-domain network, and use the generated pseudo-labels to supervise the target-domain network. When compared to several unsupervised domain adaptation approaches, our approach yields substantial improvements, that consistently carry over to the semi-supervised and supervised learning settings.

Unsupervised domain adaptation for segmentation of urban scenes

In Chapter 6, we proposed an unsupervised domain adaptation approach for semantic segmentation of urban scenes. As in Chapter 4, 5, we rely on stochastic generative modelling to capture inherent translation ambiguities. This allows us to (i) train more accurate target networks by generating multiple outputs conditioned on the same source image, leveraging both accurate translation and data augmentation for appearance variability, (ii) impute robust pseudo-labels for the target data by averaging the predictions of a source network on multiple translated versions of a single target image and (iii) train and ensemble diverse networks in the target domain by modulating the degree of stochasticity in the translations. We report improvements over strong recent baselines, leading to state-of-the-art UDA results on two challenging semantic segmentation benchmarks.

7.2 Future work

Biopsy-based or prostatectomy-based ground-truth. In the present work, we rely on labels corresponding to PI-RADS scores. However, it would be very valuable from a clinical perspective to train and evaluate the performance of the proposed approaches on biopsy-based or prostatectomy-based ground-truth. Currently, the clinical problem lies in accurately differentiating between clinically significant and non-clinically significant prostate cancer in-vivo. Recent studies [92, 95, 133, 134, 135, 136, 94] focus on the development of deep learning methods for discriminating between clinically significant and non-clinically significant cancer on mp-MRI. In these studies biopsy-based or prostatectomy-based annotations serve as ground truth and the results indicate that deep learning models can achieve high diagnostic accuracy. Future work should focus on obtaining biopsy-based ground-truth for pairs of VERDICT MRI and DW data from mp-MRI acquisitions. This would allow to examine more thoroughly whether FCNs trained on VERDICT MRI can better discriminate between clinically significant and non-clinically significant cancer than models trained on standard DW data from mp-MRI acquisitions. In addition, it would be interesting to evaluate the performance of the proposed methods on both lesion segmentation and Gleason grading. Methods that provide accurate Gleason grade classification could eliminate biopsies and play an important role in clinical practice.

Using stochastic translation in other medical image analysis application. Scarcity of high-quality annotated data and mismatch between the development dataset and the target environment is a major challenge in machine learning for medical imaging. To address this challenge, several domain adaptation methods have been recently proposed. In addition, several datasets have become publicly available, facilitating the evaluation of domain adaptation methods in different medical image analysis applications. For instance, there are publicly available datasets for cross-modality adaptation between MRI and CT images for cardiac substructure segmentation and abdominal multi-organ segmentation. In addition, a recent cross-modality domain adaptation challenge (CrossModa) focuses on cross-modality adaptation between contrast-enhanced T1 (ceT1) MRI and high-resolution T2 (hrT2) MRI for segmenting brain structures. Future work could focus on extending and evaluating the methods proposed in this thesis using the aforementioned publicly available medical imaging dataset.

Bibliography

[1] Donald F. Gleason. Histologic grading and clinical staging of prostate carcinoma. *Urologic pathology: the prostate*, 171(98), 1977.

[2] Dwight G. Nishimura. *Principles of magnetic resonance imaging*. Stanford Univ., Stanford, Calif., 1996.

[3] Bas Israël, Marloes van der Leest, Michiel Sedelaar, Anwar R. Padhani, Patrik Zámecnik, and Jelle O. Barentsz. Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: what urologists need to know. part 2: interpretation. *European Urology*, 77(4):469–480, 2020.

[4] Thais C. Mussi, Ronaldo H. Baroni, Ronald J. Zagoria, and Antonio C. Westphalen. Prostate magnetic resonance imaging technique. *Abdominal Radiology*, 45(7):2109–2119, 2020.

[5] Cher H. Tan, Jihong Wang, and Vikas Kundra. Diffusion weighted imaging in prostate cancer. *European Radiology*, 21(3):593–603, 2011.

[6] Eleftheria Panagiotaki, Rachel W. Chan, Nikolaos Dikaios, Hashim U. Ahmed, James O'Callaghan, Alex Freeman, David Atkinson, Shonit Punwani, David J. Hawkes, and Daniel C. Alexander. Microstructural characterization of normal and malignant human prostate tissue with vascular, extracellular, and restricted diffusion for cytometry in tumours magnetic resonance imaging. *Investigative Radiology*, 50(4):218–227, 2015.

[7] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, and Wenqiang Zhang. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9082–9091, 2021.

[8] Geert Litjens, Thijs Kooi, Babak E. Bejnordi, Arnaud A. A. Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[9] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017.

[10] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[11] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[12] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *The Journal of the American Medical Association*, 316(22), 2016.

[13] Simon L.F. Walsh, Lucio Calandriello, Mario Silva, and Nicola Sverzellati. Deep learning for classifying fibrotic lung disease on high-resolution com-

puted tomography: a case-cohort study. *The Lancet Respiratory Medicine*, 6(11):837–845, 2018.

[14] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998, 2018.

[16] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[17] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S. Mageras, Joseph O. Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 777–785, 2018.

[18] Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to X-ray image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pages 599–607, 2018.

[19] Jinzheng Cai, Zizhao Zhang, Lei Cui, Yefeng Zheng, and Lin Yang. Towards cross-modal organ translation and segmentation: A cycle and shape consis-

tent generative adversarial network. *Medical Image Analysis*, 52:174–184, 2019.

[20] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle and shape consistency generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9242–9251, 2018.

[21] Eletheria Panagiotaki, Simon Walker-Samuel, Bernard Siow, Peter S. Johnson, Vineeth Rajkumar, Barbara R. Pedley, Mark F. Lythgoe, and Daniel C. Alexander. Noninvasive quantification of solid tumor microstructure using VERDICT MRI. *Cancer Research*, 74(7):1902–1912, 2014.

[22] Edward W. Johnston, Elisenda Bonet-Carne, Uran Ferizi, Ben Yvernault, Hayley Pye, Dominic Patel, Joey Clemente, Wivijin Piga, Susan Heavey, Harbir S. Sidhu, et al. VERDICT-MRI for prostate cancer: Intracellular volume fraction versus apparent diffusion coefficient. *Radiology*, 291(2):391–397, 2019.

[23] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118, 2016.

[24] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, June 2016.

[25] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3213–3223, 2016.

[26] Eleni Chiou, Francesco Giganti, Elisenda Bonet-Carne, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Prostate cancer classification on VERDICT DW-MRI using convolutional neural networks. In *Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9*, pages 319–327, 2018.

[27] Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Automatic classification of benign and malignant prostate lesions: A comparison using VERDICT DW-MRI and ADC maps. In *International Society for Magnetic Resonance in Medicine*, 2019.

[28] Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Harnessing uncertainty in domain adaptation for MRI prostate lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 510–520, 2020.

[29] Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Domain adaptation for prostate lesion segmentation on VERDICT-MRI. In *International Society for Magnetic Resonance in Medicine*, 2020.

[30] Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Unsupervised domain adaptation with semantic consistency across heterogeneous modalities for MRI prostate lesion segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 90–100. Springer, 2021.

[31] Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Prostate lesion segmentation on VERDICT-MRI driven by unsupervised domain adaptation. In *International Society for Magnetic Resonance in Medicine*, 2020.

[32] Eleni Chiou, Eleftheria Panagiotaki, and Iasonas Kokkinos. Beyond deterministic translation for unsupervised domain adaptation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, 2022.

[33] Eleni Chiou, Vanya Valindria, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Synthesizing verdict maps from standard DWI data using GANs. In *International Workshop on Computational Diffusion MRI*, pages 58–67. Springer, 2021.

[34] Vanya Valindria, Marco Palombo, Eleni Chiou, Saurabh Singh, Shonit Punwani, and Eleftheria Panagiotaki. Synthetic q-space learning with deep regression networks for prostate cancer characterisation with VERDICT. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 50–54, 2021.

[35] Vanya Valindria, Saurabh Singh, Eleni Chiou, Thomy Mertzanidou, Baris Kanber, Shonit Punwani, Marco Palombo, and Eleftheria Panagiotaki. Non-invasive gleason score classification with VERDICT-MRI. In *International Society for Magnetic Resonance in Medicine*, 2021.

[36] Marco Palombo, Vanya Valindria, Saurabh Singh, Eleni Chiou, Francesco Giganti, Hayley Pye, Hayley C Whitaker, David Atkinson, Shonit Punwani, Daniel C Alexander, et al. Joint estimation of relaxation and diffusion tissue parameters for prostate cancer grading with relaxation-VERDICT MRI. *medRxiv*, 2021.

[37] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer

statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[38] Klaus Eichler, Susanne Hempel, Jennifer Wilby, Lindsey Myers, Lucas M. Bachmann, and Jos Kleijnen. Diagnostic value of systematic biopsy methods in the investigation of prostate cancer: A systematic review. *The Journal of Urology*, 175(5):1605–1612, 2006.

[39] Axel Heidenreich, Gunnar Aus, Michel Bolla, Steven Joniau, Vsevolod B. Matveev, Hans Peter Schmid, and Filliberto Zattoni. EAU guidelines on prostate cancer. *European Urology*, 53(1):68–80, 2008.

[40] Peter A. Humphrey. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern pathology*, 17(3):292–306, 2004.

[41] Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G. Hindley, Alex Freeman, et al. Diagnostic accuracy of multiparametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *The Lancet*, 389(10071):815–822, 2017.

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

[43] Masanori Noguchi, Thomas A. Stamey, John E. Mcneal, and Cheryl M. Yemoto. Relationship between systematic biopsies and histological features of 222 radical prostatectomy specimens: lack of prediction of tumor significance for men with nonpalpable prostate cancer. *The Journal of Urology*, 166(1):104–110, 2001.

[44] Philippe Puech, Olivier Rouvière, Raphaele Renard-Penna, Arnauld Villers, Patrick Devos, Marc Colombel, Marc-Olivier Bitker, Xavier Leroy, Florence Mège-Lechevallier, Eva Comperat, et al. Prostate cancer diagnosis: multiparametric mr-targeted biopsy with cognitive and transrectal us–mr fusion guidance versus systematic biopsy—prospective multicenter study. *Radiology*, 268(2):461–469, 2013.

[45] Pepijn Brocken, Judith B. Prins, Richard P.N. Dekhuijzen, and Henricus F.M. van der Heijden. The faster the better?—A systematic review on distress in the diagnostic phase of suspected cancer, and the influence of rapid diagnostic pathways. *Psycho-Oncology*, 21(1):1–10, 2012.

[46] Arjun Sivaraman and Rafael Sanchez-Salas. Transperineal template-guided mapping biopsy of the prostate. *Technical Aspects of Focal Therapy in Localized Prostate Cancer*, pages 101–114, 2015.

[47] A.V Taira, G.S. Merrick, R.W. Galbreath, H. Andreini, W. Taubenslag, R. Curtis, W.M. Butler, E. Adamovich, and K.E. Wallner. Performance of transperineal template-guided mapping biopsy in detecting prostate cancer in the initial and repeat biopsy setting. *Prostate cancer and prostatic diseases*, 13(1):71–77, 2010.

[48] Hashim Uddin Ahmed, Yipeng Hu, Tim Carter, Nimalan Arumainayagam, Emilie Lecornet, Alex Freeman, David Hawkes, Dean C. Barratt, and Mark Emberton. Characterizing clinically significant prostate cancer using template prostate mapping biopsy. *The Journal of Urology*, 186(2):458–464, 2011.

[49] Gregory S. Merrick, Walter Taubenslag, Hugo Andreini, Sarah Brammer, Wayne M. Butler, Edward Adamovich, Zachary Allen, Richard Anderson, and Kent E. Wallner. The morbidity of transperineal template-guided prostate mapping biopsy. *BJU international*, 101(12):1524–1529, 2008.

[50] Marc A. Bjurlin, Peter R. Carroll, Scott Eggener, Pat F. Fulgham, Daniel J. Margolis, Peter A. Pinto, Andrew B. Rosenkrantz, Jonathan N. Rubenstein, Daniel B. Rukstalis, Samir S. Taneja, et al. Update of the standard operating procedure on the use of multiparametric magnetic resonance imaging for the diagnosis, staging and management of prostate cancer. *The Journal of Urology*, 203(4):706–712, 2020.

[51] Stacy Loeb, Annelies Vellekoop, Hashim U. Ahmed, James Catto, Mark Emberton, Robert Nam, Derek J. Rosario, Vincenzo Scattoni, and Yair Lotan. Systematic review of complications of prostate biopsy. *European Urology*, 64(6):876–892, 2013.

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

[53] Jeffrey C. Weinreb, Jelle O. Barentsz, Peter L. Choyke, Francois Cornud, Masoom A. Haider, Katarzyna J. Macura, Daniel Margolis, Mitchell D. Schnall, Faina Shtern, Clare M. Tempany, Harriet C. Thoeny, and Sadna Verma. PI-RADS prostate imaging – reporting and data system: 2015, version 2. *European Urology*, 69(1):16 – 40, 2016.

[54] C. M. A. Hoeks, J. O. Barentsz, T. Hambrock, D. Yakar, D. M. Somford, S. W. T. P. J. Heijmink, T. W. J. Scheenen, P. C. Vos, H. Huisman, I. M. Oort, J. A. Witjes, A. Heerschap, and J. J. Fütterer. Prostate cancer: Multiparametric MR imaging for detection, localization, and staging. *Radiology*, 261(1):46–66, 2011.

[55] H. Hricak, R.D. Williams, D.B. Spring, K.L. Moon, M.W. Hedgcock, R.A. Watson, and L.E. Crooks. Anatomy and pathology of the male pelvis by mag-

netic resonance imaging. *American Journal of Roentgenology*, 141(6):1101–1110, 1983.

[56] Hedrig Hricak, Georges C. Dooms, John E. McNeal, Alexander S. Mark, Miljenko Marotti, Antony Avallone, Mark Pelzer, Evelyn C. Proctor, and Emil A. Tanagho. MR imaging of the prostate gland: normal anatomy. *American Journal of Roentgenology*, 148(1):51–58, 1987.

[57] John E. McNeal. The zonal anatomy of the prostate. *The prostate*, 2(1):35–49, 1981.

[58] Oguz Akin, Evis Sala, Chaya S. Moskowitz, Kentaro Kuroiwa, Nicole M. Ishill, Darko Pucar, Peter T. Scardino, and Hedvig Hricak. Transition zone prostate cancers: features, detection, localization, and staging at endorectal mr imaging. *Radiology*, 239(3):784–792, 2006.

[59] C. H. Tan, W. Wei, V. Johnson, and V. Kundra. Diffusion-weighted MRI in the detection of prostate cancer: Meta-analysis. *American Journal of Roentgenology*, 199(4):822–829, 2012.

[60] R. T. Gupta, B. Spilseth, N. Patel, A. F. Brown, and J. Yu. Multiparametric prostate MRI: focus on T2-weighted imaging and role in staging of prostate cancer. *Abdominal Radiology*, 41(5):831–843, 2016.

[61] Jelle O. Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J. Fütterer. ESUR prostate MR guidelines 2012. *European Radiology*, 22(4):746–757, 2012.

[62] S. Verma, B. Turkbey, N. Muradyan, A. Rajesh, F. Cornud, M. A. Haider, P. L. Choyke, and M. Harisinghani. Overview of dynamic contrast-enhanced MRI in prostate cancer diagnosis and management. *American Journal of Roentgenology*, 198(6):1277–1288, 2012.

[63] C. H. Tan, B. P. Hobbs, W. Wei, and V. Kundra. Dynamic contrast-enhanced MRI for the detection of prostate cancer: Meta-analysis. *American Journal of Roentgenology*, 204(4):439–448, 2015.

[64] Peter Carmeliet and Rakesh K. Jain. Angiogenesis in cancer and other diseases. *Nature*, 407(6801):249–257, 2000.

[65] P.S. Tofts, D.A.G. Wicks, and G. Barker. The MRI measurement of NMR and physiological parameters in tissue to study disease process. In *Information Processing in Medical Imaging*, pages 313–326. Wiley-Liss Inc, USA, 1991.

[66] Gunnar Brix, Wolfhard Semmler, Rüdiger Port, Lothar R. Schad, Günter Layer, and Walter J. Lorenz. Pharmacokinetic parameters in CNS Gd-DTPA enhanced MR imaging. *Journal of Computer Assisted Tomography*, 15(4):621–628, 1991.

[67] Sofie Isebaert, Laura Van den Bergh, Karin Haustermans, Steven Joniau, Evelyne Lerut, Liesbeth De Wever, Frederik De Keyzer, Tom Budiharto, Pieter Slagmolen, Hendrik Van Poppel, et al. Multiparametric MRI for prostate cancer localization in correlation to whole-mount histopathology. *Journal of Magnetic Resonance Imaging*, 37(6):1392–1401, 2013.

[68] Gregory J. Metzger, Chaitanya Kalavagunta, Benjamin Spilseth, Patrick J. Bolan, Xiufeng Li, Diane Hutter, Jung W. Nam, Andrew D. Johnson, Jonathan C. Henriksen, Laura Moench, et al. Detection of prostate cancer: quantitative multiparametric mr imaging models developed using registered correlative histopathology. *Radiology*, 279(3):805–816, 2016.

[69] Roger Bourne and Eleftheria Panagiotaki. Limitations and prospects for diffusion-weighted MRI of the prostate. *Diagnostics*, 6(2):21, 2016.

[70] Dow-Mu Koh and David J. Collins. Diffusion-weighted MRI in the body: applications and challenges in oncology. *American Journal of Roentgenology*, 188(6):1622–1635, 2007.

[71] Denis Le Bihan. Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience*, 4(6):469–480, 2003.

[72] Chan Kyo Kim, Byung Kwan Park, and Bohyun Kim. Diffusion-weighted MRI at 3 T for the evaluation of prostate cancer. *American Journal of Roentgenology*, 194(6):1461–1469, 2010.

[73] G. Lematre, R. Marta, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. *Computers in Biology and Medicine*, 60:8–31, 2015.

[74] G. J. S. Litjens, J. O. Barentsz, N. Karssemeijer, and H. J. Huisman. Automated computer-aided detection of prostate cancer in MR images: from a whole-organ to a zone-based approach. *Proceedings of SPIE–the International Society for Optical Engineering*, 8315, 2012.

[75] S. Klein, U. A. van der Heide, I. M. Lips, M Vulpen, M. Staring, and J. P. W. Pluim. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Medical Physics*, 35(4):1407–1417, 2008.

[76] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

[77] R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Transactions on Medical Imaging*, 29:2000–2008, 2010.

[78] S. E. Viswanath, B. N. Bloch, M. Rosen, J. Chappelow, N. M. Rofsky, R. E. Lenkinski, E. M. Genega, A. Kalyanpur, and A. Madabhushi. Integrating structural and functional imaging for computer assisted detection of prostate

cancer on multi-protocol in vivo 3 Tesla MRI. In *Proceedings of SPIE–the International Society for Optical Engineering*, volume 7260, 2009.

[79] R. Toth, J. Chappelow, M. Rosen, S. Pungavkar, A. Kalyanpur, and A. Madabhushi. Multi-attribute non-initializing texture reconstruction based active shape model (MANTRA). In *Medical Image Computing and Computer-Assisted Intervention*, pages 653–661. Springer Berlin Heidelberg, 2008.

[80] I. Reda, A. Shalaby, M. Elmogy, A. Aboulfotouh, F. Khalifa, M. A. El-Ghar, G. Gimelfarb, and A. El-Baz. Image-based computer-aided diagnostic system for early diagnosis of prostate cancer. In *Medical Image Computing and Computer-Assisted Intervention*, pages 610–618, Cham, 2016. Springer International Publishing.

[81] I. Reda, A. Shalaby, M. Elmogy, A. A. Elfotouh, F. Khalifa, M. A. El-Ghar, E. Hosseini-Asl, G. Gimel'farb, N. Werghi, and A. El-Baz. A comprehensive non-invasive framework for diagnosing prostate cancer. *Computers in Biology and Medicine*, 81:148–158, 2017.

[82] I. Reda, A. Shalaby, M. A. El-Ghar, F. Khalifa, M. Elmogy, A. Aboulfotouh, E. Hosseini-Asl, A. El-Baz, and R. Keynton. A new NMF-autoencoder based CAD system for early diagnosis of prostate cancer. In *International Symposium on Biomedical Imaging*, pages 1237–1240, 2016.

[83] V. Giannini, A. Vignati, S. Mazzetti, M. De Luca, C. Bracco, M. Stasi, F. Russo, E. Armando, and D. Regge. A prostate CAD system based on multiparametric analysis of DCE T1-w, and DW automatically registered images. In *Proceedings of SPIE–the International Society for Optical Engineering*, volume 8670, 2013.

[84] A. P. Kiraly, C. A. Nader, A. Tuysuzoglu, R. Grimm, B. Kiefer, N. El-Zehiry, and A. Kamen. Deep convolutional encoder-decoders for prostate cancer detection and classification. In *Medical Image Computing and Computer-*

Assisted Intervention, pages 489–497, Cham, 2017. Springer International Publishing.

[85] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35– 51, 1996.

[86] X. Yang, Z. Wang, C. Liu, H. M. Le, J. Chen, K.-T. Cheng, and L. Wang. Joint detection and diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. In *Medical Image Computing and Computer-Assisted Intervention*, pages 426–434, Cham, 2017. Springer International Publishing.

[87] Y. Artan, M. A. Haider, D. L. Langer, T. H. van der Kwast, A. J. Evans, Y. Yang, M. N. Wernick, J. Trachtenberg, and I. S. Yetik. Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields. *IEEE Transactions on Image Processing*, 19(9):2444–2455, 2010.

[88] E. Niaf, O. Rouvière, F. Bratan F Mège-Lechevallier, and C. Lartizien. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. *Physics in Medicine & Biology*, 57(12):2444–2455, 2012.

[89] A. Mehrtash, A. Sedghi, M. Ghafoorian, M. Taghipour, C. M. Tempany, W. M. Wells, T. Kapur, P. Mousavi, P. Abolmaesumi, and A. Fedorova. Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. *Proceedings of SPIE–the International Society for Optical Engineerings*, 10134, 2017.

[90] Y. K. Tsehay, N. S. Lay, H. R. Roth, X. Wang, J. T. Kwak, B. I. Turkbey, P. A. Pinto, and R. M. Summers B. J. Wood. Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric

magnetic resonance images. *Proceedings of SPIE–the International Society for Optical Engineerings*, 10134, 2017.

[91] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015.

[92] Z. Wang, C. Liu, D. Cheng, L. Wang, X. Yang, and K. T. Cheng. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. *IEEE Transactions on Medical Imaging*, 37(5):1127–1139, 2018.

[93] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

[94] Ruiming Cao, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman, and Kyunghyun Sung. Joint prostate cancer detection and gleason score prediction in mp-MRI via focalnet. *IEEE Transactions on Medical Imaging*, 38(11):2496–2506, 2019.

[95] Pritesh Mehta, Michela Antonelli, Saurabh Singh, Natalia Grondecka, Edward W. Johnston, Hashim U. Ahmed, Mark Emberton, Shonit Punwani, and Sébastien Ourselin. Autoprostate: Towards automated reporting of prostate MRI for prostate cancer assessment using deep learning. *Cancers*, 13(23):6138, 2021.

[96] Lucy A.M. Simmons, Abi Kanthabalan, Manit Arya, Tim Briggs, Dean Barratt, Susan C. Charman, Alex Freeman, James Gelister, David Hawkes, Yipeng Hu, et al. The picture study: diagnostic accuracy of multiparametric mri in men requiring a repeat prostate biopsy. *British Journal of Cancer*, 116(9):1159–1165, 2017.

[97] Praful Hambarde, Sanjay Talbar, Abhishek Mahajan, Satishkumar Chavan, Meenakshi Thakur, and Nilesh Sable. Prostate lesion segmentation in MR images using radiomics based deeply supervised U-Net. *Biocybernetics and Biomedical Engineering*, 40(4):1421–1435, 2020.

[98] Yizheng Chen, Lei Xing, Lequan Yu, Hilary P. Bagshaw, Mark K. Buyyounouski, and Bin Han. Automatic intraprostatic lesion segmentation in multiparametric magnetic resonance images with proposed multiple branch UNet. *Medical Physics*, 47(12):6421–6429, 2020.

[99] Yatong Liu, Yu Zhu, Wei Wang, Bingbing Zheng, Xiangxiang Qin, and Peijun Wang. Multi-scale discriminative network for prostate cancer lesion segmentation in multiparametric mr images. *Medical Physics*, 2022.

[100] Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin, and Carole Lartizien. Prostattention-net: A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans. *Medical Image Analysis*, 77:102347, 2022.

[101] Shirin Sabouri, Silvia D. Chang, Richard Savdie, Jing Zhang, Edward C. Jones, S. Larry Goldenberg, Peter C. Black, and Piotr Kozlowski. Luminal water imaging: a new MR imaging T2 mapping technique for prostate cancer diagnosis. *Radiology*, 284(2):451–459, 2017.

[102] Shirin Sabouri, Ladan Fazli, Silvia D. Chang, Richard Savdie, Edward C. Jones, S. Larry Goldenberg, Peter C. Black, and Piotr Kozlowski. MR measurement of luminal water in prostate gland: Quantitative correlation between MRI and histology. *Journal of Magnetic Resonance Imaging*, 46(3):861–869, 2017.

[103] Aritrick Chatterjee, Carla Harmath, and Aytekin Oto. New prostate MRI techniques and sequences. *Abdominal Radiology*, 45(12):4052–4062, 2020.

[104] Shiyang Wang, Yahui Peng, Milica Medved, Ambereen N. Yousuf, Marko K. Ivancevic, Ibrahim Karademir, Yulei Jiang, Tatjana Antic, Steffen Sammet,

Aytekin Oto, et al. Hybrid multidimensional T2 and diffusion-weighted MRI for prostate cancer detection. *Journal of Magnetic Resonance Imaging*, 39(4):781–788, 2014.

[105] Aritrick Chatterjee, Roger M. Bourne, Shiyang Wang, Ajit Devaraj, Alexander J. Gallan, Tatjana Antic, Gregory S. Karczmar, and Aytekin Oto. Diagnosis of prostate cancer with noninvasive estimation of prostate tissue composition by using hybrid multidimensional MR imaging: a feasibility study. *Radiology*, 287(3):864–873, 2018.

[106] Nathan S. White, Carrie R. McDonald, Niky Farid, Josh Kuperman, David Karow, Natalie M. Schenker-Ahmed, Hauke Bartsch, Rebecca Rakow-Penner, Dominic Holland, Ahmed Shabaik, et al. Diffusion-weighted imaging in cancer: physical foundations and applications of restriction spectrum imaging. *Cancer research*, 74(17):4638–4652, 2014.

[107] Nathan S. White, Trygve B. Leergaard, Helen D'Arceuil, Jan G. Bjaalie, and Anders M. Dale. Probing tissue microstructure with restriction spectrum imaging: histological and theoretical validation. *Human Brain Mapping*, 34(2):327–346, 2013.

[108] Ileana O. Jelescu, Marco Palombo, Francesca Bagnato, and Kurt G. Schilling. Challenges for biophysical modeling of microstructure. *Journal of Neuroscience Methods*, 344:108861, 2020.

[109] Hassan Bagher-Ebadian, Kourosh Jafari-Khouzani, Panayiotis D. Mitsias, Mei Lu, Hamid Soltanian-Zadeh, Michael Chopp, and James R. Ewing. Predicting final extent of ischemic infarction using artificial neural network analysis of multi-parametric MRI in patients with stroke. *PloS one*, 6(8):e22626, 2011.

[110] Vladimir Golkov, Tim Sprenger, Jonathan Sperl, Marion Menzel, Michael Czisch, Philipp Samann, and Daniel Cremers. Model-free novelty-based dif-

fusion MRI. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1233–1236. IEEE, 2016.

[111] Vladimir Golkov, Alexey Dosovitskiy, Jonathan I. Sperl, Marion I. Menzel, Michael Czisch, Philipp Sämann, Thomas Brox, and Daniel Cremers. q-space deep learning: Twelve-fold shorter and model-free diffusion MRI scans. *IEEE Transactions on Medical Imaging*, 35(5):1344–1351, 2016.

[112] Vladimir Golkov, Alexey Dosovitskiy, Philipp Sämann, Jonathan I. Sperl, Tim Sprenger, Michael Czisch, Marion I. Menzel, Pedro A. Gómez, Axel Haase, Thomas Brox, and Daniel Cremers. q-space deep learning for twelve-fold shorter and model-free diffusion MRI scans. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 37–44, Cham, 2015. Springer International Publishing.

[113] Davood Karimi, Lana Vasung, Camilo Jaimes, Fedel Machado-Rivas, Simon K. Warfield, and Ali Gholipour. Learning to estimate the fiber orientation distribution function from diffusion-weighted MRI. *Neuroimage*, 239:118316, 2021.

[114] Eric K. Gibbons, Kyler K. Hodgson, Akshay S. Chaudhari, Lorie G. Richards, Jennifer J. Majersik, Ganesh Adluru, and Edward V.R. DiBella. Simultaneous NODDI and GFA parameter map generation from subsampled q-space imaging using deep learning. *Magnetic Resonance in Medicine*, 81(4):2399–2411, 2019.

[115] Eric Aliotta, Hamidreza Nourzadeh, Jason Sanders, Donald Muller, and Daniel B. Ennis. Highly accelerated, model-free diffusion tensor MRI reconstruction using neural networks. *Medical Physics*, 46(4):1581–1591, 2019.

[116] E. Johnston, H. Pye, E. Bonet-Carne, E. Panagiotaki, D. Patel, M. Galazi, S. Heavey, L. Carmona, A. Freeman, G. Trevisan, C. Allen, A. Kirkham, K. Burling, N. Stevens, D. Hawkes, M. Emberton, C. Moore, H. U. Ahmed, D. Atkinson, M. Rodriguez-Justo, T. Ng, D. Alexander, H. Whitaker, and

S. Punwani. INNOVATE: A prospective cohort study combining serum and urinary biomarkers with novel diffusion-weighted magnetic resonance imaging for the prediction and characterization of prostate cancer. *BMC Cancer*, 16(816), 2016.

[117] Eleftheria Panagiotaki, Andrada Ianus, Edward Johnston, R. Chan, David Atkinson, D. Alexander, et al. Optimised VERDICT MRI protocol for prostate cancer characterisation. In *International Society for Magnetic Resonance in Medicine*, 2015.

[118] Sébastien Ourselin, Alexis Roche, Gérard Subsol, Xavier Pennec, and Nicholas Ayache. Reconstructing a 3D structure from serial histological sections. *Image and Vision Computing*, 19(1-2):25–31, 2001.

[119] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[120] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241, 2015.

[121] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[122] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[123] K. He and et al. Deep residual learning for image recognition. In *CVPR*, 2016.

[124] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

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

[126] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, pages 807–814, 2010.

[127] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[128] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[129] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[130] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop Autodiff Decision*, 2017.

[131] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron

Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242, 2017.

[132] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[133] Nader Aldoj, Steffen Lukas, Marc Dewey, and Tobias Penzkofer. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. *European Radiology*, 30(2):1243–1253, 2020.

[134] Jose M.T. Castillo, Muhammad Arif, Martijn P.A. Starmans, Wiro J. Niessen, Chris H. Bangma, Ivo G. Schoots, and Jifke F. Veenland. Classification of clinically significant prostate cancer on multi-parametric MRI: A validation study comparing deep learning and radiomics. *Cancers*, 14(1):12, 2021.

[135] Coen De Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim, and Mitko Veta. Deep learning regression for prostate cancer detection and grading in bi-parametric MRI. *IEEE Transactions on Biomedical Engineering*, 68(2):374–383, 2020.

[136] Danyan Li, Xiaowei Han, Jie Gao, Qing Zhang, Haibo Yang, Shu Liao, Hongqian Guo, and Bing Zhang. Deep learning in prostate cancer diagnosis using multiparametric magnetic resonance imaging with whole-mount histopathology referenced delineations. *Frontiers in Medicine*, 8, 2021.

[137] Jian Ren, Ilker Hacihaliloglu, Eric A. Singer, David J. Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 201–209, 2018.

[138] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, pages 597–609, 2017.

[139] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[140] Cheng Ouyang, Konstantinos Kamnitsas, Carlo Biffi, Jinming Duan, and Daniel Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 669–677, 2019.

[141] Junlin Yang, Nicha C. Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S. Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 255–263, 2019.

[142] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017.

[143] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[144] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.

[145] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[146] S. Rebuffi, A. Vedaldi, and H. Bilen. Efficient parametrization of multi-domain deep neural networks. In *CVPR*, 2018.

[147] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[148] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248, 2017.

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

[150] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017.

[151] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[152] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of The Thirty-Third Conference on Artificial Intelligence (AAAI)*, pages 865–872, 2019.

[153] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-relaxed domain adaptation for image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 490–499, 2020.

[154] Guodong Zeng, Florian Schmaranzer, Till D. Lerch, Adam Boschung, Guoyan Zheng, Jürgen Burger, Kate Gerber, Moritz Tannast, Klaus Siebenrock, Young-Jo Kim, Eduardo N. Novais, and Nicolas Gerber. Entropy guided unsupervised domain adaptation for cross-center hip cartilage segmentation from MRI. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 447–456, 2020.

[155] Kang Li, Shujun Wang, Lequan Yu, and Pheng-Ann Heng. Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 418–427, 2020.

[156] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[157] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020.

[158] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 480–498, 2020.

[159] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, pages 195–204, 2018.

[160] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018.

[161] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV*, pages 642–659, 2020.

[162] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, 2019.

[163] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-

center regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, 2021.

[164] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 415–430, 2020.

[165] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019.

[166] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020.

[167] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[168] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019.

[169] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[170] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wenmei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for

stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020.

[171] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[172] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

[173] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.

[174] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 532–548, 2020.

[175] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021.

[176] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019.

[177] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12435–12445, 2021.

[178] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *International Workshop on Artificial Intelligence and Statistics*, pages 57–64, 2005.

[179] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2017.

[180] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S. Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018.

[181] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[182] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[183] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[184] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020.

[185] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, 2021.

[186] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9092–9101, 2021.