Robust White Matter Hyperintensity Segmentation

Orbes Arteaga, Mauricio

Awarding institution:
King's College London

Robust White Matter Hyperintensity Segmentation

H. Mauricio Orbes

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

King's College London.

Life Sciences & Medicine

King's College London

August 31, 2020

I, H. Mauricio Orbes, 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

Quantification of white matter hyperintensities (WMH) is necessary to understand their role in several neurological diseases. Measurements such as volumetric estimations need to be accurate enough to provide valid insights. Consequently, robust and accurate segmentation of WMH is needed. Modern machine learning techniques such as Deep Learning (DL) have made important advances in this field, showing unprecedented performance in segmentation. However, their applicability in realistic clinical scenarios is still questioned due to their lack of generalization capabilities when trained on limited amounts of data. This problem could be more notable in segmentation of anomalies such WMHs. Therefore in this thesis, I focus on the three main challenges that make the task of WMHs segmentation more difficult: *i)* Inter-modality heterogeneity –the number and type of modalities to analyzed WMHs can vary from center to center, where less expensive to acquired modalities could not be always available for every single patient; *ii)* morphological heterogeneity –WMHs present diverse shapes and appearances, therefore, learn invariant features to these variations requires large and morphological diverse datasets; *iii)* intra-modality heterogeneity –scanners or acquisition protocols can be different from one center to another, therefore, segmenting images acquired at a different center than the one used for model training can result in an inaccurate segmentation.

In this thesis, I proposed different approaches to tackle the aforementioned challenges. Firstly, I address inter modality-heterogeneity; I proposed an efficient strategy to leverage information from all available modalities (at the training stage) with the ultimate end of improving segmentation performance on models that only can use more simplistic modalities as input. Next, I proposed a data augmentation approach to increase morphological variability in the training sets; Thus, models can learn features that are robust to this type of variation. To deal with intra-modality heterogeneity, I proposed two different approaches for unsupervised domain adaptation. The first approach proposed a simple but effective strategy based on Knowledge Distillation (KD) to transfer information from labeled (source domain) to unlabeled data (target domain). The second approach extends this idea by introducing data augmentation and consistency training to encourage robustness to different levels of noise on the data. I believe that the proposed methodologies offer an important contribution to the medical imaging field by providing solutions that improve the performance of Deep Learning approaches in realistic clinical settings.

Acknowledgements

Foremost I would like to express my gratitude to my primary supervisor Jorge Cardoso, for his guidance and support throughout my PHD, and for providing me the environment to develop ideas and encourage me to start fruitful collaborations.

I would also like to thank Biomediq A/S, the company that accepted me to be part of the DEMO project and provided the resources for my research. I want to extend my gratitude to Mads Nielsen, Lauge Sørensen, Mostafa, and Akshay Pai, for their guidance during my time in Denmark.

Next, I would like to thank the AMIGOs, for creating a welcoming office environment and their contagious good energy. Huge thanks to Mark, Walter, and Chin for taking the time to make corrections to this thesis. Special thanks to Tom Varsavsky for the fruitful collaboration, the discussions about science, and his patience with my lack of organization and software engineering skills.

I also owe thanks to my friends for their encouragement and support during this last stage of my research, special thanks to my very good friend Sergio Garcia for taking the time to call almost every day during the lockdown and check that I had not lost my mind.

Finally, all my gratitude to my family for their unconditional support, for understanding my absence during these 3 years, and encouraging me to do my best every day.

This project has received funding from the EU H2020 under the Marie Sklodowska-Curie grant agreement No 721820

List of publications

- **Orbes-Arteaga, M**, Cardoso, M. J., Sørensen, L., Modat, M., Ourselin, S., Nielsen, M., & Pai, A. (2018). Simultaneous synthesis of FLAIR and segmentation of white matter hypointensities from T1 MRIs. *In 1st International Conference on Medical Imaging with Deep Learning MIDL2018.*

- **Orbes-Arteaga M**, Lauge Sørensen, Jorge Cardoso, Marc Modat, Sebastien Ourselin, Stefan Sommer, Mads Nielsen, Christian Igel , and Akshay Pai "PADDIT: Probabilistic Augmentation of Data using Diffeomorphic Image Transformation", Proc. SPIE 10949, Medical Imaging 2019: Image Processing,

- **Orbes-Arteaga, M**, Cardoso, J., Sørensen, L., Igel, C., Ourselin, S., Modat, M., ... & Pai, A. (2019).. "Knowledge distillation for semi-supervised domain adaptation." in Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging, MICCAI 2019.

- **Orbes-Arteaga, M.**, Varsavsky, T., Sudre, C. H., Eaton-Rosen, Z., Haddow, L. J., Sørensen, L., ... & Nachev, M. Jorge Cardoso . "Multi-domain adaptation in brain MRI through paired consistency and adversarial learning." *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, MICCAI 2019.

- Varsavsky, T., **Orbes-Arteaga, M**.,Sudre, C. H, Haddow, L. J, Nachev, P, M. Jorge Cardoso. "Test-time Unsupervised Domain Adaptation". In Conference on Medical Image Computing & Computer Assisted Intervention MICCAI 2020

In Preparison

- **Orbes-Arteaga, M.**, Varsavsky, T., Sudre, C, Haddow, L. J., Sørensen, L., ... Nachev, M. Jorge Cardoso "Augmentation based Unsupervised Domain Adaptation".

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

White Matter Hyperintensities (WMHs) are constantly analyzed in medical images because of their correlation with aging and several neurological disorders such as cognitive impairment, stroke, and dementia. Obtaining quantitative measures of WMHs is a crucial step in the study of these relationships and requires substantial effort from radiologists which have to manually delineate those anomalies on images.

Computer vision and machine learning techniques have shown to be essential in the development of automatic delineation (a.k.a segmentation) solutions. For an automatic method to be accepted in clinical settings, it has to be computationally efficient, accurate, and robust under different conditions of resolution, noise, image artifacts, and the inherent variability in the presentations of the pathology of study.

Several challenges make segmentation of WMHs difficult. WMHs are highly diverse in morphology, appearance, and locations in the brain. Moreover, acquisition protocols, scanners, or modalities chosen to imaging WMHs change from case to case producing several image appearances. In addition, current public WMHs datasets used to train segmentation models are usually composed of a few samples which come from the similar populations or were acquired with the same scanners/acquisition protocols.

In the past few years, modern techniques such as Deep Neural Networks have proved to be the state-of-the-art for several medical imaging tasks including registration, image classification, and segmentation [Lundervold and Lundervold, 2019]. However, despite the impressive success of such methods, in the absence of large annotated datasets, most current deep learning-based methods lack generalization capabilities when deployed on data which differs from the one used on training sets. This is a critical barrier that limits their applicability in clinical practice, where it is very likely these methods would be deployed on data with different characteristics (e.g. due to differences in acquisition protocols, modalities, scanner type, pathology, and demographics).

The focus of this thesis is the development of methodologies that improve the applicability of deep learning segmentation methods to clinical settings by making them robust to variations in WMH morphology and generalizable to realistic shifts in the data distribution.

1.2 Challenges in WMH segmentation

Deep neural networks (DNN) have become the first choice for the segmentation of WMH. The top-three performing methods in the 2017 MICCAI WMH segmentation challenge[1] relied on some form of DNNs. However, most segmentation methods lack the ability to generalize to shifts in the data distribution. They are designed under the assumption that training and testing sets come from the same domain (e.g. scanner, modality, machine protocol etc.). However, this kind of homogeneity on data does not represent a realistic scenario, as in practice it is uncommon to have access to labeled data from a new center to retrain a model. Thus, a pre-trained model from a given domain can in many cases, have its performance reduced drastically when deployed on data which differs significantly from the training set. Therefore, achieving this generalization depends mostly on the available training data's ability to represent the expected heterogeneity in test cases. However, as opposed to other visual applications, acquiring medical images is time-consuming, expensive, and requires special equipment. Consequently, current retrospective datasets hold a limited number of training samples with homogeneous properties. Besides those

[1] http://wmh.isi.uu.nl

datasets deficiencies, the segmentation of WMH faces additional challenges that need to be addressed to achieve more robust and generalizable models.

- **Inter-Modality heterogeneity:** The segmentation of WMHs benefits from complementary information provided by different sequences. Consequently, segmentation approaches ideally aim to combine sequences such as T1-w and fluid-attenuated inversion recovery (FLAIR) –Those modalities are able to encode most of the information provided for other modalities such as T2-w and PD-w for WMHs analysis. However, in clinical practice, the combination of modalities for each case of study is heterogeneous with some modalities missing. For example, for a given patient only the T1-w scan could be acquired, either because it is faster or because they are generally acquired at a higher resolution. The heterogeneity of available models limits the practicality of systems where cases with missing modalities have to be discarded for evaluation or training.

- **Morphological heterogeneity** Particularly, WMHs vary largely in shape, locations, and appearance. For proper generalization performance of Deep neural networks, the learned features should be invariant under particular morphology variations of the input. However, learning invariant features to cope with these variations is difficult if the data used to train the models is composed of few samples with insufficient variability among them.

- **Intra-Modality heterogeneity** Deep learning models that assume independent identically distributed data require that training and test sets come from a similar domain (e.g., scanner, machine protocol) in order to guarantee good learning [Perone et al., 2019]. However, this assumption does not hold in practice as acquisition parameters and/or protocols vary from clinic to clinic producing several image appearances. Such differences result in shifts between the training (source) and test (target) sets distributions, which leads to a performance decrease when the model is applied to unseen domains. Although this problem can be alleviated by retraining on labeled data from the new domain, this solution is neither practical nor scalable. Therefore, there is a need to develop strategies that learn to generalize well to the target data without the need of additional labeling work. Achieving this objective requires efficient domain-adaptation strategies between labeled and unlabeled training sets.

1.3 Thesis Overview and Contributions

As previously described there are critical challenges that make it difficult to train models that generalize well to unseen data. This thesis aims to provide solutions that overcome those challenges with the ultimate aim of obtaining robust segmentation models for WMHs. From the technical point of view, some hyper-parameters such as model architecture, optimizer selection, loss and/or regularization functions influence to some degree the model capability of the generalization. Besides the above mentioned variables, this work highlight however, that generalization relies in big proportion on the data which the model learns from. Therefore, I argue that model generalization can be strongly benefited by smart utilization of the data at hand. Consequently, most of the contribution of this work focuses on leveraging the available training data to extract the maximum amount of information that allows us to overcome the aforementioned challenges on the segmentation of WMHs.

In Chapter 2, I summarize the background required to put the later work in context with medical and technical domains, so readers who are not familiar with these topics can follow the remaining Chapters.

The remaining Chapters present the contributions of this thesis. Approaches presented in Chapter 3 and Chapter 4 are considered as fully supervised as those only leverage the initial training set which is manually annotated. The approaches in Chapter 5 and Chapter 6 are considered unsupervised domain adaptation strategies, which also leverage unlabeled data from the clinic/set at the deployment stage.

Chapter 3 addresses inter-modality heterogeneity. The combination of modalities used to analyze WMHs can vary from clinic to clinic or even patient to patient. Moreover, some clinical settings rely only on commonly used but less time consuming or expensive modalities.

A new strategy is proposed to learn efficiently from available multi-modal information in such a way that the model can be deployed on commonly used imaging modalities. The proposed method learns simultaneously FLAIR synthesis and WMH segmentation from T1-w scans . Once the training is done, the model can be deployed on unpaired data where FLAIR images are missing. I demonstrate that the joint optimization of synthesis and segmentation tasks induce a regularization effect which results in an improvement in segmentation performance. The proposed method produces more realistic synthetic FLAIR images compared to the traditional synthesis strategies. Most importantly, it outperforms segmentation of WMHs from T1-w scans only as well as traditional missing modalities imputation methodologies.

Chapter 4 addresses morphological heterogeneity. The shape of lesions can vary largely among subjects. To achieve proper generalization, models should learn features that are invariant to the particular shape variations of the input. However, in small training sets the amount of morphological patterns present in the training data is typically limited. Therefore models overfit to the shape variations that are present in the training set. To encourage morphological invariance in DNNs, I introduce PADDIT (Probabilistic Augmentation of Data using Diffeomorphic Image Transformation). The main advantage of PADDIT is the ability to produce transformations that capture the morphological variability in the training data. Therefore, PADDIT is able to generate more realistic and diverse samples that promote learning of more invariant features. I show that DNNs trained with PADDIT have significant improvement in segmentation performance when compared with random deformations based augmentations.

Chapter 5 deals with intra-modality heterogeneity, i.e the distribution shift between scans of a single modality when they come from different scanners or centres. I tackle this issue by formulating it as an unsupervised domain adaptation problem. Here, annotated training samples are referred to as the source domain whereas data at the deployment stage is referred to as the target domain. The overall idea is to leverage unlabeled data from the target domain to refine a model pre-trained on the source domain, making it more robust to variation in the inputs. To this end, I modify the technique known as knowledge distillation [Hinton et al., 2015](initially conceived to transfer knowledge from a large model, referred to as the teacher, to a simpler model referred to as the student) to transfer knowledge from source to target data. A teacher model is initially trained on source data using ground truth labels. Then, unlabeled samples from the source and target domain go through the teacher to get label probability maps (soft-labels). Finally, a student model is trained on those soft-labels to learn discriminative features for both domains. I demonstrate that this simple yet effective method allows adaptation from source to target domain without the need for annotations nor specific hyper-parameter tuning, making it generally applicable. Moreover, the proposed method outperforms adversarial domain adaptation on the segmentation of WMHs.

In Chapter 6, I extend the domain adaptation idea presented in Chapter 5. While acknowledging the benefits of learning from soft-labels, I also identify the following drawbacks: *i)* The success of adaptation is conditioned on the initial performance of the teacher on the target data (soft-labels are obtained once, off-line) *ii)* generalization capability depends on the diversity of the available training data which can be limited if it is composed of few samples. These problems were addressed through a consistency training approach. Specifically, I introduce a paired consistency loss (PC) to enforce consistency on the model predictions of a given input and its corresponding augmented version. Because the input and augmented predictions act as soft-labels for each other, the soft-predictions are iteratively updated during the optimization process without the need of a teacher/student scheme. In addition, It is shown how the type of augmentation applied has a significant effect on segmentation performance. Robustness to a wider range of heterogeneities is shown to be maximized through a combination of different augmentations operations. I explored different transformations to account for: geometric variations, MR-artifacts, and acquisition differences. The adaptation is supplemented with an adversarial loss that encourages the model to be domain agnostic, preventing the model from getting stuck in bad local minima. Finally, the proposed method outperforms other training consistency methods that rely on teacher/student strategy.

Chapter 2

Background

Although the contributions of this thesis are primarily technical, it is also important to understand the clinical aspects that were the base for the development of the contributions of this thesis. In Section 2.1 I first introduce the biological and clinical context related to white matter hyperintensities. In Section 2.2 I highlight the importance of neuroimaging and I review the existing solutions for the visualization of WMHs. In Section 2.3 I highlight the need for automatic segmentation and I give an overview of the available methods for WMH segmentation. Finally, in Section 2.4 I mention the solutions that have been proposed to deal with the lack of annotated data. Note this topic has not been investigated too much in the context of WMH segmentation, so the mentioned reviewed approaches were in the context of medical image segmentation in general.

2.1 White Matter

The white matter (WM) represents around 40-50% of the brain volume of young healthy adults. This tissue is composed mainly of the neural axon and their supportive ganglia [Malloy et al., 2007], which controls the delivery of nutrients and prevents the entry of damaging elements, and is also responsible for producing and maintaining the myelin sheaths that cover the axons. The myelin, which is mostly composed of lipids (80%), acts as an insulator and allows the electrical conduction along the axons, thereby speeding up the neuronal signal transmission [Deber and Reynolds, 1991]. Damage to the white matter that affects the myelin or loss of axons produce a detrimental effect in the signal transmission and can be associated with cognitive impairment. The causes for such damage can be diverse and range from genetic disorders such as leukodystrophies, traumatic events, infectious diseases, demyelinative disorders (multiple sclerosis), or vascular pathologies [Schmahmann et al., 2008].

2.1.1 White Matter Disease

The term white matter disease (WMD) refers to pathological observations in white matter with ischaemic origin as opposed to multiple sclerosis. White matter lesions associated with cerebral small vessel disease were classified as lacunes, cerebral microbleeds (CMB), infarcts, and enlarged perivascular spaces (EPVS), and leukoaraiosis (LA) or white matter hyperintensities (WMH) [Wardlaw et al., 2013b]

2.1.1.1 White matter hyperintensities / leukoaraiosis

The terms Leukoaraiosis and White matter hyperintensities (WMHs) are interchangeably used to account for changes observed in the white matter of the aging population. WMHs have been considered as a consequence of ischemia of the tissue. Degradation of the myelin occurs as a consequence of neurons, and oligodendrocytes decay due to the inability to obtain the required survival components [Wharton et al., 2015]. There is evidence that shows the relationship between the occurrence of WMH and impaired vascular endothelium [Wardlaw et al., 2013a]. The term WMH is indefinite enough that can make reference to a large variety in appearance, spatial distribution, and histopathological explanations. Several directions have been suggested to address the spatial distribution, however, this task has been challenging due to difficulty in obtaining robust and reproducible characterization of locations. Lesions located continuously to the ventricular surface have been referred to as periventricular WMH or PVWMH whereas the remaining ones are referred to as deep (WMH or DWMH) [Kim et al., 2008]. DWMH has been commonly associated with an ischaemic explanation. For its part, PVWMH can be associated to pathophysiological explanations according to their appearance. In the case of periventricular (PV) caps, the areas with hyperintensities seem to be due to a sort of discontinuity in the ependymal lining delimiting the ventricles. These changes do not affect the signal transmission directly through myelin deterioration [Kim et al., 2008]. As a consequence, the observed hyperintensities in periventricular regions are not discriminative for myelin degradation and should not be associated with myelin pathology [Haller et al., 2013]. Another challenge for proper characterization of lesion spatial distribution is due to the evolution over the time of white matter damage [Yoshita et al., 2005]. When studying pathologies, longitudinal evolution plays an important role to clarify causal relationships. A consistent pattern has been found regarding the evolution of white matter disease. WMHs seem to appear first at the horns of lateral ventricles, then they spread around them before reaching the deep white matter and the basal ganglia. Moreover, it has been found that the rate of appearance of WMH changes is associated with the volume of the lesion at the beginning [Maillard et al., 2014]. Thus more severe cases will develop faster [Patel and Markus, 2011].

2.2 Medical Imaging

2.2.1 Imaging the brain

Medical imaging has allowed the understanding of the structural organization and functioning of the brain as well as the anomalies that can be present in it. In terms of imaging, abnormalities can refer to changes in structural appearance such as enlarged ventricles or shrinkage of the hippocampus as well as the unexpected intensity presentations as observed in tumor, necrosis, or white matter lesions. Regarding to abnormalities, it is expected those observations are related to some pathological processes that affect the functioning of a given organ. It is important to note, that finding causality relationships is a difficult task, consequently, most of the conclusions are based on correlations that have proven to be useful and clinically relevant along time. Imaging the brain relies on the differential properties of the tissues to respond to stimulus such as X-rays, magnetic fields, or the way they behave respect to a source of stimulus e.g., positron emission radiotracers.

Magnetic resonance imaging (MRI) has proven to be effective to differentiate the soft tissues that compose the brain, becoming the modality of choice for its analysis. MRI works by analyzing how tissues respond to different types of magnetic excitations in time. The MRI acquisition process consists in applying continuously a magnetic field and the proton's spin naturally aligns with it. A magnetic pulse oriented in a perpendicular plane is applied to force the spins to align with the new field and rotate in phase at a determined frequency of resonance. When the excitation pulse decreases, its respected spins return to their original stated, according with two tissue properties namely T1 and T2 relaxation properties. The T1 relaxation property or spin-lattice relaxations makes reference to the level of energy that is released on the intersection between protons and the molecules around it. On the other hand, the T2 relaxation property or spin-spin relaxation, reflects the velocity with the exited spins dephases as a result of the magnetic interactions between protons. Different contrast can be obtained by different designs of pulse sequences. T1-weighted images (T1-w or T1), were designed to enhance structural contrast between tissues. Here T1 relaxation properties have more prominence that T2 characteristics which accentuate the energy iterations levels between protons and neighboring molecules. In T1-weighted images, tissues with short T1 relaxation time such fat tissues will appear with a higher signal compared with tissues in which protons exhibit lower energy interaction with surrounding molecules such as CSF which is displayed dark. On the other hand, T2-weighted (T2-w or T2) pulse sequences highlight the property of interactions between protons whereas PD-weighted images rely on the proportions for protons in tissues. Fluid Attenuated inversion Recovery (FLAIR) is able to provide the same contrast as T2-weighted while nulling the free water signal.

2.2.2 Imaging white matter hyperintensities

White matter hyperintensities appear as bright in some MRI pulse sequences. Deterioration of myelin in white matter damage due to age results in an increment of water in the extracellular volume close to the ventriculus that contributes to change in the ratio between tissue and water so between molecules and free protons. When the water/fat [Malloy et al., 2007] ratio changes, it proportionally affect the T1 and T2 relaxation time in such a way that tissue properties resemble to CSF properties. Therefore, in T2-weighted contrast image, such damage appears brighter than healthy tissue. However, when the water proportion increases in such lesions , it is not entirely fluid-filled which makes it still visible as bright on FLAIR scans except in most severe cases. Therefore white matter hyperintensities have mostly been visualized using T2-weighted, PD-weighted and FLAIR sequences.

As mentioned in Section 2.1.1.1, WMH start around ventricular regions. Therefore, on T2-w and PD-w images distinction between CSF and WMHs is challenging as both appear bright. As a result, FLAIR images have a big advantage as they are able to null the CSF signal while keeping the bright signal for lesions. On the other hand, in T1-w depending on the severity of the lesion, the signal goes from iso to hypointense. T1-w images are usually not suitable for the quantification and assessment of WMH, however, they provide rich structural information [Olsson et al., 2013]. Figure 2.1 illustrates the presence of WMH on T1, T2, PD and FLAIR images.

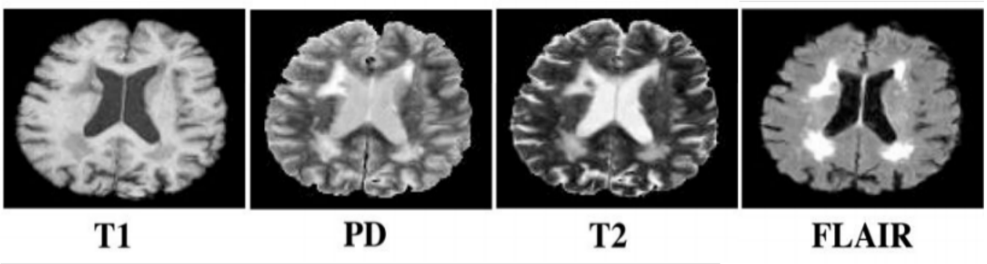

Figure 2.1: Example of WMH on T1, T2 and FLAIR MR images (the figure is adapted from [Schwarz et al., 2009])

2.2.2.1 FLAIR disadvantages

Although the FLAIR sequence is usually the modality of choice for WMH assessment, it suffers from some drawbacks. It has been demonstrated that FLAIR images tend to overestimate the extent of lesions [Rovaris et al., 1999]. Moreover, there is no direct relation between the intensity depicted on FLAIR and the severity of the lesion [Uhlenbrock and Sehlen, 1989] as opposed to T1-w or T2-w. In addition, the hyperintense signals are non-discriminative for the underlying pathological process. Consequently, different diseases exhibit similar presentation e.g., MS lesions and age-related WMH. FLAIR images tend to be prone to be affected by artifacts that can affect their prediction [Lavdas et al., 2014]. Moreover, shine through effects due to the increased signal intensity at the boundaries between the cortical ribbon and external CSF can make difficult delineation of lesions. 3D FLAIR images, which are high-resolution 3D volumes with isotropic voxel dimensions and the absence of inter-slice gap [Chagla et al., 2008], can solve most of the artifacts that are mostly observed in 2D acquired FLAIR [Kakeda et al., 2012]. However 3D FLAIR images are not widely used in clinical settings [Lavdas et al., 2014].

2.3 WMHs Identification

2.3.1 The need for automatic segmentation

In order to establish associations between white matter damage and any neurological disorder, a quantitative analysis is required to determine the extent of the damage. Two directions can be taken to accomplish that task: Visual grading scales and volumetric measurements.

Grading scales such as the Scheltens scale or the Fazekas scale [Scheltens et al., 1998] allows quantification of the severity of the damage over the whole brain. What makes visual grading attractive is that they are fast to obtain. However, they can suffer from non-linearity, poor sensitivity to small changes and they are susceptible to flooring and ceiling effects [Kim et al., 2008, van Straaten et al., 2007].

On the other hand, quantitative measurement based on delineation of tissues seems to provide an alternative for better assessment of WM abnormalities [van Straaten et al., 2007]. Measures such as volumetric estimations however, need to be robust and accurate to ascertain valid clinical correlations. However, manual segmentation is cumbersome and time-consuming, and shows inter and intrarater variability [Ashton et al., 2003, Filippi et al., 1995] as a consequence it is not a feasible solution in large clinical studies. Thus, the development of reliable and robust automated WMHs segmentation methods is crucial. For a segmentation method to be reliable and robust ideally it would aim high sensitivity, i.e. be able to find all the lesions, but without compromising specificity i.e. no produce overestimation or find more lesions than there actually are. And most important, predictions should be consistent in the borders definitions under several conditions of resolutions, artifacts, noise. Several challenges make this task difficult, WMHs can occur everywhere and present high variability in morphology. Moreover, lesion intensity can be confused with healthy gray matter specially at the boundaries of lesions which leads to missclassification.

2.3.2 Overview of segmentation methods

Several techniques have been designed for automatic segmentation of WMHs. The literature is presented in two groups: First, I consider unsupervised methods, the methods that do not require explicitly a database with manual annotations. Next I present supervised methods, I first review non-deep learning based method and finally I present segmentation methods based on deep neural networks.

2.3.2.1 Unsupervised methods

Generally, methods in this category use clustering techniques. Among those, Fuzzy-C means (FCM) [Bezdek et al., 1984] is one of the most common approaches. In [Boudraa et al., 2000], a two-stage classification approach is proposed, in the first stage a 3 class FCM is applied on PD images enhanced by histogram equalization. Intensities from corresponding T2-w images are used to select potential lesions and CSF classes. In the next stage a second FCM is applied on the selected elements to provide more a accurate selection. Finally correction for false positives based on minimum size and brain border adjacency to border is performed. In [Admiraal-Behloul et al., 2005] an adaptative level classifier is first performed where FCM with a different number of clustering cases is applied to T2-w, PD-w, and FLAIR images and the output is combined with a fuzzy inference system which is based on intensity linguistic rules such us brig, dark, etc. To improve false-positive correction they initialize the centroids of the FCM with spatial priors that come from template atlases. Wu et al. [2006] computed seeds automatically based on the intensity histogram of FLAIR images. A threshold is set to be the mean plus 3 standard deviations and is used to label the seeds. Next, a fuzzy connected algorithm is employed to carry out the segmentation of WMHs while the seeds are updated iteratively. The process stops when the algorithm can no longer identify any seed, this point clusters are combined and the final segmentation is produced. One approach that successfully includes spatial variability into the FCM framework is presented in [Anitha et al., 2012], they show improvement in lesion segmentation on images from elderly population. In [Gao et al., 2014] they used a non-local mean regularizer guided by spatial consistency information which is introduced in the energy function. For their part, in [Shiee et al., 2010] a bias field modeling is included in the FCM energy function aiming to ensure spatial consistency guided for both statistical and topological atlas.

In conclusion, segmentation methods based on unsupervised methodologies have proven to be effective in the segmentation of WMH. The simplicity in their implementation, and the no need for annotated training sets make them attractive to some clinical scenarios where annotated data is scarce. Despite the decent segmentation that can be obtained from only intensity information, more recent works have stressed out the need to include spatial information [Wang et al., 2012, Shi et al., 2013]. Including spatial information can improve methods robustness to noise especially required for accurate segmentation of lesion boundaries. Moreover, spatial information enables modeling of voxel intensity dependency making models less prone to false positives.

2.3.2.2 Supervised Methods

Before the boom of deep learning, early supervised methods for WMH segmentation relied on common classification algorithms. On those approaches, segmentation was voxel-wise, where each voxel was represented by a feature vector and constituted the inputs for traditional supervised learning algorithms. Anbeek et al. [2004], used multi-modal information from T1-w, PD, T2-w and FLAIR images to train a K-nearest neighbors (KNN) classifier. For a randomly selected set of voxels intensity and 3D spatial information, features were extracted and used to train the classifier. They found that the combination of intensity and 3D spatial features yields the best performance. In [Lao et al., 2008] they replaced KNN for a support vector machine (SVM) classifier. In addition, they change the voxel feature representation to account for neighborhood information. Specifically, for each voxel, the feature vector is composed of spatial and intensity information from the small neighborhood around it. They demonstrated that introducing neighborhood information makes the classifier more robust to misregistration. Instead of performing voxel-wise WMHs segmentation, the method proposed by Beare et al. [2009] aims to segment regions. In their method region-based features are extracted and combined with an adaptative boosting statistical classifier. Another group of approaches aims to model spatial and intensity distributions via parametric models. In [Simões et al., 2013] a 3-component Gaussian mixture model with CSF, GM+WM, and WMH classes is used to model FLAIR images based on histogram analysis. In addition, a context-sensitive penalty term is introduced in the Expectation-Maximization algorithm to enforce model regularity in segmentation. In [Khayati et al., 2008], an adaptative mixture model (AMM) is combined with a Bayesian classifier which includes a Markov random field to encourage neighborhood consistency. If the current model can not model properly an observed sample, a new gaussian component is added, those components are grouped on three classes whose parameters are optimized using the MRF and AMM. Schmidt et al. [2012] leverage information of different modalities in a sequential methodology. In the initial step, beliefs maps are created using T1-w images and then a threshold is applied to get seeds that are used to initialize the segmentation. In the next step, a 3-component GMM is used to model FLAIR images for healthy tissues, and a gamma distribution for the lesion class.

In the presence of manually annotated training sets, supervised classification methods have proven to be successful in solving automatic segmentation problems. However, the selection of the right feature extraction methodologies adds complexity to the problem that has shown to be solvable by modern supervised strategies based on neural networks.

2.3.2.3 Deep learning segmentation methods

Deep neural networks established a significant change regarding the traditional machine learning approaches for segmentation. Instead of relying on hand-crafted features, deep neural networks are able to learn hierarchical features automatically [Perone et al., 2019]. Since Deep learning models appeared their inclusion in medical image applications has been progressively increasing. They have been successfully applied in several applications such as registration, object detection and segmentation [Suzuki, 2017]. When applying deep learning techniques to the medical domain, one of the main challenges is how to adapt architectures to different input formats such as three-dimensional data. A common strategy that is still commonly used is to divide a Volume of Interest (VOI) into slices which are then fed into a network. These approaches avoid using a large number of parameters as 3D convolutions are replaced by their 2D counterpart.

Prasoon et al. [2013] was one of the pioneers using this approach for the segmentation of knee cartilage. Patches or slices from different orientations (sagittal, axial, coronal) can be extracted to feed a network in a multi-stream fashion. Examples of these methods also referred to as 2.5D classification are presented in [Roth et al., 2015b, Setio et al., 2016]. In early approaches for segmentation, convolutional neural networks were used to classify every voxel in the image individually. For each voxel, a patch is extracted and used to feed the network. However, a critical disadvantage of this 'sliding-window' approach is that the same convolutions are computed many times due to the huge overlap from input patches from neighboring voxels. A more efficient architecture well known as Fully Convolutional Neural Networks (FCNN) [Long et al., 2015] was proposed to overcome this problem. In this scheme, fully connected layers are rewritten as a convolution to output a likelihood map, instead of an output for a single voxel. Therefore FCNN can be applied to an entire input image/volume more efficiently. One problem that arises due to pooling layers is that outputs may present a lower resolution than the input. This problem is solved by applying a FCNN to differently shifted versions of the input image. Then the final resolution is obtained by stitching the result of the different versions. On the other hand, an elegant architecture well known as U-net, was proposed for Ronneberger et al. [2015]. This architecture comprises a FCNN followed by an upsampling part where up-sampling operators are used for increasing the size image. Moreover, skip connections are used to connect opposing contracting and expanding convolutional layers. A 3D implementation called V-net was proposed for Milletari et al. [2016] which is more naturally applicable for 3D volumes and enables capturing full context information. Regarding WMH segmentation, several approaches have been proposed. The MICCAI 2017 segmentation challenge [Kuijf et al., 2019] evaluated 20 methods using a dataset of 170 images (60 for training and 110 for testing) from three different centers (Amsterdam, Singapore, Utrecht). Among the presented methods, 15 relied on some form of convolutional neural network. Where the best performing team [Li et al., 2018] relies on an ensemble of three U-net models that use 2D slices extracted from T1-w and FLAIR images.

2.4 Lack of annotated data

The lack of annotated data is a critical problem when applying supervised deep learning methods on medical applications. In this section, I review some of the traditional solutions to this problem such a Data Augmentation, Transfer Learning, and Domain Adaptation (DA).

2.4.1 Data augmentation

Data augmentation provides and effective solution to reduce the effect of over-fitting in scenarios where the data is scarce. The Data Augmentation process involves expanding the training set by introducing new samples derived from the available data. Approaches vary in the type of operations applied to the available data to produce new training samples.

2.4.1.1 Traditional Data augmentation

The data augmentation methods applied in medical imaging can be divided according to the image property they manipulated [Zhang et al., 2019]. Among these properties, I can find: image quality, image appearance, and image spatial arrangement (geometry).

- **Image quality:** Different types of filters changing sharpness or adding blurriness or noise can be applied to images to change its quality. Gaussian noise was applied to CT images in [Christ et al., 2016] for data augmentation. Sirinukunwattana et al. [2017] blurs colon histology images using a gaussian filter. Zhang et al. [2019] show the benefits of augmenting the data by adjusting image quality. They demonstrated that image sharpening leads the largest improvement in the application of unsharp masking.

- **Image appearance:** New samples for augmentation can be obtained by manipulating brightness, saturation, and contrast of images. Dong et al. [2017] proposed the enhancement of brightness in 3D MR volumes. In [Fu et al., 2017] and [Alex et al., 2017] it was demonstrated that contrast-based augmentation is helpful in data images with inhomogeneous intensities. The former applies a contrast transformation to microscopy images for the segmentation of nuclei. The latter uses histogram matching between images in the dataset and a reference image chosen randomly from the training data.

- **Geometry :** Data augmentation can be performed by applying spatial transformations such as rotation, scaling, and shearing. More, sophisticated methods apply non-linear transformations such as elastic deformations [Ronneberger et al., 2015] , dense deformation field [Milletari et al., 2016] , b-spline deformations [Çiçek et al., 2016].

2.4.1.2 Synthetic augmentation

This type of augmentation strategy has been governed mostly by Generative Adversarial Networks (GAN) [Goodfellow et al., 2014] and its variations. Fu et al. [2018] improve CycleGAN by including a spatial information that enables CycleGAN to generate synthetic images with the object of interest appearing in desired locations. They demonstrated that augmented samples generated with spatially constrained CycleGAN improve segmentation model performance. Guibas et al. [2017] propose a sequential algorithm composed of both a GAN and a conditional GAN (cGAN) to generate synthetic fundus images. Firstly, a GAN takes a random vector sampled from a normal distribution to generate a segmentation mask. This mask goes through a cGAN to generate a synthetic fundus image. Tang et al. [2019] train a stacked

GAN (SGAN) with a pair of GANs that aim to generate both denoised and high-resolution image pairs. Shin et al. [2018] provide a powerful method based on a cGAN to generated synthetic customized MR-images with tumors. The cGAN takes as input a tumor mask and a brain mask, where the size and location of tumors can be predefined by the user. The output is a MR image with a tumor in accordance to the size and location defined by the user. Segmentation models that are trained with both synthetic and real real images achieve a significant improvement in performance compared to the models trained with real images only. Jin et al. [2018] aim to genereate synthesize pleural nodules from nodule free CT slices. To this end, they proposed an inpainting model based on a conditional GAN.

Another group of approaches aim to generate synthetic images via Transformation networks. The method proposed by Zhao et al. [2019] is able to generate a training set from only one labeled image and a given set of unlabeled images. To this end, they proposed an hybrid spatial-intensity transformation model. A spatial transformation network is firstly used to deform the labeled image to match the shape of a given unlabeled image. Then the intensity at each voxel in the labeled images is change using a intensity transformation network to match the appearance of the given unlabeled image. A similar approach was proposed by Chaitanya et al. [2019], a conditional spatial and intensity generators are trained in an adversarial fashion to generate images that resemble the appearance of both, labeled and unalbeled images in a training set.

2.4.2 Leveraging external datasets

Techniques like Transfer Learning (TL) and Domain Adaptation (DA) have also been developed aiming to alleviate the problem of lack of annotated data by leveraging data that can be from a different domain or nature. In Transfer Learning, the general idea is to pre-train a model with a large external database. Then, this model can be fine-tuned using the target dataset. For its part, Domain Adaptation techniques aim to bridge the distribution gap between Training and testing sets.

2.4.2.1 Transfer Learning

For medical imaging applications where data is scarce the idea of transfer learning is to leverage the potential of large non-medical image datasets. However, the adoption of transfer learning in medical image segmentation has been limited due to the 3D nature of medical images. The majority of large datasets are composed of natural images that are 2D which impede straightforward transferability.

As previously mentioned in Section 2.3.1, 3D segmentation can be carried out by splitting the 3D images into 2D slices to train 2D networks. This makes it easier for the transfer of knowledge from natural images to 2D medical image segmentation models. In [Ma et al., 2019] the authors take as a base an autoencoder that has been trained on natural images and subsequently is fine-tuned with medical images. Qin [2019] proposed a slightly different strategy. They initialize and encoder randomly that is appended to a decoder that has been trained for the classification of natural images. Then the entire network is fine-tuned with medical images.

On the other hand, there are few works that explore transfer from 2D pre-trained models to models targeted at 3D medical applications. Yu et al. [2018] makes an attempt to transfer models trained on natural scenes, where the third dimension of medical images is used as a temporal axis. However, this approach can not guarantee the learning of 3D context of medical scans. Liu et al. [2018] proposed a sophisticated approach for transforming 2D models into a 3D network. This is possible by extending 2D convolution into 3D separable anisotropic filters. Consequently, this approach enables the initialization of 3D models from 2D models.

2.4.2.2 Domain adaptation

Besides the scarcity of large manually annotated training sets, another challenge is the distribution shift between available training data and the data faced in clinical practice. Different scanners or protocols used during the image acquisition process or even different patient population and demographics, can produce different distributions. Due to training datasets typically come from the same center, there is an inherent bias and the resulting deep learning models tend to decrease their performance when deployed in data which is different from the data used during training. Domain adaptation methods aim to bridge the gap between different domains distributions. Common directions to accomplish this task are: *i)* learning a representation invariant to domain signal or *ii)* learning to translate images from the training domain to the testing domain. Domains can be due to different imaging modalities or different distributions on data with the same modality. The vast majority of domain adaptation techniques rely on some sort of adversarial learning using GANs or CycleGANs [Zhu et al., 2017]. The original GAN proposed by Goodfellow et al. [2014] is based on a dual network scheme: A discriminator is training to differentiate between real and synthetic images whereas the generator is trained to generate realistic synthetic images that fool the discriminator. In the context of domain translation, the idea is the generator learns to map images from one domain to another. CycleGANs can achieve this objective by a dual mapping using a double pair of generators and discriminators. One pair performs the translation from source to target and the other does the inverse mapping. Methods performing domain adaptation can be grouped in supervised domain adaptation methods, which require labels for the target domain or unsupervised domain adaptation methods, which may only need unlabeled data from the target domain.

Unsupervised domain adaptation: In this type of approaches, a set of annotated data (Source Domain) is available, but there are not labels in the target domain. Under this condition some authors approached domain adaptation by transforming images from the source domain to have the style of the target domain while preserving the anatomical structure. Once the transformation is done, a segmentation network is trained using the transformed images, note the labels are preserved. One example of this approach was presented by Huo et al. [2018] which uses a joint synthesis and segmentation framework using unlabeled target images and labeled images from a source domain. In Chen et al. [2018] a CycleGan is used to perform domain translation from MR to CT images. They use a shared encoder for both the segmentation and the synthesis networks, this multitask setting also prevents the model from over-fitting. As an alternative to the source to target domain translation, a group of authors propose to perform the adaptation in the opposite direction. A synthesis network is trained to perform translation from target to source domain, so during inference, target images are converted to source domain style and then used as input in a segmentation network previously trained in source domain images [Zhao et al., 2017, Zhang et al., 2018, Yang et al., 2018] .

Another group of approaches aim to minimizing the discrepancy between feature representations from the source and the target domain. An adversarial approach to learn domain invariant features was proposed in [Kamnitsas et al., 2017] for brain tumor segmentation. Dong et al. [2018] proposed a similar approach but in their method, the discriminator aims to distinguish if the segmentation mask is a ground truth or if it is a predicted mask. In [Wang et al., 2019b] a patch-based adversarial learning is proposed to encourage a segmentation network generating similar prediction for source and target regions of interest. In [Wang et al., 2019a] a boundary and entropy-driven adversarial learning is proposed to encourage boundary predictions and probability entropy maps of the source and target domain to be similar. As a result, the generated boundaries are more accurate and the uncertainty in the predictions is reduced.

Supervised domain adaptation: When labels are available for both domains, there is no need to make a distinction between source and target domains. The adaptation is then achieved by learning a shared feature representation which leads to a more robust prediction independently of the input domain. Learning from multiple datasets enables more efficient training, as models can learn discriminative features from different distributions. Moreover, enhanced regularization, as data from multiple sources can provide further supervision. Harouni et al. [2018] propose a modality agnostic model that is trained using data from different modalities. The authors showed that the jointly trained network achieved similar performance than individually trained networks for each modality. Dmitriev and Kaufman [2019] use a similar approach by training a segmentation model with images from different single organ datasets. An additional channels is introduced to condition segmentation predictions according to the desired organ for segmentation.

Chapter 3

Dealing with Inter-Modality Heterogeneity

3.1 Derived publication

- **Orbes-Arteaga, M**, Cardoso, M. J., Sørensen, L., Modat, M., Ourselin, S., Nielsen, M., & Pai, A. (2018). Simultaneous synthesis of FLAIR and segmentation of white matter hypointensities from T1 MRIs. *In 1st International Conference on Medical Imaging with Deep Learning (MIDL 2018)*

3.2 Preface

The analysis of WMHs is benefited from the information provided by multiple modalities. WMHs are better highlighted on T2-W, PD-W, and FLAIR whereas T1-w provides more rich structural and lesion severity information. Similarly to clinicians, neural networks can learn more information from multiple modalities. This fact motivated the development of multi-modal segmentation models. In deep learning methods, the simplest yet common approach to learn from multi-modal data is by feeding a neural network using each modality as one input channel. In this setting, training data is paired meaning each sample holds a set of scans from different modalities that have been co-registered to have one-to-one-spatial correspondence. During the deployment stage however, those approaches require for new patient to get the same modalities used during training. This is a huge disadvantage in real clinical scenarios where there is no guarantee that every sequence will be available.

Moreover, In many circumstances only less involved sequences are acquired. As explained in Section 2.2 , although FLAIR sequences are the modality of choice for WMH analysis, this is a more involved pulse sequence if acquired at high resolution compared to T1-w (it requires additional time, adding cost to standard clinical imaging protocol). In order to provide segmentation solutions that can have a real impact on clinical practice, algorithms should be able to work on simplistic settings which rely on most commonly used modalities.

In this chapter, we propose a systematic training strategy to learn from multimodal-information with the ultimate goal of performing segmentation on commonly used sequences. The proposed method relies on a dual optimization strategy to learn efficiently from available multi-modal data by simultaneously learning synthesis of FLAIR and segmentation of WMH segmentation using only T1-w as input. We demonstrated that the proposed method is able to generate realistic synthetic FLAIR sequences from T1-w scans. Most important, segmentation from available modalities is enhanced through the optimization process.

3.3 Related work

Given the presence of limited data with desired multiple modalities, data imputation methods are used to learn the synthesis of missing modality using T1-W scans. The intention of imputing data is to guide the optimization using prior information, i.e., the available FLAIR sequence. As stated in van Tulder and de Bruijne [2015], synthetic data helps the segmentation because of two reasons. Firstly, the flexibility of synthesis model allows finding features that can not be seen by the classifier in an otherwise single-modality model. Secondly, the size of the training set is synthetically increased which is useful in the training process.

Among CNN-based imputation methods, the most popular ones using a flavor of generative adversarial networks (GANs) Goodfellow et al. [2014]. For instance, Nie et al. [2017] use GANs to generate CT images from MRI images. However, most of the current implementations treat synthesis as a preprocessing step [Ben-Cohen et al., 2018, Zhang et al., 2018, Huo et al., 2017]. This restricts the network, and the features may not be particularly useful for the final segmentation.

Instead, we proposed a simultaneous training based synthesis method that combines generation of the missing modality and segmentation – inspired from Tran et al. [2017]. We show with experiments that using the proposed method to synthesize FLAIR images, we not only obtain higher quality synthetic flair images (when compared to treating synthesis a preprocessing step) but also improve the segmentation of WMH using T1-w images only.

3.4 Methods

Let $\mathcal{X}$=$\{\boldsymbol{X^n},\boldsymbol{L^n} : 1,\ldots,N\}$ be an annotated training set which have N subjects . Here, $\boldsymbol{X} = \{X_a,X_b\}$, is a pair of MRI images from two different modality sources for a given subject, and $\boldsymbol{L}$ is a volume with the manual annotation for WMH. The goal in multi modal segmentation task is to find a mapping $C(\boldsymbol{X},\theta_c)$ from a pair of available modalities to a corresponding segmentation.

$$C : \{X_a,X_b\}, \rightarrow L \tag{3.1}$$

Here, C is a function represented by a CNN with parameters θ_c. We then train C to maximize:

$$\max_{\theta_c} \mathbb{E}[\log \ p(L|X_a,X_b,\theta_c)] \tag{3.2}$$

It is evident that to train, and subsequently test such a scheme, both modalities are needed. This is a restriction, specially when the network is used to test retrospective data with missing modalities. One common approach to deal with missing modalities is to impute them. Formally, a function G (a CNN) is trained to learn a mapping between the available modality and the missing modality., i.e $G(X_a) \approx X_b$. Subsequently, the synthesized modality is used in conjunction with the available modality to train a classifier for segmentation. The optimization function for the classifier in Equation (3.2) can be re-written as:

$$\max_{\theta_c} \mathbb{E}[\log\ p(L|X_a, G(X_a), \theta_c)] \tag{3.3}$$

Note that in this scheme, the generation and the classification are different optimizations. No complementary information is taken into account. Therefore, in this work, we aim to learn the generation and classification (respectively performed by G and C) simultaneously so that C reinforces the generation G to produce not only realistic images but also relevant features that help in the optimization of C.

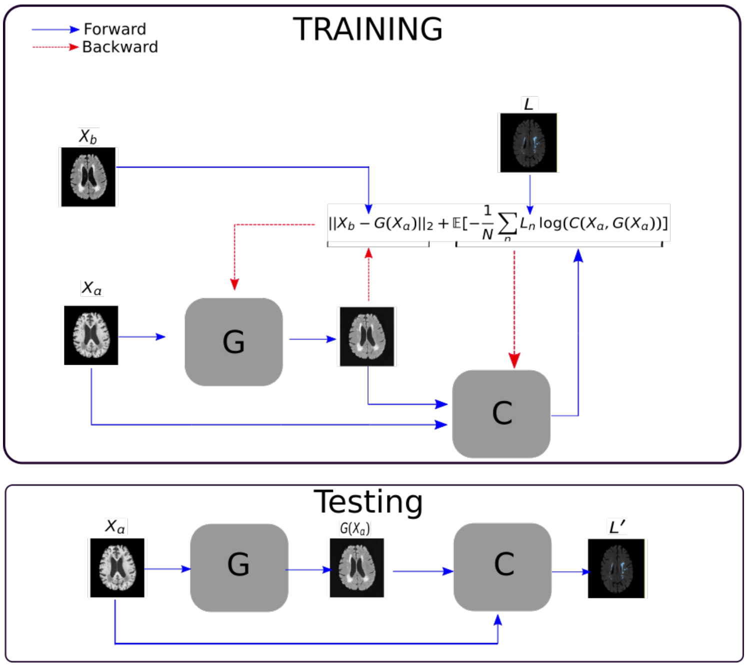
Forward
Backward
TRAINING
X_b
L
$||X_b - G(X_a)||_2 + \mathbb{E}[-\frac{1}{N}\sum_n L_n \log(C(X_a, G(X_a))]$
X_a
G
C
Testing
X_a
G
$G(X_a)$
C
L'

Figure 3.1: Illustration of process follow for training and testing of our method.

The scheme is basically composed of two networks, a generator G and a classifier C where both networks are trained end to end iteratively, see Figure 3.1. The classifier training is linked to the generator by taking both the real T1 image denoted by X_a and the generated image $G(X_a)$ to produce a segmentation $\boldsymbol{L'}$. The loss function of the classifier network is:

$$\mathcal{L}_C = \mathbb{E}[-\frac{1}{N}\sum_n L_n \log(C(X_a^n, G(X_a^n))] \tag{3.4}$$

L2 distance is typically used as a loss function for the synthesis and reconstruction of images using FCN. L2 is attractive due to its easy implementation, clear physical interpretation, and is equivalent to maximizing other metrics such as the peak signal-to-noise rate (PSNR) [Nie et al., 2018]. In addition, L2 has been the auxiliary loss of choice for reconstruction or synthesis methods that are based on adversarial learning or that involve a combination of different losse [Dong et al., 2015, Huang et al., 2017, Yang et al., 2021, Lahiri et al., 2020]. Finally, although L2 shares similar properties to L1, L2 has outperformed L1 in some reconstruction tasks [Ganguli et al., 2019]. Consequently, in order to train G we use L2 as a reconstruction error between the real missing modality image and its corresponding generation. Thus, one may then view the classifier to be a regularization term to the generator or vice versa. The $\mathcal{L}_G$ loss for the generator is given by:

$$\mathcal{L}_G = ||X_b - G(X_a)||^2 + \mathbb{E}[-\frac{1}{N}\sum_n L_n \log(C(X_a^n, G(X_a^n))] \tag{3.5}$$

3.4.1 Network architectures

We use U-Nets Li et al. [2018] (winner in 2017 MICCAI- WMHs segmentation challenge) as the segmentation network, and a modification of it as a generation network. The changes involve changing the number of inputs channels from two to one which corresponds to the T1 modality, we also change the `Sigmoid` function in the final layer by `LeakyRelu`. We use Adam optimizer with learning rate `0.0002` for both the networks, and batch normalization. The classifier and generator are trained iteratively with the same frequency. We do not use any data augmentation.

3.5 Experiments and results

3.5.1 Data and Experiments

We validated our proposed method on the training dataset from the 2017 White Matter Hyperintensity Segmentation Challenge (`http://wmh.isi.uu.nl`). This dataset is composed of T1 and FLAIR scans for 60 subjects from three different clinics (Utrecht, Singapore, and AmsterdamGE3T, 20 subjects for each one), the data is complemented with manual annotations of WMH from presumed vascular origin. FLAIR images have been used as a reference for label annotations, so, T1 images have been registered to this space. The images were also corrected for bias field inhomogeneities using SPM12. As a further preprocessing we use only two of three stages performed in Li et al. [2018], which include *i)* cropping or padding of axial slices *ii)* Gaussian normalization of voxel intensities. We did not perform data augmentation as these did not show significant improvement in segmentation.

All the methods were evaluated using a 6-fold cross validation. The dataset was split in such a way that all the 60 images are tested at least once. For each fold, we pick 10 subjects for test, 5 for validation, and the remaining 45 are used for training. For evaluation, dice scores (DSC), false positive rates (FPR), and false negative rates (FNR) are used.

3.5.2 Results

We evaluated our method in segmenting WMH from T1-w images using: a) Synthesized FLAIR images by treating the synthesis as a preprocessing step – we will refer to this method as *offline synthesis*; b) Synthesized FLAIR images using the proposed method, and c) without any synthesis – we will refer to this method as *Unimodal*. Baseline methods are illustrated in Figure 3.2

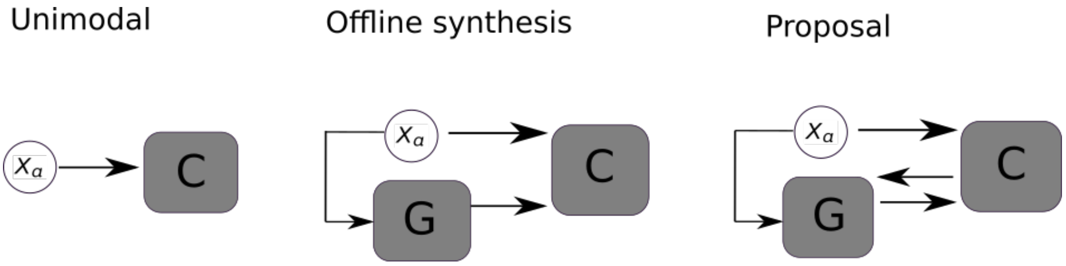

Figure 3.2: Illustration of methods of comparison, X_a represent a T1 image.

Table 3.1 shows the mean of each measure for all considered methods. As we can see, our method achieves higher dice scores than baseline methods. A mean dice improvement of nearly three percent is obtained using our proposed method when compared the baseline method without any imputation. In addition, the proposed method also improves segmentation when compared to an offline synthesis.

Table 3.1: Average of performance measures for all comparison methods, results in bold are significantly different (p<0.005) from the baseline *Unimodal* method (top row)

	Evaluation Metric		
Method	DSC(%)	FPR(%)	FNR(%)
Unimodal	55.99	78.67	38.06
Offline synthesis	54.42	63.50	**43.39**
Proposal	**57.81**	**58.20**	**41.33**

It is important to note, that our proposed method shows a FPR 20.47% lower than *Unimodal* and 5.3% lower than *offline synthesis* method, showing the effectiveness of our method to reduce the number of false positives. On the other hand, *Unimodal* method shows the lower rates in terms of FN.

In order to better understand the above results, we visually analyzed the output segmentation performed for each method. Figure 3.3 shows the results for three different slices (one slice per column). As illustrated, the proposed method is able to produce less false positives. It is also important to note that, unimodal segmentation is the one that produces more false positives, showing the advantage of using synthetic data. Regarding the nature of false positives, it can be easy to see in the third column a large number of false positives are on the border of periventricular lesions for the *Unimodal* method in comparison to the proposed method. Also from the first and second column, it can be observed that *Unimodal* tend to produce more small regions of false positives near to cortical areas. Removing such false positives requires additional post-processing steps, therefore, it is of value avoid this kind of over-segmentation. It can also be noted that synthesis methods tend to produce the same kind of false negatives, this may be due to the blurring effects in synthesized images since the information available during testing is limited – which otherwise is available from a FLAIR sequence.

3.5.3 Results of Generation

Here we compare the generate FLAIR images obtained for the generator using our optimization strategy against the generated images obtained for using *off-line* *synthesis*. Firstly, images were quantitatively evaluated in terms of reconstruction using two well known measures, namely mean absolute error (MAE) and peak-signal-to-noise-ratio (PSNR). Results of reconstruction measures are shown in Table 3.2 , as we can see our proposal outperforms the baseline approach in both MAE and PSNR. Specifically, images generated for our proposed method achieve an average PSNR of 11.01 which is considerably higher compared with 9.65 obtained for images generated *Offline*. Reconstruction superiority of our methods is confirmed by the MEA results, 0.26 and 0.31 for our proposal and the baseline respectively.

In order to analyze qualitatively the results of our generator, we extract slices with different WMHs loads, Fig. 3.4 shows the reconstruction results for three different levels of loads. As we can see in the first row, both methods produce a similar response in regions with a low load of lesions, it can be observed that generated images are similar to the real FLAIR images in the left, and these not present evident structural distortions. However, it can be noted images exhibit blurred effects, which can be due to L2 based optimization, more complex generative networks with adversarial loss optimization as GANs tend to eliminate blurred effect but at the expense to produce structural distortions. In the application presented in this work it is important to preserve the structural information, thus, our L2 based optimization present a good balance between preserve structural information and blurred effects. In the second and third column, it can be observed the performance of both methods when facing the presence of lesions, as can be seen, both methods have a good response to large and contiguous lesions. It also can be noted both methods tend to produce poor performance in small and diffuse WMHs marked in red, note, these lesion do not exhibit identifiable patterns in T1 images, however it can be seen that our proposed method is more sensitive to these patterns which enable to highlight some small regions as those marked in green.

3.6 Discussion and concluding remarks

In this chapter, a new CNN-based method to improve WMH segmentation from T1-w images alone is proposed. The method jointly performs imputation and segmentation in such a way that both tasks are mutually benefited. To this end, FLAIR sequences are used to drive the optimization, which reflects in the results where joint optimization of synthesis and segmentation yield better segmentation from T1-only images.

From segmentation results in Section 3.5.2, it is evident that the T1-based segmentation tends to have excessive over-segmentation of images. By using prior information from FLAIR images through a generator, we are able to reduce the number of false positives. However, it could be observed that, if imputation comes from an independent synthesis model, images tend to be under segmented (high FNR) reducing the overall segmentation accuracy. The proposed joint optimization strategy better adapts to capture small lesions, which leads to significantly better overall segmentation performance.

In addition to an improved segmentation performance, we can see in Section 3.5.3 that the proposed method also produces better synthetic FLAIR images when compared to networks that trained to only specialize in generation. This may be due to the complementary information available through a joint optimization with the segmentation network. Specially, lesions that are barely visible in T1 images are seen in synthetic images produced by the proposed method.

One of the disadvantages of our method is using L2 as a loss function can produce blurring effect on the images. Using adversarial training by the use of a discriminative network as a loss function may overcome this issue. However, with an introduction of an additional network and the availability of limited training data, the optimization may be prone overfitting. Therefore the proposed method with L2 loss provides a good compromise between the complexity of the model and segmentation performance.

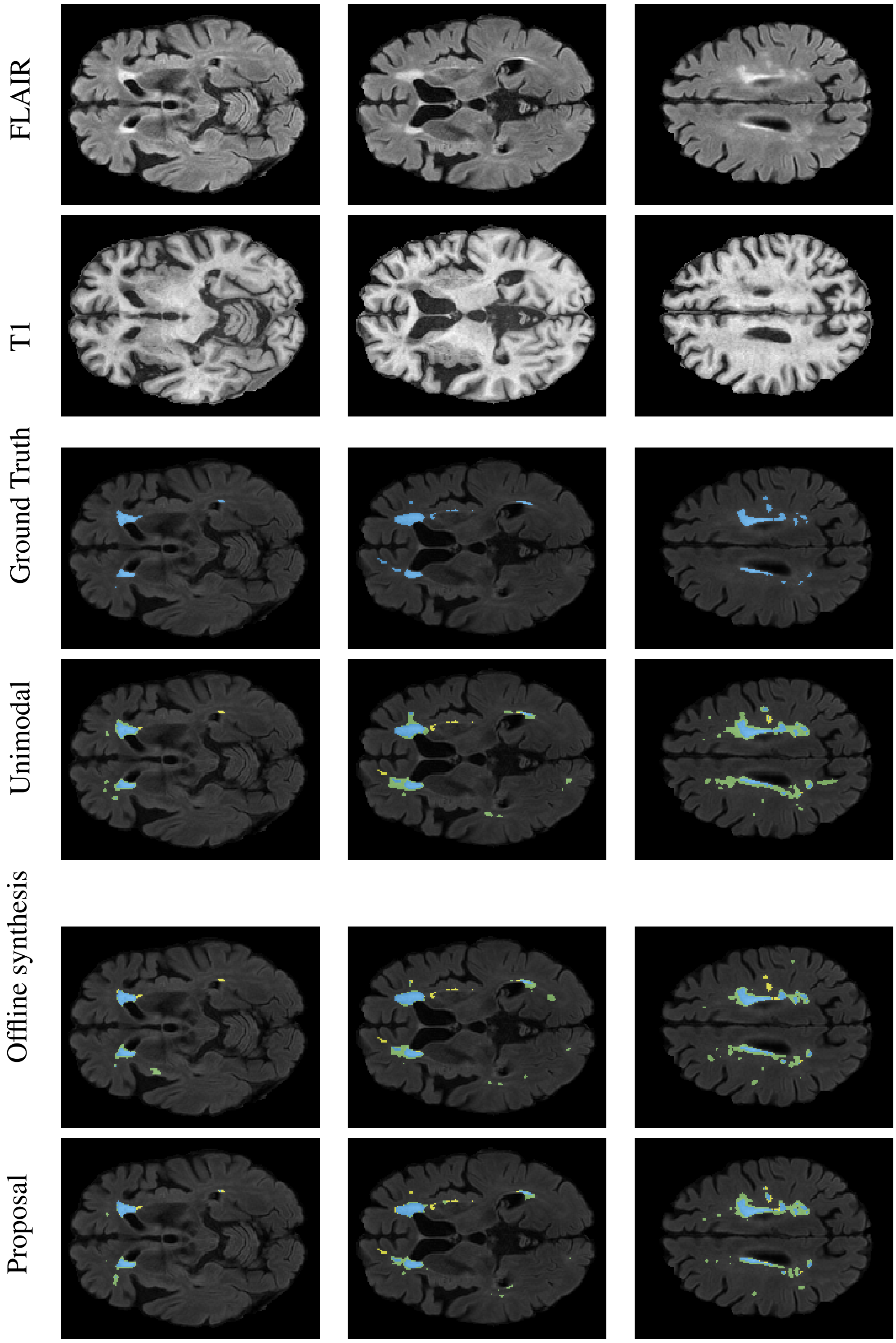

Figure 3.3: Segmentation results for all proposed methods. Each column represents a different slide in the image, blue areas are regions that were correctly labeled (Ground truth annotations were done using FLAIR images), false positives are shown in green, and false negatives in yellow . The first and second rows show the original FLAIR and T1 as references. Ground truth labels, false positives, and false negatives maps are overlaid on the synthetic FLAIR images (from the fourth to the sixth row). The third column shows that unimodal segmentation produces more false negatives, compared to offline synthesis and the proposal, especially on the border of periventricular lesions. The second and third columns show that the unimodal segmentation produces more small false lesions near cortical areas. False negatives are similar for all the compared methods.

Table 3.2: Average MAE and PSNR between real FLAIR images and the synthetic images generated for each method

	Method	
Measure	Offline synthesis	proposal
MAE	0.3153	0.2566
PSNR(DB)	9.65	11.01

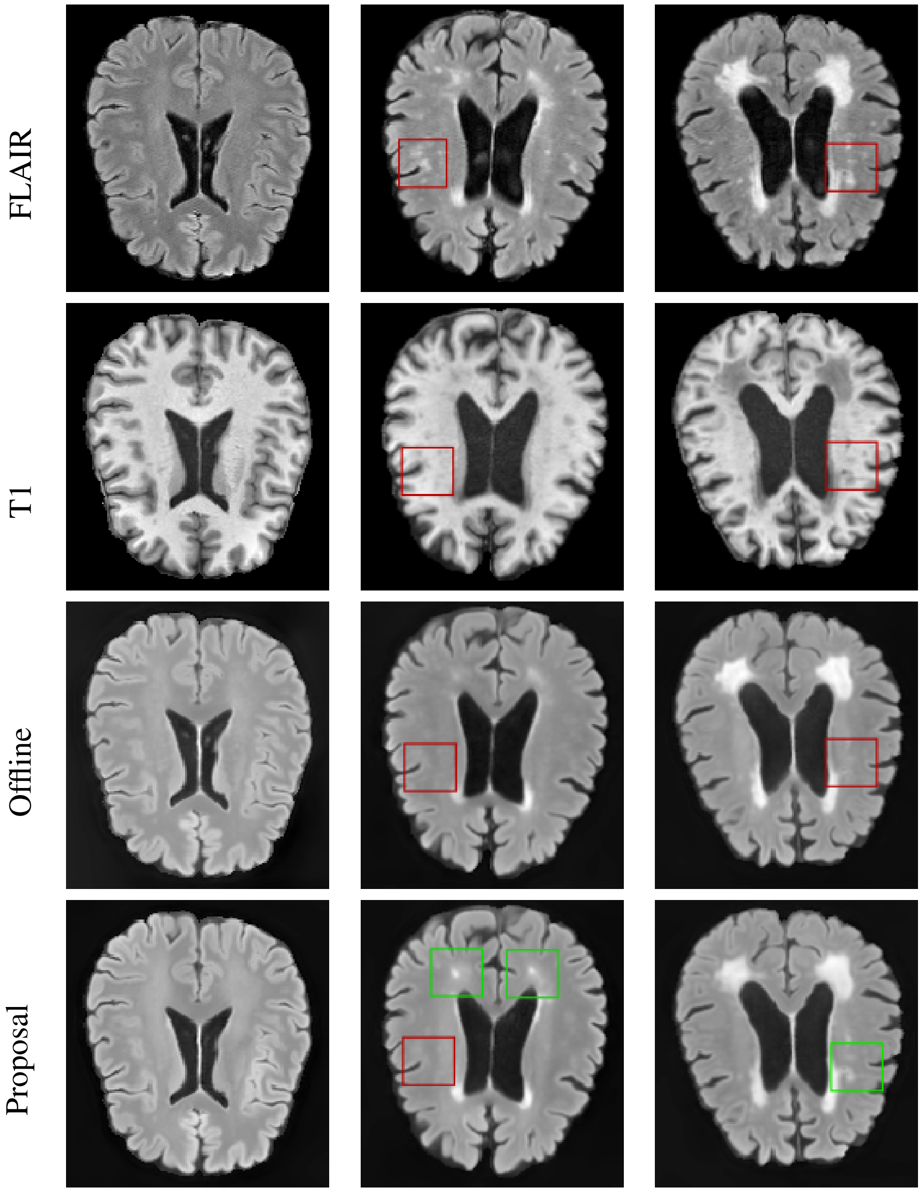

Figure 3.4: Results of Generation for all the proposed methods,

Chapter 4

Dealing with morphological variability

4.1 Derived publication:

- **Mauricio Orbes-Arteaga**, Lauge Sørensen, Jorge Cardoso, Marc Modat, Sebastien Ourselin, Stefan Sommer, Mads Nielsen, Christian Igel , and Akshay Pai "PADDIT: Probabilistic Augmentation of Data using Diffeomorphic Image Transformation", Proc. SPIE 10949, Medical Imaging 2019: Image Processing,

4.2 Preface

Brain abnormalities such WMHs present irregular morphology with large variations among patients. A segmentation model should be robust enough to be unaffected by those variations. However, learning those invariant properties depends directly on the shape variability presented on the available training samples. Publicly available WMHs databases hold a small amount of training cases, which also mean limited morphological variety.

A common yet effective methodology to increase variability on training samples is through data augmentation. New images can be generated by applying a series of transformations to the available training images. For those new samples to be useful they should meet these criteria: *i)* They should be valid, so they don't generate unrealistic morphologies. *ii)* They should be relevant so the generated samples can improve generalization. Although applying deformation is intuitively the methodology of choice to alternate anatomically samples, it can also lead to the generation of unrealistic samples.

In order to obtain a model that produces transformations that capture shape variations in training data, we propose Probabilistic Augmentation of Data using Diffeomorphic Image Transformation (PADDIT). PADDIT involves an unsupervised approach to learn shape variations that naturally appear in the training dataset. This is done by first constructing an unbiased template image that represents the central tendency of shapes in the training dataset. We sample – using a Hamiltonian Monte Carlo (HMC) scheme Duane et al. [1987], Neal [2011] – transformations that warp the training images to the generated mean template. The main advantage of PADDIT is that it encourage the generation of transformations that does not alter the anatomical plausibility of the images. That because, those transformations capture the shape variations in the training data. On the other hand, on generic augmentation the deformation follow a random pattern (e.g using random b-splines) which can result in augmented images that can not be seen in real world scenarios. The sampled transformations are used to perturb training data which is then used for augmentation. We show that DNNs trained with PADDIT outperforms DNNs trained without augmentation and with generic augmentation (using b-spline transformations) in segmenting white matter hyperintensities from T1 and FLAIR brain MRI scans.

4.3 Related work

In order to address generalization, one has to find models that generate features equivariant or invariant under different transformations of the input. Equivariance of feature maps generated by CNNs to certain transformations can be obtained by using group convolutions [Cohen and Welling, 2016] where different orientations of the features maps are learnt by kernels with shared weights. While group convolutions are very efficient due to weight sharing in learning multiple orientations for same feature maps, they are restricted to a limited set of transformations, i.e., symmetric, linear transformations. In order to reach generalization across a large group of transformations one has to rely on data augmentation. Data augmentation is commonly achieved by applying transformations that generate warped versions of the available training data. Accessing a larger group of transformations for augmentation is specially important in the field of medical image analysis because features related to the human anatomy need to maintain their identity under non-linear transformations. For instance, cortical surfaces of brain, structures with arbitrary shapes such as tumors, or structures subject to atrophy such as hippocampus in the brain have large variations in their expected morphology. The choice of the transformation in literature so far has been fairly arbitrary – often restricted to rotations, translations, reflections, and very small nonlinear deformations [Roth et al., 2015a, Jaderberg et al., 2015a, Hauberg et al., 2016]. Some degree of learning the right kind of transformations needed to improve the network performance was introduced in Jaderberg et al. [2015a]. Hauberg et al. Hauberg et al. [2016] propose to learn a particular group of transformations. The authors suggest to use the space of transformations called diffeomorphisms, which are well-behaved in the sense of being differentiable and invertible. In order to learn the kind of diffeomorphisms needed to account for all shape variations in the training data, the authors propose to measure relative shape changes by using non-linear image registration. From the resulting set of transformations, a distribution is constructed from which new transformations for augmentation are sampled using a Metropolis Markov chain Monte Carlo scheme (MMCMC). While the performance on MNIST LeCun et al. [1998] improved significantly, digits are simpler shapes compared to the more complex brain images considered in this study. Given the size of each brain image, it is computationally intensive to randomly register sufficient pairs of images. In addition, since the posterior distribution of transformations is not a trivial space, MMCMC tend to get stuck in local isolated modes of distribution. Therefore, images that can not be plausibly registered may induce transformations that are not meaningful, as those transformations will generated augmented images that do not represent the testing population. Feeding the model with augmented images that will not be seen in real life, can result in a decrease in the model performance.

To cope with the above mentioned problems we introduced and an effective and logical strategy to generate realistic deformable transformations based on shape tendencies in the training set.

4.4 Methods

Probabilistic Bayesian models for template estimation in registration was introduced by Zhang et al. [2013], albeit using a different class of transformations. In short, the method views image registration as a maximum a posteriori (MAP) problem where the similarity between two images (I_1, I_2) is the likelihood. The transformations are (lie group exponential of a time-constant velocity field $\mathbf{v}$) regularized by a prior which is in the form of a norm attached to velocity field. Formally, it is a minimization of the energy

$$E(I_1, I_2, \mathbf{v}) = \|I_1 \circ \mathrm{Exp}(\mathbf{v}) - I_2\|^2 + \lambda \|\mathbf{v}\|^2. \tag{4.1}$$

The norm on the vector field is generally induced by a differential operator $\mathcal{L}$. Under certain conditions on $\mathcal{L}$ [Younes, 2010], there exist a Hilbert space $V \subset L_2$ with norm $\|\cdot\|_V$ so that $\|\mathcal{L}\mathbf{v}\|^2 = \|\mathbf{v}\|_V^2 = \langle \mathbf{v}, \mathbf{v} \rangle_V$, where $\langle \cdot, \cdot \rangle_V$ is the inner product. In this work however, we directly choose a kernel inducing a reproducing kernel Hilbert space to parameterize the velocity field [Pai et al., 2016]. Let Ω be the spatial domain of I_1, with $x \in \Omega$ the spatial location. Let Diff(Ω) the diffeomorphic transformations $\varphi : \Omega \times \mathbb{R} \to \Omega$, and V the tanget space of Diff(Ω) containing the velocity fields $\mathbf{v}$. There exist reproducing kernels $K : \Omega \times \Omega \to \mathbb{R}^{dxd}$, which induce a norm $\|\cdot\|_W$ (W:reproducible kernel Hilbert space) which there need not be a corresponding differential operator. One advantage of reproducing kernels is that their inner product can be calculated directly as:

$$\langle a, b \rangle_W = \langle K(\cdot, x)a, K(\cdot, y)b \rangle = a^T K(x, y) b$$

For all vectors $a, b \in \mathbb{R}^d$, for all kernels centers $x, y \in \mathbb{R}^d$. The norm and inner product on this space are defined from the kernel. This space is approximate this space by $\sum_i K(\cdot, x_i) a_i$, and by linearity of the inner and the reproducing property, the norm of the kernel can be evaluated by

$$\begin{aligned}\left\|\sum_i K(\cdot,x_i)a_i\right\|^2 &= \left\langle \sum_i K(\cdot,x_i)a_i, \sum_j K(\cdot,x_j)a_j \right\rangle \\ &= \sum_{i,j} \langle K(\cdot,x_i)a_i, K(\cdot,x_j)a_j \rangle \\ &= \sum_j a_i^T K(x_i,x_j)a_j,\end{aligned}$$

Where a are the vectors attached to each spatial kernel, and (x_i,x_j) is the spatial position of each kernel.

If we chose the regularization term $||\mathbf{v}||^2$ to be a *RKHS* norm, with the vectors fields parametrized using the reproducing kernels, the evaluation of the regularization energy will be just the evaluation of the double sum ($||\mathbf{v}||^2 = \sum_j a_i^T K(x_i,x_j)a_j$). Thus, the optimal fields will be linear combinations of the reproducing kernels. Therefore, using the reproducing kernel in the parametrization ensures that the optimization of the norm takes place in a space that contains optimal solutions.

Using the $L2$ distance metric between two images (minimization of (4.1)), one can formulate template estimation as a Fréchet mean estimation problem. In other words, given a set of N images (or observations) $I_1,\ldots,I_N$, the atlas $\hat{I}$ is the minimization of the sum-of-squared distances function

$$\hat{I} = \arg\min_{I_T} \frac{1}{N}\sum_{k=1}^{N} ||I_T - I_k||^2. \tag{4.2}$$

Since (4.1) is viewed as a MAP problem, the velocity fields are considered as latent variables, i.e., $a \sim \mathcal{N}(0,K)$, a normal distribution with zero mean and covariance K derived from a kernel function. In the presence of latent variables, the template estimation is posed as an expectation maximization (EM) problem. Further, for simplicity, we assume an i.i.d. noise at each voxel, with a likelihood term (for each k^{th} observation) given by

$$p(I_k|\mathbf{v}_k, I_T, \sigma) = \frac{1}{(2\pi)^{M/2}\sigma^M}\exp\left(-\frac{||I_T - I_k \circ \text{Exp}(\mathbf{v}_k)||^2}{2\sigma^2}\right), \tag{4.3}$$

where $\theta = \{\sigma, I_T\}$ are the parameters to be estimated via MAP; σ is the noise variance, I_T is the mean template, and M is the number of voxels. Each observation can be viewed as a random variation around a mean ($I_T \circ \mathrm{Exp}(-\mathbf{v})$). The prior on the velocity field may be defined in terms of the norm as

$$p(\mathbf{v}_k) = \frac{1}{(2\pi)^{M/2}|K|^{\frac{1}{2}}} \exp\left(-\frac{\|\mathbf{v}_k\|^2}{2}\right) \tag{4.4}$$

Estimating the posterior distribution involves the marginalization of it over the latent variables as

$$p(\theta|I_k) = p(I_k|\theta)p(\theta) = \int_{\mathbf{v}} p(I_k|\mathbf{v},\theta)p(\mathbf{v})d\mathbf{v} \tag{4.5}$$

This is computationally intractable due to the dimensionality of $\mathbf{v}$. To solve this, Hamiltonian Monte Carlo (HMC) Neal [2011] is employed to sample velocity field for marginalization. The posterior distribution to draw S number of samples from is

$$\begin{aligned} \log \prod_{k=1}^{N} p(\mathbf{v}_k|I_k;\theta) &= \log \prod_{k=1}^{N} p(\theta|I_k) = \sum_{s=1}^{S} \log \prod_{k=1}^{N} p(I_k|\mathbf{v}_{ks},\theta)p(\mathbf{v}_{ks}), \\ = \sum_{s=1}^{S} &\left(-\frac{N}{2}\log|K| - \frac{1}{2}\sum_{k=1}^{N} a^T K a - \frac{MN}{2}\log\sigma - \frac{1}{2\sigma^2}\sum_{k=1}^{N} \|I_k \circ \mathrm{Exp}(\mathbf{v}_k) - I_T\|^2 \right). \end{aligned} \tag{4.6}$$

The sampled velocity fields ($\mathbf{v}_{ks}$ of the k^{th} image) are used in an EM algorithm to estimate an optimal θ. The two steps are as follows:

- **E-Step**: We draw samples from the posterior distribution (post1) using HMC with the current estimate θ_t. Given S sampled velocity fields, let v^{kj}, $j = 1,\dots,S$, denote the $j-th$ point in this sample for the $k-th$ velocity field.

The sample mean is taken to approximate the Q function.

$$\begin{aligned} Q(\theta|\theta^t) &= E_{\mathbf{v}_k|I_k,\theta_t}\left[\log\prod_{k=1}^{N} p(\mathbf{v}_k|I_k;\theta)\right] \\ &\approx \frac{1}{S}\sum_{j}^{S}\sum_{j}^{N}\log p(\mathbf{v}_{kj}|I_k;\theta) \end{aligned} \tag{4.7}$$

- **M-Step**: Update the parameters by maximizing $Q(\theta|\theta_t)$. The maximization is a close form for I_T and θ_t and is given by:

  $$\sigma^2 = \frac{1}{MNS}\sum_{s=1}^{S}\sum_{k=1}^{N}\|I_T - I_k \circ \mathrm{Exp}(\mathbf{v}_{ks})\|^2 \tag{4.8}$$

  In order to update the atlas image I, we set the derivative of the Q function approximation with respect to I to zero. The solution I gives a closed-form update.

  $$I_T = \frac{\sum_{s=1}^{S}\sum_{k=1}^{N} I_k \circ \mathrm{Exp}(\mathbf{v}_{ks})|D\mathrm{Exp}(\mathbf{v}_{ks})|}{\sum_{s=1}^{S}\sum_{k=1}^{N}|D\mathrm{Exp}(\mathbf{v}_{ks})|} \tag{4.9}$$

  Where D denotes the Jacobian matrix.

A single-scale Wendland kernel [Pai et al., 2016] is used to parameterize the velocity field and construct the covariance matrix for regularization. Once a template is estimated, the posterior distribution is sampled for a set of velocity fields for each training data. To induce more variations, the velocity fields are randomly integrated between 0 and 1. The training samples are deformed with cubic interpolation for the image, and nearest neighbor interpolation for the atlas to create the new set of synthetic data. The input (for one image as an example) to the deep-learning network will be of the form

$$\langle\langle I_n, L_n\rangle, \langle I_n \circ \mathrm{Exp}(\mathbf{v}_{n1}), L_n \circ \mathrm{Exp}(\mathbf{v}_{n1})\rangle, \ldots \langle I_n \circ \mathrm{Exp}(\mathbf{v}_{nA}), L_n \circ \mathrm{Exp}(\mathbf{v}_{nA})\rangle\rangle, \tag{4.10}$$

Where A is the number of augmentations and L_n is the label of input image I_n. Note that the label is a segmentation assigning a class to each voxel and is transformed using the same transformation accordingly. Algorithm 1 summarizes the workflow of PADDIT in pseudo-code.

Algorithm 1 PADDIT

```
Generate template using Equations (4.7) and (4.8).
for number of training epochs do
    Sample $A = 2$ velocity fields per training image using HMCNeal [2011] from the distribution (4.6).
    Integrate the sampled velocity field upto a randomly chosen time $t \leq 1$ to warp the training image and its corresponding label image.
    Extract slices from the warped images and add them to the slices extracted from the original images, see (4.10).
    Train the convolutional neural network to classify each voxel.
end for
```

4.5 Experiments and Results

We considered CNNs based on a U-net architecture in our experiments. To evaluate the proposed method, the performance of CNNs trained with data augmentation using PADDIT was compared to training without augmentation and training with augmentation using deformations based on random B-splines – we call this method the baseline. The above-mentioned strategies were applied to White Matter Hyperintensities (WMH) segmentation from FLAIR and T1 MRI scans. To this end, we use the training dataset from the 2017 WMH segmentation MICCAI challenge [1]. The set is composed of T1/FLAIR MRI scans and manual annotations for WMH from 60 subjects. Manual notations were performed in FLAIR space, therefore T1 modalities have been registered to such space. The images were also corrected for bias field inhomogeneities using SPM12. As further preprocessing images were cropped or padded to 200 × 200 × 200 voxels. Also, images were subtracted by its mean and divided by its variance, to normalized voxel intensities. The dataset was split into a training(30), validation(5) and testing(10) set. For each method two different deformed versions of each training case were created, i.e the training set size was tripled.

[1] http://wmh.isi.uu.nl

The Random deformations for the baseline were obtained by using a deformation field defined on a grid with Cp number of control points and B-spline interpolation. The size of deformation was controlled by adding Gaussian noise with 0 mean and standard deviation Sd. We evaluate the impact of Cp and Sd hyperparameters, specifically we tried: Cp $= [4 \times 4 \times 4, 8 \times 8 \times 8, 16 \times 16 \times 16]$ and Sd $= [2, 4, 6]$.

Figure 4.1 shows examples of the obtained deformed versions of a FLAIR scan from one subject from the training dataset. As can be observed, both methods generated new shapes for WMHs regions. It is worth noting, however, that images provided by PADDIT look more realistic and without drastic alterations to the Brain. In contrast, those obtained using random B-spline deformations exhibit some aberrations in cortical and ventricular structures depending on the size of the deformation used.

As we mentioned before, we split the data into validation and training set. In this case the validation set is used to tune the best configuration for the number of control points Cp and size of deformation Sd . Figure 4.2, shows the dice performance at each epoch on the validation and testing set. It is worth noting that PADDIT achieved higher accuracy than training with random B-spline deformations as well as training without augmentation. Also, it can be noted that random B-spline deformations did not provide a consistent improvement compared to the training without data augmentation.

For the final assessment of PADDIT, the validation data was used for early stopping. The final evaluation of each method is carried out on the testing set using the network configuration at the epoch where it showed the highest accuracy on the validation set. The best configuration for random deformations was achieved using Cp $= 8 \times 8 \times 8$ and Sd $= 4$. For PADDIT, the control points were placed every 8 voxels. Results for evaluation on the testing set are summarized in Table 4.1. Our proposed method PADDIT achieved ≈ 0.2 higher dice accuracy compared to the network performance without data augmentation and ≈ 0.15 compared to the baseline data augmentation approach (best configuration). (both differences where statistically significant ($p < 0.5$))

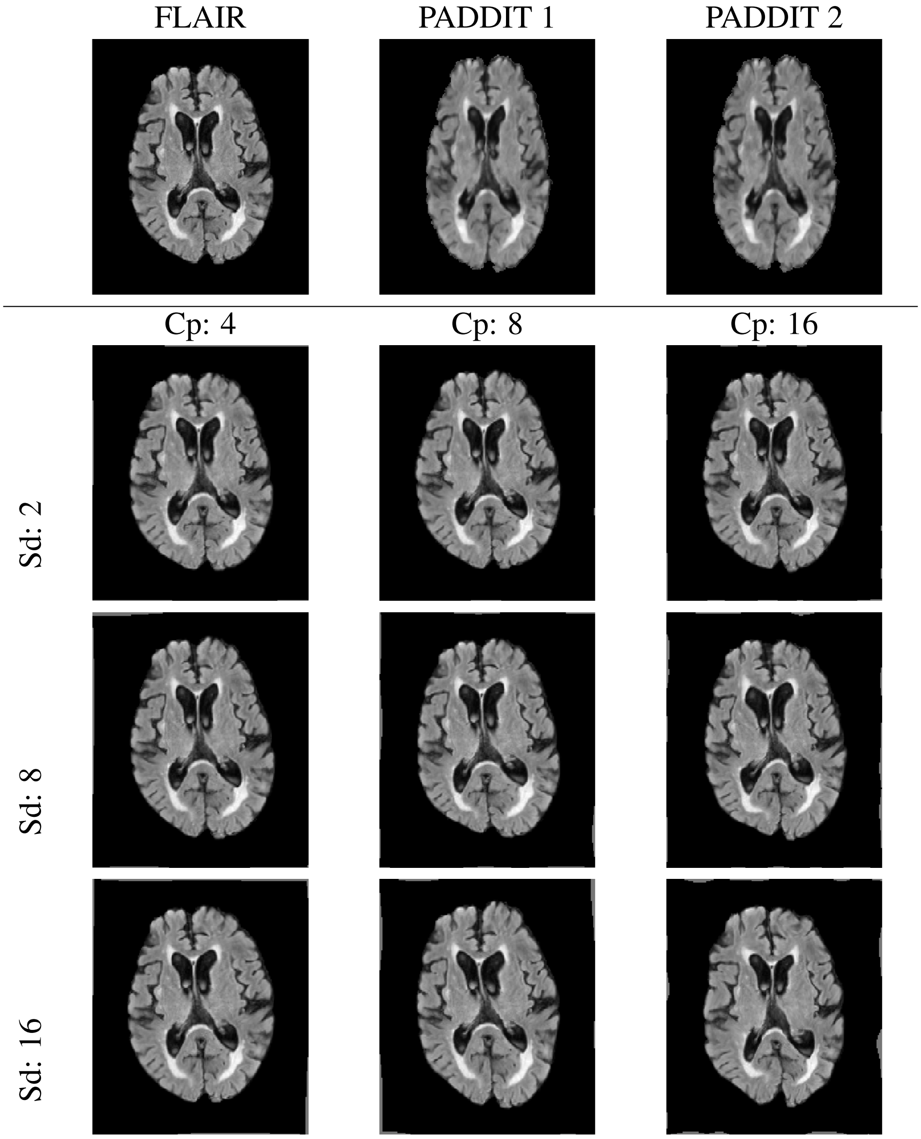

Figure 4.1: Example of generated deformations. The first row shows the original FLAIR image, and the two deformed versions using PADDIT. The remaining rows show examples of images generated by random B-spline-based deformation using different values for the number of control points (Cp), and the standard deviation (SD) which controls the amount of deformation. Thus for example the image in the right bottom corner was generated using Cp:16 and Sd:16.

4.6 Discussion and concluding remarks

In this chapter, we presented a probabilistic data augmentation approach using diffeomorphic image transformations. Contrary to traditional augmentation strategies (based on aleatory transformations), the proposed method can learn transformations that better capture the anatomical variations of the training dataset, while preserving

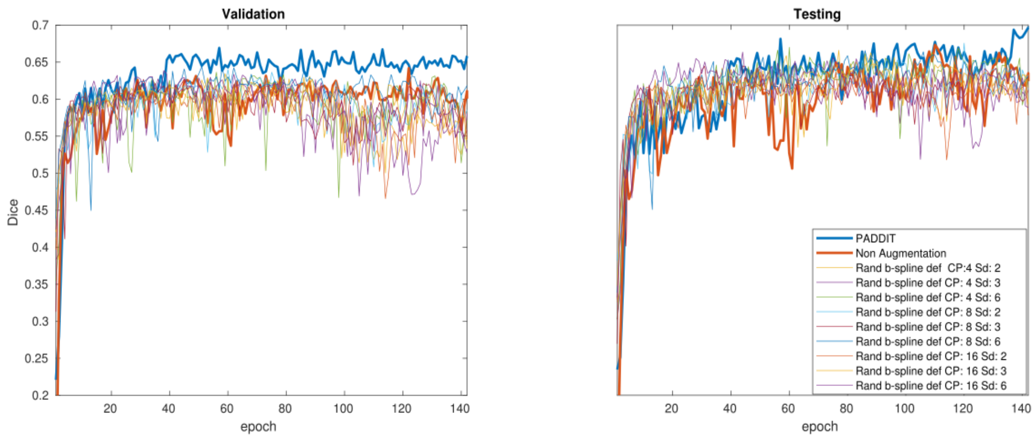

Figure 4.2: Performance on the validation and testing set for each method. Dice is computed at each epoch

	Methods	Dice (mean)	Dice (std)
	No Data Aug	66.32	24.82
Random Deformations	Cp: 4, Sd: 2	66.28	22.60
	Cp: 4, Sd4:	63.47	24.66
	Cp: 4, Sd6:	66.61	22.74
	Cp: 8, Sd2:	64.52	23.47
	Cp: 8, Sd4:	64.38	24.03
	Cp: 8, Sd6:	65.66	23.27
	Cp: 16, Sd2:	63.58	24.57
	Cp: 16, Sd4:	65.87	22.38
	Cp: 16, Sd6:	65.35	23.41
	PADDIT	**68.13**	**21.85**

Table 4.1: Segmentation accuracy for all the assessed strategies, the highest dice score achieved by the random B-spline deformation approach is underlined

the structural topology. According with the experiments, Even though several configurations of random transformations generated realistic looking images, they were not necessarily useful in DNNs training. On the other hand, the best configuration of random transformations generated images that were not necessarily biologically plausible. We hypothesize that such noisy data may help the optimization to find better minimums. However, one has to be careful in choosing the configuration of transformations since other configurations with a higher magnitude of deformations had a negative effect on the training. That because large deformations can generate images that does not represent real brain morphologies which only introduce bias into the model. This can be seen on the results on Table 4.1, where the performance of the majority of the different configurations for random b-spline deformations achieved lower performance than the model without any augmentations. In the case of PADDIT, one need not worry about the transformation configuration too much since the method learns the right transformation needed to capture the shape variations in the data set.

The proposed probabilistic augmentation approach PADDIT, proved to be an effective way to increase the training set by generating new training images which improve the segmentation performance of DNN's based approaches.

From the results, it is evident that the network trained with PADDIT is performed statistically significantly better than the networks with either no data augmentation or random B-splines based augmentation.

Chapter 5

Dealing with Inter-Modality Heterogeneity: Knowledge Distillation

5.1 Derived Publication

- **Orbes-Arteaga, M**, Cardoso, J., Sørensen, L., Igel, C., Ourselin, S., Modat, M., ... Pai, A. (2019).. "Knowledge distillation for semi-supervised domain adaptation." in Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging, MICCAI 2019.

5.2 Preface

Methods proposed in Chapter 3 and Chapter 4 leverage only annotate training data to cope with morphology and modality heterogeneity on segmentation of WMHs. However, achieving generalization to variations in image appearance due to differences in acquisition settings among center needs more powerful solutions, which also can leverage unlabeled data. Despite advancements of DNNs in segmentation, their performance always degrades when algorithms are applied on images draw from a different domains (scanner type, parameters, patient pool etc) as the one used in training the model. This performance gap is a critical barrier to the safe implementation and widespread adoption of these techniques in clinical practice. The process of adapting a model from a ‘source’ domain to a ’target’ domain is called domain adaptation. In particular unsupervised domain adaptation methodologies provide a way to learn from unlabeled data in new domains. This is particularly useful as in real practice is more feasible to have access to unlabeled data from the new domains. In this Chapter, we presented a modified knowledge distillation method for unsupervised domain adaptation. Knowledge distillation as originally proposed in [Hinton et al., 2015, Lopez-Paz et al., 2015] aims to transfer the learned knowledge from a large network (Teacher) to a simpler network (student). Instead, we aim to transfer knowledge learned on source domain data to unlabeled data from the target domain. The training strategy occurs in two phases. In the first phase, a teacher is trained in a supervised way using the source data. In the second phase, the teacher is applied to samples from source and target domains to generate soft-labels (posterior probabilistic maps for each class) which are used to train the student. The proposed scheme enables learning discriminative features from the target domain without need of ground truth labels. Moreover, soft-labels encode co-label similarities that provide more rich information in boundaries of lesions. Finally, optimization with smoother targets speeds up convergence and prevents overfitting in a better way than early stopping [Hinton et al., 2012]. Note that the knowledge is only transferred forward (from source to target domain). Our primary objective is to improve the performance on data of the target domain while keeping the performance on the source domain. However, we don’t encourage backward transfer knowledge (from target to source domain) due to the lack of labels on the target domain, which can result in the learning of inaccurate predictions that can decrease the model performance on the source data.

We present experiments on data coming from three different centers to assess the adaptation performance of the proposed method. We show that the proposed knowledge distillation based method is able to perform domain adaptation achieving higher performance than other methodologies such as adversarial approaches.

5.3 Related work

Among the recent works on domain adaptation, several methods rely on using a small amount of data (*annotated*) to fine-tune a baseline model [Hoffman et al., 2013, Karani et al., 2018]. The performance of this approach not only relies on a new – albeit small – set of annotations but also on the choice of the set. In contrast, unsupervised domain adaptation (UDA) do not use data annotations on new target domains. Adversarial training is a popular UDA method [Tzeng et al., 2017, Sun and Saenko, 2016, Hoffman et al., 2017]. In those approaches, networks are trained in such a way that the generated features are agnostic to the data domain with respect to a domain discriminator. A similar solution, ADA, was employed by Kamnitsas et al. [2017] to adapt networks to be agnostic to domain changes.

Another class of methods use knowledge distillation (KD) to transfer representations between data domains. For instance, Gupta et al. [2016] proposed using KD to transfer knowledge between different modalities of the same scene. Closely related to our work is Huang et al. [2018], where the authors propose to use omni-supervised learning (OSL) to include unlabelled data in the learning process. Here, data distillation is used to generate an ensemble of predictions from multiple transformations of unlabeled data, using a teacher model, to generate new training annotations. The proposed method differs from this method on two accounts: a) we only use soft labels to train the single student network, where the idea is to improve segmentation by learning label similarities from unannotated data b) the data included in the training of the student involves data from new domains in small amounts in contrast to OSL.

5.4 Methods

In UDA methods, we assume the source domain images and their annotations, $(x_s, y_s) \in \mathbf{X}_s$, are drawn from a distribution $p_s(x_s, y)$. The target domain images $x_t \in \mathbf{X}_t$, are drawn from a distribution $p_t(x_t, y)$ where there are no annotations available. We consider classification into K classes. In an ideal scenario, where p_s and p_t are sufficiently similar, the goal is to find a feature representation mapping f that maps an input to K scores, where the i^{th} score models (up to a constant) the logarithm of the probability that the input belongs to class K. These scores can then be mapped by $\sigma : \mathbb{R}^K \rightarrow \mathbb{R}^K$ to probability maps over the classes. UDA first finds a function f_s performing well on a source domain and then finds a new f_t based on f_s that performs well on the target domain. Vanilla supervised learning methods rely on including annotations from both $\mathbf{X}_s$ and $\mathbf{X}_t$.

In the popular ADA method, the goal is to minimize the distance between the empirical distributions of $p_s(f_s(\mathbf{X}_s)|y)$ and $p_t(f_t(\mathbf{X}_t)|y)$. Here, a discriminator D is a neural network that distinguishes between the two domains. Therefore, the discriminator acts as a discrepancy measure that brings the two distributions together. Overall, adversarial training involves train a network that generates f in a standard supervised manner that is indistinguishable by a discriminator Tzeng et al. [2017], Kamnitsas et al. [2017].

5.4.1 Knowledge distillation for Domain adaptation

KD [Hinton et al., 2015] was originally intended to compress neural networks with high number of parameters with networks of lower complexity. The objective is to teach a simpler student network to imitate a more complex trained teacher network, through a loss function called the distillation loss. To perform unsupervised domain adaptation, we proposed to use the teacher/student learning strategy. Specifically, the data from the source domain is used to train a teacher model in a supervised fashion. Then, the trained teacher is used to generate posterior probability maps or soft labels on the union of source and target data. These posterior probabilities are used instead of usual hard labels to train the student or target model. Note, this approach can take advantage of large amounts of unlabeled data acquired from any number of domains. An attractive feature of distillation loss is the soft representation of one-hot encoded label vectors which allow the student to be optimized over a smoother optimization landscape. Moreover, the smooth representation of labels also allows the learning of label similarities, which is particularly useful in learning boundaries in semantic segmentation tasks. The proposed unsupervised learning method is formulated below.

Training the teacher or source domain model: Consider a set of N manually annotate images from a domain $\mathbf{X}_s = \{(x_i, y_i), i = 1 \dots N\}$, where $x_i \in \mathbb{R}^d$ represent a d-dimensional MR scan, with $v = 1 \dots V$ voxels, and $y_i \in [0,1]^K$ with $\|y_i\|_1 = 1$ its correspondent label. Assuming there is a set F_s that holds functions $f : \mathbb{R}^d \rightarrow \mathbb{R}^K$ we aim to learn a feature representation f_s (teacher model) which follows the optimization of a loss function, l, according to Equation (5.1)

$$\arg\min_{f \in F_s} \frac{1}{N} \sum_{x_i \in \mathbf{X_s}} l(y_i, \sigma(f_s(x_i))) \tag{5.1}$$

$$[\sigma(z)]_k = \frac{\mathrm{e}^{[z]_k}}{\sum_{l=1}^{K} \mathrm{e}^{[z]_l}} \tag{5.2}$$

In a standard supervised learning way, the teacher network is optimized using the cross-entropy loss function (or any differentiable loss function of choice).

Training the student or target model: Even though f_s is suitable to segment the images from the source domain $\mathbf{X}_s$, it may not be suitable for data coming from a different data distribution $\mathbf{X}_t$. Our goal is find a function $f_t \in F_t$, which is suitable to segment data from $\mathbf{X}_t$. Assuming, we have access to a limited set of unlabeled scans in the target domain $\mathbf{X}_t = \{x_i, i = 1 \dots M\}$, we can then create a set

$$\begin{aligned} \mathbf{X}_U = \{(x_i, y_i) \,|\, x_i \in \mathbf{X}_s, y_i = f_s(x_i), 1 \leq i \leq N\} \cup \\ \{(x_i, y_i) \,|\, x_i \in \mathbf{X}_t, y_i = f_s(x_i), 1 \leq i \leq M\} \end{aligned}$$

that may be used to optimize a student using the distillation loss. Through soft-representations of this union dataset, the student is expected to learn a better mapping to the labels than the teacher network. When training the student network, we consider probability distributions (soft-labels) over the classes given by the teacher network as the learning target. This representation reflects the uncertainty of the prediction by the teacher network. The function f_t is found by (approximately) solving,

$$\arg\min_{f \in F_t} \frac{1}{(N+M)} \sum_{x_i \in \mathbf{X_U}} l(\sigma(T^{-1} f_s(x_i)), \sigma(f_t(x_i))) \ , \tag{5.3}$$

Here, $T > 1$ is the temperature parameter which controls the softness of the class probability predictions given f_t. A higher value of T produces a softer probability distribution over the classes. As pointed in [Hinton et al., 2015], for models with high confidence in the predictions, much of the information about the learned function resides in the radios of every sample probability in the soft targets. For example, for some lesions, a model will also predict a small probability of 10^{-6} of being a non-lesion, whereas for other lesions it can be the other way around. This information defines a rich similarity structure over the data (i.e inform which lesions look like non-lesions) but this has little influence on the coss-entropy cost function during the transfer because the probabilities are close to zero. A higher temperature of the final soft-max will increase the relevance of all the class probabilities on the cost function.

5.5 Experiments and Results

5.5.1 Databases

The **WMH segmentation challenge**(https://wmh.isi.uu.nl/) dataset is a public database that contains T1-weighted and FLAIR scans for 60 subjects from three different clinics. The data also consists of manual annotations of WMH from presumed vascular origin. T1-weighted images have been registered to FLAIR since annotations were performed in this space. The images were also corrected for bias field inhomogenities using SPM12. An important feature of this dataset is that the scanners and demographics have variance as show in the Table 5.1.

Table 5.1: Summary of data characteristics in the WMH challenge database

Clinic	Scanner Name	Voxel Size(m^3)	Size	# scans
Utrech	3T Philips Achieva	$0.96 \times 0.95 \times 3.00$	$240 \times 240 \times 48$	20
Singapore	3T Siemens TrioTim	$1.00 \times 1.00 \times 3.00$	$252 \times 232 \times 48$	20
Amsterdam	3T GE Signa HDxt	$1.20 \times 0.98 \times 3.00$	$132 \times 256 \times 83$	20

5.5.2 Experimental setup

One of the main objectives of the paper is to use semi-supervised learning to perform domain adaptation. We use the WMH challenge dataset to perform cross-clinical experiments in segmenting WMH on FLAIR images. We consider several scenarios to establish the performances of ADA and KD. The scenarios are described below.

Note that, to evaluate the performance of the algorithms, dice overlap measures are used throughout.

- Lower bound baseline, **L-bound**: Here a baseline DNN model is trained on the source dataset to establish a lower bound performance. The DNN is trained on the source domain images henceforth referred to as **S**, and tested on 20 subjects from a target dataset **T**.

- Upper bound baseline, **U-bound**: Here, a baseline DNN model is trained like L-Bound, however, the training dataset is a union of images from both **S** and a subset of **T** (10 subjects, with annotations). The network is evaluated on the remaining 10 subjects in **T**.

- Adversarial domain adaptation, **ADA**: Following Kamnitsas et al. [2017], we attempt at training a DNN model that is invariant to data domains. In this paper, to be consistent with KD, we train the domain discriminator based on the final layer of the baseline, in contrast to what was proposed in Kamnitsas et al. [2017]. We use a discriminator composed of 4 convolutional layers with 8, 16 32, 64 number of filters, followed by 3 fully connected layers with 64, 128 and 2 neurons. For this experiment, like U-bound, the training dataset is a union of images from both **S** and a subset of **T** (10 subjects, without annotations). The network is evaluated on the remaining 10 subjects in **T**.

- Knowledge distillation, **KD**: The experimental setup for KD is the same as ADA. A temperature of 2 is used in the softmax for the distillation loss. The student network trained is identical to the teacher network whose architecture is a standard UNet (like L-bound, U-bound, and ADA) optimized with an ADAM loss function and a learning rate of 10^{-4} with is gradual decrease after epoch 150. The network is trained for 400 epochs.

- Adaptation on-the-fly: A clinically relevant scenario is adapting to a small set of test images on the fly by keeping the teacher/baseline model constant. To validate this scenario, we apply ADA and KD on the same 10 unannotated **T** that are included in the training, but subject-wise. In other words, separate adaptation is performed on each instance of **T**, instead of including them together.

5.5.3 Results

As mentioned in section 5.2, we are mainly interested in achieving forward knowledge transfer. Consequently, we compare the performance of the above-mentioned methods based on the performance in the target domain. To this end, various combinations of mismatched (in terms of clinics) training and testing data were used. For instance, if the training data is from clinic 1 (Utrecth), the testing data is from either clinic 2 (Singapore), or clinic3 (Amsterdam). We did not test on two different clinics even though this scenario is practical. Table 5.2 illustrates mean dice coefficients (two folds) for each of the scenarios mentioned in Section 5.5.2 except for adaptation on the fly which is illustrated in Table 5.3. KD outperformed ADA in nearly all scenarios except for domain adaptation from Singapore clinic to Utrecht clinic and vice versa. For domain adaptation from Utrecht clinic to Singapore clinic, ADA was significantly better than KD. In the vice-versa situation, KD achieved a better mean which is statistically not significant. In all other scenarios, KD yielded statistically better dice overlaps compared to ADA. Note that the statistical comparison are made only between ADA and KD. In the adaptation-on-the-fly scenario, KD yields significantly better dice overlaps on a majority of the scenarios, the superior performance of ADA remains in the experiment that involves domain adaptation from Utrecht clinic to Singapore clinic. However, in the vice-versa scenario, KD performance better than ADA. To illustrate the differences in segmentations between KD and ADA, we plot the segmentations (scenario, Utrecht clinic to Amsterdam clinic) in Figure 5.1. As illustrated, both the methods perform quite well in segmenting lesions with relatively larger volume, however, the main difference is evident in segmenting smaller lesions, specially in the deep white matter regions. It is interesting to note that the adaptation-on-the-fly and the classical scenarios yield nearly the same dice indicating a good generalisability and less dependency on the choice of the small dataset coming from the target domain.

Table 5.2: Illustrates dice overlaps (with variance). Bold fond indicates statistical significance at 5%, p-values (paired-sample t-test at was used to computed p-values, which were 0.0002 $< p <$ 0.02). Only ADA and KD methods are considered in the statistical comparison.

Test Training	Method	Utrech	Singapore	Amsterdam
Utrech	L-bound		0.6126 (0.1092)	0.7207 (0.0793)
	ADA		**0.7004 (0.1057)**	0.7144 (0.0968)
	KD		0.6456 (0.0905)	**0.7548 (0.0755)**
	U-bound		0.8031 (0.1148)	0.7704 (0.0787)
Singapore	L-bound	0.6693 (0.2271)		0.7368 (0.0931)
	ADA	0.6859 (0.2036)		0.7337 (0.0912)
	KD	0.6924 (0.2103)		**0.7499 (0.0877)**
	U-bound	0.7063 (0.2016)		0.7699 (0.0851)
Amsterdam	L-bound	0.6471 (0.2086)	0.6811 (0.1172)	
	ADA	0.6800 (0.2128)	0.7202 (0.1154)	
	KD	**0.6909 (0.2135)**	**0.7482 (0.0975)**	
	U-bound	0.7208 (0.1851)	0.7988 (0.0869)	

Table 5.3: Mean dice overlaps from the adaptation-on-the-fly scenario. Bold fond indicates statistical significance at 5%, p-values (paired-sample t-test at was used to computed p-values, which were 0.0003 $< p <$ 0.04). Only ADA and KD methods are considered in the statistical comparison.

Test Training	Method	Utrech	Singapore	Amsterdam
Utrech	KD		**0.6285 (0.097**	**0.7465(0.0855)**
	ADA		0.7075 (0.095)	0.7220(0.0995)
Singapore	KD	**0.6945(0.1825)**		0.7425(0.0805)
	ADA	0.6680(0.1945)		0.7370(0.0880)
Amsterdam	KD	**0.6745 (0.2005)**	**0.7395 (0.1165)**	
	ADA	0.6625 (0.1890)	0.7100 (0.1125)	

5.6 Discussion and concluding remarks

We have evaluated a modified knowledge distillation approach and compared it to the popular adversarial approach under different clinical scenarios. Overall, the knowledge distillation approach gave better results and is relatively simpler to design when compared to the more architecture-dependent adversarial approaches. Adversarial approaches require extensive tuning of DNN architectures, especially for the discriminator, in order to achieve reasonable performances. In contrast, KD only involves choosing the temperature parameter which can be chosen only based on the performances on the source domain. One of the interesting outcomes is the inferior

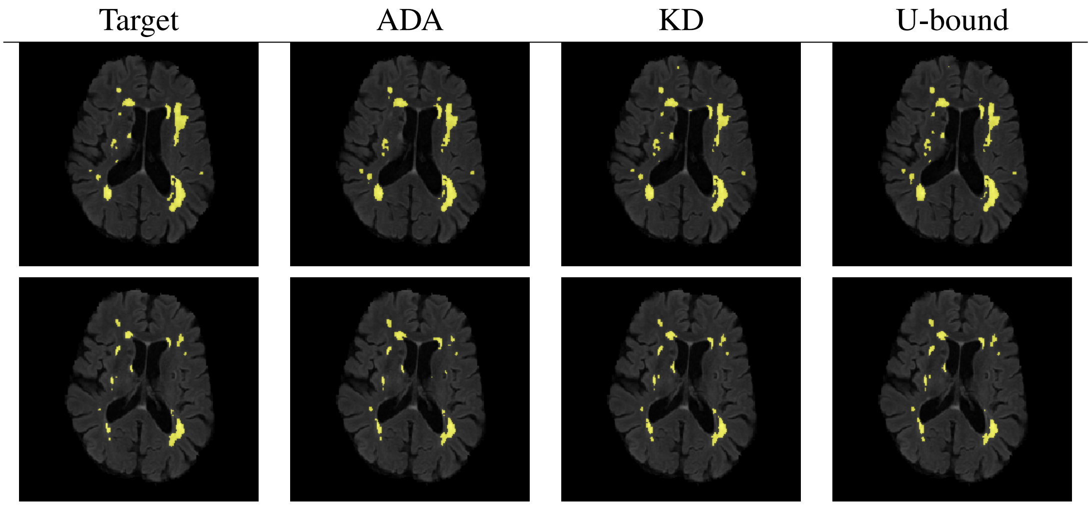

Figure 5.1: Illustration of the segmentation's obtained with different methods trained on the Utrecht dataset and tested on the Amsterdam dataset. The top and bottom row illustrate segmentations on two different subjects.

performance of KD on domain adaptation in scenario, Utrecht clinic to Singapore clinic. One of the reasons may be attributed to not just scanner differences but also differences in demographics. This may have led to an inferior teacher performance that the student network relies on. To verify this, we used the improved network from domain adaptation using ADA as a teacher and then trained a student based on it. We observed that the mean dice overlap improved from 0.65 $\rightarrow$ 0.69.

In future work, we will consider combining the adversarial approaches with knowledge distillation to improve the generalisability of DNNs across domains without the need for large annotated datasets.

Chapter 6

Dealing with Intra-Modality Heterogeneity: Paired Consistency under Data Augmentation

6.1 Derived publication

- Orbes-Arteaga, M., Varsavsky, T., Sudre, C. H., Eaton-Rosen, Z., Haddow, L. J., Sørensen, L., ... & Nachev, M. Jorge Cardoso. Multi-domain adaptation in brain MRI through paired consistency and adversarial learning. *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, MICCAI 2019.

6.2 Preface

In the previous chapter, we demonstrated that learning from soft labels improves model generalization. That improvement is mainly because soft-labels not only help to prevent over-fitting but also enable the model to learn discriminative features for the target domain. Despite those advantages, there are still some aspects that need attention. The knowledge distillation requires a teacher/student scheme, where the student learns from the teacher "soft predictions" that are obtained once before the student training. Therefore, the adaptation success will depend on the initial teacher performance on the target data without any chance of rectification during the student learning process. Moreover, the generalization properties of the learned features depend on the amount of variability on the unlabeled training set. Data from the target domain could only hold a few cases, with very little variability among them. This could prevent the model learning invariant enough features that extrapolate well to new cases. In this chapter, we refine the proposed method in Chapter 5 aiming to maximize generalization to a wider range of heterogeneities. Firstly, we dispose of the Teacher/Student learning scheme. Instead, we use consistency training to iteratively refine the model predictions on the target domain. We introduce a paired consistency loss (PC) which guides the adaptation. The proposed (PC) method enforces the output consistency between the predictions obtained on a given input and its augmented counterpart (they act as soft-labels each other). We enforce the model to produce consistent predictions for a given input and its augmented counterpart (they act as a soft-labels each other). We explore different types of augmentation functions aiming to encourage model robustness to a greater range of expected variability on test data. Specifically, we use synthetic augmentations which could be geometric (rotations, scale, shearing) but also MRI-specific augmentations like k-space and bias field. Moreover, we use paired scans (with different acquisition of the same patient) which are considered as realistic augmented samples.

We study two scenarios: First, we consider the typical case where we want to adapt to unpaired data coming only from one domain. Then, we study the case where we have access to paired data from different sequences. Because those sequences belong to two different domain distributions, this can also be seen as a multi-domain adaptation scenario.

We compared the proposed adaptation against adversarial unsupervised domain adaptation, and Mean Teacher (as in [Perone et al., 2019]) which is another self-learnig method for domain adaptation.

6.3 Methods

Let L a set of annotated data from a source domain, and U a set from unlabeled data from the target domain. The proposed training strategy for domain adaptation occurs in two phases. In the first phase, the network is trained only on labeled data until convergence. At the end of the first phase, the network can achieve a good performance on images from the source domain but its performance reduces on data from the target domain. During the second phase of the training, unlabeled data is presented in addition to the labeled data and a consistency term is added to the loss function. This consistency term is inspired from the loss proposed by Xie et al. [2019] that aims at minimizing the Kullback-Leibler divergence $\mathcal{D}_{KL}$ between the output probability distribution y when conditioned on the unlabeled input x from the set U or its augmented counterpart $\hat{x}$ drawn from $q(\hat{x}|x)$.

$$\min_{\theta} \mathcal{L}_{PC} = \mathop{\mathbb{E}}_{x \in U} \mathop{\mathbb{E}}_{\hat{x} \sim q(\hat{x}|x)} [\mathcal{D}_{KL}(p_{\tilde{\theta}}(y|x) || p_{\theta}(y|\hat{x}))] \tag{6.1}$$

6.3.1 Adapting consistency training for segmentation task:

KL divergence suits well with classification task as the one in Xie et al. [2019]. Given we are aiming a segmentation task we use dice loss as this adapts better to the particularities of the task (while preserves the soft-label nature). In this sense, we use the dice loss [Milletari et al., 2016] which has proven to be efficient on problems with high class imbalance. In the following, we denote as y_l the labeled ground truth, $\hat{y}_l$ the prediction over labeled images, $\hat{y}_u$ the prediction over unlabeled input and $\hat{y}_{\hat{u}}$ the prediction over its augmented/paired counterpart. The losses used in our framework are thus expressed as follows:

$$\mathcal{L}_S = dice(\hat{y}_l, y), \quad \mathcal{L}_{PC} = dice(\hat{y}_u, \hat{y}_{\hat{u}}), \quad \mathcal{L}_{tot} = \mathcal{L}_S + \alpha \mathcal{L}_{PC} \tag{6.2}$$

We trained networks using $\mathcal{L}_{tot}$ as specified in (6.2) and denote them as PC. These networks $f_{\theta}(h|x)$ produce a feature representation h from which $\hat{y}$ is calculated.

6.3.2 Preventing trivial solutions:

Early in our experiments, we encountered a specific degenerate solution: our network was able to produce one solution for source images (a good lesion mask) while producing a trivial result on the target domain (in this case, a mask of the foreground). Because there is not labels in the target domain to guide the network to predict lesions, the network found the prediction of large masks as a solution to minimize the dice loss (increasing consistency). This meant that there was good agreement between $\hat{y}_u$ and $\hat{y}_{\hat{u}}$ because they simply segmented foreground — ignoring the lesions altogether because it is easier achieve higher dice between larger masks than smaller ones such the lesions. This means that the network was identifying the domain of the images and using this to inform its solution: undesirable behaviour. We introduced an additional adversarial term to avoid these 'solutions'. Inspired by the domain adversarial literature (see Section 2.4.2.2) we propose an adversarial loss to minimize the amount of information about domain contained in h. We introduce a discriminator d_Ω which takes h as input and outputs a domain prediction $\hat{d}$. The adversarial loss, $\mathcal{L}_{adv}$ is given by the cross-entropy, $\mathcal{L}_{adv} = -\sum_{i=1}^{n} \mathcal{L}_{ce}^{i}(d_i, \hat{d}_i)$ where n is the number of domains , $\mathcal{L}_{ce}^{i}$ is the multi-class cross entropy loss, d is a one-hot encoded vector of the domain label and $\hat{d}$ is the model's domain prediction as in [Schoenauer-Sebag et al., 2019]. We use a gradient reversal layer as in Kamnitsas et al. [2017] in order to minimize L_{tot} whilst maximizing L_{adv}. Figure 6.1 presents the diagram of the proposed method with the combination of different losses, where β controls the strength with which the model is adapting its features whereas α controls the weights the consistency effect.

6.3.3 Augmentation:

The augmented function employed to perturb the inputs plays an important role in consistency training. By applying augmentation we are maximizing the robustness of models to wider range o variations. In Xie et al. [2019] the authors suggest various properties of augmented samples necessary for performing Unsupervised Data Augmentation. Samples should be realistic, valid (meaning they should not alter the underlying label), smooth, diverse and make use of targeted inductive biases (domain knowledge). Consequently we explore three groups of augmentation.

- **Geometric augmentation:** transforming the spacial relations in the images leads to generating a more diverse set of samples which also improves sample efficiency. Therefore, encouraging consistency to geometric augmentations can significantly improve robustness to such geometrical variations.

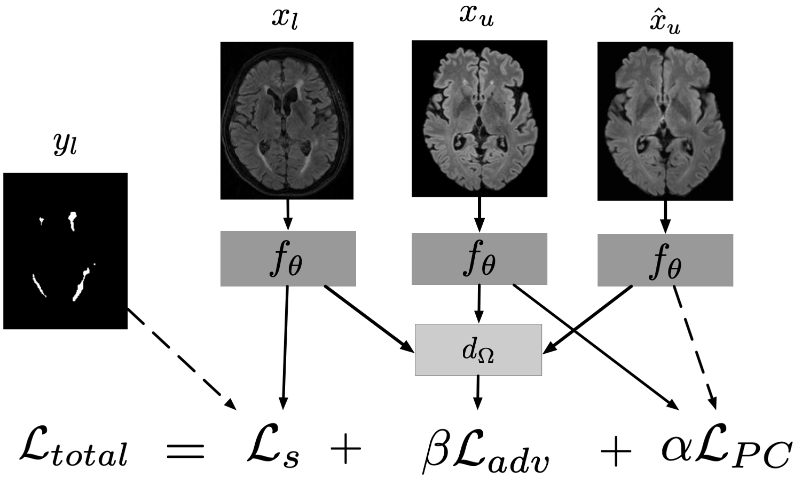
x_l
x_u
$\hat{x}_u$
y_l
f_θ
f_θ
f_θ
d_Ω
$\mathcal{L}_{total} = \mathcal{L}_s + \beta\mathcal{L}_{adv} + \alpha\mathcal{L}_{PC}$

Figure 6.1: Diagram of proposed method. At training time, x_u, x_l and y_l are supplied to the network. x_u is an image from the unlabeled target domain and $\hat{x}_u$ is the result of applying some augmentation function to x_u. A labeled image, x_l, is passed through the network, f_θ before combining with a label y_l to form the segmentation loss, $\mathcal{L}_s$. The image representations are fed to a domain discriminator d_Ω which attempts to maximise the cross-entropy between predicted domain and actual domain, $\mathcal{L}_{adv}$. Finally, similarity is promoted between the network predictions on x_u and $\hat{x}_u$ using $\mathcal{L}_{PC}$.

We sample independently geometric augmentations which include, random rotations (all axis ranging from -10 to 10 degrees), random shears ($[0.5, 0.5]$), and random scaling ($[0.75, 1.5]$) and combined them as one affine transform. We did not consider non-linear deformations as it increase the complexity of implementation when performed online, and also it could produce not realistic results. Moreover, the augmentation method PADDIT presented in Chapter 4 it does not have neither Differentiable (needed for backpropagation) or online implementation at the moment. Differentiable and online implementations will be considered as a future work.

MRI-specific augmentation: Common artefacts caused during the acquisition process of MRIs could produce errors in segmentation models. Therefore, augmenting the data with synthetically generated artefacts will increase the image appearance variability and improve the robustness to those anomalies commonly observed in MRIs. Specifically, We apply k-space motion artefact augmentation as described in Shaw et al. [2019] and bias field augmentation as implemented in Gibson et al. [2017].

- **Realistic :** Lastly we consider the case where we have access different separate acquisition of the same subject (e.g 2D and 3D acquisitions). If the samples are paired (one-to-one spatial correspondence) they can be considered as realistic augmented versions of each other. Given each sequence belongs to a different domain distribution, encouraging invariance on their predictions makes the model robust to the distribution shift of their domains. This can be also seen as a multi-domain adaptation approach, where we get the adaptation from one source to multiple target domains. However, taking the sequences as they are, makes for a discrete augmentation function with discontinuous jumps. In order to encourage continuity we also applied to each acquisition a combination of both, geometric and MRI specific augmentations.

6.4 Data

In this work we focus on white matter hyperintensity segmentation. The data comes from two separate studies. As a source domain we use the White Matter Hyperintensity challenge data presented in MICCAI 2017 Kuijf et al. [2019]. The other dataset was used as target domain and comes from a sub-study within the Pharmacokinetic and Clinical Observations in PeoPle over fiftY (POPPY) Haddow et al. [2019]. In this study two different FLAIR sequences were acquired during the same MR session for all 72 subjects on a Philips 3T scanner. The in-plane FLAIR was an axial acquisition with 3mm slice thickness and 1mm^2 planar resolution (Repetition time (TR) 8000ms, Inversion time (TI) 2400 ms and echo time (TE) 125 ms) while the volumetric FLAIR was of resolution $1.04 \times 1.04 \times 0.56$ mm^3 (TR=8000ms TI=1650ms TE=282ms). Both images were rigidly coregistered to the 1mm^3 T1 sequence acquired during the session. All individuals were male with mean age of 59.1 ± 6.9 yrs, including HIV-positive subjects and population-matched controls.

6.5 Experiments

In this section we aim to evaluate the Consistency Training method (PC) (Section 6.3.1) and the effects of the proposed augmentation based strategies (Section 6.3.3) on two domain adaptation scenarios: *i)* We consider the case where we have access to paired data with different sequences per sample. As we mentioned before, in order to avoid discontinuities, here we combine real augmentations together with the geometric and MR- specific augmentations.

ii) We consider the classical domain adaptation setting where we do not have paired data in the target domain. With this experiment we want to demonstrate the proposed method is useful to perform domain adaptation in more a commonly seen scenario. Also we want to explore the effects of each time of augmentation in the segmentation performance. So consequently we referred to the above scenarios as Adaptation with paired data and adaptation with unpaired data respectively.

6.5.1 Domain adaptation on paired data

In this set of experiments, we use as a source domain the whole MICCAI dataset, whereas for the target domain we use a set of 72 images that come from a sub-study within the Pharmacokinetic and Clinical Observations in PeoPle over fiftY (POPPY). We split the MICCAI dataset with a train:validation:test assignment of 40:10:10. For the POPPY dataset, the split was 38:15:20.

Implementation details: Training was done using 2d axial slices of size 256×256 with inference carried out by concatenating the predictions across all slices to form a 3d volume. The segmentation network uses the U-Net architecture Ronneberger et al. [2015] with depth of 4 and a maximum number of filters of 256 at the deepest layer, with ReLU as the activation function. Phase one of training is done on the dataset assigned as source. We use use the the Adam optimizer with an initial learning rate 10^{-3} and a learning rate decay schedule decaying with $\gamma = 0.1$ (γ is a multiplicative factor of learning rate decay) at epoch 300 and 350. The source validation set is used for early stopping, thus the baseline model takes the network configuration at the epoch where it showed the highest accuracy on the validation set. All adaptation models and adversarial models were initialized with the weights of this trained baseline model.

The choice of α parameter balancing the segmentation and the consistency loss in the domain adaptation runs proved to be important. Generally, high values of α led to degenerate solutions, where predictions on the target dataset were no longer capturing lesions. Since scheduling a slowly increasing α did not help, α was fixed at 0.2 in all experiments.

In case of an adversarial setting, empirical assessment of the best choice of architecture for the discriminator led to the following choice: four 2D convolutional layers with a kernel size of 3×3 and a stride of 2 followed by batch normalisation and leaky ReLU activation. The number of output channels is 4 to begin with and doubles at each layer to a total of 32. Finally, there are three fully connected layers with output sizes of 64, 32, and 2 with relu activations and dropout applied ($p = 0.5$).

Points of comparison: we compared the proposed PC with adversarial setting and augmentation (PC+Adv+Aug) to the version without adversarial setting (PC+Aug) and the simplest version removing also the augmentation (PC). In addition, we trained mean-teacher framework (MT) as well as the adversarial domain adaptation (Adv+Aug) and without augmentation (Adv). Finally, we compared to the baseline U-Net model trained only on the MICCAI dataset with (Baseline+Aug) and without (Baseline) augmentation. For all experiments we augment the data using a combination of geometric and task-specific augmentations.

For the final results table checkpoints were chosen for each of the experiments by looking at the performance across the validation set.

Reported metrics: As the first metric of consistency, we compute the Dice score overlap between the two volumes. However, high dice agreement may arise without predicting lesions, for instance with the segmentation of the foreground or of another anatomical structure. Such degenerate solutions can indeed occur as the consistency term in the loss can be minimized for any consistent prediction between volumes. As there is no lesion segmentation for the POPPY dataset, we use the known association between age and white matter hyperintensity load reported for this dataset [Haddow et al., 2019] as surrogate evaluation that the segmented elements

Table 6.1: Performance of different methods on the target (POPPY) and the source domain (MICCAI 2017 WMH Challenge). We report the dice between our models' predictions and the ground truth annotations in the source domain as well as the HD95. The evaluation on target domains is done with the Dice, the HD95, the volume difference (VD) and the recall. A significative rank measure is calculated across all metrics. Results are reported with the format median (IQR) in percentages for all metrics except the HD95 in mm. Best results are in bold andunderlined when significantly better than all others ($p<0.05$ paired Wilcoxon tests).

	POPPY				MICCAI		
	Dice	HD	VD	Recall	Dice	HD	Rank
PC+Adv+Aug	**54.5 (10.6)**	32.7 (9.8)	**15.2 (22.8)**	**52.4 (14.4)**	81.4 (9.6)	28.5 (8.6)	2.5
PC+Aug	53.2 (15.1)	39.2 (15.5)	25.4 (15.6)	43.5 (12.5)	81.6 (15.5)	18.6 (4.8)	3.3
PC	50.7 (17.0)	35.1 (11.9)	16.6 (21.4)	43.6 (11.0)	81.4 (22.6)	**17.2 (3.6)**	3.4
Perone et al 2018	48.6 (12.3)	33.6 (14.8)	33.7 (19.0)	40.9 (5.0)	80.0 (18.2)	20.0 (7.3)	4.3
Baseline+Aug	42.8 (14.6)	34.9 (11.1)	39.3 (22.3)	33.5 (12.6)	80.6 (14.8)	17.8 (4.9)	4.9
Baseline	43.0 (16.2)	33.3 (15.1)	40.3 (24.8)	33.3 (14.8)	81.1 (16.9)	17.5 (3.3)	5.6
Adv	41.8 (15.4)	**32.6 (6.1)**	25.2 (24.0)	33.5 (12.7)	**82.5 (12.0)**	17.6 (5.2)	5.7
Adv+Aug	41.4 (16.4)	36.6 (9.0)	38.0 (16.0)	33.6 (13.9)	81.9 (11.1)	19.7 (11.0)	6.3

are lesions. The effect size is a useful metric for determining whether the lesion loads predicted by the various models agree with the reported literature. For the eight compared models, the effect size ranged from 1.2-fold to 1.5-fold increase in lesion load normalized by total intracranial volume per decade. This compares well with the reported effect size on the POPPY dataset of 1.4-fold with a 95th confidence interval of $[1.0;2.0]$. Predictions from in-plane POPPY and volumetric POPPY were compared using the dice overlap, the 95th percentile Hausdorff distance measured in mm (HD95), the recall (or sensitivity), the ratio of difference in volume between the two predictions (VD) as in [Kuijf et al., 2019].

The results, gathered in Table 6.1, reporting median and interquartile range are ordered according to the average significance ranking, follows the guidelines of the MICCAI Decathlon challenge 2018 [1].

6.5.2 Domain adaptation on unpaired data:

Since MICCAI dataset contains data from 3 clinics, we take advantage of this splitting to perform cross-clinical domain adaptation. The clinic that acts as a source domain is split with a train:validation:test assignment 10:5:5 whereas the clinic that acts as the target domain is split with a train:test: 10:10. Training details are the

[1] http://medicaldecathlon.com/files/MSD-Ranking-scheme.pdf

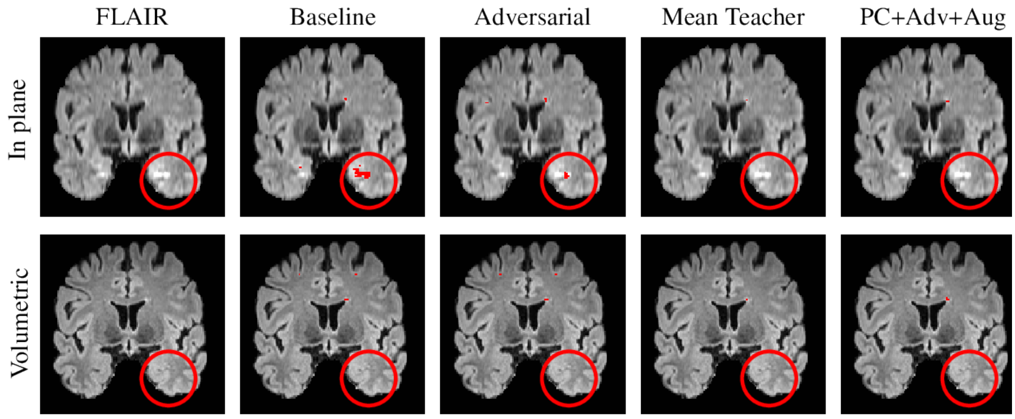

Figure 6.2: Qualitative results on a single slice from a single subject in the POPPY dataset. The top row shows a slice from the in-plane FLAIR acquisition whilst the bottom row shows a slice from the volumetric FLAIR acquisition. Each column shows a model's predictions on that row's image. This slice is used to highlight an example of an artifact (shown in the red circle) introduced by the in-plane acquisition. As can be seen in the first row, the in-plane acquisition depicts a high signal which is not a lesion. In the second column, it can be that the baseline method introduces a false positive in this region. Although adversarial domain adaptation introduces fewer false positives compared to the baseline, it still wrongly recognizes the artifact as a lesion. On the other hand, the mean teacher method and our proposal (columns fourth and fifth) are better at ignoring this artifact.

same as in 6.5.1. Note, this splited is different than the reported on Chapter 5.

Reported Metrics: As measures of comparison, we compute dice overlap, the 95$^{\text{th}}$ percentile Hausdorff distance measured in mm (HD95), the recall (or sensitivity), the ratio of difference in volume between the two predictions (VD) as was used in Kuijf et al. [2019]. In addition, we provide a single rank score comparing all methods as in Section 6.5.1. Note this ranking provides is a single summary metric that uses a per-metric non-parametric statistical significance model.

6.5.2.1 Assessing effects of different augmentation :

We evaluate performance training consistency when combined with different types of augmentations. When we use consistency training without any augmentation, this is denoted as **PC no-aug**, Training consistency combined with geometric augmentation is denoted as **PC + Geo**, when combined with task-specific augmentation it is denoted with **PC + MRI-aug**, Finally, we combined geometric and artifacts augmentation, which we refer to as **PC + all-aug.**. Table 6.2 Summarize the results

Method	Dice	HD95	VD	Recall	Rank
clinic1 → clinic2					
PC No-aug	66.89 (13.0)	6.29 (9.82)	0.43 (0.09)	52.76 (12.38)	3.625
PC Geo	70.6 (13.02)	**5.61 (7.6)**	0.37 (0.07)	57.82 (12.14)	2.75
PC + MRI-aug	73.74 (12.2)	5.67 (8.35)	0.34 (0.1)	61.67 (12.54)	2.125
PC + all-aug	**73.92 (13.58)**	6.82 (9.6)	**0.32 (0.1)**	**62.6 (13.51)**	**1.5**
clinic1 → clinic3					
PC No-aug	70.45 (11.04)	5.74 (3.44)	0.32 (0.17)	60.0 (14.45)	3.375
PC Geo	**73.14 (9.47)**	**4.82 (3.61)**	**0.27 (0.16)**	**64.04 (13.62)**	**1.375**
PC + MRI-aug	68.77 (12.19)	6.43 (4.07)	0.34 (0.17)	57.77 (15.24)	3.625
PC + all-aug	71.65 (11.01)	5.76 (4.08)	0.26 (0.18)	63.14 (14.83)	1.625
clinic2 → clinic1					
PC No-aug	62.3 (23.17)	8.51 (6.21)	1.28 (2.06)	**81.52 (14.87)**	2.625
PC Geo	61.03 (24.7)	8.09 (6.71)	1.37 (2.24)	80.38 (18.05)	3.125
PC + MRI-aug	**62.51 (23.14)**	**7.93 (6.31)**	**1.12 (1.81)**	79.61 (15.07)	**2**
PC + all-aug	61.99 (23.69)	8.04 (6.42)	1.18 (1.8)	80.55 (15.8)	2.25
clinic2 → clinic3					
PC No-aug	73.24 (10.18)	5.03 (4.79)	0.18 (0.14)	69.81 (14.97)	2.625
PC Geo	**74.94 (9.45)**	**3.95 (3.86)**	0.18 (0.12)	**73.29 (15.04)**	**1.375**
PC + MRI-aug	72.33 (10.99)	5.31 (5.18)	0.2 (0.16)	67.89 (15.85)	3.375
PC + all-aug	73.39 (10.49)	5.02 (4.7)	**0.17 (0.14)**	70.74 (15.57)	2.625
clinic3 → clinic1					
PC No-aug	61.71 (24.21)	12.69 (7.61)	1.43 (3.33)	69.85 (14.05)	2.875
PC Geo	61.74 (24.92)	13.04 (7.54)	1.52 (3.6)	**70.8 (15.06)**	2.375
PC + MRI-aug	**61.94 (22.97)**	**10.33 (6.42)**	**1.15 (2.49)**	68.99 (14.38)	2.375
PC + all-aug	61.86 (24.02)	10.57 (6.0)	1.26 (2.88)	68.98 (14.88)	2.375
clinic3 → clinic2					
PC No-aug	72.85 (11.81)	5.78 (8.04)	0.29 (0.06)	64.8 (9.28)	4
PC Geo	76.29 (11.44)	5.6 (8.11)	0.24 (0.08)	69.58 (9.96)	2.25
PC + MRI-aug	76.66 (8.91)	**4.55 (6.16)**	0.23 (0.08)	69.64 (9.66)	2.25
PC + all-aug	**76.9 (10.58)**	5.18 (7.74)	**0.21 (0.1)**	**71.12 (9.72)**	**1.5**

Table 6.2: Performance of different augmentation combinations for each cross-clinical setting, We report Dice, HD95, volume difference(VD) and Recall, that were computed between our prediction and the ground truth labels. A significance rank is calculated across all metrics. Results are reported with the format median (IQR) in percentages for all metrics except the HD95 in mm. Best results are in bold

for the aforementioned methods. We report median and interquartile range for each metric as well as the ranking for each method.

In addition, we computed the mean significance rank for each method , where the mean is computed from the six cross-clinical experiments. This, give the overall performance of each method. Results are summarized in Table 6.3

Method	Rank
PC + all-aug	1.979167
PC Geo	2.208333
PC + MRI-aug	2.625000
PC No-aug	3.187500

Table 6.3: Average Significance ranking computed for PC with different augmentations, the lower the rank score the higher the overall performance of the methods

6.5.2.2 Methods comparison

For final assessment we present result for five methods as follows:

- **No adaptation:** this is a lower bound where the model is trained only on data from the source domain and applied to data from the target domain.
- **Adversarial:** we perform adversarial domain adaptation following the guideline on Kamnitsas et al. [2017]. This is equivalent to train the model using the loss $\mathcal{L}_s + \beta\mathcal{L}_{adv}$
- **Mean Teacher (MT) :** we train a classical self-learning based domain adaptation method with a mean-teacher framework following the guidelines in Perone et al. [2019]. For fair comparison, we complemented mean-teacher with geometric and task specific augmentations.
- **PC + adv + all-aug (PC + AD):** Our framework for supervised domain adaptation which combines paired consistency and adversarial learning, augmentations are also a combination of geometric and task specific augmentations.
- **Supervised:** We use images and labels from the target domain to fine-tune a pre-trained model on source data. This is an upper bound as we expect higher performance when included annotations from the target domain.

Performance results along with the respective significance ranking scores are shown in Table 6.4. The mean significance ranking for each method across all domain adaptation experiments is shown in Table 6.5, which summarizes the overall performance across all cross-domain experiments. From results on Table 6.4 we can see that

PC ADV outperforms (lower rank) the remaining unsupervised domain adaptation methods on three out of six cross-clinical experiments (clinic1 → clinic2, clinic2 → clinic3, and clinic3 → clinic2), whereas Mean teacher perform the best on two(clinic2 → to clinic1, and clinic3 → to clinic1. Adversarial learning only performs the best on clinic1 → clinic3. Table 6.5 confirms that on average PC-ADV perform the best among the unsupervised domain adaptation methods and that means teacher outperform adversarial domain adaptation. Nonetheless, all domain adaptation methods outperformed the model without domain adaptation.

6.6 Discussion and remarks

In this work, we presented a novel method of performing unsupervised domain-adaptation which takes advantage of different augmentation functions to improve consistency training approaches. A pre-trained model on the source domain is retrained to encourage consistent predictions on two augmented versions of the same input.

The proposed approach was evaluated against existing unsupervised domain adaptation strategies including representation learning approaches using domain adversarial training [Kamnitsas et al., 2017], and the 'Mean Teacher' algorithm for unsupervised domain adaptation [Xie et al., 2019] as well as a supervised baseline for WMH segmentation.

Overall, our method was able to produce more consistent predictions across target domains while retaining similar performance on its original training domain.

More specifically, adaptation techniques optimizing pairwise consistency not only outperformed baseline models not benefiting from any adaptation but also other domain adaptation strategies.

Furthermore, it appeared that the PC method while closest to the mean teacher algorithm, outperformed this approach potentially thanks to differences in the optimisation strategies. Understanding the reasons for these differences also reported by Xie et al. [2019] could be an interesting avenue of future investigation.

Method	Dice	HD95	VD	Recall	Rank
clinic1 → clinic2					
No adaptation	67.61 (14.02)	7.57 (8.16)	0.37 (0.1)	55.14 (13.22)	5.375
Mean Teacher	68.34 (15.71)	7.79 (9.28)	**0.28 (0.14)**	58.35 (13.07)	4.125
Adversarial	71.41 (13.25)	**5.5 (7.66)**	0.35 (0.08)	58.99 (12.08)	4.625
PC ADV	**75.72 (9.84)**	3.81 (3.46)	0.33 (0.1)	**63.73 (10.87)**	**2.625**
Supervised	82.2 (8.88)	3.01 (5.16)	0.16 (0.11)	75.99 (11.31)	1
clinic1 → clinic3					
No adaptation	66.77 (13.1)	10.41 (8.04)	0.25 (0.15)	62.46 (16.12)	5
Mean Teacher	67.58 (11.89)	8.98 (6.28)	0.24 (0.16)	65.34 (14.67)	4.5
Adversarial	**72.41 (11.48)**	6.02 (5.31)	**0.21 (0.15)**	**66.35 (15.51)**	**2**
PC ADV	72.3 (11.58)	**5.24 (4.32)**	0.25 (0.17)	64.19 (15.86)	3.75
Supervised	74.98 (9.27)	3.89 (2.61)	0.22 (0.16)	67.16 (13.17)	1.375
clinic2 → clinic1					
No adaptation	59.27 (24.28)	9.51 (6.74)	1.6 (2.73)	80.07 (16.65)	4.5
Mean Teacher	58.9 (22.05)	8.8 (5.96)	1.49 (1.95)	**84.47 (12.84)**	3.5
Adversarial	65.53 (12.16)	**7.78 (4.75)**	**0.47 (0.6)**	72.37 (16.63)	4.25
PC ADV	**65.68 (19.11)**	8.21 (5.43)	0.6 (0.73)	75.74 (14.78)	3.625
Supervised	65.88 (21.68)	7.99 (5.97)	0.65 (0.9)	76.88 (15.1)	2.25
clinic2 → clinic3					
No adaptation	68.67 (12.19)	8.66 (8.95)	0.22 (0.16)	64.93 (16.4)	4.75
Mean Teacher	69.31 (11.94)	8.69 (6.2)	0.22 (0.21)	**74.27 (14.92)**	3.75
Adversarial	67.59 (12.06)	5.75 (4.59)	0.2 (0.15)	64.18 (17.33)	4.25
PC ADV	72.05 (9.63)	**4.75 (4.67)**	**0.17 (0.11)**	72.76 (15.3)	**2.375**
Supervised	74.02 (9.45)	4.85 (4.74)	0.19 (0.15)	67.52 (12.85)	3
clinic3 → clinic1					
No adaptation	58.56 (26.45)	10.89 (6.55)	1.32 (3.03)	0.21 (0.1)	3.75
Mean Teacher	**62.3 (23.47)**	10.16 (5.62)	1.28 (2.79)	71.04 (13.24)	3.75
Adversarial	61.77 (22.95)	**9.0 (6.14)**	**1.08 (2.48)**	64.14 (13.99)	3.875
PC ADV	62.17 (22.89)	9.16 (5.94)	1.08 (2.48)	64.26 (13.52)	4.25
U-bound	67.99 (19.0)	7.4 (5.9)	0.32 (0.33)	71.8 (14.62)	1.125
clinic3 → clinic2					
No adaptation	72.36 (9.69)	5.69 (7.41)	0.3 (0.08)	62.88 (9.25)	4.625
Mean Teacher	73.37 (11.82)	7.03 (8.98)	0.26 (0.08)	66.1 (9.96)	3.875
Adversarial	64.55 (10.5)	7.14 (8.65)	0.38 (0.19)	53.33 (11.39)	6
PC ADV	74.55 (8.56)	6.93 (9.0)	0.26 (0.1)	66.32 (7.38)	3.5
Supervised	82.86 (7.39)	2.4 (2.08)	0.15 (0.08)	78.41 (10.16)	1

Table 6.4: We report Dice, HD95, volume difference(VD) and Recall, that were computed between our prediction and the ground truth labels. A significance rank is calculated across all metrics.Results are reported with the format median (IQR) in percentages for all metrics except the HD95 in mm. Best results are in bold.

Method	Rank
Supervised	1.625000
PC ADV	3.354167
Mean Teacher	3.916667
Adversarial	4.166667
No adaptation	4.666667

Table 6.5: Average Significance ranking for each method, the mean is computed from the six cross clinical experiments, the lower the rank score the higher the overall performance of the methods

The type of augmentations demonstrated to play an important role in consistency training. A significant improvement in performance can be observed when combining training consistency with any type of augmentation as it promoted a good label distribution in our target images. According to the mean significance ranking in Table 6.3, the best performance is achieved when combining the maximum amount of augmentations. However, according to results in Table 6.2 the rankings vary in each coss-clinical experiment. It remains unclear what makes one type of augmentation perform better on certain scenarios. Future work will focus on identifying which type of augmentation to apply based only on the available unlabeled data.

Regarding the adversarial results, the observed inferior performance on paired experiments suggests that depending on the adaptation problem, the learning of a latent space invariant to domain (as enforced in the adversarial approach) may cause an information loss detrimental to the segmentation task. In addition, the multi-domain setting requires to train a discriminator in a multi-label classification task, which makes it hard to find the point of equilibrium that determines if the discriminator is being fooled.

In conclusion, PC is a promising method to adapt automated image segmentation tools to different scanner manufacturers, MR sequences, and other confounds. This adaptation is critical to the clinical translation of these tools notably in the context of scanner upgrades and multi-center trials.

Chapter 7

Conclusions

7.1 Summary

The aim of this thesis was the development of methodologies to increase the robustness and generalization of deep learning segmentation models for white matter hyperintensities. The need for robust and generalizable segmentation models stems from the desire to provide solutions that can be integrated into real-world scenarios.

In Chapter 3, We developed a simultaneous synthesis and segmentation method that aims to improve segmentation performance under more common and simplistic image acquisition settings. The proposal was used to produce WMHs segmentation and synthesis of FLAIR images from T1-w only scans. We demonstrated the joint optimization strategy yields a more accurate segmentation but also is able to generate more realistic synthetic FLAIR images. We acknowledge that improvement to the regularization effect induced by the synthesis in the segmentation optimization process.

In Chapter 4, we introduced PADDIT, a data augmentation methodology to increase morphology diversity on training sets. The proposed method aids optimization, forcing models to learn shape invariant features that can generalize better to morphological variations. The main strength of PADDIT is the ability to learn how morphological variations occur in the training set and use this knowledge to produce more realistic augmentations. We demonstrated that DNN trained with PADDIT produces more accurate segmentation compared to other augmentation strategies.

To cope with more complex variations as those coming from differences in acquisition protocols, parameters, or scanner, we proposed domain adaptation strategies that are able to leverage unlabeled data. In Chapter 5, we present a knowledge distillation based algorithm to transfer knowledge from labeled to unlabeled data. We demonstrated how a teacher/student learning scheme optimized on soft-labels, enables learning discriminative features on the target domain data without needing ground truth.

This idea was extended further in Chapter 6 where we incorporate data augmentation and consistency training in order to increase robustness to geometric, MR-artificats as well as implicit differences among paired scans from different acquisitions. We show that optimization guided by consistency loss enables models to learn from their own predictions without the need for a teacher/student scheme. Moreover, we demonstrated that a combination of different augmentation operations led to an improvement in segmentation performance. Finally, the main advantage of methods in Chapter 5 and Chapter 6 relies on the simplicity to take advantage of unlabeled data to carry out the adaptation. This feature favors their impact in real-world scenarios.

7.2 Limitations and future directions

In this section, we describe main limitations of the presented methods, and provide some research directions in order to tackle these limitations.

7.2.1 Multi-modal learning from unpaired data

The jointly synthesis and segmentation learning proposed in Chapter 3 proved to be effective in learning from available multi-modal information in order to improve segmentation accuracy on one of the modalities (the less involved or more commonly used one). The limitation of this method is the need for paired data (for each sample different images are acquired and co-registered across modalities) to work. This limitation prevents the proposed method to take advantage of data which do not fulfill this condition.

One direction to overcome this is to explore synthesis with cycleGAN [Zhu et al., 2017], which has been successful in achieving translation on unpaired data. However, including cycleGAN adds complexity to the optimization process as it requires the additional learning of two discriminators. Another direction to take could be learning surrogate tasks [Noroozi and Favaro, 2016, Zhang et al., 2016, Gidaris et al., 2018] with combined data with different modalities. It has been demonstrated that by learning a surrogate task the model will be forced to learn a shared representation for both tasks [Tajbakhsh et al., 2019], which also can be seen as a way of regularization.

Providing solutions that work with unpaired data will will enable us to leverage existing large datasets with different modalities.

7.2.2 Learning deformable augmentations online

Despite PADDIT can provide anatomically meaningful augmentations, the main limitation of the current implementation is that the augmented images are computed offline. Consequently, the size of the augmented set has to be predefined before training and it could be conditioned by storage requirements.

As a feature work, we would like to develop an online implementation for PADDIT where different augmentations can be applied to each batch of images in each iteration during the network learning. This will maximize the amount of variability that the network can see during training.

One of the features that can be implemented is deformable registration thorough spatial transformer networks [Jaderberg et al., 2015b]. With the addition of this feature, we don't have to worry about complex deformable registration algorithms that are time consuming. Moreover, spatial transformer networks can be accelerated by GPU which would potentially assist the adoption of PADDIT in several DNNs segmentation frameworks.

7.2.3 Towards domain generalization

Proposed methodologies in Chapter 5 and Chapter 6 were effective leveraging unlabeled data for domain adaptation. Although getting unlabeled data from the center is feasible, the development of methodologies that can generalize to different domains without any knowledge about domain statistics could have a bigger impact on clinical practice.

The challenge is how to learn models agnostic to domains signal, but without affect the discriminative power of learned features.

Although there is not too much work on domain generalization for medical imaging applications, some directions for future research can take inspiration from domain generalization work in other visual applications. In this sense, three different avenues can be considered. *i) Data augmentation:* previous works such us [Shankar et al., 2018, Volpi et al., 2018] proved to be effective for domain generalization on visual applications based on some way of data augmentation. In addition domain generalization has also been benefited from self-learning [Carlucci et al., 2019]. Therefore one direction to follow is extending data augmentation strategies previously developed in Chapter 4 and Chapter 6 in order to simulate domain variability through data perturbation. The advantage of data augmentation is that it can be easily integrated into a self-learning scheme. *ii) Harmonizing of feature spaces.* Alignment of multi-domain distributions can help to produce domain invariant features. However the explicit harmonization is carried out on the training domains, therefore it would be of great value to explore how to extend harmonization to unseen domains. *iii) Meta learning:* meta-training and meta-test procedures using the available multi-domain training data can be used to expose the optimization to domain-shift making it robust to this phenomenon.

7.2.4 Application to large-scale real clinical datasets and challenges

The methods proposed in this thesis can be potentially benefited from large-scale real datasets. That is because the unsupervised nature of the methods enables to leverage a big amount of data even when it does not have manual annotations (which is usually the case for large sets of medical data). Large-scale datasets usually encode a large amount of variability or even can hold different domains (e.g demographic subgroups, or images acquired with different parameters). Domain adaptation strategies based on self-learning as proposed in Chapter 5, and Chapter 6 can generate pseudo labels on those data enabling extracting relevant information from those domains. This will increase model generalization capabilities enabling to deploy models directly on unseen target domain data. Avoiding the single-source, single-target setting is of clinical significance as models can be directly applied to different clinics without the need for retraining.

However, there are additional challenges in the presence of large sets. The self-learning approaches depends on the model performance to generate reliable pseudo labels on the unseen data. Therefore it is difficult to guarantee that models can have an initial satisfactory performance these data. That is mainly due to the limited amount of currently available labeled training data, which can not be able to cope with all the variability presented on large sets. Because of the 3D volumetric structure of medical images, it is also complicated to take advantage of a large set of natural 2D images such ImageNet to improve model initial performance. That is because 3D images contain rich structural information that is significant to representing medical images. Getting a significant amount of annotations is cumbersome. Therefore, current models need to adapt to take maximum advantage of the very limited amount of labeled data, which is always a complicated task.

Bibliography

Faiza Admiraal-Behloul, DMJ Van Den Heuvel, Hans Olofsen, Matthias JP van Osch, Jeroen van der Grond, Mark A van Buchem, and Johan HC Reiber. Fully automatic segmentation of white matter hyperintensities in mr images of the elderly. *Neuroimage*, 28(3):607–617, 2005.

Varghese Alex, Kiran Vaidhya, Subramaniam Thirunavukkarasu, Chandrasekharan Kesavadas, and Ganapathy Krishnamurthi. Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. *Journal of Medical Imaging*, 4(4):041311, 2017.

Petronella Anbeek, Koen L Vincken, Matthias JP Van Osch, Robertus HC Bisschops, and Jeroen Van Der Grond. Probabilistic segmentation of white matter lesions in mr imaging. *NeuroImage*, 21(3):1037–1044, 2004.

M Anitha, P Tamije Selvy, and V Palanisamy. Automated detection of white matter lesions in mri brain images using spatio-fuzzy and spatio-possibilistic clustering models. *Computer Science & Engineering*, 2(2):1, 2012.

Edward A Ashton, Chihiro Takahashi, Michel J Berg, Andrew Goodman, Saara Totterman, and Sven Ekholm. Accuracy and reproducibility of manual and semi-automated quantification of ms lesions by mri. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 17(3):300–308, 2003.

Richard Beare, Velandai Srikanth, Jian Chen, Thanh G Phan, Jennifer Stapleton, Rebecca Lipshut, and David Reutens. Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities. *Neuroimage*, 47(1):199–203, 2009.

Avi Ben-Cohen, Eyal Klang, Stephen P Raskin, Shelly Soffer, Simona Ben-Haim, Eli Konen, Michal Marianne Amitai, and Hayit Greenspan. Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection. *arXiv preprint arXiv:1802.07846*, 2018.

James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984.

Abdel-Ouahab Boudraa, Sidi Mohammed Réda Dehak, Yue-Min Zhu, Chahin Pachai, Yong-Gang Bao, and Jérôme Grimaud. Automated segmentation of multiple sclerosis lesions in multispectral mr imaging using fuzzy clustering. *Computers in biology and medicine*, 30(1):23–40, 2000.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

Gulzar Hameed Chagla, Reed F Busse, Ryan Sydnor, Howard A Rowley, and Patrick A Turski. Three-dimensional fluid attenuated inversion recovery imaging with isotropic resolution and nonselective adiabatic inversion provides improved three-dimensional visualization and cerebrospinal fluid suppression compared to two-dimensional flair at 3 tesla. *Investigative radiology*, 43(8):547, 2008.

Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. In *International conference on information processing in medical imaging*, pages 29–41. Springer, 2019.

Cheng Chen, Qi Dou, Hao Chen, and Pheng-Ann Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation.

In *International workshop on machine learning in medical imaging*, pages 143–151. Springer, 2018.

Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. *Proceedings of the International Conference on Machine Learning*, 48, 2016.

Charles M Deber and Steven J Reynolds. Central nervous system myelin: structure, function, and pathology. *Clinical biochemistry*, 24(2):113–134, 1991.

Konstantin Dmitriev and Arie E Kaufman. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9501–9511, 2019.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In *annual conference on medical image understanding and analysis*, pages 506–517. Springer, 2017.

Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In *International conference on medical image computing and computer-assisted intervention*, pages 544–552. Springer, 2018.

Simon Duane, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222, 1987.

M Filippi, MA Horsfield, S Bressi, V Martinelli, C Baratti, P Reganati, A Campi, DH Miller, and G Comi. Intra-and inter-observer agreement of brain mri lesion volume measurements in multiple sclerosis: a comparison of techniques. *Brain*, 118(6):1593–1600, 1995.

Chichen Fu, David Joon Ho, Shuo Han, Paul Salama, Kenneth W Dunn, and Edward J Delp. Nuclei segmentation of fluorescence microscopy images using convolutional neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 704–708. IEEE, 2017.

Chichen Fu, Soonam Lee, David Joon Ho, Shuo Han, Paul Salama, Kenneth W Dunn, and Edward J Delp. Three dimensional fluorescence microscopy image synthesis and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2221–2229, 2018.

Swetava Ganguli, Pedro Garzon, and Noa Glaser. Geogan: A conditional gan with reconstruction and style loss to generate standard layer of maps from satellite images. *arXiv preprint arXiv:1902.05611*, 2019.

Jingjing Gao, Chunming Li, Chaolu Feng, Mei Xie, Yilong Yin, and Christos Davatzikos. Non-locally regularized segmentation of multiple sclerosis lesion from multi-channel mri data. *Magnetic resonance imaging*, 32(8):1058–1066, 2014.

Eli Gibson et al. NiftyNet: a deep-learning platform for medical imaging. *arXiv preprint arXiv:1709.03485*, 2017.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

John T Guibas, Tejpal S Virdi, and Peter S Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

Lewis J Haddow et al. Magnetic resonance imaging of cerebral small vessel disease in men living with HIV and HIV-negative men aged 50 and above. *AIDS research and human retroviruses*, 2019.

Sven Haller, Enikö Kövari, François R Herrmann, Victor Cuvinciuc, Ann-Marie Tomm, Gilbert B Zulian, Karl-Olof Lovblad, Panteleimon Giannakopoulos, and Constantin Bouras. Do brain t2/flair white matter hyperintensities correspond to myelin loss in normal aging? a radiologic-neuropathologic correlation study. *Acta neuropathologica communications*, 1(1):1–7, 2013.

Ahmed Harouni, Alexandros Karargyris, Mohammadreza Negahdar, David Beymer, and Tanveer Syeda-Mahmood. Universal multi-modal deep network for classification and segmentation of medical images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 872–876. IEEE, 2018.

Søren Hauberg, Oren Freifeld, Anders Boesen Lindbo Larsen, John W. Fisher III, and Lars Kai Hansen. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics (AISTATS)*, 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29 (6):82–97, 2012.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*, 2013.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

Ruobing Huang, J Alison Noble, and Ana IL Namburete. Omni-supervised learning: Scaling up to large unlabelled medical datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 572–580. Springer, 2018.

Yawen Huang, Ling Shao, and Alejandro F Frangi. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6070–6079, 2017.

Yuankai Huo, Zhoubing Xu, Shunxing Bao, Albert Assad, Richard G Abramson, and Bennett A Landman. Adversarial synthesis learning enables segmentation without target modality ground truth. *arXiv preprint arXiv:1712.07695*, 2017.

Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015a. URL `http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf`.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015b.

Dakai Jin, Ziyue Xu, Youbao Tang, Adam P Harrison, and Daniel J Mollura. Ct-realistic lung nodule simulation from 3d conditional generative adversarial networks for robust lung segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 732–740. Springer, 2018.

Shingo Kakeda, Yukunori Korogi, Yasuhiro Hiai, Norihiro Ohnari, Toru Sato, and Toshinori Hirai. Pitfalls of 3d flair brain imaging: a prospective comparison with 2d flair. *Academic radiology*, 19(10):1225–1232, 2012.

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 *International conference on information processing in medical imaging*, pages 597–609. Springer, 2017.

Neerav Karani et al. A lifelong learning approach to brain mr segmentation across scanners and protocols. *arXiv:1805.10170*, 2018.

Rasoul Khayati, Mansur Vafadust, Farzad Towhidkhah, and Massood Nabavi. Fully automatic segmentation of multiple sclerosis lesions in brain mr flair images using adaptive mixtures method and markov random field model. *Computers in biology and medicine*, 38(3):379–390, 2008.

Ki Woong Kim, James R MacFall, and Martha E Payne. Classification of white matter lesions on magnetic resonance imaging in elderly persons. *Biological psychiatry*, 64(4):273–280, 2008.

Hugo J Kuijf et al. Standardized assessment of automatic segmentation of white matter hyperintensities; results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 2019.

Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior guided gan based semantic inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13696–13705, 2020.

Zhiqiang Lao, Dinggang Shen, Dengfeng Liu, Abbas F Jawad, Elias R Melhem, Lenore J Launer, R Nick Bryan, and Christos Davatzikos. Computer-assisted segmentation of white matter lesions in 3d mr images using support vector machine. *Academic radiology*, 15(3):300–313, 2008.

Eleftherios Lavdas, Ioannis Tsougos, Stella Kogia, Georgios Gratsias, Patricia Svolos, Violetta Roka, Ioannis V Fezoulidis, and Eftychia Kapsalaki. T2 flair artifacts at 3-t brain magnetic resonance imaging. *Clinical imaging*, 38(2):85–90, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *NeuroImage*, 183:650–665, 2018.

Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Chunwei Ma, Zhanghexuan Ji, and Mingchen Gao. Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 128–136. Springer, 2019.

Pauline Maillard, Evan Fletcher, Samuel N Lockhart, Alexandra E Roach, Bruce Reed, Dan Mungas, Charles DeCarli, and Owen T Carmichael. White matter hyperintensities and their penumbra lie along a continuum of injury in the aging brain. *Stroke*, 45(6):1721–1726, 2014.

Paul Malloy, Stephen Correia, Glenn Stebbins, and David H Laidlaw. Neuroimaging of white matter in aging and dementia. *The Clinical Neuropsychologist*, 21(1): 73–109, 2007.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

R.M. Neal. *MCMC using Hamiltonian dynamics*. Handbook of Markov Chain Monte Carlo, 2011.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.

Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12): 2720–2730, 2018.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

Erik Olsson, Niklas Klasson, Josef Berge, Carl Eckerström, Åke Edman, Helge Malmgren, and Anders Wallin. White matter lesion assessment in patients with cognitive impairment and healthy controls: reliability comparisons between visual rating, a manual, and an automatic volumetrical mri method—the gothenburg mci study. *Journal of Aging Research*, 2013, 2013.

Akshay Pai, Stefan Sommer, Lauge Sørensen, Sune Darkner, Jon Sporring, and Mads Nielsen. Kernel bundle diffeomorphic image registration using stationary velocity fields and wendland basis functions. *IEEE Transactions on Medical Imaging*, 35 (6), 2016.

Bhavini Patel and Hugh S Markus. Magnetic resonance imaging in cerebral small vessel disease and its use as a surrogate disease marker. *International Journal of Stroke*, 6(1):47–59, 2011.

Christian S Perone et al. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 2019.

Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 246–253. Springer, 2013.

Xiangxiang Qin. Transfer learning with edge attention for prostate mri segmentation. *arXiv preprint arXiv:1912.09847*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Holger R Roth, Christopher T Lee, Hoo-Chang Shin, Ari Seff, Lauren Kim, Jianhua Yao, Le Lu, and Ronald M Summers. Anatomy-specific classification of medical images using deep convolutional nets. In *IEEE Symposium on Biomedical Imaging*, 2015a.

Holger R Roth, Le Lu, Jiamin Liu, Jianhua Yao, Ari Seff, Kevin Cherry, Lauren Kim, and Ronald M Summers. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE transactions on medical imaging*, 35(5):1170–1181, 2015b.

Marco Rovaris, Giancarlo Comi, Maria A Rocca, Mara Cercignani, Bruno Colombo, Giuseppe Santuccio, and Massimo Filippi. Relevance of hypointense lesions on fast fluid-attenuated inversion recovery mr images as a marker of disease severity in cases of multiple sclerosis. *American journal of neuroradiology*, 20(5):813–820, 1999.

Philip Scheltens, Timo Erkinjunti, Didier Leys, Lars-Olaf Wahlund, Domenico Inzitari, Theodoro del Ser, Florence Pasquier, Frederik Barkhof, Riita Mäntylä, John Bowler, et al. White matter changes on ct and mri: an overview of visual rating scales. *European neurology*, 39(2):80–89, 1998.

Jeremy D Schmahmann, Eric E Smith, Florian S Eichler, and Christopher M Filley. Cerebral white matter: neuroanatomy, clinical neurology, and neurobehavioral correlates. *Annals of the New York Academy of Sciences*, 1142:266, 2008.

Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage*, 59(4):3774–3783, 2012.

Alice Schoenauer-Sebag et al. Multi-domain adversarial learning. *arXiv preprint arXiv:1903.09239*, 2019.

Christopher Schwarz, Evan Fletcher, Charles DeCarli, and Owen Carmichael. Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without flair. In *International Conference on Information Processing in Medical Imaging*, pages 239–251. Springer, 2009.

Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169, 2016.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

Richard Shaw et al. MRI k-space motion artefact augmentation: Model robustness and task-specific uncertainty. In *MIDL*, pages 427–436, 2019.

Lin Shi, Defeng Wang, Shangping Liu, Yuehua Pu, Yilong Wang, Winnie CW Chu, Anil T Ahuja, and Yongjun Wang. Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction. *Journal of neuroscience methods*, 213(1):138–146, 2013.

Navid Shiee, Pierre-Louis Bazin, Arzu Ozturk, Daniel S Reich, Peter A Calabresi, and Dzung L Pham. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535, 2010.

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski.

Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*, pages 1–11. Springer, 2018.

Rita Simões, Christoph Mönninghoff, Martha Dlugaj, Christian Weimar, Isabel Wanke, Anne-Marie van Cappellen van Walsum, and Cornelis Slump. Automatic segmentation of cerebral white matter hyperintensities using only 3d flair images. *Magnetic resonance imaging*, 31(7):1182–1189, 2013.

Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.

Nima Tajbakhsh, Yufei Hu, Junli Cao, Xingjian Yan, Yi Xiao, Yong Lu, Jianming Liang, Demetri Terzopoulos, and Xiaowei Ding. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1251–1255. IEEE, 2019.

Youbao Tang, Yuxing Tang, Jing Xiao, and Ronald M Summers. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. *arXiv preprint arXiv:1904.09229*, 2019.

Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems*, pages 2794–2803, 2017.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

Detlev Uhlenbrock and S Sehlen. The value of t1-weighted images in the differentiation between ms, white matter lesions, and subcortical arteriosclerotic encephalopathy (sae). *Neuroradiology*, 31(3):203–212, 1989.

ECW van Straaten, F Fazekas, E Rostrup, P Scheltens, R Schmidt, L Pantoni, D Inzitari, G Waldemar, T Erkinjuntti, R Mäntylä, et al. Impact of white matter hyperintensities scoring method on correlations with clinical data. *MRI correlates of vascular cerebral lesions and cognitive impairment*, 37:51, 2007.

Gijs van Tulder and Marleen de Bruijne. Why does synthesized data improve multi-sequence classification? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–538. Springer, 2015.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pages 5334–5344, 2018.

Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 102–110. Springer, 2019a.

Shujun Wang, Lequan Yu, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE transactions on medical imaging*, 38(11):2485–2495, 2019b.

Yanbo Wang, Joseree Ann Catindig, Saima Hilal, Hock Wei Soon, Eric Ting, Tien Yin Wong, Narayanaswamy Venketasubramanian, Christopher Chen, and Anqi Qiu. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. *Neuroimage*, 60(4):2379–2388, 2012.

Joanna M Wardlaw, Colin Smith, and Martin Dichgans. Mechanisms of sporadic cerebral small vessel disease: insights from neuroimaging. *The Lancet Neurology*, 12(5):483–497, 2013a.

Joanna M Wardlaw, Eric E Smith, Geert J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838, 2013b.

Stephen B Wharton, Julie E Simpson, Carol Brayne, and Paul G Ince. Age-associated white matter lesions: The mrc c ognitive f unction and a geing s tudy. *Brain pathology*, 25(1):35–43, 2015.

Minjie Wu, Caterina Rosano, Meryl Butters, Ellen Whyte, Megan Nable, Ryan Crooks, Carolyn C Meltzer, Charles F Reynolds III, and Howard J Aizenstein. A fully automated method for quantifying and localizing white matter hyperintensities on mr images. *Psychiatry Research: Neuroimaging*, 148(2-3):133–142, 2006.

Qizhe Xie et al. Unsupervised data augmentation. *arXiv:1904.12848*, 2019.

Guang Yang, Jun Lv, Yutong Chen, Jiahao Huang, and Jin Zhu. Generative adversarial networks (gan) powered fast magnetic resonance imaging–mini review, comparison and perspectives. *arXiv preprint arXiv:2105.01800*, 2021.

Xin Yang, Haoran Dou, Ran Li, Xu Wang, Cheng Bian, Shengli Li, Dong Ni, and Pheng-Ann Heng. Generalizing deep models for ultrasound image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 497–505. Springer, 2018.

Mitsuhiro Yoshita, Evan Fletcher, and Charles DeCarli. Current concepts of analysis of cerebral white matter hyperintensities on magnetic resonance imaging. *Topics in magnetic resonance imaging: TMRI*, 16(6):399, 2005.

Laurent Younes. *Shapes and diffeomorphisms*, volume 171. Springer, 2010.

Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8280–8289, 2018.

Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

Miaomiao Zhang, Nikhil Singh, and Thomas Fletcher. Bayesian estimation of regularization and atlas building in diffeomorphic image registration. *In Proceedngs of Information Processing in Medical Imaging*, 2013.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

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.

Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8543–8553, 2019.

Can Zhao, Aaron Carass, Junghoon Lee, Yufan He, and Jerry L Prince. Whole brain segmentation and labeling from ct using synthetic mr images. In *International Workshop on Machine Learning in Medical Imaging*, pages 291–298. Springer, 2017.

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.