POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Deep learning methods for MRI spinal cord gray matter segmentation

CHRISTIAN SAMUEL PERONE

Institut de génie biomédical

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie biomédical

Mars 2019

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Deep learning methods for MRI spinal cord gray matter segmentation

présenté par **Christian SAMUEL PERONE**

en vue de l'obtention du diplôme de Maîtrise ès sciences appliquées

a été dûment accepté par le jury d'examen constitué de :

Pierre BELLEC, président
Julien COHEN-ADAD, membre et directeur de recherche
Nikola STIKOV, membre

DEDICATION

This thesis is dedicated to my wife Thais and my family.

"A rationalist is simply someone for whom it is more important to learn than to be proved right; someone who is willing to learn from others - not by simply taking over another's opinions, but by gladly allowing others to criticize his ideas and by gladly criticizing the ideas of others."

— Karl Popper, seminar paper given at Alpach on 25 August 1958.

ACKNOWLEDGEMENTS

I'm thankful to my research advisor Prof. Julien Cohen-Adad who gave me the opportunity to join NeuroPoly lab and for his diligent guidance during all this period on this interdisciplinary field of neuroscience and medical imaging. It was a pleasure to learn countless concepts and I easily became an admirer of his hard work towards a greater good.

I would like to thank Prof. Nikola Stikov and Prof. Pierre Bellec for gently accepting be part of the committee of my thesis defense and reading my work. Your criticism will be greatly appreciated.

I'm also very grateful to Ryan Topfer for all the extensive revisions and the time invested into his careful analysis of all the articles presented in this work.

I'm also very thankful to Prof. Rodrigo C. Barros, Prof. Evandro Luis Viapiana, Roberto Silveira and Thomas Paula for all the recommendations and incentive before coming to Montreal.

I would also like to acknowledge the financial support from IVADO, RBIQ/QBIN and for the Polytechnique Montreal financial exemptions. These fundings were paramount for this research.

This entire adventure wouldn't be possible without the support from all my colleagues of NeuroPoly who were very kind and helpful to me since the day I arrived. I would like to thank Aldo, Maxime, and Charley for the very fruitful collaboration and discussions on Machine Learning. I would also like to thank Harris, Atef for helping me with the abstract translation and Agah, Tommy, and Francisco for all the friendly help during my masters at NeuroPoly. I'm also thankful to Pedro Ballester for the amazing collaboration on domain adaptation and solid friendship.

I would also like to thank Nikola for the help with the organization of journal club, it was an amazing experience.

It is also important to mention all friends from the NeuroPoly lab that helped me during this masters as well: Alexandru, Nicolas, Souad, Gabriel, George, Nibardo, Stephanie, Sara, Dominique, Benjamin, Tanguy, Tung, Alexandros, François, Mathieu, Jerôme, Ariane, Mélanie and Oumayma.

RÉSUMÉ

La moelle épinière humaine, qui fait partie du système nerveux central, est la principale voie responsable de la connexion du cerveau et du système nerveux périphérique. On sait que la matière grise présente dans la moelle épinière est associée à de nombreux troubles neurologiques tels que la sclérose en plaques et la sclérose latérale amyotrophique.

L'IRM est souvent utilisée pour étudier les maladies neurologiques et surveiller leur évolution. À cette fin, la morphométrie extraite de la substance grise de la moelle épinière, telle que le volume de la substance grise, peut être utilisée pour identifier et comprendre les modifications tissulaires associées aux troubles neurologiques comme ceux mentionnés précédemment.

Pour extraire des mesures morphométriques de la matière grise de la moelle épinière, une annotation (label) par voxel est requise pour chaque tranche du volume IRM. L'annotation manuelle ne peut donc pas être facilement implémenté dans la pratique en raison non seulement des efforts fastidieux nécessaires pour annoter manuellement chaque tranche d'un volume d'IRM, mais aussi du désaccord et des biais introduits par différents annotateurs humains.

Toutefois, il existe de nombreuses méthodes semi-automatiques ou entièrement automatiques pour annoter chaque voxel, mais la plupart d'entre elles sont composées d'approches en plusieurs étapes pouvant propager des erreurs dans le pipeline, s'appuient sur des dictionnaires de données ou ne généralisent pas bien lorsqu'il y a des changements anatomiques. Il est bien connu que les techniques modernes basées sur l'apprentissage par la représentation et l'apprentissage en profondeur ont obtenu d'excellents résultats dans un large éventail de tâches allant de la vision par ordinateur à l'imagerie médicale.

Le programme de recherche de ce projet consiste à améliorer les résultats les plus récents des méthodes existantes au moyen de techniques modernes d'apprentissage en profondeur grâce à la conception, la mise en œuvre et l'évaluation de ces méthodes pour la segmentation de la substance grise de la moelle épinière. Dans ce projet, trois techniques principales ont été développées: en open source, comme décrit ci-dessous.

La première technique consistait à concevoir une architecture d'apprentissage en profondeur pour segmenter la matière grise de la moelle épinière et a permis d'obtenir de meilleures résultats comparé à six autres méthodes développées précédemment pour la segmentation de la matière grise. Cette technique a également permis de segmenter un volume *ex vivo* avec plus de 4000 tranches en fournissant au préalable et moins de 30 échantillons annotés du même volume.

La deuxième technique a été développée pour tirer profitnon seulement des données anotées, mais aussides données qui ne le sont pas (données non anotées) au moyen d'une méthode d'apprentissage semi-supervisée étendue aux tâches de segmentation. Cette méthode a apporté des améliorations significatives dans un scénario réaliste sous un régime de données réduit en ajoutant des données non annotées au cours du processus de formation du modèle.

La troisième technique développée est une méthode d'adaptation de domaine non supervisée pour la segmentation. Dans ce travail, nous avons abordé le problème du décalage de distribution présent sur les données IRM, qui est principalement causé par différents paramètres d'acquisition. Dans ce travail, nous avons montré qu'en adaptant le modèle à un domaine cible présenté au modèle sous forme de données non annotées, il est possible d'améliorer de manière significative la segmentation de la matière grise pour le domaine cible invisible.

Conformément aux principes de la science ouverte pour tous (open science), nous avons ouvert toutes les méthodes sur des référentiels publics et en avons implémenté certaines sur la Spinal Cord Toolbox (SCT) [1], une bibliothèque complète et ouverte d'outils d'analyse pour l'IRM de la moelle épinière. Nous avons également utilisé uniquement des ensembles de données accessibles au public pour toutes les évaluations et la formation de modèles, ainsi que pour la publication de tous les articles sur les revues en libre accès, avec une disponibilité gratuite sur les serveurs d'archives pré-imprimées.

Dans ce travail, nous avons pu constater que les modèles d'apprentissage en profondeur peuvent en effet fournir des progrès considérables par rapport aux méthodes précédemment développées. Les méthodes d'apprentissage en profondeur sont très flexibles et robustes. Elles permettent d'apprendre de bout en bout l'ensemble des pipelines de segmentation tout en permettant de tirer profit de données non annotées pour améliorer les performances du même domaine dans un scénario d'apprentissage semi-supervisé ou en tirant parti de données non étiquetées pour améliorer les performances des modèles dans des domaines cibles non vus.

Il est également clair que l'apprentissage en profondeur n'est pas une panacée pour l'imagerie médicale. De nombreux problèmes demeurent en suspens, tels que le décalage de généralisation toujours présent lors de l'utilisation de ces modèles sur des domaines non vus. Un futur axe de recherche inclut le développement en cours de techniques pour éclairer les modèles d'apprentissage automatique avec paramétrisation d'acquisition IRM afin par exemple d'améliorer la généralisation du modèle à différents contrastes, ainsi que d'améliorer la variabilité inhérente de ces images due aux différentes machines et aux changements anatomiques. L'estimation de l'incertitude liée à la distillation des connaissances au cours des phases de formation des approches décrites dans ce travail constitue un autre domaine de recherche

[1] disponible à https://github.com/neuropoly/spinalcordtoolbox.

potentiel. Cependant, les mesures d'incertitude font partie d'un domaine de recherche en cours d'évolution dans le Deep Learning. En effet la plupart des méthodes fournissant une approximation médiocre ou une sous-estimation de l'incertitude épistémique présente dans ces modèles.

L'imagerie médicale reste un domaine très difficile pour les modèles d'apprentissage automatique en raison des fortes hypothèses d'identité distributionnelle formulées par les algorithmes d'apprentissage statistique ainsi que de la difficulté à incorporer de nouveaux biais inductifs dans ces modèles pour tirer parti de la symétrie, de l'invariance de rotation, entre autres. Néanmoins, avec la quantité croissante de données disponibles, elles offrent de grandes promesses et gagnent lentement en robustesse pour pouvoir entrer dans la pratique clinique.

ABSTRACT

The human spinal cord, part of the Central Nervous System (CNS), is the main pathway responsible for the connection of brain and peripheral nervous system. The gray matter present in the spinal cord is known to be associated with many neurological disorders such as multiple sclerosis and amyotrophic lateral sclerosis.

Magnetic Resonance Imaging (MRI) is often used to study diseases and monitor the disease burden/progression during the course of the disease. To that goal, morphometrics extracted from the spinal cord gray matter such as gray matter volume can be used to identify and understand tissue changes that are associated with the aforementioned neurological disorders.

To extract morphometrics from the spinal cord gray matter, a voxel-wise annotation is required for each slice of the MRI volume. Manual annotation becomes prohibitive in practice due to the time-consuming efforts required to manually annotate each slice of an MRI volume voxel-wise, not to mention the disagreement and bias introduced by different human annotators.

Many semi-automatic or fully-automatic methods exist but most of them are composed by multi-stage approaches that can propagate errors in the pipeline, rely on data dictionaries, or doesn't generalize well when there are anatomical changes. It is well-known that modern techniques based on representation learning and Deep Learning achieved excellent results in a wide range of tasks from computer vision and medical imaging as well.

The research agenda of this project is to advance the state-of-the-art results of previous methods by means of modern Deep Learning techniques through the design, implementation, and evaluation of these methods for the spinal cord gray matter segmentation. In this project, three main techniques were developed an open-sourced, as described below.

The first technique is the design of a Deep Learning architecture to segment the spinal cord gray matter that achieved state-of-the-art results when evaluated by a third-party system and compared to other 6 independently developed methods for gray matter segmentation. This technique also allowed to segment an *ex vivo* volume with more than 4000 slices by just providing less than 30 annotated samples from the same volume.

The second technique was developed to take leverage not only of labeled data but also from unlabeled data by means of a semi-supervised learning method that was extended to segmentation tasks. This method achieved significant improvements in a realistic scenario under a small data regime by adding unlabeled data during the model training process.

The third developed technique is an unsupervised domain adaptation method for segmentation.

In this work, we addressed the problem of the distributional shift present on MRI data that is mostly caused by different acquisition parametrization. In this work, we showed that by adapting the model to a target domain, presented to the model as unlabeled data, it is possible to achieve significant improvements on the gray matter segmentation for the unseen target domain.

Following the open science principles, we open-sourced all the methods on public repositories and implemented some of them on the Spinal Cord Toolbox (SCT) [2], a comprehensive and open-source library of analysis tools for MRI of the spinal cord. We also used only public available datasets for all evaluations and model training, and also published all articles on open-access journals with free availability on pre-print archive servers as well.

In this work, we were able to see that Deep Learning models can indeed provide huge steps forward when compared to the previously developed methods. Deep Learning methods are very flexible and robust, allowing end-to-end learning of entire segmentation pipelines while being able to take leverage of unlabeled data to improve the performance for the same domain on a semi-supervised learning scenario, or by taking leverage of unlabeled data to improve the performance of models in unseen target domains.

It is also clear that Deep Learning is not a panacea for medical imaging. Many problems remain open, such as the generalization gap that is still present when using these models on unseen domains. A future line of research includes the on-going development of techniques to inform machine learning models with MRI acquisition parametrization to improve the generalization of the model to different contrasts, to the inherent variability of these images due to different machine vendors and anatomical changes, to name a few. Another potential area of research is the uncertainty estimation for knowledge distillation during training phases of the approaches described in this work. However, uncertainty measures are still an open area of research in Deep Learning with most methods providing a poor approximation or under-estimation of the epistemic uncertainty present in these models.

Medical imaging is still a very challenging field for machine learning models due to the strong assumptions of distributional identity made by statistical learning algorithms as well as the difficulty to incorporate new inductive biases into these models to take leverage of symmetry, rotation invariance, among others. Nevertheless, with the amount of data availability growing, they show great promises and are slowly gaining robustness enough to be able to enter in clinical practice.

[2] Available at https://github.com/neuropoly/spinalcordtoolbox.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

CNS	Central Nervous System
MRI	Magnetic Resonance Imaging
OSI	Open Systems Interconnection
CSA	Cross-Sectional Area
EDSS	Expanded Disability Status Scale
SCGM	Spinal Cord Gray Matter
NLP	Natural Language Processing
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
IID	Independend and Identically Distributed
ERM	Empirical Risk Minimization
CPG	Central Pattern Generator
ALS	Amyotrophic Lateral Sclerosis
MS	Multiple Sclerosis
ERM	Empirical Risk Minimization
GAN	Generative Adversarial Networks
MLP	Multi-Layer Perceptron
SGD	Stochastic Gradient Descent
FCN	Fully Convolutional Network
ASPP	Atrous Spatial Pyramid Pooling

CHAPTER 1 INTRODUCTION

Neuroscientists usually divide the CNS into brain and spinal cord. The human spinal cord, responsible for connecting the brain and peripheral nervous system, is nearly as thick as an adult's little finger and has two basic types of nervous tissues: *gray matter* and *white matter* [5]. It is known, from histopathological studies [6], that tissue changes on the Spinal Cord Gray Matter (SCGM) and white matter are related to a wide spectrum of neurological conditions.

Non-invasive imaging techniques such as MRI, that takes leverage of the nuclear magnetic resonance phenomenon through the use of strong magnetic fields and magnetic field gradients to provide spatial signal location, are usually employed to assess the spinal cord tissues such as the aforementioned gray matter and white matter.

In the last two decades, many semi-automated segmentation methods have been proposed for the estimation of the cord Cross-Sectional Area (CSA), however, individual gray matter tissue analysis cannot be individually assessed using only the CSA [7]. Given the significance of the spinal cord gray matter tissue analysis, which was found to be the strongest predictor of the Expanded Disability Status Scale (EDSS) in multiple sclerosis among many other metrics such as brain gray matter, brain white matter, FLAIR lesion load, T1-lesion load, and other metrics [8], the segmentation of the spinal cord gray matter became of greater importance due to its clinical relevance.

The manual annotation of the spinal cord gray matter is, however, very time-consuming even for a trained expert. The main properties that make the SCGM area difficult to segment are: inconsistent intensities of the surrounding tissues, image artifacts and pathology-induced changes in the image contrast [9]. There are also many other factors contributing to the complexity of the task, such as disagreement between different annotators, bias introduced by different annotators, different voxel sizes, lack of standardization protocols, among others. Therefore, a fully-automated procedure for the SCGM segmentation is paramount to provide means for studies that require automatic metrics extraction, tissue analysis, lesion detection, disorder detection, among others.

Recently, many methods have been proposed for the spinal cord gray matter segmentation [2,7, 10–16]. The scientific community, including our laboratory, recently organized a collaboration effort called "Spinal Cord Gray Matter Segmentation Challenge" (SCGM Challenge) [2], to assess the state-of-the-art and compare six independently developed methods on a public dataset created through the collaboration of four internationally recognized research groups

(University College London, Polytechnique Montreal, University of Zurich and Vanderbilt University), providing a ground basis for method comparison that was previously unfeasible due to the lack of standardized datasets.

In the past few years, we were able to witness the unprecedented pace to which Deep Learning [17] methods had evolved. Since the seminal work of the AlexNet [18], the research community embraced the successful Deep Learning methods and developed many state-of-the-art techniques that became pervasive across a wide range of tasks such as image classification [19], semantic segmentation [20], speech recognition [21] and Natural Language Processing (NLP) [22], to name a few.

A recent survey [23] that analyzed more than 300 papers from the field of medical imaging, showed that Deep Learning techniques became pervasive in the entire field of medical image analysis, with a rapid increase in the number of publications between the years of 2015 and 2016. The survey also found that Convolutional Neural Network (CNN) were more prevalent in the medical image analysis, with Recurrent Neural Network (RNN) gaining more popularity.

Although the large success of Deep Learning has attracted a lot of attention from the community, it is clear that Deep Learning also poses some unique challenges such as high sample complexity, meaning that the amount of labeled data that these techniques usually require to train a reasonable classifier is very high. Another challenge that is still open is how to handle the *domain shift* that is present in many domains and especially in medical imaging due to the variability of protocols, acquisition devices, and human anatomy. Machine Learning techniques that follow the Empirical Risk Minimization (ERM) principle, often show a poor generalization performance when a trained model is evaluated on data from a different distribution, mainly because of the strong Independend and Identically Distributed (IID) assumption held by the ERM learning principle.

The first goal of this work is to show how Deep Learning methods can improve on the current state-of-the-art for the SCGM segmentation, through an extensive evaluation and comparison with other independently developed methods. The second goal is to show that even though Deep Learning has a high sample complexity, this can be alleviated through the use of semi-supervised learning techniques. The third goal is to show how *Domain Adaptation* techniques can be used to partially mitigate the poor generalization performance on unseen data. The fourth and last goal, in the spirit of Open Science principles, is to implement, document and test all developed software and make it available to the general public under a permissive open-source license at zero cost.

This thesis is organized as follows. In Chapter 2, we present a short critical literature review of previous works as well as a review of some concepts important to the themes that will

be developed later. In Chapter 3 we present an overall methodology and the main research questions guiding the research agenda. In Chapter 4, we present the first article where we develop a supervised end-to-end approach to the spinal cord gray matter segmentation using dilated convolutions; in Chapter 5 we present the second article where we develop a semi-supervised approach to the spinal cord gray matter segmentation by leveraging unlabeled data; in Chapter 6 we present the third and last article where we develop an unsupervised domain adaptation technique to address the generalization gap of Deep Learning models when applied to unseen domains. In Chapter 7 we present a general discussion and in Chapter 8 we present the conclusion, limitations and recommendations.

CHAPTER 2 LITERATURE REVIEW

2.1 Medical review

In this section, a brief literature review of the medical concepts linked to the spinal cord and the clinical relevance of the spinal cord gray matter are presented. This section provides the basic concepts to help the reader understand the main motivation, rationale and methods developed, however, it is far from a comprehensive introduction to the Spinal Cord or MRI concepts.

2.1.1 Spinal Cord

The CNS is often divided by neuroscientists in brain and spinal cord. The human spinal cord, nearly as thick as an adult's little finger has two basic types of nervous tissues: *gray matter* and *white matter* [5]. The gray matter forms an H-shape (also called "butterfly shape") surrounding the central canal of the spinal cord and consists of mainly neuronal cell bodies and neuropil, while the white matter surrounds the gray matter and consists of axons collected into overlapping fiber bundles. Many axons in the white matter have a myelin sheath that allows the rapid nerve impulse conduction and gives the white matter the pale appearance [5].

The three main segments of the spinal cord are shown in the Figure 2.1. Bilateral pairs of dorsal and ventral roots emerge along its length and form five different sets: cervical (in the neck above the rib cage), thoracic (associated with the rib cage), lumbar (near the abdomen), sacral (near the pelvis), and coccygeal (associated with tail vertebrae) [5]. In human, these spinal nerve pairs sum to 31, and are named according to the intervertebral foramen the pass through, however, this enumeration can vary between different species.

The spinal cord is the main information channel connecting the brain and the peripheral nervous system. Information (from nerve impulses) that reaches the spinal cord through sensory neurons are transmitted up to the brain. In the other direction, signals arising in the motor areas of the brain travel back down the cord. The spinal cord also contains the Central Pattern Generator (CPG), neuronal circuits (networks of interneurons) that can produce self-sustained patterns of behavior, *independent* of sensory input [24, 25]. The spinal cord is also surrounded by layers of meninges and has a central canal running through it filled with cerebrospinal fluid.

In the Figure 2.2, a graphical description of a cross-sectional slice of the spinal cord is shown.

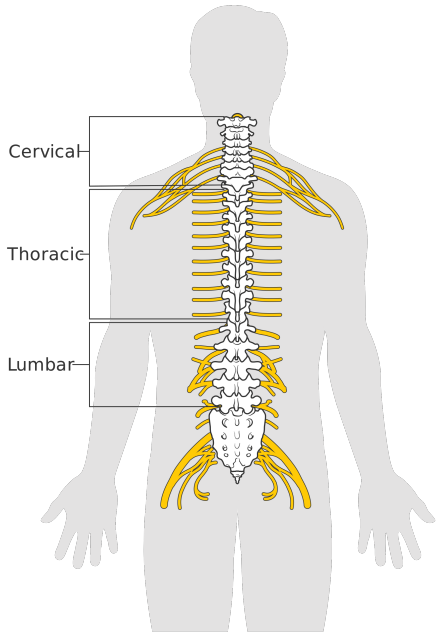


Figure 2.1 Diagram of the human spinal cord showing its segments. *Source: Cancer Research UK / Wikimedia Commons. CC BY-SA license.*

If a neuronal tissue is viewed under a microscope without proper histological procedures (such as fixing and staining), the tissue will appear almost transparent [26]. For that reason, most tissues prepared for microscopy, are usually stained. Under the microscope one can observe densely packed neuronal cell bodies (the gray matter) and unmyelinated and myelinated axons (the white matter).

In the Figure 2.3, a cross-sectional histology slice (microscopy) of the spinal cord is shown. The distinction between gray matter and the white matter tissue is clear in this image.

Tissue changes in the spinal cord gray matter and white matter has an important clinical relevance in many neurological disorders. For that reason, spinal cord imaging has come to play a vital role in the study of disorders such as Multiple Sclerosis (MS) [8, 27], Amyotrophic Lateral Sclerosis (ALS) [28], and traumatic injury [29]. Metrics extracted from the spinal cord may help to model the clinical outcomes, help to understand disorders, provide detection mechanisms and be used to monitor the disease progression.

2.1.2 Relevance of the Spinal Cord Gray Matter

As mentioned earlier, the involvement of the spinal cord gray matter was found to have an important clinical relevance on many neurological disorders. In [8], an *in vivo* study with 113

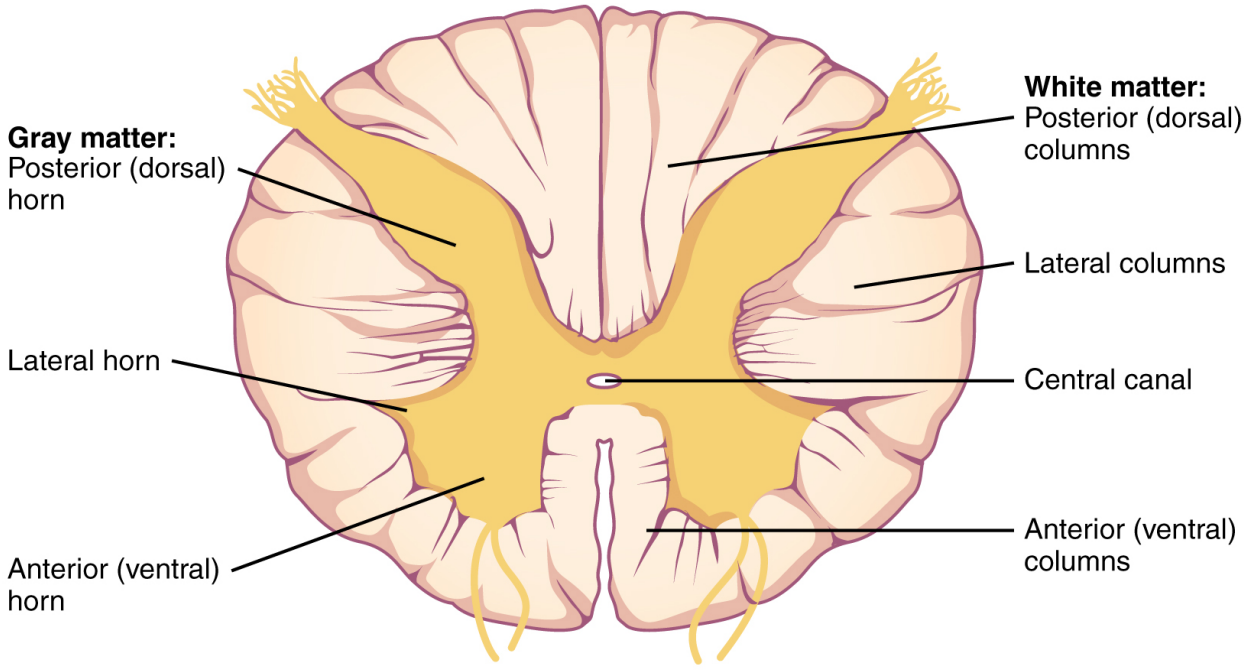


Figure 2.2 Anatomical cross-section of the spinal cord. *Source: OpenStax Anatomy and Physiology, CC-BY license.*

multiple sclerosis patients, found that on a regression analysis that the spinal cord gray matter area was the strongest correlate of disability (using the EDSS scores) in multivariate models including brain gray matter and white matter volumes, FLAIR lesion load, T1-lesion load, spinal cord white matter area, number of spinal cord T2 lesions, age, sex, disease duration.

In [28], a study with 29 ALS patients showed evidence that the use of the spinal cord gray matter as an MR imaging structural biomarker can be used to monitor the evolution of amyotrophic lateral sclerosis.

These studies, however, depend on manual segmentation of the spinal cord gray matter, which is a very time-consuming task that requires a trained expert and might introduce the expert's biases into the gold standard, not to mention the disagreement between experts (also present on other tasks such as the manual annotation of MS lesions in the spinal cord, as found by [30]) and their lack of reproducibility.

While recent cervical cord cross-sectional area (CSA) segmentation methods have achieved near-human performance [31], the accurate segmentation of the gray matter remains a challenge [2]. The main properties that make the gray matter area difficult to segment are: inconsistent intensities of the surrounding tissues, image artifacts and pathology-induced changes in the image contrast [9].

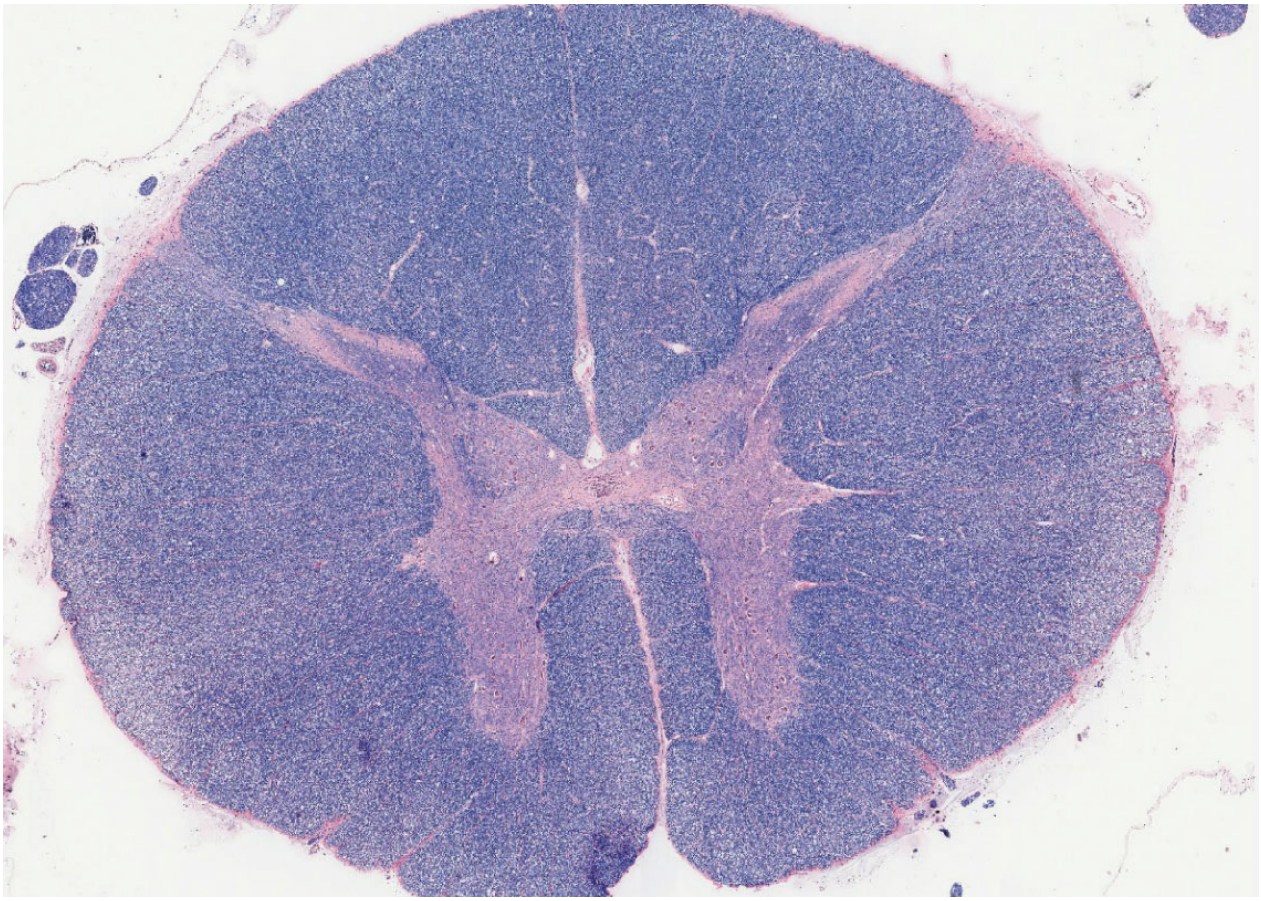


Figure 2.3 A histology slice of the spinal cord showing a clear tissue differentiation between myelinated white matter and gray matter. Best viewed in color. *Source: OpenStax Anatomy and Physiology, CC-BY license.*

Recently, given the importance of the spinal cord gray matter segmentation, the scientific community organized a challenge called "Spinal Cord Gray Matter Segmentation Challenge" (SCGM Challenge) [2] to characterize the state-of-the-art and compare six independent developed methods [2, 7, 10–16] on a public available standard dataset created through the collaboration of four recognized spinal cord imaging centers (University College London, Polytechnique Montreal, University of Zurich and Vanderbilt University), providing therefore a basis for method comparison that was previously unfeasible.

In this work, the same aforementioned dataset and evaluation procedures are used to evaluate the developed method against other previously-developed methods.

2.1.3 Magnetic Resonance Imaging (MRI)

Medical magnetic resonance imaging (MRI) uses the signal from the nuclei of the hydrogen atoms (H) for image generation [1]. Apart from the positive charge, the proton has a *spin*, an intrinsic property of elementary particles. The proton has two important properties: the *angular momentum* and the *magnetic moment* [1]. The angular momentum is due to the rotating mass as the proton acts like a spinning top. Since the rotating mass has an electrical charge, the magnetic momentum acts like a small magnet and therefore is affected by external magnetic fields and electromagnetic waves [1].

When the hydrogen nuclei are exposed to an external magnetic field (B_0), the magnetic moments do not only align with the field but, undergo precession. This precession of the nuclei occurs at a speed that is proportional to the strength of the applied magnetic field. This is called the *Larmor frequency* and is given by the following equation:

$$w_0 = \gamma_0 * B_0 \tag{2.1}$$

where w_0 is the Larmor frequency (MHz), y_0 is the gyromagnetic ratio and B_0 is the strength of the magnetic field in Tesla (T).

An MRI machine explores these properties to generate a spatial image volume of the human body. The MRI machine will produce the main magnetic field B_0 that will cause the protons to align parallel (low-energy) or anti-parallel (high-energy) to the primary field, resulting in a net magnetic vector M which is in the direction of the primary magnetic field.

The MRI machine also uses a secondary magnetic field that is generated by the gradient coils in the **x**, **y** and **z** axes. The gradients will perturb the magnetic field and therefore change the precession rate. The key aspect of the gradients is that they distort the primary magnetic field in a predictable way, causing the resonance frequency of protons to vary as a function of position in space, allowing the spatial encoding for the MRI images.

A radio-frequency (RF) pulse (B_1) is also applied with the same precession frequency by means of an antenna coil. All of the longitudinal magnetization is rotated into the transverse plane by an RF pulse that is strong enough to tip the magnetization by exactly 90° (90° RF pulse) [1]. Immediately after excitation, the magnetization rotates in the **xy**-plane, being called *transverse magnetization*. It is this transverse magnetization that produces the MR signal in the RF receiver coil. This MR signal fades very quickly due to two different processes: *T1 relaxation* and *T2 relaxation* [1].

The T1 relaxation happens in the longitudinal axis and is parallel to B_0 field (**z**-axis), while

the T2 relaxation happens perpendicular to B_0 field (**xy**-axis). Three main *intrinsic features* of biological tissues can contribute to the signal intensity on a MR image: the *proton density*, which is the number of excitable spins per unit volume, the *T1 time*, which is the time it takes for the excited spins to recover and be available for the next excitation and the *T2 time*, that mostly determines how quickly an MR signal fades after excitation [1].

These parameters, depending on which an MR sequence is emphasized, may cause the MR images to differ in its tissue-tissue contrast. This mechanism is the basis for the soft-tissue discrimination on MR imaging [1]. In Table 2.1, we can see intrinsic properties of some important tissue types.

Table 2.1 The relative proton densities in % and intrinsic T1 and T2 times (msec). Adapted from [1].

Tissue	**Proton Density**	**T1 (at 1.5 Tesla)**	**T2 (at 1.5 Tesla)**
CSF	100	>4000	>2000
White Matter	70	780	90
Gray Matter	85	920	100
Metastasis	85	1800	85
Fat	100	260	80

Apart from the traditional challenges present in the medical imaging domain, such as anatomy variability across different subjects, MRI images pose multiple additional challenges for machine learning models, such as noise, variability across machine vendors, acquisition parameters, artifacts, to name a few. In the Figure 2.4, the different distribution of the voxel intensities among different centers shows one of the problems that MRI poses to statistical learning techniques that shows a tendency to rely on surface statistical regularities [32], such as deep learning methods.

2.1.4 Magnetic Resonance Imaging of the Spinal Cord

The human anatomy makes it difficult to "see" the spinal cord without highly invasive and risky surgical procedures. Therefore, non-invasive techniques such as MRI are paramount for successful research studies, diagnostic biomarkers detection, and disease progression monitoring.

In the past, MRI of the spinal cord has been limited due to the poor white and gray matter contrast differentiation, artifacts induced by physiological processes such as cord motion [33]. According to [33], the main inherent challenges present in the spinal cord imaging are:

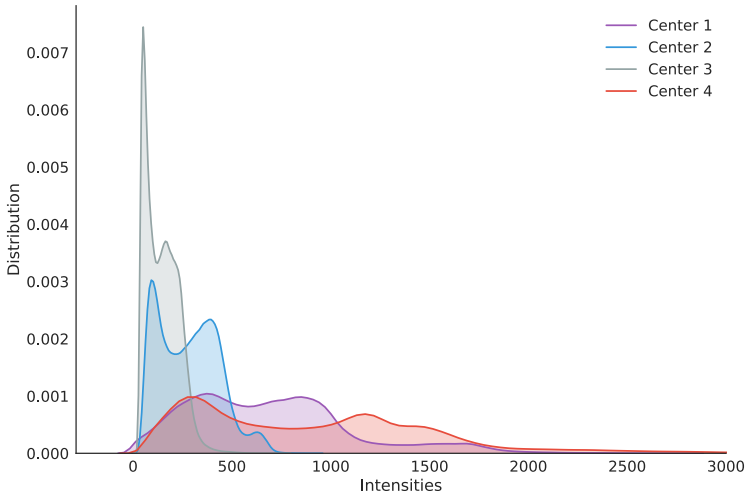


Figure 2.4 Unnormalized MRI image intensity distributions for each center using axial slices from the Spinal Cord Gray Matter Segmentation Challenge [2] dataset. The **x**-axis represent the MRI intensities and the **y**-axis represents the intensity distribution. Best seen in color.

spatially non-uniform magnetic field environment when in an MRI system, the small physical dimensions of the cord cross-section and the physiological motion.

Tissue segmentation methods that were developed in the past for brain MR images, when applied for spinal cord images, were largely unsuccessful [15]. Recently, thanks to sequences such as T2* weighted MRI [34, 35], that were able to get higher quality images in reasonably short acquisition time, they opened the door for the feasibility of tissue segmentation of these spinal cord structures.

In the Figure 2.5, an axial slice from a 3D multi-echo gradient-echo sequence acquisition is shown.

Although the human manual segmentation of the spinal cord gray matter is usually easy, the agreement between different raters is usually nearly 0.90 DSC (Dice-Sørensen coefficient), a score that measures the voxel-wise agreement between two binary masks, in average when compared with the majority voting mask. Before this work, the best algorithm for gray matter segmentation in terms of the DSC score and as evaluated on the Spinal Cord Gray Matter Segmentation Challenge [2], had a DSC score of 0.80 [2, 16].

Although it is not clear if the agreement between human raters measured on the work [2] included the rater in the majority voting, the state-of-the-art methods were still far from the human performance for the task of the spinal cord gray matter tissue segmentation.

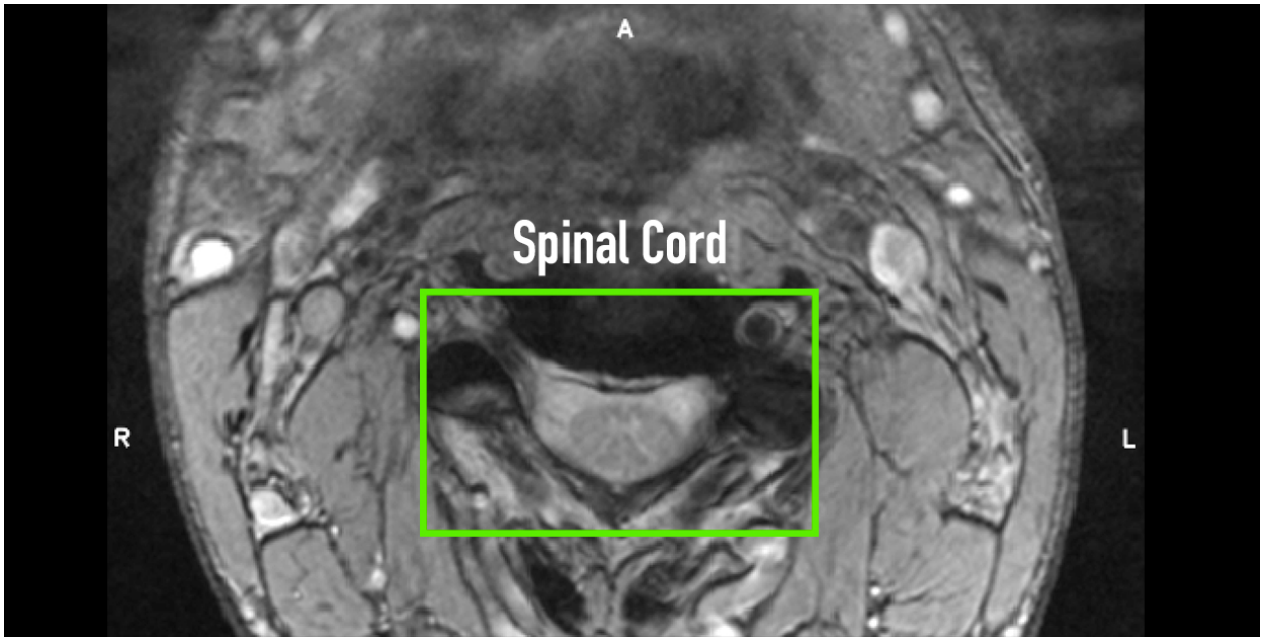


Figure 2.5 A random axial slice from a random selected subject of the Spinal Cord Gray Matter Segmentation Challenge [2]. This image was produced by a 3D multi-echo gradient-echo sequence using a resolution of 0.25x0.25x2.5 mm on a 3T Siemens Skyra machine. The spinal cord is shown inside the green rectangle.

2.2 Machine Learning Review

In this section, the main concepts related to the machine learning and Deep Learning domains are introduced. This review is far from an exhaustive review and only describe concepts required for the understanding of the present work.

2.2.1 Supervised Learning

Machine learning can be described as a sub-domain of Artificial Intelligence where learning algorithms can learn with data. A widely quoted, and formal definition of the algorithms studied in machine learning can be found in [36]:

Definition 2.2.1. Learning algorithm. A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P** if its performance at tasks in **T**, as measured by **P**, improves with experience **E** [36].

Machine learning and Artificial Intelligence are moving targets and its definition changed in the past few years. As an example, some algorithms that were employed in the past for Artificial Intelligence are no longer nowadays considered learning algorithms by the community,

therefore, a precise definition of these fields is out of the scope of this work.

Before being able to apply machine learning, one must assume that there is a pattern in the data and that data is available. The assumption of a pattern is a circular concept, given that it is evident that is really difficult for a human to evaluate patterns in data, so learning algorithms are usually applied even before knowing that a pattern is present in the data.

Machine learning tasks are usually categorized as supervised learning, semi-supervised learning, unsupervised learning, reinforcement learning or even hybrid approaches. Recently, a new term called *self-supervised* also emerged to describe unsupervised tasks where a supervised sub-task is created to learn a representation or solve a learning problem.

The learning problems can also be categorized depending on the main goal of the task. When an estimated response is a continuous dependent variable, the task is called a *regression*. When this variable is related to the identification of group membership, this task can be called as a *classification* task. When a density function is required to be estimated, this problem is often called a *density estimation*.

For this present work, we are mostly interested in the supervised learning problem. Where given a dataset:

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \tag{2.2}$$

where $\mathcal{D}$ is a collection of input samples x_n with their respective labels y_n. We want to find a model $f_\theta(x)$ parametrized by the parameters θ that describes the relationship between the random variable X and the target label Y, therefore we assume a joint distribution $p(X, Y)$. In order to evaluate how good the model is, we define a loss function $\mathcal{L}$, evaluated at $\mathcal{L}(f_\theta(x), y)$ that gives us a penalization for the difference between predictions of f_θ and the true label y.

To evaluate the loss $\mathcal{L}$ on all datapoints, we take the expectation of the loss under the distribution $p(X, Y)$:

Definition 2.2.2. Risk.

$$\mathbb{E}_{x,y\sim p}[\mathcal{L}(f_\theta(x), y)] = \int \mathcal{L}(f_\theta(x), y) dp(x, y) \tag{2.3}$$

Under the ERM framework, this is known as $R(f)$, the *risk* of the hypothesis f. However, given that we don't have access to the entire joint distribution $p(x, y)$ but only to a sample of this distribution by our dataset $\mathcal{D}$, we define the *empirical risk* as:

Definition 2.2.3. Empirical risk.

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f_\theta(x_i), y_i) \tag{2.4}$$

The main idea behind the ERM principle, is to minimize the empirical risk $R_{\text{emp}}(f_\theta)$:

$$\hat{f}_\theta = \arg\min_\theta R_{\text{emp}}(f_\theta) \tag{2.5}$$

Even though we're minimizing the empirical risk, we know through the law of the large numbers, that $R_{\text{emp}}(f) \to \mathbb{E}_{x,y\sim p}[\mathcal{L}(f_\theta(x), y)]$ as $n \to \infty$, which means that the empirical risk will converge to the risk as the number of samples grows to infinity. Given this formulation, it is easy to see that ERM can easily overfit the data, however, in practice, the hypothesis space is constrained to a particular class of hypotheses (such as linear models) or an additive regularization term is added to the loss, such as L_2 regularization.

Supervised learning is perhaps one of the most successful approaches in machine learning due to the leverage of the supervision signal, however, in medical imaging, providing annotation for images is seldom easily done as providing labels for natural images [37].

2.2.2 Semi-supervised Learning

In medical imaging, the small data regime is the norm for many tasks. As opposite to tasks involving natural images, the process of acquiring annotations/labels in medical imaging is very expensive and time-consuming because it involves the time of experts such as radiologists and it usually involves dense pixel-wise annotations as well. In the case of the SCGM Challenge [2], after slicing all volumes in 2D axial plane, the total amount of slices are less then 3000. When compared to the ImageNet size with millions of images, it is clear that these over-parametrized Deep Learning models would require extensive regularization and will suffer with a higher generalization gap.

On the other hand, unlabeled data is usually available, and it is often ignored due to the fact that the loss for unlabeled samples is undefined for supervised learning. The semi-supervised learning paradigm is halfway between supervised and unsupervised learning [38], where in addition to unlabeled data, the learning algorithm is provided with some supervision information for some samples.

Formalizing the semi-supervised learning paradigm, we are given a dataset:

$$X = \{x_i\}_{i \in [n]} \tag{2.6}$$

that can be split into two disjoint sets as:

$$X_l = \underbrace{\{x_1, \ldots, x_l\}}_{\text{Labeled set}} \tag{2.7}$$

$$X_u = \underbrace{\{x_{l+1}, \ldots, x_{l+u}\}}_{\text{Unlabeled set}} \tag{2.8}$$

where X_l set represents the set of points where we have the corresponding label set:

$$Y_l = \{y_1, \ldots, y_l\} \tag{2.9}$$

and the set X_u is the set of points where we don't have labels. If the knowledge available in $p(x)$ that we can obtain from the unlabeled set X_u contains information that can help the inference problem $p(y|x)$, then it is evident that semi-supervised can bring improvements to the learning problem [38].

Many assumptions can be made by semi-supervised learning algorithms, and these assumptions must hold for these learning algorithms to work. One common assumption is the *semi-supervised smoothness assumption* [38], that can be defined as:

Definition 2.2.4. Semi-supervised smoothness assumption. If two points x_1, x_2 in a high-density region are close, then so should be the corresponding outputs y_1, y_2.

In the Figure 2.6, we can see a graphical explanation for the motivation behind this smoothness assumption. In this figure, we can see how the decision boundary represented by a dashed line can change by adding unlabeled data and assuming the smoothness hypothesis. Therefore, in some cases, semi-supervised learning can radically change the decision boundary by incorporating unlabeled data.

There is a large body of published articles on semi-supervised learning methods [39], however, most of the previous work was developed in the context of classification or regression tasks, with only a few of them focused on segmentation or even less common, for segmentation using deep learning methods. Similar to the methods developed in this present work, we have the Ladder Networks [40] that introduced the later connections into an encoder-decoder architecture with two branches in parallel where one of the branches take the original input data, whereas the other branch is fed with the same input but corrupted with noise.

Recently, in [41], the authors expanded the work by [40] where it differs from it by removal of the parametric nonlinearity and denoising, having two corrupted paths, and comparing the outputs of the network instead of pre-activation data of the final layer, which was named *temporal ensembling*.

In the temporal ensembling [41] technique, at each training step, all the exponential moving average (EMA) predictions of the samples in the mini-batch are updated based on new predictions. Therefore, the EMA prediction for each sample is formed by an ensemble of the model's current state and those previous states that evaluated the same example. Given that each target is updated only once per epoch, the information is incorporated into the training process at a very slow pace. In [42], the authors expanded the work by [41] by overcoming the limitations of the technique by proposing averaging model weights instead of predictions. This technique demonstrated significant improvements upon the previous state-of-the-art semi-supervised methods.

Generative Adversarial Networks (GAN) were also employed for semi-supervised learning with promising results [43].

It is also important to note that recently, a critical evaluation study [44] demonstrated that the performance of simple baselines which do not use unlabeled data was often underreported and that semi-supervised learning methods differ in sensitivity to the amount of labeled and unlabeled data, with a significant performance degradation when in presence of out-of-class examples. It is interesting to note that unlabeled out-of-class examples can change the decision boundaries of the model towards the unlabeled data domain, this shows evidence of a strong link between semi-supervised learning and domain adaptation as seen in [42].

Only a few works were developed for semi-supervised learning in the context of segmentation in medical imaging. Only recently, a U-Net was employed in that context, however, as an auxiliary embedding [45] and for domain adaptation using a private dataset. In [46] they used GANs for that purpose, but employed unrealistic dataset sizes when compared to medical imaging domain datasets, along with ImageNet pre-trained networks. In [46] a technique using adversarial training was proposed, but with the focus on knowledge transfer between natural images with pixel-level labels and weakly-labeled images.

2.2.3 Domain Adaptation

The ERM principle has well-known learning guarantees when the training and test data come from the same domain [47], however, in real-world applications, and especially in the medical imaging domain, a shift in the distribution of data is very common due to many factors

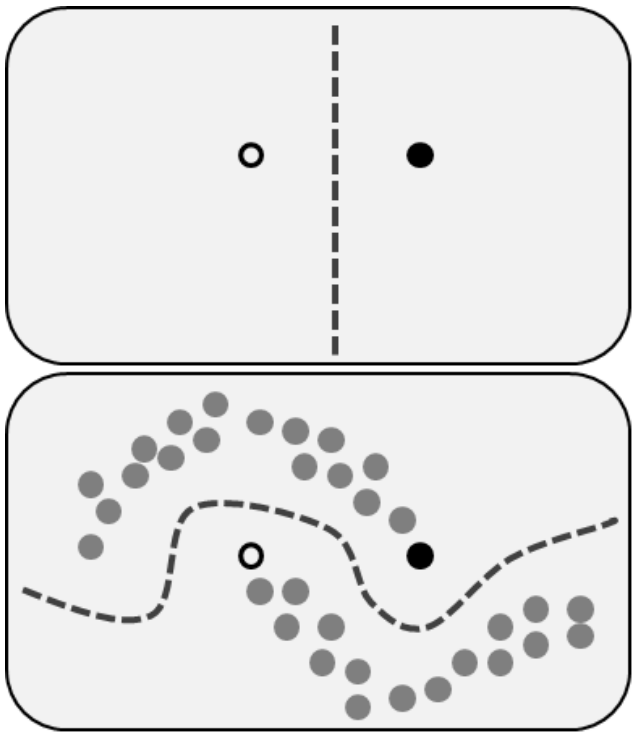


Figure 2.6 Top panel: decision boundary based on only two labeled examples (white vs. black circles). Bottom panel: decision boundary based on two labeled examples plus unlabeled data (gray circles). *Source: Techerin, Wikimedia Commons, CC-BY-SA license.*

such as natural anatomy variability, different imaging acquisition parametrization, different machine vendors, to name a few. Therefore, machine learning models trained under the ERM framework usually shows a poor generalization when applied on domains that are different than the domains where the model was trained. This constatation provides the motivation behind *domain adaptation* techniques, that are responsible for providing ways to mitigate these distributional shifts. In the Figure 2.7 we can see a graphical representation of the positioning of domain adaptation among other kinds of transfer learning.

Definition 2.2.5. Domain A domain can be defined as the combination of an input space $\mathcal{X}$, and output space $\mathcal{Y}$, and an associated probability distribution p. Given any two domains $\mathcal{D}_1$ and $\mathcal{D}_2$, we say they are different if at least one of their components $\mathcal{X}$, $\mathcal{Y}$ or p are different [48].

In medical imaging, examples of realizations from this difference among domains are multi-center studies, different acquisition parameters, different machine vendors, etc. The variability inherent in medical imaging usually violates the fundamental statistical learning assumptions that data comes from identical distributions.

Although this is one of the major problem holding machine learning models from the robustness required for practical applications [49], it is usually ignored in many studies and machine learning challenges organized by many entities. The evaluation scheme often used, especially in multi-center studies, is to have samples from all centers both in training and test sets, which consequently leads to over-optimistic evaluations of machine learning models, because in real practice, what happens is that these models are used for new centers where labeled data isn't available.

Unsupervised domain adaptation is the field of study where this exact problem is addressed. In this scenario, we have a source domain $\mathcal{D}_S$ where we have labeled data and a target domain $\mathcal{D}_T$ for which we don't have labels but want the model to generalize as well, therefore, the goal is to create a model that can generalize not only on the source domain $\mathcal{D}_S$ but also on the unlabeled target domain $\mathcal{D}_T$.

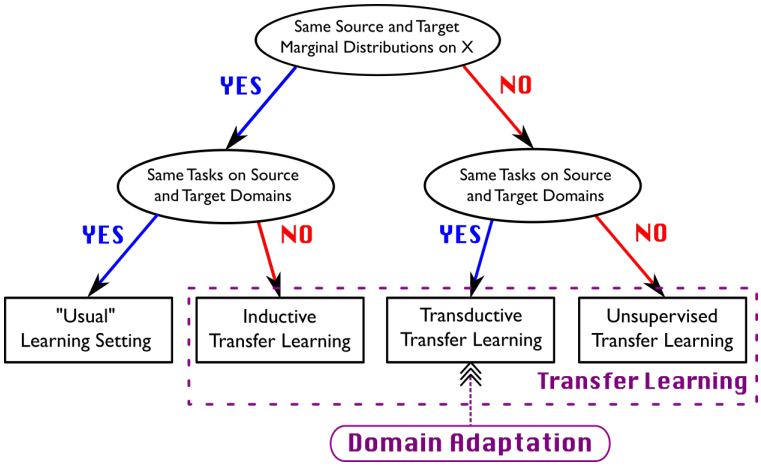


Figure 2.7 Distinction between usual machine learning setting and transfer learning, and positioning of domain adaptation. *Source: Emilie Morvant, Wikimedia Commons, CC-BY-SA license.*

While most of the techniques developed for domain adaptation in the past focused mostly on classification tasks [50,51], recently there was a surge of interest to expand these techniques to semantic segmentation as well. An example is the work by [52] where the authors expanded the domain-adversarial training [53] for segmentation tasks and applied it to medical imaging tasks.

2.2.4 Deep Learning

Deep Learning methods [17] can be characterized as a major shift from the traditional handcrafted feature engineering to a hierarchical representation learning approach. After

the seminal work from the AlexNet [18], the research community embraced Deep Learning techniques. Nowadays, Deep Learning is pervasive and had achieved state-of-the-art results in many fields such as NLP [54], computer vision [55], speech recognition [56], machine translation [57], to name a few.

The adoption of Deep Learning techniques in medical imaging also increased significantly in the past years. According to a recent survey [23] that analyzed more than 300 contributions to the field, there was a fast growth of the number of papers in 2015 and 2016, with CNNs being the most used model. The survey also showed that the topic became dominant at major conferences as well.

Deep Learning techniques are mostly based on neural network models, which are a type of statistical learning algorithms comprised by neurons (or units), that uses composition of functions. The basic building block of a neural network is the activation a that can be defined as:

$$a = \sigma(w^T x + b) \tag{2.10}$$

where a is the activation, b is the bias term and $\sigma(\cdot)$ is the activation function such as a rectified linear unit (ReLU) [58] or sigmoid. This equation is also often written as $a = \sigma(\theta^T x)$ where the bias term is collapsed into the parameter θ.

Definition 2.2.6. The Multi-Layer Perceptron (MLP) network uses several layers of these basic building blocks to form a recursive application of these functions:

$$p(y|x;\theta) = \sigma(\theta^L \sigma(\theta^{L-1} \dots \sigma(\theta^0 x))) \tag{2.11}$$

In Figure 2.8 we can see a graphical depiction of the MLP.

For a regression problem, the last activation function is usually just a linear identity, while for classification problem, the last layer is usually a softmax layer that squeezes the activations into a distribution over classes $p(y|x;\theta)$.

The problem of learning these parameters is usually posed into a frequentist optimization framework where the maximum likelihood estimator is maximized, which for convenience is treated as a minimization problem by minimizing the negative of the log likelihood:

$$\arg\min_{\theta} \sum_{i=1}^{n} -\log p(y|x_i;\theta) \tag{2.12}$$

In practice, these frequently over-parametrized networks are analytically intractable due

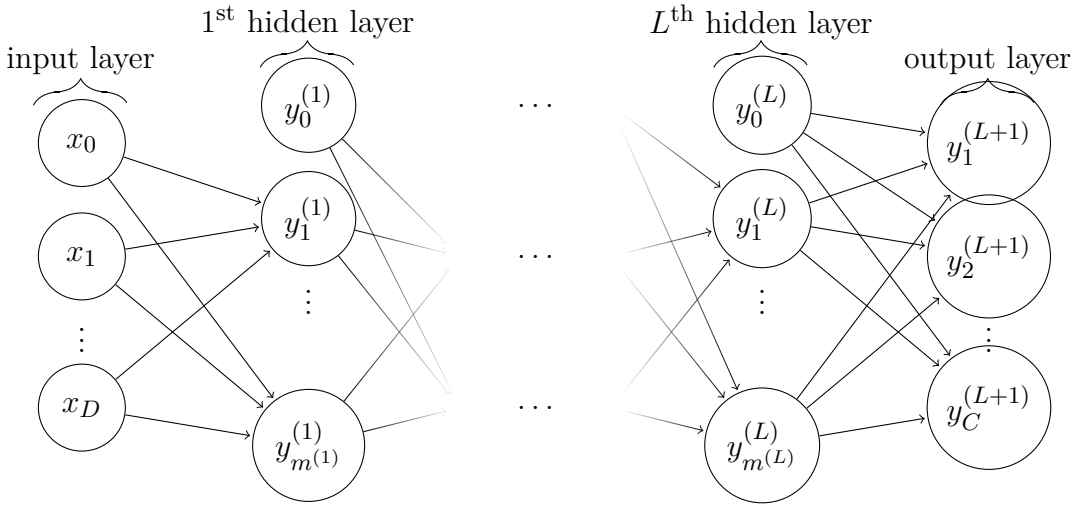


Figure 2.8 Network graph of a $(L+1)$-layer perceptron with D input units and C output units. The l^{th} hidden layer contains $m^{(l)}$ hidden units. *Source: David Stutz, BSD 3-Clause license.*

to the amount of data and model complexity, therefore a mini-batch Stochastic Gradient Descent (SGD) is used to optimize the parameters using only a portion of the data at each time. SGD works by iteratively updating the parameters according to the Algorithm 2.2.1.

Algorithm 2.2.1 The general gradient descent algorithm; different choices of the learning rate γ and the estimation technique for $\nabla\mathcal{L}(\theta)$ may lead to different implementations.

Input: initial weights $\theta^{(0)}$, number of iterations T
Output: final weights $\theta^{(T)}$
1. **for** $t = 0$ **to** $T-1$
2. estimate $\nabla\mathcal{L}(\theta^{(t)})$
3. compute $\Delta\theta^{(t)} = -\nabla\mathcal{L}(\theta^{(t)})$
4. select learning rate γ
5. $w^{(t+1)} := w^{(t)} + \gamma\Delta\theta^{(t)}$
6. **return** $w^{(T)}$

For neural networks, the gradient $\nabla\mathcal{L}(\theta^{(t)})$ is computed using backpropagation as described in the Algorithm 2.2.1.

2.2.5 Convolutional Neural Networks (CNN)

Convolutional Neural Networks [59], also known as CNNs, are a class of specialized models that achieved enormous success in many practical applications. CNNs were inspired by the *neocognitron* approach from Kunihiko Fukushima back in 1980, which in turn, were inspired

Algorithm 2.2.1 Error backpropagation algorithm for a layered neural network represented as computation graph.

(1) For a sample (x_n, y_n^*), propagate the input x_n through the network to compute the outputs $(v_{i_1}, \ldots, v_{i_{|V|}})$ (in topological order).

(2) Compute the loss $\mathcal{L}_n := \mathcal{L}(v_{i_{|V|}}, y_n^*)$ and its gradient

$$\frac{\partial \mathcal{L}_n}{\partial v_{i_{|V|}}}. \tag{2.13}$$

(3) For each $j = |V|, \ldots, 1$ compute

$$\frac{\partial \mathcal{L}_n}{\partial w_j} = \frac{\partial \mathcal{L}_n}{\partial v_{i_{|V|}}} \prod_{k=j+1}^{|V|} \frac{\partial v_{i_k}}{\partial v_{i_{k-1}}} \frac{\partial v_{i_j}}{\partial w_j}. \tag{2.14}$$

where w_j refers to the weights in node i_j.

by the biological mechanism that was unveiled by Hubel and Wiesel in the 1950s and 1960s where two basic visual cell types were identified in the brain: the *simple cells* that were fired by straight edges having particular orientations within their receptive field and *complex cells* that have larger receptive fields and are insensitive to the exact position of the edges in the field.

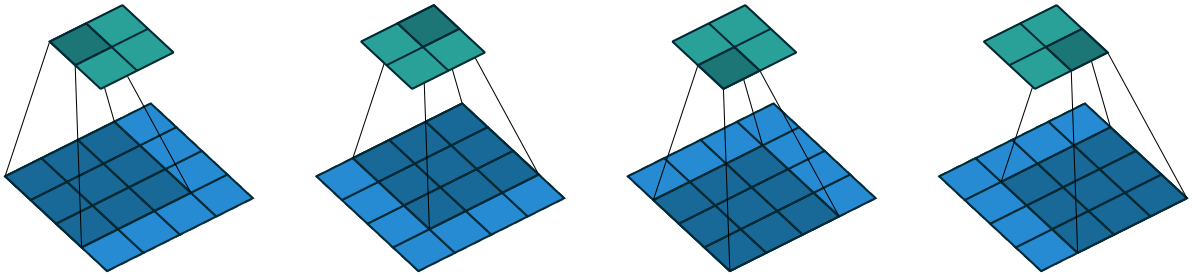


Figure 2.9 (No padding, unit strides) Convolving a 3×3 kernel over a 4×4 input using unit strides (i.e., $i = 4$, $k = 3$, $s = 1$ and $p = 0$). *Source: Vincent Dumoulin et al. [4], MIT license.*

The main difference between a CNN and a vanilla MLP is that the CNN uses shared weights due to the convolutional component that uses a sliding window to apply the same weights, as seen in Figures 2.9 and 2.11. Another difference is the introduction of pooling layers that can be seen in Figure 2.12, that acts as a subsampling mechanism that can yield certain levels of rotation invariance to the network, although recently architectures can work well without pooling as well, especially for segmentation tasks, that will be discussed in the next section.

In Figure 2.10 we show the architecture of a traditional CNN with convolutional layers followed

by non-linearities and subsampling layers. At the end of the CNN there is a fully-connected network, however, this fully-connected layer at the end isn't that common anymore on modern architectures such as ResNets [55], that employ a global average pooling before the softmax activation.

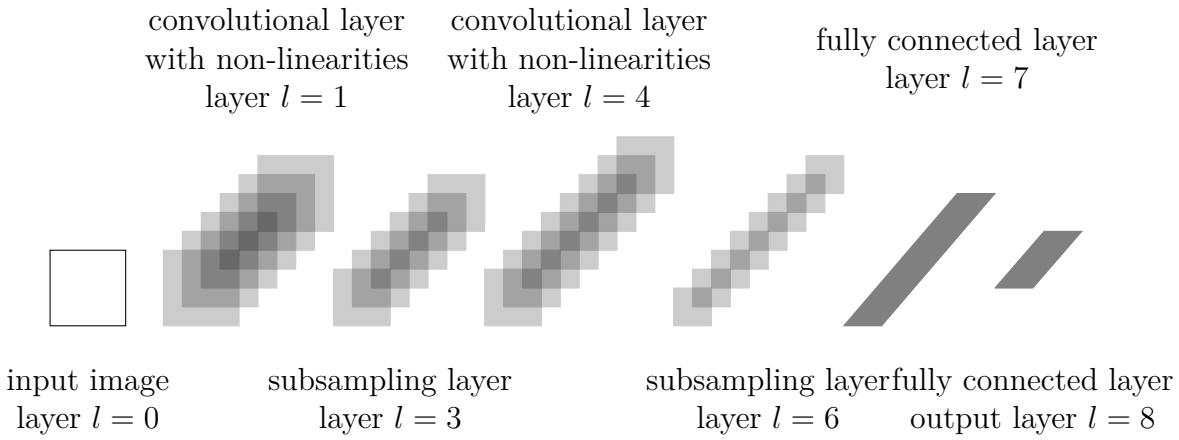


Figure 2.10 The architecture of the original convolutional neural network, as introduced by LeCun et al. (1989), alternates between convolutional layers including hyperbolic tangent non-linearities and subsampling layers. The feature maps of the final subsampling layer are then fed into the actual classifier consisting of an arbitrary number of fully connected layers. The output layer usually uses softmax activation functions. *Source: David Stutz, BSD 3-Clause license.*

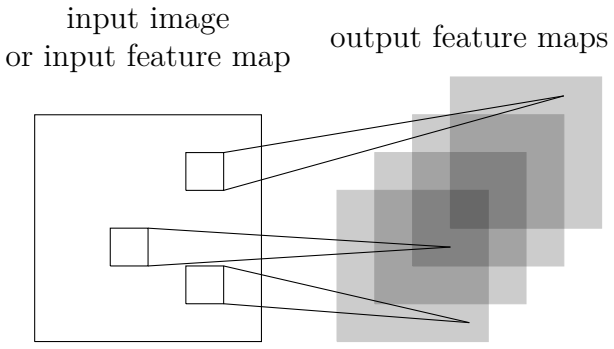


Figure 2.11 Illustration of a single convolutional layer. If layer l is a convolutional layer, the input image (if $l = 1$) or a feature map of the previous layer is convolved by different filters to yield the output feature maps of layer l. *Source: David Stutz, BSD 3-Clause license.*

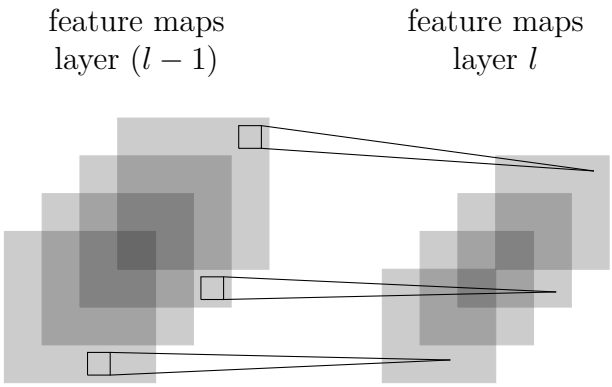


Figure 2.12 Illustration of a pooling and subsampling layer. If layer l is a pooling and subsampling layer and given $m_1^{(l-1)} = 4$ feature maps of the previous layer, all feature maps are pooled and subsampled individually. Each unit in one of the $m_1^{(l)} = 4$ output feature maps represents the average or the maximum within a fixed window of the corresponding feature map in layer $(l-1)$. *Source: David Stutz, BSD 3-Clause license.*

2.2.6 Convolutional Neural Networks for Semantic Segmentation

Several works [60–62] applied convolutional neural networks for semantic segmentation, also called *dense prediction*. In dense prediction, the network usually outputs a prediction map with the same size of the input of the network with a prediction per each pixel of this output map.

Majority of the literature in the past were constrained by small models, patch-wise training due to the memory limitations, post-processing with superpixels, to name a few. One of the most important works that recently spawned a series of important developments for semantic segmentation is the work by [63] called Fully Convolutional Network (FCN), where the authors demonstrated that a fully-convolutional architecture trained end-to-end exceeded the state-of-the-art results when compared to its predecessors.

The main insight of the FCN was to combine coarse high layer information with fine, low layer information before up-sampling and producing the final predictions. The coarse features, coming from high layers contains semantic information that are merged together (simple summation) with the low layer features that contained the local information related to the fine spatial grid. By repurposing pre-trained networks into FCNs, the authors were also able to do transfer learning from networks trained on classification tasks such as ImageNet to semantic segmentation tasks.

After FCNs [63], a significant amount of follow-up works were developed. In medical imaging, one of the most prominent models that were developed based on the insights from FCNs is

the U-Net [64], where two different paths are combined with skip-connections to concatenate the feature maps. The first part is a downsampling path that uses traditional convolutional and pooling layers and the upsampling part where 'up'-convolutions are used to increase the image size, creating an architectural shape of a U, hence the "U-Net" name.

In the Table 2.2 we show a summary of the currently available methods for the segmentation of the spinal cord gray matter. These methods are later expanded and described in the Chapter 4.

Table 2.2 Summary of the available methods for spinal cord gray matter segmentation. These are the methods that participated into the SCGM Challenge, it doesn't cover all previously developed methods.

Method nane	**Year**	**Initialization**	**Training**	**External data**	**Time p /slice**	**Summary**
JCSCS [7]	2016	Auto.	No	Yes	4-5 min	Uses OPAL [12] to detect the spinal cord and then STEPS to do segmentation propagation and consensus segmentation using best-deformed templates.
DEEPSEG [16]	2017	Auto.	Yes (4h)	No	<1 s	U-Net with pre-trained weights using a restricted Boltzmann Machine, uses a weighted loss function with two terms to balance sensitivity and specificity. Uses two models, one for cord segmentation and another for GM segmentation.
MGAC [15]	2017	Auto.	No	No	1 s	Uses external tool "Jim" (from Xinapse Systems) to provide a initial guess for an active contour algorithm.
GSBME [2]	2017	Manual	Yes (<1m)	No	5 - 80 s	Semi-automatic method that uses Propseg for cord segmentation with manual initialization followed by thresholding and outlier detection with image moments.
SCT [11]	2017	Auto.	No	Yes	8 - 10 s	Atlas-based approach using a dictionary of manually segmented WM/GM volumes projected into a PCA space, segmentation is done fusing labels.
VBEM [10]	2016	Auto.	No	No	5 s	Semi-supervised, model intensities as a Gaussian Mixture trained with Expectation-Maximization (EM).

CHAPTER 3 OVERALL METHODOLOGY

This article based thesis is organized in the following section with the articles described below:

- **Article 1**: Perone, C. S., Calabrese, E., & Cohen-Adad, J. (2018). Spinal cord gray matter segmentation using deep dilated convolutions. Nature Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-24304-3
- **Article 2**: Perone, C. S., & Cohen-Adad, J. (2018). Deep semi-supervised segmentation with weight-averaged consistency targets. DLMIA MICCAI, 1–8. https://doi.org/10.1007/978-3-030-00889-5
- **Article 3**: Perone, C. S., Ballester, P., Barros, R. C., & Cohen-Adad, J. (2018). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. Submitted to Elsevier NeuroImage, under review. Short version presented at NIPS 2018 in the Medical Imaging Workshop.

In Chapter 4 we present the Article 1 where it is shown the design a Deep Learning methods for fully-automated segmentation of the spinal cord gray matter tissue from MRI volumes. In Chapter 5 we present the Article 2 where a semi-supervised learning method is developed to attack the problem of the low data regime present in medical imaging and finally in Chapter 6 we show the Article 3 where a unsupervised domain adaptation technique was developed for segmentation tasks in order to address the problem of poor generalization due to the distributional shift from different domains.

The main research questions we answer with present articles are the following:

- 1. How can deep learning methods improve on the current state-of-the-art results for segmentation of the spinal cord gray matter tissue ?
- 2. How can these segmentation methods be extended to take leverage of unlabeled data as well ?
- 3. How can we mitigate the generalization gap when applying these models to realistic scenarios where we have only unlabeled data from a new, different and unseen domain ?

CHAPTER 4 ARTICLE 1: SPINAL CORD GRAY MATTER SEGMENTATION USING DEEP DILATED CONVOLUTIONS

Although the U-Net [64] achieved excellent results in many different tasks, given that the network architecture uses two distinct paths with contraction and expansion of the feature maps, the network also suffers from the increased number of parameters. The U-Net also uses pooling layers, which can be detrimental for translational equivariance that is important for segmentation tasks.

In this chapter we show that a network architecture based on the Atrous Spatial Pyramid Pooling (ASPP) network can achieve better or similar results than the U-Net [64] with 6 times fewer parameters. Furthermore, we show that this network, when applied to the spinal cord gray matter segmentation task (both *ex vivo* and *in vivo*), achieves state-of-the-art results in 8 out of 10 different evaluation metrics when compared to other 6 previously developed methods on the Spinal Cord Gray Matter Challenge [2] dataset by an external third-party system.

To the best of our knowledge, these results remain the state-of-the-art single-model results for the spinal cord gray matter segmentation task on the challenge [2] dataset.

The evaluations presented in this work are from the third-party evaluation system from the SCGM Challenge [2] and were evaluated on a private holdout set by the competition organizers in order to reduce chance of overfitting. Hyper-parameters of the model were ajusted using a validation split.

The dataset [2] used by this work is publicly available and contained volumes acquired by 4 independent centers, a summary of the acquisition parameters are described in Table 4.1.

Table 4.1 A summary of the acquistion parameters from each site. Adapted from [2].

	Site 1 - UCL	**Site 2 - Montreal**	**Site 3 - Zurich**	**Site 4 - Vanderbilt**
Scanner	3T Philips Achieva	3T Siemens TIM Trio	3T Siemens Skyra	3T Philips Achieva
Sequence	3D Gradient echo	2D spoiled gradient multi-echo	3D multi-echo gradient-echo	3D multi-echo gradient-echo
TE	5	5.41, 12.56, 19.16	19	7.2, 16.1, 25
TR	23	539	44	700
Flip Angle	7	35	11	28
Resolution (mm)	0.5 x 0.5 x 5	0.5 x 0.5 x 5	0.25 x 0.25 x 2.5	0.3 x 0.3 x 5

My contribution to this work was to conceive the method, implement it, conduct the experiments, provide manual segmentations and write the paper.

4.1 Article metadata

- **Title**: Spinal cord gray matter segmentation using deep dilated convolutions
- **Authors**: Christian S. Perone [1], Evan Calabrese [2,3], Julien Cohen-Adad [1,4]
- **Publisher**: Nature Scientific Reports
- **DOI**: 10.1038/s41598-018-24304-3
- **Citation**: Perone, C. S., Calabrese, E., & Cohen-Adad, J. (2018). Spinal cord gray matter segmentation using deep dilated convolutions. Nature Scientific Reports, 8.

4.2 Abstract

Gray matter (GM) tissue changes have been associated with a wide range of neurological disorders and were recently found relevant as a biomarker for disability in amyotrophic lateral sclerosis. The ability to automatically segment the GM is, therefore, an important task for modern studies of the spinal cord. In this work, we devise a modern, simple and end-to-end fully-automated human spinal cord gray matter segmentation method using Deep Learning, that works both on *in vivo* and *ex vivo* MRI acquisitions. We evaluate our method against six independently developed methods on a GM segmentation challenge. We report state-of-the-art results in 8 out of 10 evaluation metrics as well as major network parameter reduction when compared to the traditional medical imaging architectures such as U-Nets.

4.3 Introduction

Gray matter (GM) and white matter (WM) tissue changes in the spinal cord (SC) have been linked to a large spectrum of neurological disorders [6]. For example, using magnetic resonance imaging (MRI), the involvement of the spinal cord gray matter (SCGM) area in multiple sclerosis (MS) was found to be the strongest correlate of disability in multivariate models including brain GM and WM volumes, FLAIR lesion load, T1-lesion load, SCWM area, number of spinal cord T2 lesions, age, sex and disease duration [8]. Another study

[1]NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, H3T 1J4, Canada.

[2]Duke University Medical Center, Department of Radiology, Center for In Vivo Microscopy, Durham, NC, 27710, USA

[3]University of California San Francisco, Department of Radiology & Biomedical Imaging, San Francisco, CA, 94143, USA

[4]Functional Neuroimaging Unit, CRIUGM, Universite de Montreal, Montreal, QC, H3C 3J7, Canada.

showed SCGM atrophy to be a biomarker for predicting disability in amyotrophic lateral sclerosis [28].

The ability to automatically assess and characterize these changes is, therefore, an important step [9] in the modern pipeline to study both the *in vivo* and *ex vivo* SC. The segmentation outcome can also be used for co-registration and spatial normalization to a common space. Moreover, the fully-automated segmentation is useful for longitudinal studies, where the delineation of gray matter is time consuming [9].

While recent cervical cord cross-sectional area (CSA) segmentation methods have achieved near-human performance [31], the accurate segmentation of the GM remains a challenge [2]. The main properties that make the GM area difficult to segment are: inconsistent intensities of the surrounding tissues, image artifacts and pathology-induced changes in the image contrast [9].

Additional factors also contribute to the complexity of the GM segmentation task, such as lack of standardized datasets, differences in MRI acquisition protocols, different pixel sizes, different methods to acquire gold standard segmentations and different performance metrics to assess segmentation results [2]. Figure 4.1 features several examples of axial MRI acquired at different centers, demonstrating image variability due variable image acquisition systems and protocols.

Despite these difficulties, there have been major improvements in acquisition and analysis methods in recent years, making it possible to obtain reliable GM segmentations. From the acquisition standpoint, the advances in coil sensitivity [65], multi-echo gradient echo sequences [66], and phase-sensitive inversion recovery sequences [67] drastically improved the contrast-to-noise-ratio between the white and gray matter in the cord. From the analysis standpoint, the scientific community recently organized a collaboration effort called "Spinal Cord Gray Matter Segmentation Challenge" (SCGM Challenge) [2] to characterize the state-of-the-art and compare six independent developed methods [7, 10, 11, 15, 16, 68] on a public available standard dataset created through the collaboration of four internationally recognized spinal cord imaging research groups (University College London, Polytechnique Montreal, University of Zurich and Vanderbilt University), providing therefore a ground basis for method comparison that was previously unfeasible.

In the past few years, we have witnessed the fast and unprecedented development of Deep Learning [17] methods, that have not only achieved human-level performance but, in many cases, have surpassed it [69], even in health domain applications [70]. After the results presented in the seminal paper of the AlexNet [18], the Machine Learning community embraced the successful Deep Learning approach for Machine Learning and, consequently, many methods

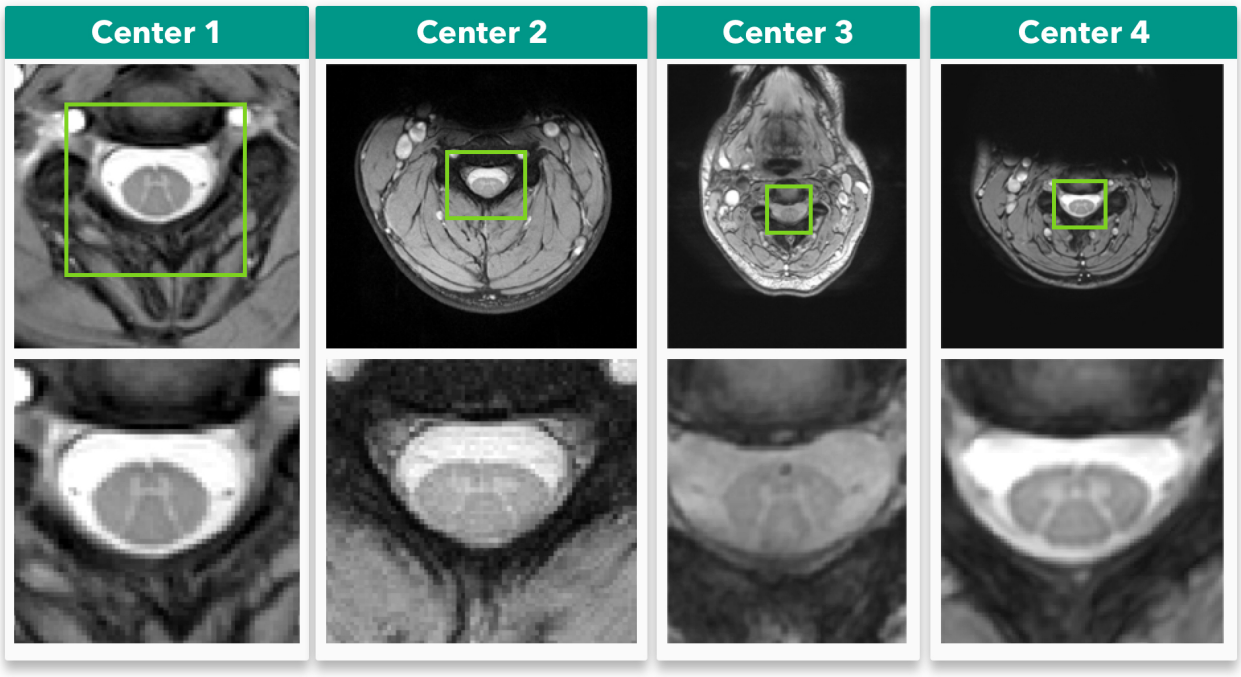


Figure 4.1 In vivo axial-slice samples from four centers (UCL, Montreal, Zurich, Vanderbilt) that collaborated to the SCGM Segmentation Challenge [2]. Top row: original MRI images. Bottom row: a crop of the spinal cord (green rectangle).

have been developed to since become the state-of-the-art and pervasive in many different fields such as image classification [19], image segmentation [20], speech recognition [21], natural language processing (NLP), among others.

Deep Learning is characterized by a major shift from traditional handcraft feature extraction to a hierarchical representation learning approach where multiple levels of automatically discovered representations are learned from raw data [17].

In a recent survey [23] of over 300 papers that used Deep Learning techniques for medical image analysis, the authors found that these techniques have spread throughout the entire field of medical image analysis, with a rapid increase in the number of publications between the years of 2015 and 2016. The survey also found that Convolutional Neural Networks (CNNs) were more prevalent in the medical image analysis, with Recurrent Neural Networks (RNNs) gaining more popularity.

Although the enormous success of Deep Learning has attracted a lot of attention from the research community, some challenges in the medical imaging domain remain open, such as data acquisition, which is usually very expensive and requires time-consuming annotation from image specialists to create the gold standards necessary for algorithm training and validation.

Standardized datasets remain also a major problem due to variability in equipment from different vendors, acquisition protocols/parameters/contrasts, especially in the MRI domain. Furthermore, data availability is limited due to concerns around ethics and regulations on patient data privacy [23].

In this work, we propose a new simple pipeline featuring an end-to-end learning approach for fully-automated spinal cord gray matter segmentation using a novel Deep Learning architecture based on the *Atrous* Spatial Pyramid Pooling (ASPP) [20, 71], where we achieved state-of-the-art results on many metrics in an *in vivo* independent dataset evaluation. We further demonstrate an excellent generalization on an *ex vivo* high-resolution acquisition dataset where only a few axial-slices were annotated to accurate segment an MRI volume with more than 4000 axial slices. Our proposed method is compared with the commonly used U-Net [64] architecture and with six other independently developed methods.

This work was implemented as the *sct_deepseg_gm* tool in the Spinal Cord Toolbox (SCT) [72] and is now freely available at SCT Github repository[1]. SCT is a comprehensive, free and open-source library of analysis tools for MRI of the spinal cord.

4.4 Related Work

Many methods for spinal cord segmentation were proposed in the past. Regarding the presence or absence of manual intervention, the segmentation methods can be separated into two main categories: semi-automated and fully-automated.

In the work [10], they propose a probabilistic method for segmentation called "Semi-supervised VBEM", whereby the observed MRI signals are assumed to be generated by the warping of an averagely shaped reference anatomy [2]. The observed image intensities are modeled as random variables drawn from a Gaussian mixture distribution, where the parameters are estimated using a variational version of the Expectation-Maximization (EM) [10] algorithm. The method can be used in a fully unsupervised fashion or by incorporating training data with manual labels, hence the semi-supervised scheme [2].

The SCT (Spinal Cord Toolbox) segmentation method [11], uses an atlas-based approach and was built based on a previous work [12] but with additional improvements such as the use of vertebral level information and linear intensity normalization to accommodate multi-site data [11]. The SCT approach first builds a dictionary of images using manual WM/GM segmentations after a pre-processing step, then the target image is also pre-processed and normalized, after that, the target image is projected into the PCA (Principal Component

[1] https://github.com/neuropoly/spinalcordtoolbox

Analysis) space of the dictionary images where the most similar dictionary slices are selected using an arbitrary threshold. Finally, the segmentation is done using label fusion between the manual segmentations from the dictionary images that were selected [2]. The SCT method is freely available as open-source software at `https://github.com/neuropoly/spinalcordtoolbox` [72].

In the work [7], a method called "Joint collaboration for spinal cord gray matter segmentation" (JCSCS) is proposed, where two existing label fusion segmentation methods were combined. The method is based on a multi-atlas segmentation propagation using registration and segmentation in 2D slice-wise space. In JCSCS, the “Optimized PatchMatch Label Fusion” (OPAL) [13] is used to detect the spinal cord, where the cord localization is achieved by providing an external dataset of spinal cord volumes and their associated manual segmentation [7], after that, the “Similarity and Truth Estimation for Propagated Segmentations” (STEPS) [14] is used to segment the GM in two steps, first the segmentation propagation, and then a consensus segmentation is created by fusing best-deformed templates (based on locally normalized cross-correlation) [7].

In the work [15], the Morphological Geodesic Active Contour (MGAC) algorithm uses an external spinal cord segmentation tool (“Jim”, from Xinapse Systems) to estimate the spinal cord boundary and a morphological geodesic active contour model to segment the gray matter. The method has five steps: first, the original image spinal cord is segmented with the Jim software and then a template is registered to the subject cord, after which the same transformation is applied to the GM template. The transformed gray matter template is then used as an initial guess for the active contour algorithm [15].

The "Gray matter Segmentation Based on Maximum Entropy" (GSBME) algorithm [2] is a semi-automatic, supervised segmentation method for the GM. The GSBME is comprised of three main stages. First, the image is pre-processed, in this step the GSBME uses the SCT [72] to segment the spinal cord using Propseg [31] with manual initialization, after which the image intensities are normalized and denoised. In the second step, the images are thresholded, slice by slice, using a sliding window where the optimal threshold is found by maximizing the sum of the GM and WM intensity entropies. In the final stage, an outlier detector discards segmented intensities using morphological features such as perimeter, eccentricity and Hu moments among others [2].

In the Deepseg approach [16], which builds upon the work [68], a Deep Learning architecture similar to the U-Net [64], where a CNN has a contracting and expanding path. The contracting path aggregates information while the expanding path upsamples the feature maps in order to achieve a dense prediction output. To recover spatial information loss, shortcuts are

added between contracting/expanding paths of the network. In Deepseg, instead of using upsampling layers like U-Net, they use an unpooling and "deconvolution" approach such as in the work [73]. The network architecture possesses 11 layers and is pre-trained using 3 convolutional restricted Boltzmann Machines [74]. Deepseg also uses a loss function with a weighted sum of two different terms, the mean square differences of the GM and non-GM voxels, thus balancing sensitivity and specificity [2]. Two models were trained independently, one for the full spinal cord segmentation and another for the GM segmentation.

We compare our method with all the aforementioned methods on the SCGM Challenge [2] dataset.

Methods and Materials

As in the *Related Work* section, the majority of the previously developed GM segmentation methods usually rely on registered templates/atlases, arbitrary distance and similarity metrics, and/or complex pipelines that are not optimized in an end-to-end fashion and neither efficient during inference time.

In this work, we focus on the development of a simple Deep Learning method that can be trained in an end-to-end fashion and that generalizes well even with a small subset of 2D labeled axial slices belonging to a larger 3D MRI volume.

4.4.1 Note on U-Nets

Many modern Deep Learning CNN classification architectures use alternating layers of convolutions and subsampling operations to aggregate semantic information and discard spatial information across the network, leading to certain levels of translation and rotation invariance that are important for classification. However, in segmentation tasks, a dense full-resolution output is required. In medical imaging, the most established architecture for segmentation is the well-known U-Net [64], where two distinct paths (encoder-decoder/contracting-expanding) are used to aggregate semantic information and recover the spatial information with the help of shortcut connections between the paths.

The U-Net architecture, however, causes a major expansion of the parameter space due to the two distinct paths that form the U-shape. As noted previously [75], the gradient flow in the high-level layers of the U-Nets (bottom of the U-shape) is problematic. Since the final low-level layers have access to the earlier low-level features, the network optimization will find the shortest path to minimize the loss, thus reducing the gradient flow in the bottom of the network.

By visualizing feature maps from the U-Net using techniques described in the work [76], we found that the features extracted in the bottom of the network were very noisy, while the features extracted in the low-level layers were the only ones that exhibited meaningful patterns. By removing the bottom layers of the network, we found that the network performed the same as, or occasionally better than, the deeper network.

4.4.2 Proposed method

Our method is based on the state-of-the-art segmentation architecture called "*Atrous* Spatial Pyramid Pooling" (ASPP) [20] that uses "*Atrous* convolutions", also called "dilated convolutions" [77]. We performed modifications to improve the segmentation performance on medical imaging by handling imbalanced data with a different loss function, and also by extensively removing decimation operations from the network such as pooling, trading depth (due to memory constraints) to improve the equivariance of the network and also parameter reduction.

Dilated convolutions allow us to exponentially grow the receptive field with a linearly increasing number of parameters, providing a significant parameter reduction while increasing the effective receptive field [78] and preserving the input resolution throughout the network, in contrast to wide stride convolutions where the resolution is lost. Dilated convolutions work by introducing "holes" [71] in the kernel as illustrated in Figure 4.2. For 1D signal $x[i]$, the $y[i]$ output of a dilated convolution with the dilation rate r and a filter $w[s]$ with size S is formulated as:

$$y[i] = \sum_{s=1}^{S} x[i + r \cdot s]w[s]. \tag{4.1}$$

The dilation rate r can also be seen as the stride over which the input signal is sampled [71]. Dilated convolutions, like standard convolutions, also have the advantage of being translationally equivalent, which means that translating the image will result in a translated version of the original input, as seen below:

$$f(g(x)) = g(f(x)) \tag{4.2}$$

Where $g(\cdot)$ is a translation operation and $f(\cdot)$ a convolution operation. However, since we don't need to introduce pooling to capture multi-scale features when using dilated convolutions, we can keep the translational equivariance property in the network, which is important for

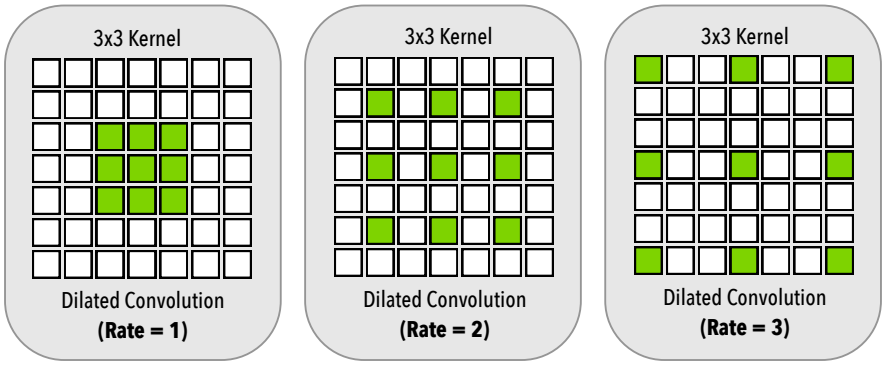


Figure 4.2 Dilated convolution. On the left, we have the dilated convolution with dilation rate $r = 1$, equivalent to the standard convolution. In the middle with have a dilation $r = 2$ and in the right a dilation rate of $r = 3$. All dilated convolutions have a 3x3 kernel size and the same number of parameters.

spatially dense prediction tasks, given that a translation of the input features should result also in an equivalent translation of outputs.

The overall proposed architecture can be seen in Figure 4.3. Our architecture works with 2D slice-wise axial images and is composed of (a) two initial layers of standard 3x3 convolutions, followed by (b) two layers of dilated convolutions with rate $r = 2$, followed by (c) six parallel branches with two layers each of a 1x1 standard convolution, 4 different dilated convolution rates (6/12/18/24) and a global averaging pooling that is repeated at every spatial position of the feature map. After that, the feature maps from the six parallel branches are concatenated and forwarded to (d) a block of 2 layers with 1x1 convolutions in order to produce the final dense prediction probability map. Each layer is followed by Batch Normalization [79] and Dropout [80] layers and we did not employ residual connections.

Figure 4.4 illustrates the pipeline of our training/inference process. An initial resampling step downsamples/upsamples the input axial slice images to a common pixel size space, then a simple intensity normalization is applied to the image, followed by the network inference stage.

Contrary to the task of natural images segmentation, the task of GM segmentation in medical imaging is usually very unbalanced. In our case, only a small portion of the entire axial slice encompasses the GM (the rest being comprised of other structures such as the white matter, cerebrospinal fluid, bones, muscles, etc.). Due to this imbalance, we employed a surrogate loss for the DSC (Dice Similarity Coefficient) called the Dice Loss, which is insensitive to imbalancing and was employed by many works in medical imaging [81, 82]. The Dice Loss can be formulated as:

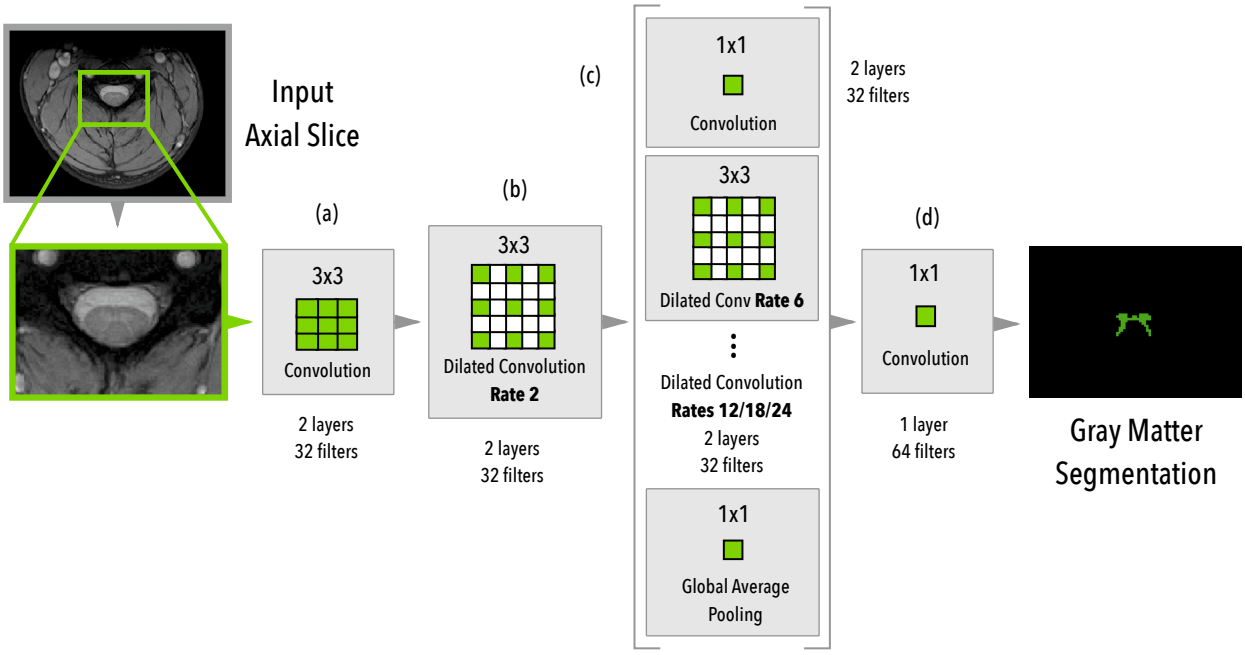


Figure 4.3 Architecture overview of the proposed method. The MRI axial slice is fed to the first block of 3x3 convolutions and then to a block of dilated convolutions (rate 2). Then, six parallel modules with different rates (6/12/18/24), 1x1 convolution, and a global average pooling are used in parallel. After the parallel modules, all feature maps are concatenated and then fed into the final block of 1x1 convolutions to produce the final dense predictions.

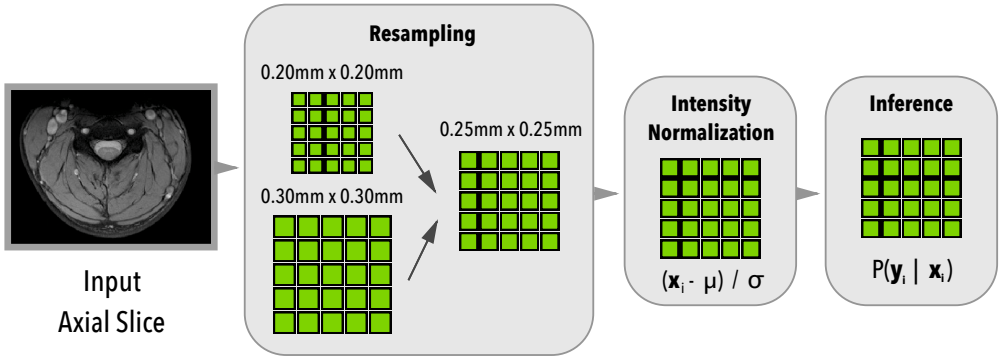


Figure 4.4 Architecture pipeline overview. During the first stage, input axial slices are resampled to a common pixel size space, then intensity is normalized, followed by the network inference.

$$\mathcal{L}_{dice} = -\frac{2\sum_{n=1}^{N} p_n r_n + \epsilon}{\sum_{n=1}^{N} p_n + \sum_{n=1}^{N} r_n + \epsilon} \tag{4.3}$$

Where p and r are the predictions and gold standard, respectively. The ϵ term is used to ensure loss stability by avoiding numerical issues. We experimentally found that the Dice Loss yielded better results when compared to the weighted cross-entropy (WCE) used by the original U-Net [64], which is more difficult to optimize due to the added weighting hyper-parameter.

Medical image datasets are usually smaller than natural image datasets by many orders of magnitude, therefore regularization and data augmentation is an important step. In this work, the following data augmentation strategies were applied: rotation, shifting, scaling, flipping, noise, and elastic deformation [83].

The main differences when we compare our proposed architecture with that of the work [20], are the following:

Initial pooling/decimation: our network does not use initial pooling layers as we found them detrimental to the segmentation of medical images;

Padding: we extensively employ padding across the entire network to keep feature map sizes fixed, trading depth to reduce memory usage of the network;

Dilation Rates: since we do not use initial pooling, we retain the parallel dilated convolution branch with the rate $r = 24$. As we found improvements by doing so, due to the large feature map size that doesn't cause filter degeneration as seen in [20];

Loss: contrary to natural images, our task of GM segmentation is highly unbalanced, therefore instead of using the traditional cross-entropy, we use the Dice Loss;

Data Augmentation: in this work we apply rotation, shifting, added channel noise, and elastic deformations [83], in addition to the scaling and flipping used previously [20].

Table 4.2 compares the setup parameters of our approach as well as the participant methods of the SCGM Segmentation Challenge [2].

U-Net architecture

For the U-Net [64] architecture model that was used for comparison, we employed a 14-layers network using standard 3x3 2D convolution filters with ReLU non-linearity activations [69].

Table 4.2 Parameters of each compared method. Time per slice is an estimated value, since different hardware were employed by the different techniques. Values replicated from the work [2].

Method	Init.	Training	External data	Time p/ slice
JCSCS	Auto.	No	Yes	4-5 min
DEEPSEG	Auto.	Yes (4 h)	No	<1 s
MGAC	Auto.	No	No	1 s
GSBME	Manual	Yes (<1 m)	No	5-80 s
SCT	Auto.	No	Yes	8-10 s
VBEM	Auto.	No	No	5 s
Proposed	Auto.	Yes (19 h)	No	<1 s

For a fair comparison, we used the same training protocol and loss function. For the data augmentation strategy, we employed a more aggressive augmentation due to overfitting issues with the U-Net (see the Discussion section). We also performed an extensive architecture exploration and used the best performing U-Net model architecture.

4.4.3 Datasets

In this subsection, we present the datasets used for evaluation in this work.

Spinal Cord Gray Matter Challenge

The Spinal Cord Gray Matter Challenge [2] (SCGM Challenge) dataset consists of 80 healthy subjects (20 subjects from each center). The demographics range from a mean age of 28.3 up to 44.3 years old. Three different MRI systems were used (Philips Achieva, Siemens Trio, Siemens Skyra) with different acquisition parameters based on a multi-echo gradient echo sequence. The voxel size range from 0.25x0.25x2.5 mm up to 0.5x0.5x5.0 mm. The dataset is split between training (40) and test (40) with the test set hidden. For each labeled slice in the dataset, 4 gold standard segmentation masks were produced by 4 independent expert raters (one per site). Examples of the datasets from each center are shown in Figure 4.1.

During the development of this work, we found some misclassified voxels in the training set. These issues were reported, however, for the sake of a fair comparison, all the evaluations done in this work used the original pristine training dataset.

Ex vivo high-resolution spinal cord

To evaluate our method on an *ex vivo* dataset, we used an MRI acquisition that was performed on an entire human spinal cord, from the pyramidal decussation to the *cauda equina* using a 7T horizontal-bore small animal MRI system.

MR images of the entire spinal cord were acquired in seven separate overlapping segments. The segment field of view was 8 x 2 x 2 cm with 1 cm of overlap on each end. Between each acquisition, the specimen was advanced precisely 7 cm through the magnet bore using a custom-machined gantry insert. T2*-weighted anatomic images were acquired using a 3D gradient echo sequence with an acquisition matrix of 1600 x 400 x 400, resulting in 50 micron isotropic resolution. Scan parameters included: TR = 50 ms, TE = 9 ms, flip angle = 60°, bandwidth = 100 kHz and number of averages = 1. Per-segment acquisition time was 2 hours 22 minutes, resulting in a total acquisition time of approximately 16 hours. Individual image segments were composed into a single volume using automated image registration and weighted averaging of overlapping segments.

Although the acquisition was obtained from a deceased adult male with no known history of neurologic disease, the review of images revealed a clinically occult SC lesion close to the 6th thoracic nerve root level, with imaging features suggestive of a chronic compressive myelopathy or possible sequela of a previous viral infection such as herpes zoster.

The volume is comprised of a total 4676 axial slices with 100 µm isotropic resolution.

The annotations (gold standard) for axial slices of this dataset were made by a researcher with the help of an expert radiologist. The annotation procedure was as follows: first, the contour of the GM was delineated using a gradient method from MIPAV [84] software. After that, a pixel-wise fine-tuning was performed using the fslview tool from FSL [85].

4.4.4 Training Protocol

Spinal Cord Gray Matter Challenge

The training protocol for the SCGM Challenge [2] dataset experiments are described in Table 4.3 and the data augmentation parameters are described in Table 4.4.

Contrary to the smooth decision boundaries characteristic of models trained using cross-entropy, the Dice Loss has the property of creating sharp decision boundaries and models with high recall rate. We found experimentally that thresholding the dense predictions with a threshold $\tau = 0.999$ provided a good compromise between precision/recall, however, no optimization was employed to choose the threshold τ value for the output predictions.

Table 4.3 Training protocol for the Spinal Cord Gray Matter Challenge dataset.

Resampling and Cropping	All volumes were resampled to a voxel size of 0.25x0.25 mm, the highest resolution found between acquisitions. All the axial slices were center-cropped to a 200x200 pixels size.
Normalization	We performed only mean centering and standard deviation normalization of the volume intensities.
Train/validation split	For the train/validation split, we used 8 subjects (2 from each site) for validation and the rest for training. The test set was defined by the challenge. We haven't employed any external data or used the vertebral information from the provided dataset. Only the provided GM masks were used for training/validation.
Batch size	We used a small batch size of only 11 samples.
Optimization	We used Adam [86] optimizer with a small learning rate $\eta = 0.001$.
Batch Normalization	We used a momentum $\phi = 0.1$ for BatchNorm due to the small batch size.
Dropout	We used a dropout rate of 0.4.
Learning Rate Scheduling	Similar to the work [20], we used the "poly" learning rate policy where the learning rate is defined by $\eta = \eta_{t_0} * (1 - \frac{n}{N})^p$ where η_{t_0} is the initial learning rate, N is the number of epochs, n the current epoch and p the power with $p = 0.9$.
Iterations	We trained the model for 1000 epochs (w/ 32 batches at each epoch).
Data augmentation	We applied the following data augmentations: rotation, shift, scaling, channel shift, flipping and elastic deformation [83]. The data augmentation parameters were chosen using random search. More details about the parameters of the data augmentation are presented in Table 4.4.

Table 4.4 Data augmentation parameters used during the training stage of the Spinal Cord Gray Matter Challenge dataset.

Augmentation	**Parameter**	**Probability**
Rotation (degrees)	[-4.6, 4.6]	0.5
Shift (%)	[-0.03, 0.03]	0.5
Scaling	[0.98, 1.02]	0.5
Channel Shift	[-0.17, +0.17]	0.5
Elastic Deformation [83]	$\alpha = 30.0$, $\sigma = 4.0$	0.3

Since the test dataset is hidden from the challenge participants, to evaluate our model we sent our produced test predictions to the challenge website for automated evaluation. Results are presented in Table 4.5 under the column "Proposed Method", alongside with the six other previously developed methods and 10 different metrics.

The training time on a single NVIDIA P100 GPU took approximately 19 hours using single-precision floating-point and TensorFlow 1.3.0 with cuDNN 6, while inference time took less than 1 second per subject.

Table 4.5 Comparison of different segmentation methods that participated in the SCGM Segmentation Challenge [2] against each of the four manual segmentation masks of the test set, reported here in the format: mean (std). For of fair comparison, the metrics are the same as used in the study [2] and the results from other methods are replicated here, where we have: Dice similarity coefficient (DSC), mean surface distance (MSD), Hausdorff surface distance (HSD), skeletonized Hausdorff distance (SHD), skeletonized median distance (SMD), true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), Jaccard index (JI) and conformity coefficient (CC). In bold font, we represent the best-obtained results on each metric. We also note that MSD, HSD, SHD and SMD metrics are in millimeters and that lower values mean better results.

	JCSCS	**DEEPSEG**	**MGAC**	**GSBME**	**SCT**	**VBEM**	**Proposed Method**
DSC	0.79 (0.04)	0.80 (0.06)	0.75 (0.07)	0.76 (0.06)	0.69 (0.07)	0.61 (0.13)	**0.85** (0.04)
MSD	0.39 (0.44)	0.46 (0.48)	0.70 (0.79)	0.62 (0.64)	0.69 (0.76)	1.04 (1.14)	**0.36** (0.34)
HSD	2.65 (3.40)	4.07 (3.27)	3.56 (1.34)	4.92 (3.30)	3.26 (1.35)	5.34 (15.35)	**2.61** (2.15)
SHD	1.00 (0.35)	1.26 (0.65)	1.07 (0.37)	1.86 (0.85)	1.12 (0.41)	2.77 (8.10)	**0.85** (0.32)
SMD	0.37 (0.18)	0.45 (0.20)	0.39 (0.17)	0.61 (0.35)	0.39 (0.16)	0.54 (0.25)	**0.36** (0.17)
TPR	77.98 (4.88)	78.89 (10.33)	87.51 (6.65)	75.69 (8.08)	70.29 (6.76)	65.66 (14.39)	**94.97** (3.50)
TNR	**99.98** (0.03)	99.97 (0.04)	99.94 (0.08)	99.97 (0.05)	99.95 (0.06)	99.93 (0.09)	99.95 (0.06)
PPV	81.06 (5.97)	**82.78** (5.19)	65.60 (9.01)	76.26 (7.41)	67.87 (8.62)	59.07 (13.69)	77.29 (6.46)
JI	0.66 (0.05)	0.68 (0.08)	0.60 (0.08)	0.61 (0.08)	0.53 (0.08)	0.45 (0.13)	**0.74** (0.06)
CC	47.17 (11.87)	49.52 (20.29)	29.36 (29.53)	33.69 (24.23)	6.46 (30.59)	-44.25 (90.61)	**64.24** (10.83)

Inter-rater variability as label smoothing regularization

The training dataset provided by the SCGM Challenge is comprised of 4 different masks that were manually and independently created by raters for each axial slice. As in the study [68], we used all the different masks as our gold standard. We also found that this approach shares the same principle of using label smoothing as seen in work [87].

Label smoothing is a mechanism that has the effect of reducing the confidence of the model by preventing the network from assigning a full probability to a single class, which is commonly evidence of overfitting. In the study [88], a link was found between label smoothing and the confidence penalty through the direction of the Kullback–Leibler divergence. Since the different gold standard masks for the same axial slices diverges usually only in the border of the GM, it is easy to see that this has a label smoothing effect on the contour of the GM, thereby encouraging the model to be less confident in the contour prediction of the GM, a kind of "empirical contour smoothing".

This interpretation suggests that one could also incorporate this contour smoothing by artificially adding label smoothing on the contours of the target anatomy, where raters usually disagree on the manual segmentation, leading to a potentially better model generalization on many different medical segmentation tasks where the contours are usually the region of raters disagreement.

We leave the exploration of this contour smoothing to future work.

Ex vivo high-resolution spinal cord

The training protocol for the *ex vivo* high-resolution spinal cord dataset experiments are described in Table 4.6 and the data augmentation parameters are described in the Table 4.7.

Table 4.6 Training protocol for the *ex vivo* high-resolution spinal cord dataset.

Cropping	All the slices were center-cropped to a 200x200 pixels size.
Normalization	We performed only mean centering and standard deviation normalization of the volume intensities.
Train/validation split	For the training set we selected only 15 evenly spaced axial slices out of 4676 total slices from the volume. For the validation set, we selected 7 (evenly spaced) axial slices and our test set was comprised of 8 axial slices (also evenly distributed across the entire volume).
Batch size	We used a small batch size of only 11 samples.
Optimization	We used Adam [86] optimizer with a small learning rate $\eta = 0.001$.
Batch Normalization	We used a momentum $\phi = 0.1$ for BatchNorm due to the small batch size.
Dropout	We used a dropout rate of 0.4.
Learning Rate Scheduling	Similar to the work [20], we used the "poly" learning rate policy where the learning rate is defined by $\eta = \eta_{t_0} * (1 - \frac{n}{N})^p$ where η_{t_0} is the initial learning rate, N is the number of epochs, n the current epoch and p the power with $p = 0.9$.
Iterations	We trained the model for 600 epochs (w/ 32 batches at each epoch).
Data augmentation	For this dataset, we used the following aforementioned augmentations: rotation, shift, scaling, channel shift, flipping and elastic deformation [83]. We didn't employed random search to avoid overfitting due to the dataset size. More details about the parameters of the data augmentation are presented in Table 4.7.

Table 4.7 Data augmentation parameters used during the training stage of the *ex vivo* high-resolution spinal cord dataset.

Augmentation	**Parameter**	**Probability**
Rotation (degrees)	[-5.0, 5.0]	0.5
Shift (%)	[-0.1, 0.1]	0.5
Scaling	[0.9, 1.1]	0.5
Channel Shift	[-0.3, +0.3]	0.5
Flipping	Horizontal	0.5
Elastic Deformation [83]	$\alpha = 30.0$, $\sigma = 4.0$	0.3

Like in the SCGM Segmentation task, we used a threshold $\tau = 0.999$ to binarize the prediction mask.

The training time on a single NVIDIA P100 GPU took approximately 2 hours using single-precision floating-point and TensorFlow 1.3.0 with cuDNN 6. While inference time took approximately 25 seconds to segment 4676 axial slices.

4.4.5 Data Availability

The SCGM Challenge dataset analyzed during the current study is available on the SCGM Challenge repository at `http://rebrand.ly/scgmchallenge`. The *ex vivo* dataset analyzed during the current study is not publicly available, but is available from the corresponding author on reasonable request.

4.5 Results

In this section, we discuss the experimental evaluation of the method in the presented datasets.

4.5.1 Spinal Cord Gray Matter Challenge

In this subsection, we show the evaluation on SCGM Challenge [2] dataset.

Qualitative Evaluation

In Figure 4.5, we show the segmentation output of our model in four different subjects, from acquisitions performed at the four different centers, on the test set of the SCGM Segmentation Challenge. The majority voting segmentation was taken from the study [2]. As we can see in Figure 4.5, our approach was able to capture many properties of the GM anatomy, providing good segmentations even in presence of blur, as seen in the samples from Site 1 and Site 3.

When compared with the segmentation results from Deepseg [16], that uses a U-Net like structure with pre-training and 3D-wise training, we can see that our method succeeds at segmenting the gray commissure of the GM structure, which was observed to pose a problem for Deepseg, as indicated in Figure 4 of the work [2].

Quantitative Evaluation

As we can see in Table 4.5 and Figure 4.6, our approach achieved state-of-the-art results in 8 out of 10 different metrics and surpassed 4 out of 6 previous methods on all metrics. A description of the metrics used in this work is given in Table 4.8.

We can also see that the Dice Loss is not only an excellent surrogate for the Dice Similarity Coefficient (DSC) but also a surrogate for distance metrics, as we note that our model not only achieved state-of-the-art results on overlap metrics (i.e. DSC) but also on distance and statistical metrics.

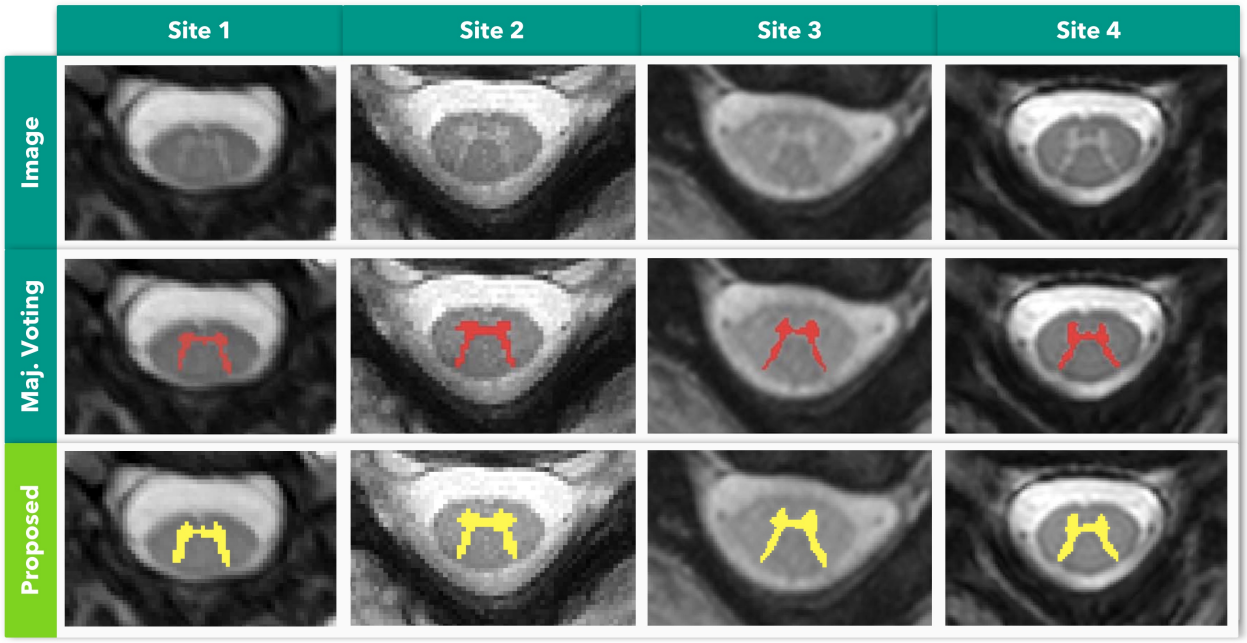


Figure 4.5 Qualitative evaluation of our proposed approach on the same axial slice for subject 11 of each site. From top to bottom row: input image, majority voting segmentation gold standard, and the result of our segmentation method. Adapted from the work [2].

The True Negative Rate (TNR) and Positive Predictive Value (PPV) or precision, were metrics for which the proposed method did not achieve the best results. However, we note that the TNR was very close to the results of other methods. We also hypothesize that the suboptimal results of the precision (PPV) are an effect of the sharp decision boundary produced by our model due to the Dice Loss. We are confident that the prediction threshold optimization can yield better results, however, this cost optimization would require further investigations.

When compared to the Deepseg [16] method, the only method using Deep Learning in the challenge, where an U-Net based architecture was employed, our proposed approach performed better in 8 out of 10 metrics, even though our method did not employ 3D convolutions, pre-training, or threshold optimization as was done in Deepseg [16].

4.5.2 *Ex vivo* high-resolution spinal cord

In this subsection, we show the evaluation on the *ex vivo* high-resolution spinal cord dataset.

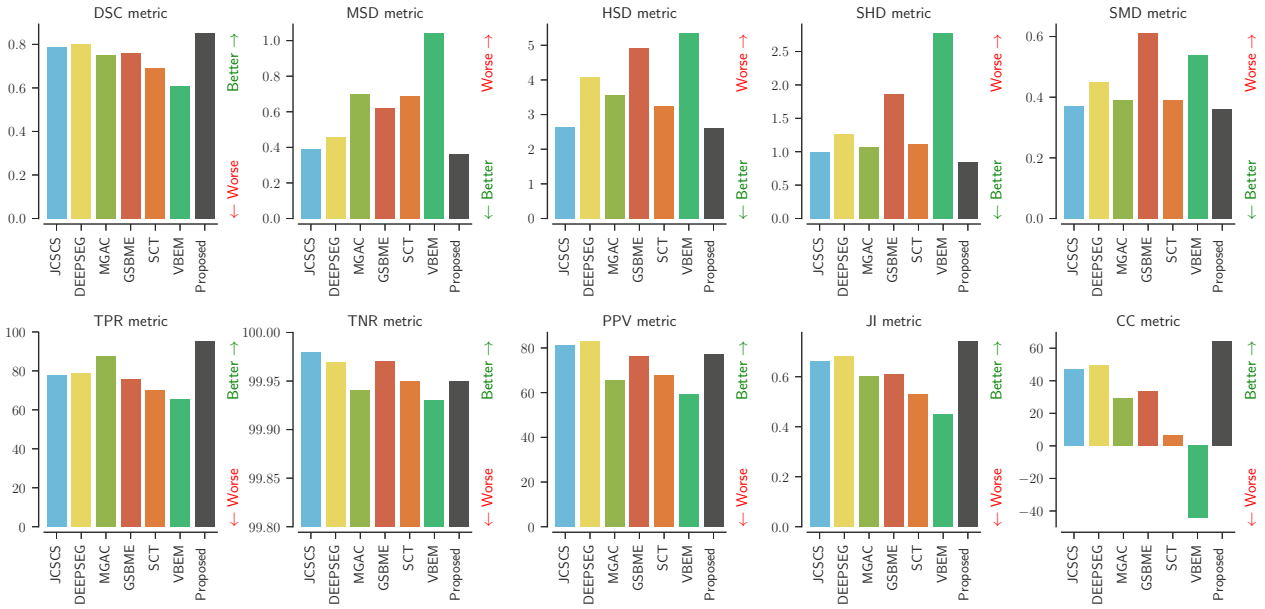


Figure 4.6 Test set evaluation results from the SCGM segmentation challenge [2] for each evaluated metric, with the Dice similarity coefficient (DSC), mean surface distance (MSD), Hausdorff surface distance (HSD), skeletonized Hausdorff distance (SHD), skeletonized median distance (SMD), true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), Jaccard index (JI) and conformity coefficient (CC). Our method is shown as "Proposed". Best viewed in color.

Table 4.8 Description of the validation metrics. Adapted from the work [2].

Metric Name	Abbr.	Range	Interpretation	Category
Dice Similarity Coefficient	DSC	0 - 1	Similarity between masks	Overlap
Jaccard Index	JI	0 - 100	Similarity between masks	Overlap
Conformity Coefficient	CC	<100	Ratio between mis-segmented and correctly segmented. Range can be between $(-\infty, 1]$ as defined in [3]	Overlap
Symmetric Mean Absolute Surface Distance	MSD	>0	Mean euclidean distance between mask contours (mean error)	Distance
Hausdorff Surface Distance	HSD	>0	Longest euclidean distance between mask contours (absolute error)	Distance
Skeletonized Hausdorff Distance	SHD	>0	Indicator of maximal local error	Distance
Skeletonized Median Distance	SMD	>0	Indicator of global errors	Distance
True Positive Rate or Sensitivity	TPR	0 - 100	Low values mean that method tends to under-segment	Statistical
True Negative Rate or Specificity	TNR	0 - 100	Quality of segmented background	Statistical
Positive Predictive Value or Precision	PPV	0 - 100	Low values mean that method tends to over-segment	Statistical

Qualitative Evaluation

In Figure 4.7, we show a qualitative evaluation of the segmentations produced by our method and those of U-Net model, contrasting the segmentations against the original and gold standard images.

As can be seen in the test sample depicted in the first column of Figure 4.7, the predictions of the U-Net "leaked" the gray matter segmentation into the cerebrospinal fluid (CSF) close to the dorsal horn (see green rectangle on first column), while our proposed segmentation was much more contained in the gray matter region only.

Also, in the third column of the Figure 4.7, the U-Net significantly oversegmented a large portion of the GM region, extending the segmentation up to the white matter close to the right lateral horn of the GM anatomy (see the green rectangle), while our proposed method performed well.

We also provide in Figure 4.8 a 3D rendered representation of the segmented gray matter using our method.

Quantitative Evaluation

As seen in Table 4.9, which shows the quantitative results of our approach, our method achieved better results on 6 out of 8 metrics. One of the main advantages that we can see from these results is that our method uses 6x fewer parameters than the U-Net architecture,

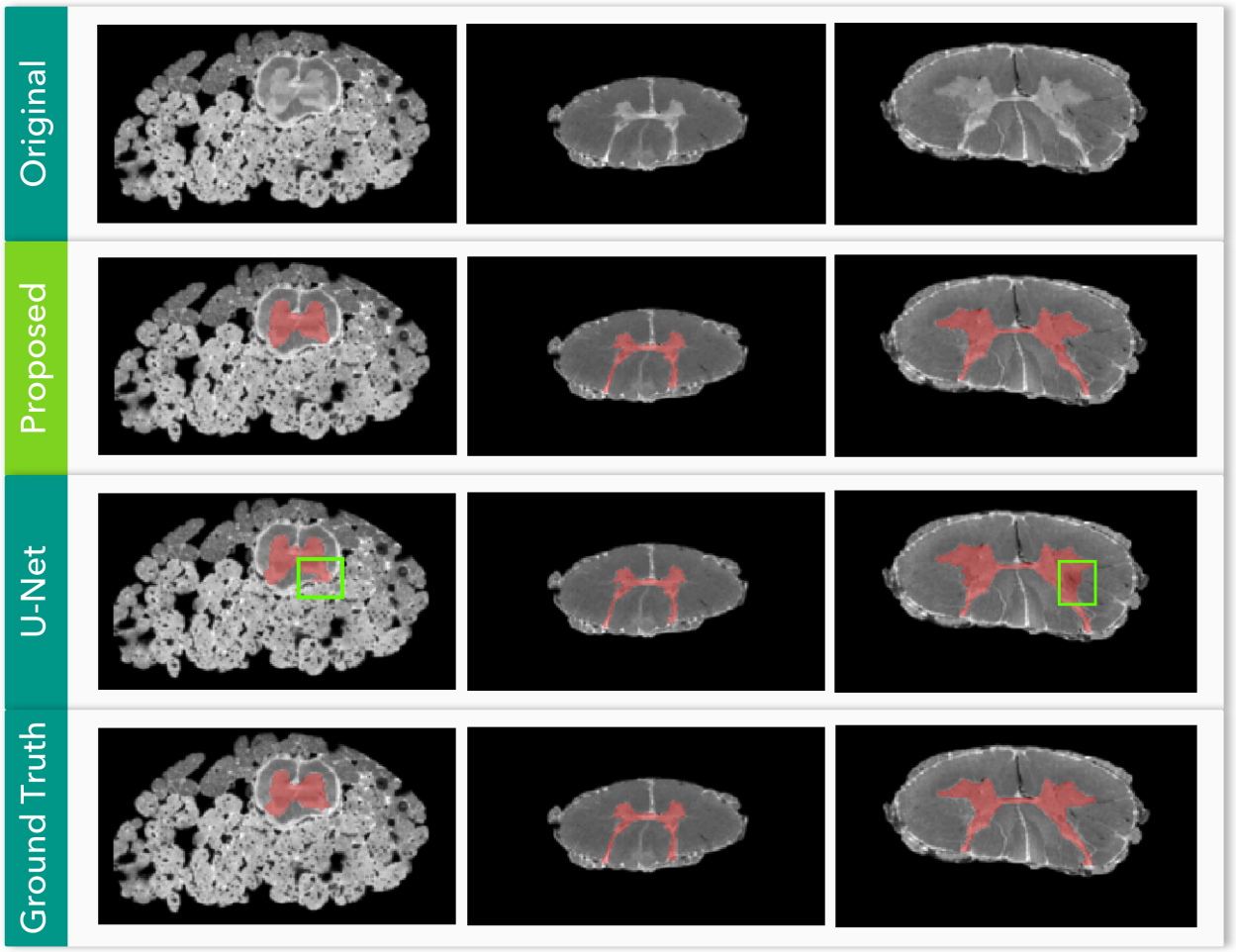


Figure 4.7 Qualitative evaluation of the U-Net and our proposed method on the *ex vivo* high-resolution spinal cord dataset. Each column represents a random sample of the test set (regions from left to right: sacral, thoracic, cervical). Green rectangles frame the oversegmentations of the U-Net model predictions.

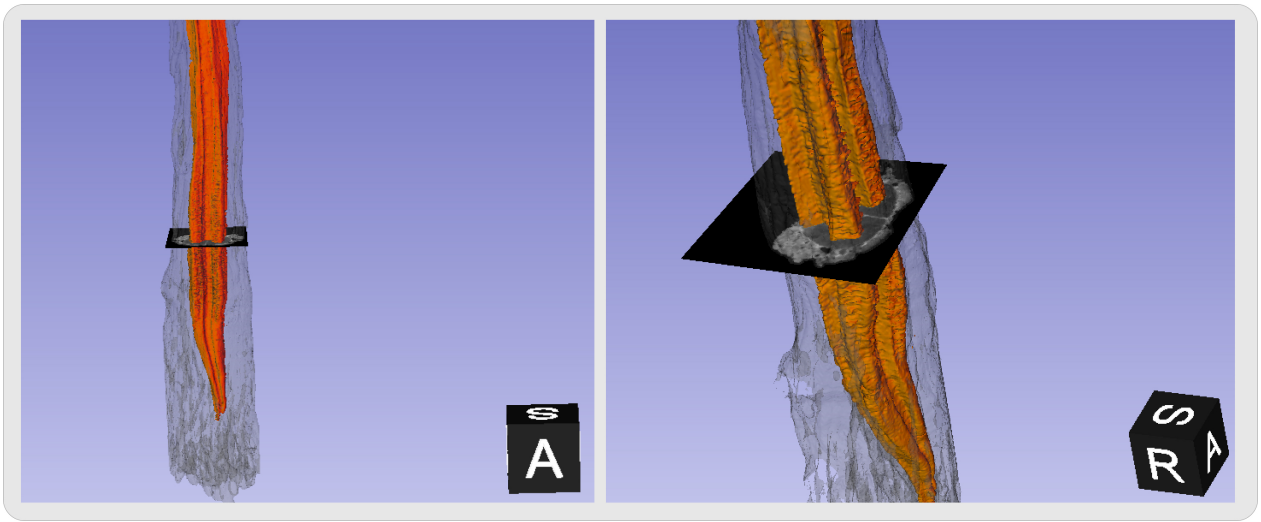


Figure 4.8 Lumbosacral region 3D rendered view of the *ex vivo* high-resolution spinal cord dataset segmented using the proposed method. The gray matter is depicted in orange color while the white matter and other tissues are represented in transparent gray color.

leading to less chance of overfitting and potentially better generalization.

Table 4.9 Quantitative metric results comparing a U-Net architecture and our proposed approach on the *ex vivo* high-resolution spinal cord dataset.

Metric name	**U-Net**	**Proposed**
Num. of Params.	776,321	**124,769**
Dice	0.9027 (0.07)	**0.9226** (0.04)
Mean Accuracy	**0.9626** (0.02)	0.9561 (0.03)
Pixel Accuracy	0.9952 (0.01)	**0.9968** (0.00)
Recall	**0.9287** (0.05)	0.9135 (0.06)
Precision	0.8831 (0.10)	**0.9335** (0.04)
Freq. Weighted IU	0.9913 (0.01)	**0.9938** (0.00)
Mean IU	0.9121 (0.06)	**0.9280** (0.04)

During the training of the two architectures (U-Net and our method), we noticed that even with a high dropout rate of 0.4, the U-Net was still overfitting, forcing us to use a more aggressive data augmentation strategy to achieve better results, especially for the shifting parameters of the data augmentation; we hypothesize that this is an effect of the decimation on the contracting path of the U-Net, that disturbs the translational equivariance property of the network, leading to a poor performance on segmentation tasks.

4.6 Discussion

In this work, we devised a simple, efficient and end-to-end method that achieves state-of-the-art results in many metrics when compared to six independently developed methods, as detailed in Table 4.5. To the best of our knowledge, our approach is the first to achieve better results in 8 out of the 10 metrics used in the SCGM Segmentation Challenge [2].

One of the main differences with other methods from the challenge is that our method employs an end-to-end learning approach, whereby the entire prediction pipeline is optimized using backpropagation and gradient descent. This is in contrast to the other methods, which generally employ separate registrations, external atlases/templates data and label fusion stages. As we can also see in Table 4.9, when we compare our method to the most commonly used method (U-Net) for medical image segmentation, our method provides not only better results for many metrics, but also a major parameter reduction (more than 6 times).

In the lens of Minimum Description Length (MDL) theory [89], which describes models as languages for describing properties of the data and sees inductive inference as finding regularity in the data [90], when two competing explanations for the data explains the data well, MDL will prefer the one that provides a shorter description of the data. Our approach using dilated filters provides more than 6x parameter reduction compared to U-Nets, but is also able to outperform other methods in many metrics, an evidence that the model is parameter-efficient and that it can capture a more compact description of the data regularities when compared with more complex models such as U-Nets.

The proposed approach has been tested on data acquired using Phase-Sensitive Inversion Recovery (PSIR) sequence, and as expected, the method did not work given that the model was not trained on PSIR samples and that these data exhibit very different contrast than the T2* images the model was trained on. This can be solved by aggregating the PSIR data into the existing datasets before training the model or even by training a specific model for PSIR data. Other techniques in the field of Domain Adaptation, which is currently a very active research area, could also be useful to generalize the existing model without even requiring annotated PSIR data, depending on the technique. These investigations will be the focus of follow-up studies.

Our approach is limited to 2D slices, however, the model does not restrict the use of 3D dilated convolutions and we believe that incorporating 3D context information into the model would almost certainly improve the segmentation results, however, at the expense of increased memory consumption.

Although we believe that this method can be extended to the GM segmentation in the presence

of many different neurological conditions such as multiple sclerosis, this will need to be further confirmed in follow-up studies in which patients would be included in the training/validation dataset.

We also believe that our method can be expanded to take advantage of semi-supervised learning approaches due to the strong smoothness assumption that holds for axial slices in most volumes, especially in *ex vivo* high-resolution spinal cord MRI.

4.7 Acknowledgments

Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging (JCA), the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [28826], the Fonds de Recherche du Québec - Nature et Technologies [2015-PR-182754], the Natural Sciences and Engineering Research Council of Canada [435897-2013], IVADO, TransMedTech and the Quebec BioImaging Network, ISRT and Wings for Life (INSPIRED project), NVIDIA Corporation for the donation of a GPU Titan X board, Compute Canada for the GPU cluster, Zhuoqiong Ren for the help with gray matter gold standard, Ryan Topfer for the paper review, organizers of the SCGM Segmentation Challenge and participant teams that invested so much effort in this challenge and the United States National Institutes of Health awards P41 EB015897 and 1S10OD010683-01 for funding the *ex vivo* study.

4.8 Author Contributions

C.S.P conceived the method, conducted the experiments, manual segmentations and wrote the paper. J.C.A. provided expert guidance and wrote the paper. E.C. provided the volume and information for the high-resolution *ex vivo* dataset. All authors reviewed the paper.

4.9 Additional Information

Competing interests: C.S.P., E.C. and J.C.A. declare no competing interests.

CHAPTER 5 ARTICLE 2: DEEP SEMI-SUPERVISED SEGMENTATION WITH WEIGHT-AVERAGED CONSISTENCY TARGETS

It is well known that Deep Learning models have a high sample complexity, which means that the required amount of data to learn a good predictor is usually large. Classical bounds don't explain generalization properties of these over-parametrized networks [91], and this is an active area of research in Deep Learning methods.

Although transfer learning [92] can be used to partially mitigate the problem in a small data regime scenario, which is the most common scenario in medical imaging, pre-training a model to do transfer learning usually involves the existence of a large and related dataset.

For natural images, models are usually pre-trained on ImageNet [37], however, in medical imaging, large datasets such as ImageNet are usually prohibitive due to the time-consuming task of annotating data (especially in segmentation tasks) and the cost to acquire data, not to mention regulations.

In the Chapter 4 we saw an excellent performance of the spinal cord gray matter segmentation on an annotated dataset, however, in practice, it is very common to have only a few samples labeled and many samples unlabeled. Recently, many methods were developed in which a learning algorithm can learn not only from labeled data but as well from unlabeled data, by taking leverage of semi-supervised learning techniques. However, most of these works, such as in [42], are only developed with classification tasks in mind and are usually evaluated on the natural image domain.

In this work, we extend the semi-supervised method from [42] to segmentation tasks and evaluate it on a realistic small data regime for the spinal cord gray matter segmentation task and show that significant performance improvements can be achieved even on a very small data regime.

My contribution to this work was to conceive the method, implement it, conduct the experiments and write the paper.

5.1 Article metadata

- **Title**: Deep semi-supervised segmentation with weight-averaged consistency targets

[1]NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, H3T 1J4, Canada.

[2]Functional Neuroimaging Unit, CRIUGM, Universite de Montreal, Montreal, QC, H3C 3J7, Canada.

- **Authors**: Christian S. Perone [1], Julien Cohen-Adad [1,2]
- **Publisher**: Lecture Notes in Computer Science (LNCS, volume 11045)
- **DOI**: 10.1007/978-3-030-00889-5
- **Citation**: Perone, C. S., & Cohen-Adad, J. (2018). Deep semi-supervised segmentation with weight-averaged consistency targets. DLMIA MICCAI, 1–8.

5.2 Abstract

Recently proposed techniques for semi-supervised learning such as Temporal Ensembling and Mean Teacher have achieved state-of-the-art results in many important classification benchmarks. In this work, we expand the Mean Teacher approach to segmentation tasks and show that it can bring important improvements in a realistic small data regime using a publicly available multi-center dataset from the Magnetic Resonance Imaging (MRI) domain. We also devise a method to solve the problems that arise when using traditional data augmentation strategies for segmentation tasks on our new training scheme.

5.3 Introduction

In the past few years, we witnessed a large growth in the development of Deep Learning techniques, that surpassed human-level performance on some important tasks [69], including health domain applications [70]. A recent survey [23] that examined more than 300 papers using Deep Learning techniques in medical imaging analysis, made it clear that Deep Learning is now pervasive across the entire field. In [23], they also found that Convolutional Neural Networks (CNNs) were more prevalent in the medical imaging analysis, with end-to-end trained CNNs becoming the preferred approach.

It is also evident that Deep Learning poses unique challenges, such as the large amount of data requirement, which can be partially mitigated by using transfer learning [93] or domain adaptation approaches [94], especially in the natural imaging domain. However, in medical imaging domain, not only the image acquisition is expensive but also data annotations, that usually requires a very time-consuming dedication of experts. Besides that, other challenges are still present in the medical imaging field, such as privacy and regulations/ethical concerns, which are also an important factor impacting the data availability.

According to [23], in certain domains, the main challenge is usually not the availability of the image data itself, but the lack of relevant annotations/labeling for these images. Traditionally,

systems like Picture Archiving and Communication System (PACS) [23], used in the routine of most western hospitals, store free-text reports, and turning this textual information into accurate or structured labels can be quite challenging. Therefore, the development of techniques that could take advantage of the vast amount of unlabeled data is paramount for advancing the current state of practical applications in medical imaging.

Semi-supervised learning is a class of learning algorithms that can take leverage not only of labeled samples but also from unlabeled samples. Semi-supervised learning is halfway between supervised learning and unsupervised learning [38], where the algorithm uses limited supervision, usually only from a few samples of a dataset together with a larger amount of unlabeled data.

In this work, we propose a simple deep semi-supervised learning approach for segmentation that can be efficiently implemented. Our technique is robust enough to be incorporated in most traditional segmentation architectures since it decouples the semi-supervised training from the architectural choices. We show experimentally on a public Magnetic Resonance Imaging (MRI) dataset that this technique can take advantage of unlabeled data and provide improvements even in a realistic scenario of small data regime, a common reality in medical imaging.

5.4 Semi-supervised segmentation using Mean Teacher

Given that the classification cost for unlabeled samples is undefined in supervised learning, adding unlabeled samples into the training procedure can be quite challenging. Traditionally, there is a dataset $\mathbf{X} = (x_i)_{i\in[n]}$ that can be divided into two disjoint sets: the samples $\mathbf{X}_l = (x_1, \ldots, x_l)$ that contains the labels $\mathbf{Y}_l = (y_1, \ldots, y_l)$, and the samples $\mathbf{X}_u = (x_{l+1}, \ldots, x_{l+u})$ where the labels are unknown. However, if the knowledge available in $p(x)$ that we can get from the unlabeled data also contains information that is useful for the inference problem of $p(y|x)$, then it is evident that semi-supervised learning can improve upon supervised learning [38].

Many techniques were developed in the past for semi-supervised learning, usually creating surrogate classes as in [95], adding entropy regularization as in [96] or using Generative Adversarial Networks (GANs) [97]. More recently, other ideas also led to the development of techniques that added perturbations and extra reconstruction costs in the intermediate representations [40] of the network, yielding excellent results. A very successful method called Temporal Ensembling [41] was also recently introduced, where the authors explored the idea of a temporal ensembling network for semi-supervised learning where the predictions of multiple

previous network evaluations were aggregated using an exponential moving average (EMA) with a penalization term for the predictions that were inconsistent with this target, achieving state-of-the-art results in several semi-supervised learning benchmarks.

In [98], the authors expanded the Temporal Ensembling method by averaging the model weights instead of the label predictions by using Polyak averaging [99]. The method described in [98] is a student/teacher model, where the student model architecture is replicated into the teacher model, which in turn, get its weights updated as the exponential moving average of the student weights according to:

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \tag{5.1}$$

where α is a smoothing hyperparameter, t is the training step and θ are the model weights. The goal of the student is to learn through a composite loss function with two terms: one for the traditional classification loss and another to enforce the consistency of its predictions with the teacher model. Both the student and teacher models evaluate the input data by applying noise that can come from Dropout, random affine transformations, added Gaussian noise, among others.

In this work, we extend the mean teacher technique [98] to semi-supervised segmentation. To the best of our knowledge, this is the first time that this semi-supervised method was extended for segmentation tasks. Our changes to extend the mean teacher [98] technique for segmentation are simple: we use different loss functions both for the task and consistency and also propose a new method for solving the augmentation issues that arises from this technique when used for segmentation. For the consistency loss, we use a pixel-wise binary cross-entropy, formulated as

$$\mathcal{C}(\theta) = \mathbb{E}_{x\in\mathbf{X}}\left[-y\log(p) + (1-y)\ log(1-p)\right], \tag{5.2}$$

where $p \in [0,1]$ is the output (after sigmoid activation) of the student model $f(x;\theta)$ and $y \in [0,1]$ is the output prediction for the same sample from the teacher model $f(x;\theta')$, where θ and θ' are student and teacher model parameters respectively. The consistency loss can be seen as a pixel-wise knowledge distillation [100] from the teacher model. It is important to note that both labeled samples from $\mathbf{X}_l$ and unlabeled samples from $\mathbf{X}_u$ contribute for the consistency loss $\mathcal{C}(\theta)$ calculation. We used binary cross-entropy, instead of the mean squared error (MSE) used by [98] because the binary cross-entropy provided an improved model

performance for the segmentation task. We also experimented with confidence thresholding as in [101] on the teacher predictions, however, it didn't improve the results.

For the segmentation task, we employed a surrogate loss for the Dice Similarity Coefficient, called the Dice loss, which is insensitive to imbalance and was also employed by [102] on the same segmentation task domain we experiment in this paper. The Dice Loss, computed per mini-batch, is formulated as

$$L(\theta) = -\frac{2\sum_i p_i y_i}{\sum_i p_i + \sum_i y_i}, \tag{5.3}$$

where $p_i \in [0,1]$ is the i^{th} output (after sigmoid non-linearity) and $y_i \in \{0,1\}$ is the corresponding ground truth. For the segmentation loss, only labeled samples from $\mathbf{X}_l$ contribute for the $\mathcal{L}(\theta)$ calculation. As in [98], the total loss used is the weighted sum of both segmentation and consistency losses. An overview detailing the components of the method can be seen in the Figure 5.1, while a description of the training algorithm is described in the Algorithm 5.4.1.

Algorithm 5.4.1 Semi-supervised segmentation algorithm.

Require: x_i = training samples
Require: y_i = labels for the labeled inputs $i \in \mathbf{Y}_l$
Require: t = global step (initialized with zero)
Require: $w(t)$ = consistency weight ramp-up function
Require: $f_\theta(\cdot)$ = neural network model with parameters θ
Require: $g_\phi(\cdot)$ = stochastic input augmentation procedure with parameters ϕ
for k in $[1, num\_epochs]$ **do**
  for each minibatch B **do**
    $z_{i\in B} \leftarrow f_\theta(g_\phi(x_{i\in B}))$ ▷ evaluate augmented inputs with student model
    $\tilde{z}_{i\in B} \leftarrow f_{\theta'}(g_{\phi'}(x_{i\in B}))$ ▷ teacher model evaluation w/ different perturbations
    $loss \leftarrow \mathcal{L}(z, y) + w(t)\frac{1}{|B|}\sum_{i\in B}\mathcal{C}(z_i, \tilde{z}_i)$ ▷ supervised and unsupervised loss components
    update θ using, e.g., Adam ▷ update student model parameters
    $t \leftarrow t + 1$ ▷ increment the global step counter
    $\theta'_t \leftarrow \alpha\theta'_{t-1} + (1-\alpha)\theta_t$ ▷ update teacher model parameters with using EMA
  end for
end for

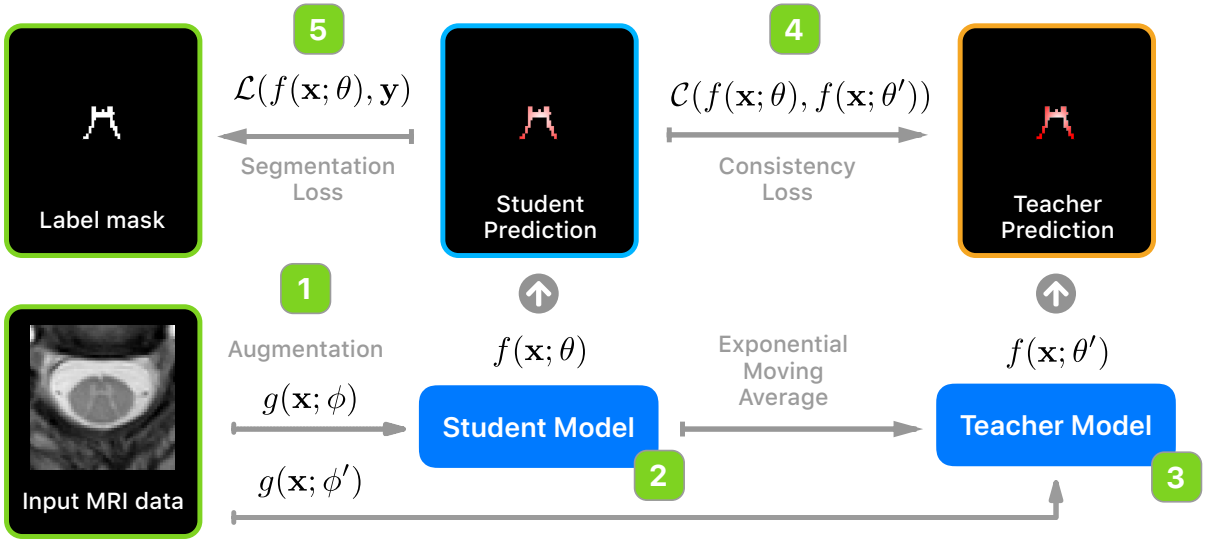


Figure 5.1 An overview with the components of the proposed method based on the mean teacher technique. **(1)** A data augmentation procedure $g(x;\phi)$ is used to perturb the input data (in our case, a MRI axial slice), where ϕ is the data augmentation parameter (i.e. $\mathcal{N}(0,\,\phi)$ for a Gaussian noise), note that different augmentation parameters are used for student and teacher models. **(2)** The student model. **(3)** The teacher model that is updated with an exponential moving average (EMA) from the student weights. **(4)** The consistency loss used to train the student model. This consistency will enforce the consistency between student predictions on both labeled and unlabelled data according to the teacher predictions. **(5)** The traditional segmentation loss, where the supervision signal is provided to the student model for the labeled samples.

5.4.1 Segmentation data augmentation

In segmentation tasks, data augmentation is very important, especially in the medical imaging domain where data availability is limited, variability is high and translational equivariance is desirable. Traditional augmentation methods such as affine transformations (rotation, translation, etc) that change the spatial content of the input data, as opposed to pixel-wise additive noise, for example, are also applied with the exact same parameters on the label to spatially align input and ground truth, both subject to a pixel-wise loss. This methodology, however, is unfeasible in the mean teacher training scheme. If two different augmentations (one for the student and another for the teacher) causes spatial misalignment, the spatial content between student and teacher predictions will not match during the pixel-wise consistency loss.

To avoid the misalignment during the consistency loss, such transformations can be applied with the same parametrization both to the student and teacher model inputs. However, this wouldn't take advantage of the stronger invariance to transformations that can be introduced

through the consistency loss. For that reason, we developed a solution that applies the transformations in the teacher in a delayed fashion. Our proposed method is based on the application of the same augmentation procedure $g(x;\phi)$ before the model forward pass only for the student model, and then after model forward pass in the teacher model predictions, making thus both prediction maps aligned for the consistency loss evaluation, while still taking leverage of introducing a much stronger invariance to the augmentation between student and teacher models. This is possible because we do backpropagation of the gradients only for the student model parameters.

5.5 Experiments

5.5.1 MRI Spinal Cord Gray Matter Segmentation

In this work, in order to evaluate our technique on a realistic scenario, we use the publicly available multi-center Magnetic Resonance Imaging (MRI) Spinal Cord Gray Matter Segmentation dataset from [2].

Dataset

The dataset is comprised of 80 healthy subjects (20 subjects from each center) and obtained using different scanning parameters and also multiple MRI systems. The voxel resolution of the dataset ranges from 0.25x0.25x2.5 mm up to 0.5x0.5x5.0 mm. A sample of one subject axial slice image can be seen in Figure 5.1. We split the dataset in a realistic small data regime: only 8 subjects are used as training samples, resulting in 86 axial training slices. We used 8 subjects for validation, resulting in 90 axial slices. For the unlabeled set we used 40 subjects, resulting in 613 axial slices and for the test set we used 12 subjects, resulting in 137 slices. All samples were resampled to a common space of 0.25x0.25 mm.

Network Architecture

To evaluate our technique, we used a very simple U-Net [64] architecture with 15 layers, Batch Normalization, Dropout and ReLU activations. U-Nets are very common in medical imaging domain, hence the architectural choice for the experiment. We also used a 2D slice-wise training procedure with axial slices.

Training procedure

For the supervised-only baseline, we used Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, mini-batch size of 8, dropout rate of 0.5, Batch Normalization momentum of 0.9 and L2 penalty of $\lambda = 0.0008$. For the data augmentation, we used rotation angle between $-4.5°$ and $4.5°$ and pixel-wise additive Gaussian noise sampled from $\mathcal{N}(0, 0.01)$. We used a learning rate $\eta = 0.0006$ given the small mini-batch size, also subject to a initial ramp-up of 50 epochs and subject to a cosine annealing decay as used by [98]. We trained the model for 1600 epochs.

For the semi-supervised experiment, we used the same parameters of the aforementioned supervised-only baseline, except for the L2 penalty of $\lambda = 0.0006$. We used an EMA $\alpha = 0.99$ during the first 50 epochs, later we change it to $\alpha = 0.999$. We also employed a consistency weight factor of 2.9 subject to a ramp-up in the first 100 epochs. We trained the model for 350 epochs.

Results

As we can see in Table 5.1, our technique not only improved the results on 5/6 evaluated metrics but also reduced the variance, showing a better regularized model in terms of precision/recall balance. The model also showed a very good improvement on overlapping metrics such as Dice and mean intersection over union (mIoU). Given that we exhausted the challenge dataset [2] to obtain the unlabeled samples, a comparison with [102] was unfeasible given different dataset splits. We leave this work for further explorations given that incorporating extra external data would also mix domain adaptation issues into the evaluation.

Table 5.1 Result comparison for the Spinal Cord Gray Matter segmentation challenge using our semi-supervised method and a pure supervised baseline. Results are 10 runs average with standard deviation in parenthesis where bold font represents the best result. Dice is the Dice Similarity Coefficient and mIoU is the mean intersection over union. Other metrics are self-explanatory.

	Dice	**mIoU**	**Accuracy**	**Precision**	**Recall**	**Specificity**
Supervised	67.915 (0.313)	53.679 (0.327)	99.745 (0.005)	57.948 (0.788)	**92.495 (0.907)**	99.775 (0.010)
Semi-supervised	**70.209 (0.229)**	**55.509 (0.253)**	**99.792 (0.003)**	**64.732 (0.773)**	86.112 (0.936)	**99.846 (0.006)**

5.6 Related Work

Only a few works were developed in the context of semi-supervised segmentation, especially in the field of medical imaging. Only recently, a U-Net was used as auxiliary embedding in [45], however, with focus on domain adaptation and using a private dataset.

In [46], they use a Generative Adversarial Networks (GAN) for the semi-supervised segmentation of natural images, however, they employ unrealistic dataset sizes when compared to the medical imaging domain datasets, along with ImageNet pre-trained networks.

In [103] they propose a technique using adversarial training, but they focus on the knowledge transfer between natural images with pixel-level annotation and weakly-labeled images with image-level information.

5.7 Conclusion

In this work we extended the semi-supervised mean teacher approach for segmentation tasks, showing that even on a realistic small data regime, this technique can provide major improvements if unlabeled data is available. We also devised a way to maintain the traditional data augmentation procedures while still taking advantage of the teacher/student regularization. The proposed technique can be used with any other Deep Learning architecture since it decouples the semi-supervised training procedure from the architectural choices.

It is evident from these results that future explorations of this technique can improve the results even further, given that even with a small amount of unlabeled samples, we showed that the technique was able to provide significant improvements.

5.8 Acknowledgements

Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging (JCA), the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [28826], the Fonds de Recherche du Québec - Nature et Technologies [2015-PR-182754], the Natural Sciences and Engineering Research Council of Canada [435897-2013], IVADO, TransMedTech, the Quebec BioImaging Network and NVIDIA Corporation for the donation of a GPU.

CHAPTER 6 ARTICLE 3: UNSUPERVISED DOMAIN ADAPTATION FOR MEDICAL IMAGING SEGMENTATION WITH SELF-ENSEMBLING

Empirical risk minimization (ERM), discussed in Section 2.2, has well-known learning guarantees [47] when training and test data come from the same domain or distribution. However, when models trained on a data distribution are applied on another domain with a different distribution during inference time, they often show a poor generalization.

In medical imaging, and especially in MRI, the variability present on the acquired images can be significant enough to make Deep Learning models generalize poorly when a different parametrization is used. A concrete example is the difference among T1 and T2 contrasts, but variability is also present among different machine vendors, natural anatomical differences, to name a few.

Although this is a very common problem in medical imaging, the fact remains usually ignored on the design of many experiments and on many challenges organized in the field, which contains often in multi-center studies, both in training and test data, samples coming from all centers. This is not an ideal evaluation of generalization because, in reality, these models will be applied on data coming from new centers, therefore, these evaluations will be over-optimistic and will not represent the performance of the model on a real scenario.

In the Section 2.2.3, we saw that the techniques that can make a model adapt to new unseen domains is called *domain adaptation*. In this work, we extend the aforementioned technique from semi-supervised learning called Mean Teacher [42] to perform unsupervised domain adaptation similar to the work [101], but extending it to segmentation tasks and evaluating it on multiple centers on a realistic evaluation scenario.

The main insight on the relationship between semi-supervised learning techniques and unsupervised domain adaptation is that given semi-supervised learning can change the decision boundary of a model, being sometimes even detrimental [44] for the semi-supervised learning scheme if data comes from a different domain. The same technique can be used with unlabeled data from another domain, in order to change the decision boundary towards the new unseen domain without requiring annotations.

My contributions to this work were the method conception, implementation, evaluation, and paper writing.

6.1 Article metadata

- **Title**: Unsupervised domain adaptation for medical imaging segmentation with self-ensembling
- **Authors**: Christian S. Perone [1], Pedro Ballester [2], Rodrigo C. Barros [2], Julien Cohen-Adad [1,3]
- **Publisher**: Submitted to Elsevier NeuroImage
- **DOI**: N/A, under review, arXiv ID: 1811.06042v1
- **Citation**: Perone, C. S., Ballester, P., Barros, R. C., & Cohen-Adad, J. (2018). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling.
- **Conference**: A short version of this work was presented at NIPS 2018 on the Medical Imaging workshop.

6.2 Abstract

Recent advances in deep learning methods have come to define the state-of-the-art for many medical imaging applications, surpassing even human judgment in several tasks. Those models, however, when trained to reduce the empirical risk on a single domain, fail to generalize when applied to other domains, a very common scenario in medical imaging due to the variability of images and anatomical structures, even across the same imaging modality. In this work, we extend the method of unsupervised domain adaptation using self-ensembling for the semantic segmentation task and explore multiple facets of the method on a small and realistic publicly-available magnetic resonance (MRI) dataset. Through an extensive evaluation, we show that self-ensembling can indeed improve the generalization of the models even when using a small amount of unlabelled data.

6.3 Introduction

In the past few years, the research community has witnessed the fast developmental pace of deep learning [17] approaches for unstructured data analysis, arguably establishing an

[1]NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada.
[2]Machine Intelligence and Robotics Research Group, School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil.
[3]Functional Neuroimaging Unit, CRIUGM, Universite de Montreal, Montreal, QC, Canada.

important scientific milestone. Deep neural networks constitute a paradigm shift from traditional machine learning approaches for unstructured data. Whereas the latter rely on hand-crafted feature engineering for improving learning over images, text, audio, and similarly unstructured inputs, deep neural networks are capable of automatically learning robust hierarchical features, in what is known as *representation learning*. Deep learning approaches have achieved human-level performance on many tasks and, indeed, sometimes even surpassing it in applications such as natural image classification [69], or arrhythmia detection [70].

Due to its popularity and compelling results in many domains, deep learning attracted a lot of attention from the medical imaging community. A recent survey by Litjens et al. [23] analyzed more than 300 medical imaging studies, and found that deep neural networks have become pervasive throughout the field of medical imaging, with a significant increase in the number of publications between 2015 and 2016. The survey also identified that the most addressed task is image segmentation, likely due to the importance of quantification of anatomical structures and pathologies [104] for disease diagnosis and prognosis, as opposed to less informative tasks such as classification of pathologies or detection of structures, which can be posed as a segmentation tasks as well, but not the opposite.

Deep neural networks are thus becoming the norm in medical imaging, though there are still several unsolved challenges that remain to be addressed. For instance, one of the most well-known problems is the high sample complexity, or how much data deep learning requires to accurately learn and perform well on unseen images, which is related to the concepts of model complexity and generalization, active areas of research in learning theory [105].

The large amount of required data to train deep neural networks can be partially mitigated with techniques such as transfer learning [93, 106]. However, transfer learning is problematic in medical imaging because a large dataset is still required so the models can benefit from the inductive transfer process. Unlike the case of natural images, where annotations can be easily provided by non-experts, medical images require careful and time-consuming analysis from trained experts such as radiologists.

Yet another challenge when deploying deep learning models to medical imaging analysis – and perhaps one of the most difficult to solve – is the so-called *data distribution shift*, wherein different imaging scenarios (e.g. parameter choices, different protocols) can result in vastly different data distributions, despite imaging a common object. Therefore, models trained under the empirical risk minimization (ERM) principle, might fail to generalize to other domains due to its strong assumptions. ERM is the statistical learning principle behind many machine learning methods, and it offers good learning guarantees and bounds if its

assumptions hold, such as the fact that the training and test datasets derive from similar domains. However, in practice, this assumption is often violated.

When a deep learning model that assumes independent and identically-distributed (iid) data is trained with images from one domain and is subsequently deployed on images from a different domain (e.g. distinct imaging center), that follow a distinct data distribution, its performance often degrades by a large margin. An example of domain shift can be seen in magnetic resonance imaging (MRI), where the same machine vendor, using the same protocol, and for the same subject, can nevertheless produce different images. Variability tends to be even greater between different centers where machine vendor, software versions, radio-frequency coils and sequence parameters (e.g., slice positioning, image resolution) often vary. Figure 6.1 illustrates those inter-center differences in data distribution, based on data from the Gray Matter (GM) segmentation challenge [2]. Figure 6.2 illustrates the associated voxel intensity distribution for the same dataset.

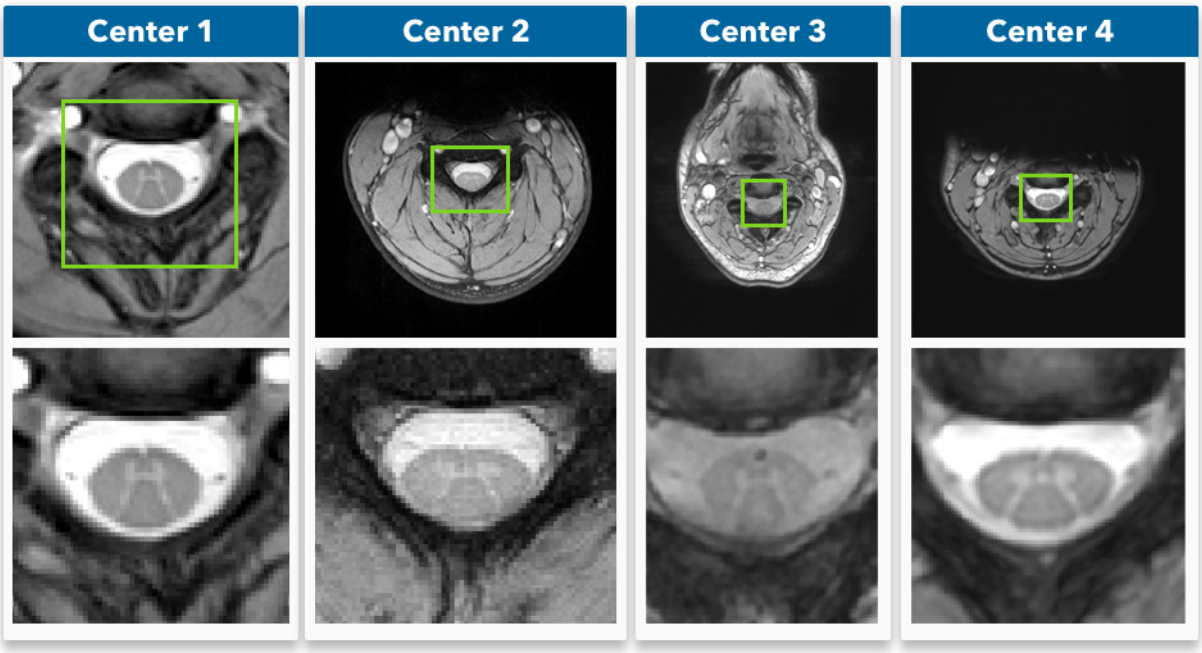


Figure 6.1 Samples of axial MRI from four different centers (UCL, Montreal, Zurich, Vanderbilt) that participated in the SCGM Segmentation Challenge [2]. **Top row**: original MRI images. **Bottom row**: crop of the spinal cord (green rectangle). Reproduced from [107]. Best viewed in color.

Although this distribution shift is common in medical imaging, the problem is surprisingly ignored during the design of many different challenges in the field. It is common to have the same domain data (same machine, protocol, etc.) on both training and test sets. However, this

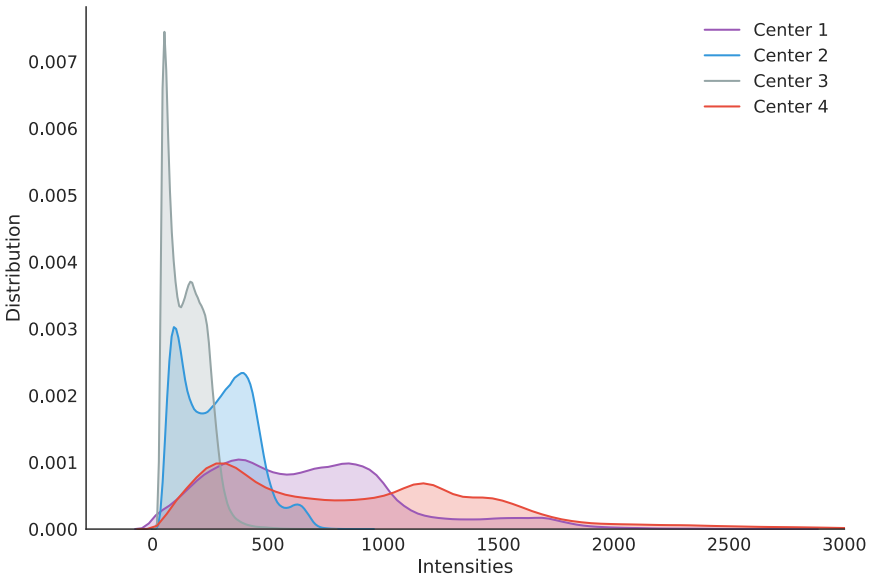


Figure 6.2 MRI axial-slice pixel intensity distribution from four different centers (UCL, Montreal, Zurich, Vanderbilt) that collaborated to the SCGM Segmentation Challenge [2].

homogeneous data split often does not represent the reality and in many cases will produce over-optimistic evaluation results. In practice, it is rare to have labeled data available from a new center before training a model, hence it is common to use a pre-trained model from a different domain on completely different data. Therefore, it is paramount to have a proper evaluation avoid contaminating the test set with data from the same domain that is present in the training set. Incurring the risk of the detrimental effects of inadequate evaluations [108]. The name given to learn a classifier model or any other predictor with a shift between the training and the target/test distributions is known as "domain adaptation" (DA). In this work we expand upon a previously-developed method [109] for DA based on the Mean Teacher [110] approach, to segmentation tasks, the most addressed task in medical imaging.

We provide the following contributions: (i) we extend the unsupervised DA method using self-ensembling for the semantic segmentation task; to the best of our knowledge, this is the first time this method is used for semantic segmentation in medical imaging; (ii) we explore some model components such as different consistency losses, and evaluate the performance of our method on a series of experiments using a realistic small MRI dataset; (iii) we perform an ablation experiment to provide strong evidence that unlabeled data is responsible for the observed performance improvement, ruling out the effects of the exponential moving average; (iv) we provide visualizations to derive insight into the model dynamics of the unsupervised DA task.

This paper is organized as follows. In Section 6.4 we present related work, in Section 6.5 we give a brief treatment to the unsupervised DA task and its connection to semi-supervised learning. In Section 6.6 we describe our method in terms of model architecture and corresponding design decisions. In Section 6.7 we describe the dataset used in our experiments and how we performed the data split for the DA scenario. In Section 6.8 we provide the experiment results, followed by an ablation study in Section 6.9. In Section 6.10 we provide visual insights from the adaptation dynamics of the model for multiple domains. Finally, in Section 6.11 we discuss our findings and limitations of our work. In the spirit of open science and reproducibility, we also provide more information regarding data and source-code availability in Section 6.12.

6.4 Related work

Deep learning methods for medical imaging has become a popular research focus in recent years [23]. Before the development of deep learning models, initial work was focused mostly on patch-based [111] segmentation. With the growing interest in deep learning for computer vision, the first attempts using Convolutional Neural Networks (CNNs) for image segmentation processed image patches through a sliding window, to yield segmented patches, which were then stitched together to yield the final segmented image [112]. The main drawbacks of this approach are computational cost (i.e., several forward passes are required to produce the segmented images) and inconsistency in predictions, the latter of which can be fixed or partially mitigated by overlapping sliding windows, depending on the network architecture.

Though patch-wise methods continue to be actively researched [113] and have led to several advances in segmentation [112], presently, the most common deep architecture for segmentation is or is based on the so-called Fully Convolutional Network (FCN) [114]. This architecture is solely based on convolutional layers with the final result not depending on the use of fully-connected layers. FCNs can provide a fully-segmented image within a single forward step, and with variable output size depending on the size of the input tensor. One of the most well-known FCNs for medical imaging is the U-net [64], which combines convolutional, downsampling, and upsampling operations with skip non-residual connections. In this work we used the U-Net architecture, although the proposed framework is decoupled from the choice of network architecture, as further discussed in Section 6.6.3.

Deep Domain Adaptation (DDA), which is a field unrelated in essence to medical imaging, has been widely studied in the recent years [115]. We can divide the literature on DDA as follows: (i) methods based on building domain-invariant feature spaces through auto-encoders [116], adversarial training [117], GANs [118, 119], or disentanglement strategies [120, 121]; (ii) methods based on the analysis of higher-order statistics [122, 123]; (iii) methods based on

explicit discrepancy between source and target domains [124]; and (iv) self-ensembling methods based on implicit discrepancy [109, 110].

In [118], the authors trained GANs with cycle-consistent loss functions [125] to remap the distribution from the source to the target dataset, thereby creating target domain specific features for completing the task. In [119], GANs were employed as a means of learning aligned embeddings for both domains. Similarly, disentangled representations for each domain have been proposed [120, 121] with the goal of generating a feature space capable of separating domain-dependent and domain-invariant information.

In [122], the authors proposed to change parameters of the neural network layers for adapting domains by directly computing or optimizing higher-order statistics. More specifically, they proposed an alternative for batch normalization called Adaptive Batch Normalization (AdaBN) that computes different statistics for the source and target domains, hence creating domain-invariant features that are normalized according to the respective domain. In a similar fashion, Deep CORAL [123] provides a loss function for minimizing the covariances between target and source domain features.

Discrepancy-based methods pose a different approach to DDA. By directly minimizing the discrepancy between activations from the source and target domain, the network learns to generate reasonable predictions while incorporating information from the target domain. The seminal work of Tzeng et al. [124] directly minimizes the discrepancy between a specific layer with labeled samples from the source set and unlabeled samples from the target set.

Implicit discrepancy-based methods such as self-ensembling [109] have become widely used for unsupervised domain adaptation. Self-ensembling is based on the Mean Teacher network [110], which was first introduced for semi-supervised learning tasks. Due to the similarity between unsupervised domain adaptation and semi-supervised learning, there are very few adjustments that need to be made to employ the method for the purposes of DDA. Mean Teacher optimizes a task loss and a consistency loss, the latter minimizing the discrepancy between predictions on the source and target dataset. We further detail how Mean Teacher works in Section 6.6.1.

There are a few studies that report results of using different data domains for medical imaging by making use of the unsupervised domain adaptation literature. The work [126] discusses the impact of deep learning models across different institutions, showing a statistically significant performance decrease in cross-institutional train-and-test protocols. A few studies have applied domain adaptation to medical imaging directly by using adversarial training [127–132], with some studies using generative models to augment training [133, 134]. Nevertheless, to the best of our knowledge, this present work is the first to address the problem of domain shift in medical image segmentation by extending the unsupervised DA self-ensembling method to

semantic segmentation tasks.

6.5 Semi-supervised learning and unsupervised domain adaptation

A common approach for improving training when few labeled examples are available is semi-supervised learning, which is defined as follows: given a labeled dataset with distribution $P(X_l)$ and unlabeled data with distribution $P(X_u)$, learn from both labeled and unlabeled data in order to improve a supervised learning task (say, classification) or an unsupervised learning task (say, clustering).

Semi-supervised learning methods tend to perform well when unlabeled data actually come from the same distribution as the labeled data. This allows the learning algorithm to leverage its knowledge using unlabeled data, which usually represents the majority of samples. As promising as semi-supervised learning is, the assumption that the distribution of unlabeled data $P(X_u)$ is similar to $P(X_l)$ often fails in real-world applications. We refer the reader to a thorough evaluation of semi-supervised learning methods and their limitations in [135].

It often happens that models are applied in situations that are largely different from those in which they were originally trained. Examples include different weather conditions for outdoor activity recognition, or different cities for training driverless vehicles. Those changes in scenario shift the data distribution $P(X)$, reducing the quality of the predictions in cases where the model was not properly adapted to the desired condition.

The difference between the distributions from the examples used in training and test sets is called *domain shift*. Consider a source dataset with input distribution $P(X_s)$ and label distribution $P(Y|X_s)$, as well as a target dataset with input distribution $P(X_t)$ and labels $P(Y|X_t)$, $P(X_s) \neq P(X_t)$. Domain adaptation can be addressed via a supervised approach where labeled data from the target domain is available, or via unsupervised learning where only unlabeled data is available for the target domain.

When a method addresses the problem of domain adaptation using unlabeled data for the target domain, which is the most common and useful scenario, the task at hand is called *unsupervised domain adaptation*. Unsupervised domain adaptation methods assume that distributions $P(X_s)$, $P(Y|X_s)$ and $P(X_t)$ are available, while $P(Y|X_t)$ is not. In other words, only the source dataset provides labeled examples. Hence, the task is to leverage knowledge from the target domain using the unlabeled data available in $P(X_t)$.

6.6 Method

This section details the base domain adaptation methods that we used for the medical image application. We further discuss the changes that are needed to enable unsupervised domain adaptation for segmentation tasks, as opposed to the typical classification scenario.

6.6.1 Self-ensembling and mean teacher

Self-ensembling was originally conceived as a viable strategy for generating predictions on unlabeled data [136]. The original paper proposes two different models for self-ensembling. The first model, called Π, employs a consistency loss between predictions on the same input. Each input from a batch is passed twice through a neural network, each time with distinct augmentation parameters, to yield two different predictions. A squared difference between those predictions is minimized along with the cross-entropy for the labeled examples. The second model, called *temporal ensembling*, works under the assumption that as the training progresses, averaging the predictions over time on unlabeled samples may contribute to a better approximation of the true labels. This pseudo-label is then considered as a target during training. The squared difference between the averaged predictions and the current one is minimized along with the cross-entropy for labeled examples. The network performs the exponential moving average (EMA) to update the generated targets at every epoch:

$$f'(x)_t = \alpha f'(x)_{t-1} + (1-\alpha) f(x)_t \tag{6.1}$$

Where t is the step, x is the data, $f(\cdot)$ is the network and α is a momentum term that controls how far the ensemble reaches training history data.

Self-ensembling was extended to directly combine model weights instead of predictions. This adaptation is called the Mean Teacher [110] model. Considering Eq. (6.1) for updating the target pseudo-labels, Mean Teacher updates the model weights at each step, thus generating a slightly improved model compared to the model without the EMA, a framework which is linked to the Polyak-Ruppert Averaging [137, 138]. In this scenario, the EMA model was named teacher, and the standard model, student. The update function is as follows:

$$\theta'_t = \alpha \theta'_{t-1} + (1-\alpha)\theta_t \tag{6.2}$$

where θ are the model parameters, t is the step and α is the hyperparameter regulating the importance of the current model's weights with respect to previous models. The best

results are achieved when α is increased later on during training, causing the model to forget more about the parameters during earlier stages of training than later when the network is performing better.

Each training step involves a loss component for both labeled and unlabeled data. All samples from a batch are evaluated by both the student and teacher models, with their respective predictions compared via the consistency loss. The labeled data, however, is also compared to its ground truth, as traditionally performed in segmentation tasks, in what we call the task loss:

$$J(\theta) = J_{task}(\theta) + \gamma J_{consistency}(\theta) + \lambda R(\theta) \tag{6.3}$$

where γ and λ are the Lagrange multipliers that represent, respectively, the consistency and regularization weights. The γ hyperparameter was empirically found to improve results when it varied through time, given that in the earlier training steps the network continues to generate poor results. The consistency weight follows a sigmoid ramp-up saturating at a given user-defined value.

Mean Teacher follows the dynamics of model distillation [139]. In this scenario, a trained model is used for predicting instances and its output is used as labels for another, smaller model. This is considered a good practice as soft labels tend to better represent the characteristics of the classes (e.g., the representation distance between a Siberian Husky and an Alaskan Malamute should arguably be smaller than the distance between a Siberian Husky and a Persian Cat). Unlike traditional distillation formulations, the Mean Teacher framework also uses the teacher model to generate labels for unlabeled data and represents a model of the same size that is simultaneously updated during training.

The Mean Teacher framework was also extended for unsupervised domain adaptation in [109]. Among the proposed changes, the authors modified the data batches such that each batch consists of images from both the source and target domains. At each step, the student model evaluates images from the source domain and computes derivatives via a task loss based on the ground truth. The target domain images, which are unlabeled, are used to compute the consistency loss by comparing predictions from both student and teacher models. It differs from its original formulation in that the teacher model only has access to unlabeled examples (in this case, examples from the target domain). Each loss function is thus responsible for improving learning at a single domain. The task loss is evaluated by comparing the predictions against the ground truth for the labeled examples (source domain). For the consistency loss, MSE is often used to evaluate the predictions from both student and teacher models for the unlabeled examples (target domain).

6.6.2 Adapting mean teacher for segmentation tasks

Both the original and adapted Mean Teacher versions for unsupervised domain adaptation rely on the cross-entropy classification cost. Given that we are not dealing with classification but with a segmentation task, we need to minimize a different loss function that takes into consideration the particularities of that task. Originally proposed in [140], the Dice loss generates reliable segmentation predictions due to its low sensitivity to class imbalance:

$$J_{task}(\theta) = -\frac{2 * \sum_i^N p_i g_i}{\sum_i^N p_i + \sum_i^N g_i} \tag{6.4}$$

where p_i and g_i are flattened predictions and ground truth values for an instance, respectively. Dice was kept as the task loss for both baseline and adaptation experiments. Note that the dice loss is computed for the entire batch at once, unlike the typical strategy of averaging when using cross-entropy, for instance.

A second problem when training the student and teacher models for segmentation tasks is the inconsistency introduced between training samples of the student and teacher models when a spatial transformation (e.g., translation, rotation, or any similar spatial transformation for the purpose of data augmentation) is applied with different parameters to both inputs of the teacher and student models. To solve that problem we used the same approach employed by [141] as shown in Figure 6.5. The spatial transformation $g(x; \phi)$, where x is the input data and ϕ are the transformation parameters (i.e., rotation angle), is applied to the student model before feeding data into the model. For the teacher model, the same transformation $g(x; \phi)$ is applied to the predictions of the teacher model, causing both predictions to be aligned for the consistency loss. This framework is possible because backpropagation only occurs for the student model and therefore there is no need for differentiation on the delayed augmentation of the teacher model. The proposed method is illustrated in Figure 6.4. Examples of images after data augmentation and their respective compensated ground truth are shown in Figure 6.3.

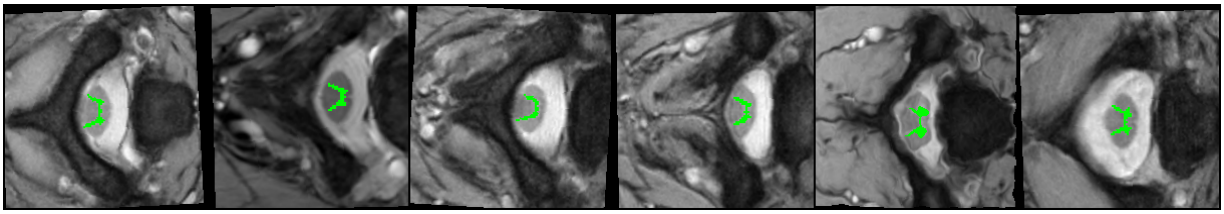


Figure 6.3 Data augmentation result of random MRI axial-slices samples from the SCGM Segmentation Challenge [2]. The ground truth is shown in green with the same transformation parameters applied. Best viewed in color.

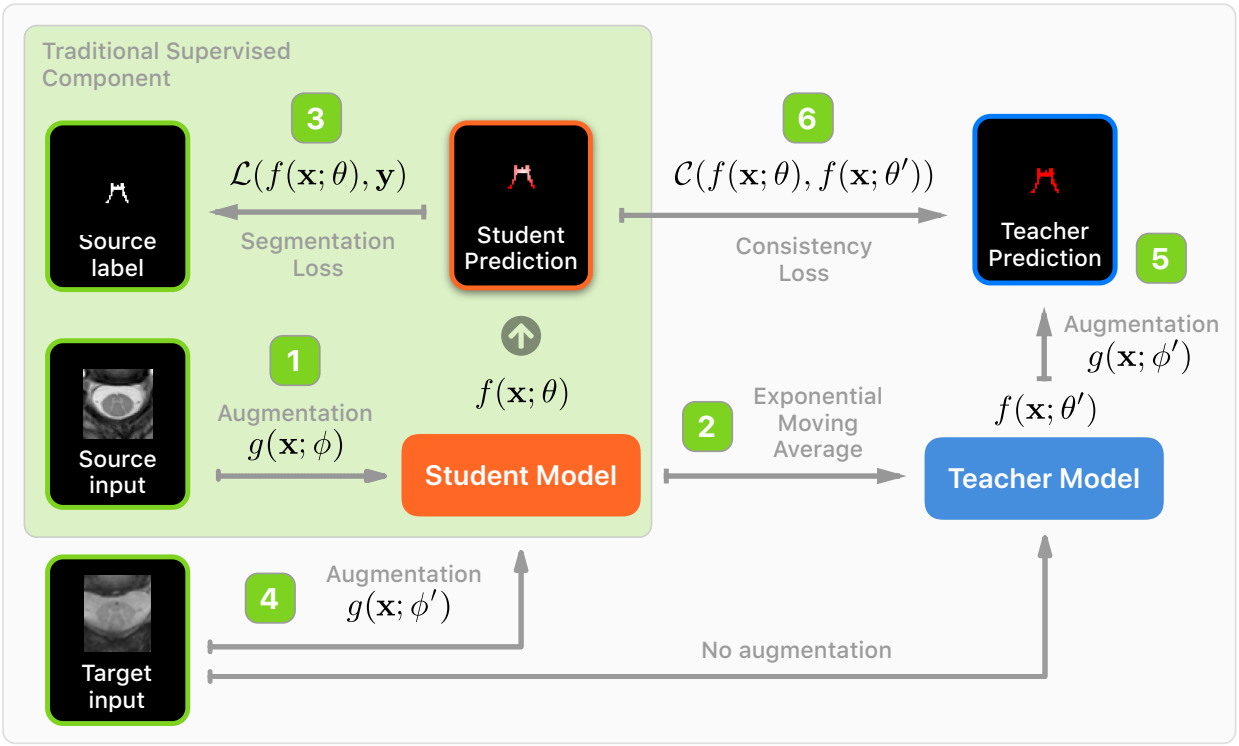


Figure 6.4 Overview of the proposed method. The green panel represents the traditional supervision framework. **(1)** The source domain input data is augmented by the $g(x;\phi)$ transformation and fed into the student model. **(2)** The teacher model parameters is updated with an exponential moving average (EMA) from the student weights. **(3)** The traditional segmentation loss, where the supervision signal is provided with the source domain labels. **(4)** The input unlabeled data from the target domain is transformed with $g(x;\phi')$ before the student model forward pass (note the different parametrization ϕ'). **(5)** The teacher model prediction is transformed with $g(x;\phi')$ (same transformation as in Step 4). **(6)** The consistency loss, which enforces consistency between student and teacher predictions.

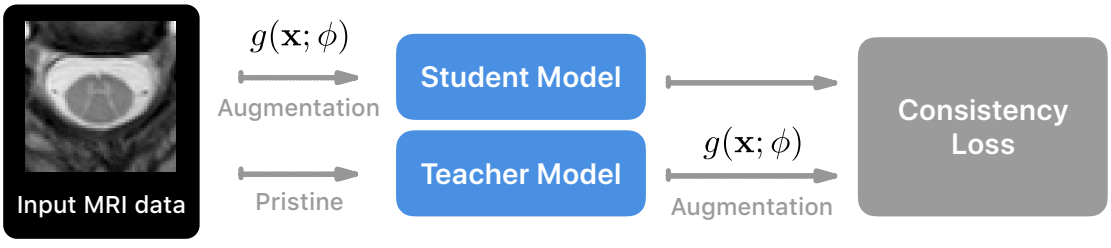


Figure 6.5 Data augmentation scheme used to overcome the spatial misalignment between student and teacher model predictions. The same augmentation parameters are used for the input data for the student model and on the teacher model predictions.

6.6.3 Model architecture

Since the U-net [64] is widely applied in medical imaging field for diverse tasks, in order to provide results that can generalize to a wide spectrum of applications, for all experiments we employed the U-net [64] model architecture with 15 layers, group normalization [142], and dropout. The rationale behind group normalization and not batch normalization is discussed later.

To provide a fair comparison, we followed the recommendations from [44] and kept the same model for the baseline and for our method. While the Mean Teacher model also acts as a regularizer, we kept the same regularization weights for all comparisons. Regularization weights can be fine-tuned, however, possibly improving even further the results of Mean Teacher.

6.6.4 Baseline employed

We conducted an extensive hyperparameter search to find a proper baseline model, yielding a mini-batch size of 12 and a dropout rate of 0.5. For training, we used the Adam optimizer [86] with L_2 penalty factor of $\lambda = 6 \times 10^{-4}$, $\beta_1 = 0.99$, and $\beta_2 = 0.999$. For learning rate, we used a sigmoid learning rate ramp-up strategy until epoch 50 followed by a cosine ramp-down until epoch 350. Eq. (6.5) shows the sigmoid ramp-up strategy:

$$R_{up}(m) = \alpha e^{-5(1-m)^2} \tag{6.5}$$

where α is the highest learning rate and m represents the ratio between current epoch and the total ramp-up epochs. Eq. (6.6) presents the cosine ramp-down strategy:

$$R_{down}(r) = \alpha \frac{\cos(\pi r) + 1}{2} \tag{6.6}$$

where α is the highest learning rate and r is the ratio between the number of epochs after the ramp-up procedure and the total number of epochs expected for training.

For a fair comparison, and to be able to assess the specific benefits of domain adaptation, no hyperparameter from the baseline model was changed in the adaptation scenario. The only change concerned the hyperparameters, which only affect the domain adaptation training procedure.

6.6.5 Consistency loss

The consistency loss is one of the most important aspects of Mean Teacher. If the difference between predictions from teacher and student is not representative enough for distilling the knowledge on the student model, the method will not properly work or training may even diverge. In the original implementation of the Mean Teacher method, the mean squared error (MSE) was proposed:

$$J_{MSE}(\theta) = \frac{\sum_i^N (p_i - g_i)^2}{N} \tag{6.7}$$

where p_i and g_i are flattened predictions from student and teacher, respectively.

As an alternative, the cross-entropy is more commonly used for classification tasks. The cross-entropy is defined as:

$$J_{CE}(\theta) = -\sum_i^N p_i \log g_i \tag{6.8}$$

where p_i and g_i are predictions from student and teacher, respectively. However, cross-entropy is also known to be sensitive to class imbalance.

Our preliminary experiments led to use MSE with different weights per class to address the problem of class imbalance. However, this approach relies on thresholding predictions from the teacher to define binary expected voxel values for the student. Defining both the correct weights and the threshold value is a difficult task that did not seem to improve overall results.

The same problem happens with more complex losses, e.g., the Focal Loss [143], due to additional hyperparameters (in this case, γ and β).

We have thus explored other losses: the Dice loss, presented in Section 6.6, and the Tversky loss [144]. The Tversky loss is a variation of the dice loss that aims at mitigating the problem of class imbalance, which is common in medical image segmentation tasks. It is defined as:

$$J_{tversky}(\theta) = -\frac{\sum_i^N p_{0i} g_{0i}}{\sum_i^N p_{0i} g_{0i} + \alpha \sum_i^N p_{0i} g_{1i} + \beta \sum_i^N p_{1i} g_{0i}} \tag{6.9}$$

where p_{0i} and g_{0i} represent the predicted probabilities and expected ground-truth of a voxel that belongs to the correct tissue, whereas p_{1i} and g_{1i} respectively represent the predicted probabilities and expected ground-truth (0 or 1) of a voxel that belongs to any other tissue. The α and β hyperparameters address the problem of class imbalance. The Tversky loss, however, is hampered by the difficulty of determining more hyperparameters alongside the

consistency weight value (same issue as noted above with the weighted MSE).

We have also noticed that both Dice and Tversky coefficients are problematic when used as consistency losses. Albeit properly representing the nature of the task, their formulation is based on multiplication and it is assumed that the ground-truth is binary, i.e. $g_i \in \{0, 1\}$. However, given that we use the teacher soft outputs (i.e., not binary), both Dice and Tversky losses do not obey the proper score orientation: $S(G, y) > S(G^*, y)$, where S is the scoring function and y is the ground truth. This relationship should hold only if G is a better probabilistic forecast, which is not the case for Tversky and Dice when using soft targets.

For example, if $p_i = 0.9$ and $g_i = 1.0$, the numerator yields 0.9. However, when $p_i = 0.9$ and $g_i = 0.9$, the score should increase (because the predicted and ground-truth are the same), but instead the numerator decreases to 0.81 and the output score also decreases.

One way to overcome this issue is to threshold the teacher's predictions such that the loss functions can work as expected. However, identifying suitable threshold values is not trivial since they drastically impact how the network adapts, and reduces the benefits of using a distillation-based [139] approach. An alternative to thresholding is to modify the formulations of the loss functions such that they can properly handle non-binary labels. A detailed analysis of such modifications falls outside the scope of this paper so we left it for future work.

6.6.6 Batch Normalization and Group Normalization for domain adaptation

Batch Normalization [79] is a method used to improve the training of deep neural networks through the stabilization of the distribution of layer inputs. Nowadays, Batch Normalization is pervasive in most deep learning architectures, enabling the use of large learning rates and helping with convergence.

Initially thought to help with the *internal covariate shift* (ICS) problem [79], Batch Normalization was recently found [145] to smooth the optimization landscape of the network due to the improvement of the Lipschitzness, or β-smoothness [145] of both loss and gradients.

Batch Normalization works differently for training and inference. During training, the normalization happens using the batch statistics, while on inference it uses the population statistics, usually estimated with moving averages on each batch during the training procedure. This strategy, however, is problematic for domain adaptation via Mean Teacher, given that there are multiple distributions being fed during training, causing the Batch Normalization statistics to be computed with both source and target data.

One possible approach to overcome that issue is to use different batch statistics for the source

and the target domains as done in AdaBN [122]. Implementing this approach within the training procedure is easily achieved using modern frameworks because it only requires to forward the batch to each domain separately [109]. However, in the implementation of French et al., both source and target domain data were used to compute the running average at inference. One should ideally perform running averages and population statistics on both domains separately, though at the expense of increased complexity on training, especially when running on a multi-GPU scenario with small batch sizes, a very common scenario in segmentation tasks where synchronization is also required.

Besides the mentioned issues, Batch Normalization also suffers from sub-optimal results when using small batch sizes [142], which are very common in segmentation tasks due to memory requirements. For those reasons, we chose Group Normalization [142], an alternative to Batch Normalization where channels are divided into groups and where mean and variance are computed within each group regardless of batch sizes. Group Normalization works consistently better than Batch Normalization with small batch sizes (typically <15) and does not require storing running averages for the population statistics, simplifying the training and inference procedures and providing better results for our scenario that involves domain adaptation and segmentation tasks.

6.6.7 Hyperparameters for unsupervised domain adaptation

A problem shared by many techniques for unsupervised domain adaptation is how to set proper hyperparameters such as the learning rate or the consistency weight. In unsupervised settings, there are no labeled data from the target domain so the estimation of hyperparameters from the source distribution alone can be completely different from those from the target distribution.

An alternative method to solve this issue is to use *reverse cross-validation* [146], which was also used in [117]. However, once again, this approach comes at the expense of increasing the complexity of the validation process. Nevertheless, we found that the estimation of hyperparameters for Mean Teacher on the source domain yielded robust results, therefore we adopted them in our experiments. We are aware that such a simple strategy is a limitation of our evaluation procedure since we could probably achieve better results for our proposed method by incorporating a more sophisticated hyperparameter estimation procedure.

6.7 Materials

The Spinal Cord Gray Matter Challenge [2] dataset is a multi-center, multi-vendor, and publicly-available MRI data collection that is comprised of 80 healthy subjects with 20 subjects from each center.

The demographics of the dataset range from a mean age of 28.3 up to 44.3 years old. Three different MRI systems were employed (Philips Achieva, Siemens Trio, Siemens Skyra) with distinct acquisition parameters. The voxel size resolution of the dataset ranges from $0.25 \times 0.25 \times 2.5$ mm up to $0.5 \times 0.5 \times 5.0$ mm and the number of axial slices ranged from 3 to 28. The dataset is split between training (40) and test (40) sets, and the test set labels are hidden (not publicly available). For each labeled slice in the dataset, 4 gold-standard segmentation masks were manually created by 4 independent experts (one per participating center). For more detailed information regarding the dataset (e.g., the MRI parameters), please see [2].

Since the Spinal Cord Gray Matter Challenge dataset contains data from all 4 centers both in the training and test sets, we used a non-standard split in order to evaluate our technique within the domain adaptation scenario, where the domain present in the test set is not contaminated by the training data domain. Therefore, we used centers 1 and 2 as the training set, center 3 as the validation set, and center 4 as the test set.

We used the unlabeled data from center 4 test set (which does not contain publicly-available labels) as the unlabeled data for the target domain, and we used the training data from center 4 (with labels) as the test set to evaluate the final performance of our model. We also slice all 3D samples into 2D axial slices and resampled each slice to 0.25×0.25 mm. An overview of the dataset is presented in Figure 6.6.

6.8 Experiments

We have designed several experiments to understand the behavior of different aspects of domain adaptation on the medical imaging domain. We have also performed ablation studies and evaluated multiple metrics for each center.

6.8.1 Adapting to different centers

We trained the network with both centers 1 and 2 in a supervised fashion. We then adapted the network to centers 3 and 4 separately. With this setup, we were able to address three related research questions on adaptation and semi-supervised learning:

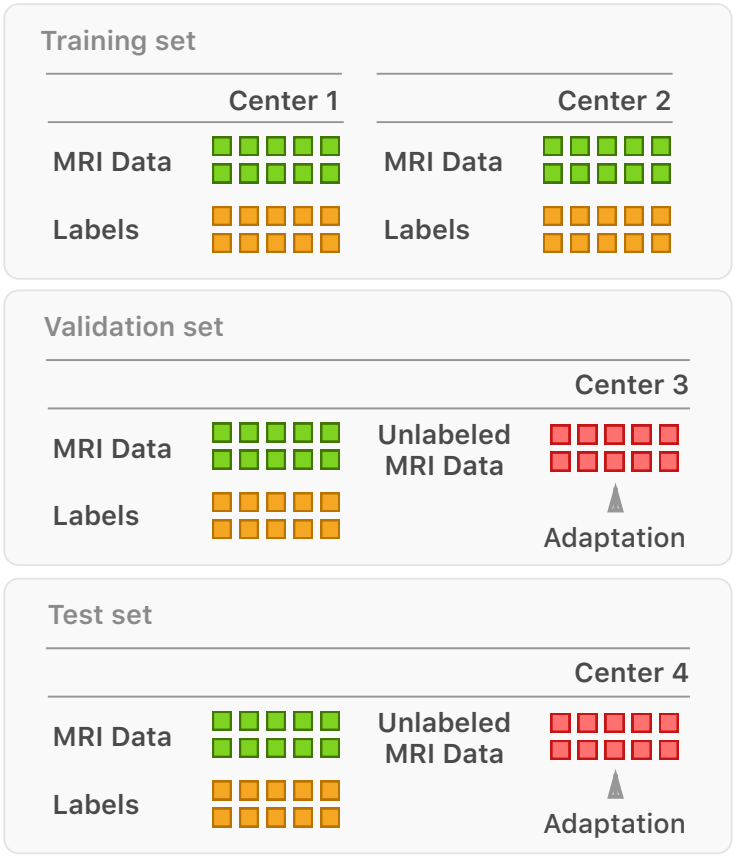


Figure 6.6 Overview of the data splitting method for training machine learning models. Each colored square represents a single subject of the dataset (containing multiple axial slices).

Table 6.1 Evaluation results in different centers. The evaluation and adaptation columns represent, respectively, the centers where testing and adaptation data were collected. Results are averages and standard deviations over 10 runs (with independent initialization of random weights). Values highlighted represent the best results at each center. All experiments were trained in both centers 1 and 2 simultaneously. Dice represents the Sørensen–Dice coefficient and mIoU represents the mean Intersection over Union.

Evaluation	Adaptation	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 1	Baseline	47.25 ± 0.10	31.46 ± 0.08	**94.90** ± 0.29	32.18 ± 0.09	99.66 ± 0.0	2.88 ± 0.01
	Center 3	47.71 ± 0.16	31.84 ± 0.14	94.18 ± 0.16	32.69 ± 0.15	99.67 ± 0.0	**2.85** ± 0.01
	Center 4	**48.42** ± 0.92	**32.47** ± 0.80	94.51 ± 0.57	**33.33** ± 0.93	**99.68** ± 0.02	2.86 ± 0.02
Center 2	Baseline	50.69 ± 0.09	34.44 ± 0.08	**94.79** ± 0.24	35.32 ± 0.10	99.61 ± 0.00	2.89 ± 0.01
	Center 3	51.05 ± 0.25	34.76 ± 0.23	93.78 ± 0.42	35.83 ± 0.31	99.62 ± 0.01	**2.87** ± 0.01
	Center 4	**51.29** ± 0.67	**34.98** ± 0.61	93.87 ± 0.91	**36.06** ± 0.82	**99.63** ± 0.02	2.87 ± 0.02
Center 3	Baseline	82.81 ± 0.33	71.05 ± 0.36	**90.61** ± 0.63	77.09 ± 0.34	99.86 ± 0.0	2.14 ± 0.02
	Center 3	**84.72** ± 0.18	**73.67** ± 0.28	87.43 ± 1.90	**83.17** ± 1.62	**99.91** ± 0.01	**2.01** ± 0.03
	Center 4	84.45 ± 0.14	73.30 ± 0.19	87.13 ± 1.77	82.92 ± 1.76	**99.91** ± 0.01	2.02 ± 0.03
Center 4	Baseline	69.41 ± 0.27	53.89 ± 0.31	**97.22** ± 0.11	54.95 ± 0.35	99.70 ± 0.00	2.50 ± 0.01
	Center 3	73.27 ± 1.29	58.50 ± 1.57	94.92 ± 1.48	60.93 ± 2.51	99.77 ± 0.03	2.36 ± 0.06
	Center 4	**74.67** ± 1.03	**60.22** ± 1.24	93.33 ± 1.96	**63.62** ± 2.42	**99.80** ± 0.02	**2.29** ± 0.05

1. How do predictions change at inference time when images from domains different than the source domain are presented?

2. How does the network change its predictions to the novel domain after performing domain adaptation?

3. How well does an adapted network generalize when presented with images that were not used during training, neither as a supervised signal nor as an unsupervised adaptation component?

Results of this first experiment are presented in Table 6.1.

Regarding *Question 1.* Both centers 1 and 2 are included in the training set and we would like to assess whether additional unsupervised data from different domains (centers 3 or 4) improve generalization on the centers 1 and 2. For both adapted centers 3 and 4, results for all metrics (except for recall) outperform the baseline, suggesting a positive change in prediction performance for the source domain after domain adaptation on unseen domains leveraging unlabelled data.

To answer *Question 2*, one can analyze the rows where both evaluation and adaptation centers are the same (3 or 4). Both rows present the highest values for almost all metrics (again, excepted for recall). This suggests that domain adaptation is working properly for that

scenario.

Regarding *Question 3*, by looking at evaluation on center 3 and adaptation using center 4 (and vice-versa), we observe gains over the baseline once again for most metrics, suggesting that domain adaptation improves generalization for unseen centers.

6.8.2 Varying the consistency loss

We executed multiple runs of the Mean Teacher algorithm by varying the consistency loss to determine which one works best. We focused just on losses that do not contain additional hyperparameters. The Tversky Loss [144], for instance, is quite similar to the Dice loss but with two additional hyperparameters (α and β).

Our choices of losses were thus limited to cross-entropy, mean squared error (MSE), and Dice, as previously described in Section 6.6. We believe, however, that a thorough analysis of distinct loss functions is of great importance for domain adaptation and should be explored in future work.

6.8.3 Behavior of Dice loss and thresholding

A well-known fact regarding the Dice loss is that it usually produces predictions concentrated around the upper and lower bounds of the probability distribution, with very low entropy. As in [107], we used a high threshold value (0.99) for the Dice predictions to produce a balanced model. We have found, however, that the domain adaptation method also regularizes the network predictions, shifting the Dice probability distribution outside of the probability bounds. For that reason, we have decreased the Dice prediction threshold to 0.9 (instead of 0.99), which produced a more balanced model in terms of precision and recall.

6.8.4 Training stability

For unsupervised domain adaptation, it is important to have a stable training procedure. Since, in the most difficult scenarios, there are no annotations for validating the adaptation, an unstable training may produce sub-optimal adaptation results.

To evaluate the training stability, we tried distinct consistency weights for each possible consistency loss and we evaluated the difference between the best values that were found and the final results after 350 epochs. Table 6.2 summarizes results of this analysis.

We can observe that cross-entropy consistently fails, even with different weights, potentially due to the class imbalance of this particular task. Though it also achieves high dice values in

Table 6.2 Results on evaluating on center 3. The training set includes centers 1 and 2 simultaneously, with unsupervised adaptation for center 3. Values within parentheses represent the best validation results for each metric. The remaining values represent the final result after 350 epochs.

Loss	Weight	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
CE	5	0.00 (85.50)	0.00 (74.91)	0.00 (95.01)	0.00 (98.90)	100.0 (100.00)	0.00 (0.00)
	10	0.00 (80.73)	0.00 (69.54)	0.00 (83.21)	0.00 (98.78)	100.0 (100.00)	0.00 (0.00)
	15	6.43 (37.03)	4.89 (26.06)	5.38 (77.05)	17.34 (65.85)	100.0 (100.00)	0.28 (0.00)
	20	2.30 (67.61)	1.86 (52.55)	2.09 (65.00)	7.94 (96.57)	100.0 (100.00)	0.12 (0.03)
Dice	5	76.76 (80.74)	62.76 (68.16)	97.88 (99.66)	63.72 (72.50)	99.71 (99.81)	2.36 (2.16)
	10	4.77 (10.55)	2.45 (5.64)	96.25 (99.99)	2.45 (5.85)	79.59 (99.75)	8.80 (2.57)
	15	2.30 (7.74)	1.16 (4.12)	99.95 (100.00)	1.16 (4.62)	55.07 (99.80)	11.75 (2.50)
	20	1.79 (4.43)	0.90 (2.27)	99.99 (100.00)	0.90 (2.30)	42.02 (99.84)	12.68 (2.43)
MSE	5	83.7 (83.88)	72.2 (72.46)	91.24 (98.19)	78.1 (78.57)	99.87 (99.93)	2.1 (2.00)
	10	84.38 (84.38)	73.19 (73.19)	90.15 (99.07)	80.12 (80.12)	99.88 (99.94)	2.05 (1.89)
	15	84.59 (84.59)	73.49 (73.50)	89.19 (98.52)	81.28 (81.28)	99.89 (99.89)	2.03 (2.03)
	20	84.5 (84.50)	73.36 (73.37)	90.36 (94.63)	80.16 (80.16)	99.88 (99.98)	2.05 (1.46)

its best scenario during training. Thus cross-entropy becomes a possible alternative to MSE when a few annotated images are available for validation in the target domain. Figure 6.7 shows how the training diverges for cross-entropy after several iterations.

We can observe that both Dice and cross entropy have trouble stabilizing the training after achieving high results. However, MSE tends to be more invariant to consistency weight, thus being a robust approach when no annotated data is available at the target center. As in [109], we also tried confidence thresholding, although we did not observe improvements.

6.9 Ablation studies

This section describes the ablation analyses, the purpose of which was to better understand the behavior of different components in the domain adaptation scenario.

6.9.1 Exponential moving average (EMA)

The improvement seen in Table 6.1 could also be explained by introducing the exponential moving average (EMA) during the training procedure, since it averages and smoothes the SGD trajectories.

To demonstrate that the improvement is specific to using unlabeled data and does not only come from the exponential average component, we performed an ablation experiment that

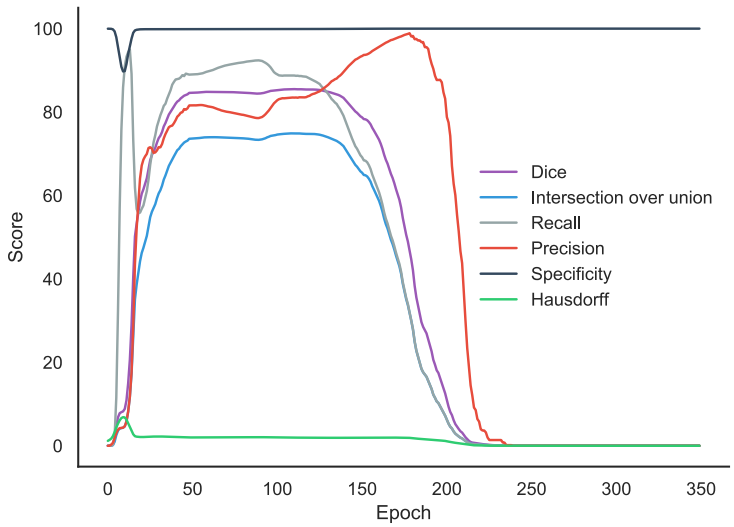


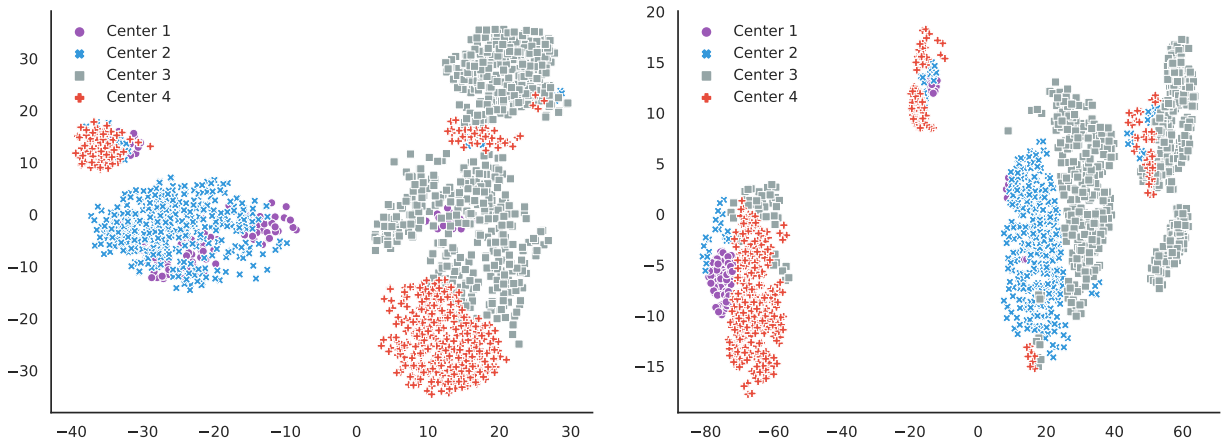
Figure 6.7 Per-epoch validation results for the teacher model at center 3 with cross-entropy as the consistency loss. Training was conducted in both centers 1 and 2 simultaneously, and adapted to center 3 with consistency weight $\gamma = 5$. Best viewed in color.

leaves the EMA active but sets the consistency weight to zero. This experiment allowed us to evaluate the impact of the exponential average in the absence of the unlabeled data used to enforce consistency.

We reproduced the same experimental setup from the Table 6.1 but with the consistency weight set to zero. Results are presented in Table 6.3 and show that the EMA model (teacher) presents no gains over the non-averaged model (the supervised baseline). This could arguably be due to a poorly chosen α. However, note that Mean Teacher, which heavily relies on the EMA model, was nevertheless able to outperform a purely-supervised method by a great margin as seen in Table 6.1.

Table 6.3 Results of the ablation experiment where the baseline model was trained and compared against its exponential moving average (EMA) model without using Mean Teacher training scheme with unlabeled data. All experiments were trained in both center 1 and 2 simultaneously. Center 3 is the validation set and Center 4 is the test set.

Evaluation	**Version**	**Dice**	**mIoU**	**Recall**	**Precision**	**Specificity**	**Hausdorff**
Center 3	Baseline	83.06	71.36	90.98	77.24	99.86	2.13
	EMA	83.09	71.40	90.97	77.30	99.86	2.13
Center 4	Baseline	69.41	53.90	97.20	54.98	99.70	2.48
	EMA	69.50	54.00	97.19	55.09	99.71	2.48



(a) A visualization of the t-SNE 2D non-linear embedding projection for the supervised learning scenario. The colors represent data from different centers.

(b) A visualization of the t-SNE 2D non-linear embedding projection for the domain adaptation scenario. The colors represent data from different centers.

Figure 6.8 Execution of t-SNE algorithm for two different scenarios. Best viewed in color.

6.10 Domain shift visualization

Next, we investigated how domain adaptation affects the prediction space of segmentation at distinct centers. By using t-SNE [147], a non-linear dimensionality reduction technique, we were able to assess changes on the predictive perception of the network regarding unsupervised data. All data presented in the following figures were not used for training.

We created two baselines for this experiment. The first model was trained in a supervised fashion following the same hyperparameters presented in Section 6.6.4. The second was an adaptation scenario where both centers 1 and 2 were used as supervised centers and 3 as adaptation target. The vectors projected with t-SNE represents the features from the network prior to the final sigmoid activation.

Both t-SNE executions had a learning rate set to 10, perplexity to 30, and were executed for about 1,000 iterations[1]. We notice that more iterations than 1,000 preserved the groups' structure but further compressed them. This made visualizing the centers harder, so 1,000 was a good trade-off between identifying emerging groups and interpretability.

Results from the supervised experiment are shown in Figure 6.8a. Note that there is a clear separation between data from centers used during training (1 and 2) and unseen centers (3

[1]We used the TensorBoard embedding projector, available at `https://github.com/tensorflow/tensorboard`

and 4). This shows that the network predictions greatly differ according to the center to which the sample belongs to.

When adapting the network with unlabelled samples from a different domain, predictions become more diffuse, at least for centers presented during training. Results from the unsupervised adaptation experiment are shown in Figure 6.8b. In that scenario, centers with labeled data (centers 1 and 2) form clusters with domains seen only in an unsupervised manner (3) or not presented to the network at all (4). A possible explanation for the close proximity of clusters (1, 4) and of clusters (2, 3) is the similarity of intensity distribution within each pair of clusters, as highlighted in Figure 6.2. See appendix for more details regarding the relationship between the data distribution and the t-SNE clusters.

6.11 Conclusion and limitations

Variability and scarcity of annotations in the medical imaging context is still challenging for machine learning. The large set of parameters that can be used to acquire image modalities and the lack of standardized protocols or industry standards are pervasive across the entire field.

In this work, we have shown that unsupervised domain adaptation, without depending on annotations, is an effective way to increase the performance of machine learning models for medical imaging across multiple centers.

Through the evaluation of multiple metrics in a large set of experiments, we have shown how self-ensembling methods can improve generalization on unseen domains through the leverage of unlabeled data from multiple domains. We also performed an ablation study that demonstrated strong evidence that the improvements come by the introduction of the unlabeled data and not only due to the exponential moving average.

We assessed how cross-entropy (when used as a consistency loss function) fails at maintaining training stability when the number of epochs progresses. We have discussed how this can lead to potential problems in more challenging scenarios for multiple centers. We also discussed issues related to the Dice loss when used as consistency loss.

We acknowledged the following limitations in our study. Firstly, we did not evaluate adversarial training methods for domain adaptation. Even considering the Mean Teacher as the current state-of-the-art method on many datasets, we believe that further analyses on the same realistic small data regime could significantly increase the importance of our contributions, and thus we leave that aspect for future work.

Secondly, the single-task evaluation of the gray matter segmentation could be extended to

other tasks in other domains. Increasing the number of centers alongside the number of tasks would be relevant for confirming results obtained in the present study.

Further work on the field could lead to methods capable of measuring the risk of adaptation to particular centers or domains. This would be an important step towards understanding the limitations of the domain adaptation methods.

We believe that the problems that arise from the variability of medical imaging modalities require rethinking some of the strong assumptions made in machine learning models and training procedures. An important step in that direction is to reassess the importance of proper multi-domain evaluation in studies and medical imaging challenges, which rarely provide a test set from different domains (such as different centers, machines, etc) that contain the variability found in real-world scenarios.

6.12 Source-code and dataset availability

In the spirit of Open Science and reproducibility, the source-code used to perform the experiments presented in this study is publicly available [2].

The dataset used for this work is also available on the Spinal Cord Gray Matter Segmentation Challenge website[3].

6.13 Acknowledgments

We are very thankful to Ryan Topfer for the sensible review and time dedicated to improve this article. Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [950-230815], the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [28826], the Fonds de Recherche du Québec - Nature et Technologies [2015-PR-182754], the Natural Sciences and Engineering Research Council of Canada [435897-2013], the Canada First Research Excellence Fund (IVADO and TransMedTech) and the Quebec BioImaging Network [5886]. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001.

[2] https://github.com/neuropoly/domainadaptation

[3] http://cmictig.cs.ucl.ac.uk/niftyweb/program.php?p=CHALLENGE

6.14 Article Appendix: Extended visualizations

In Figure 6.9 we show an extended visualization of the t-SNE embeddings from the domain adaptation scenario where the underlying raw intensity distribution is described together with their respective clusters.

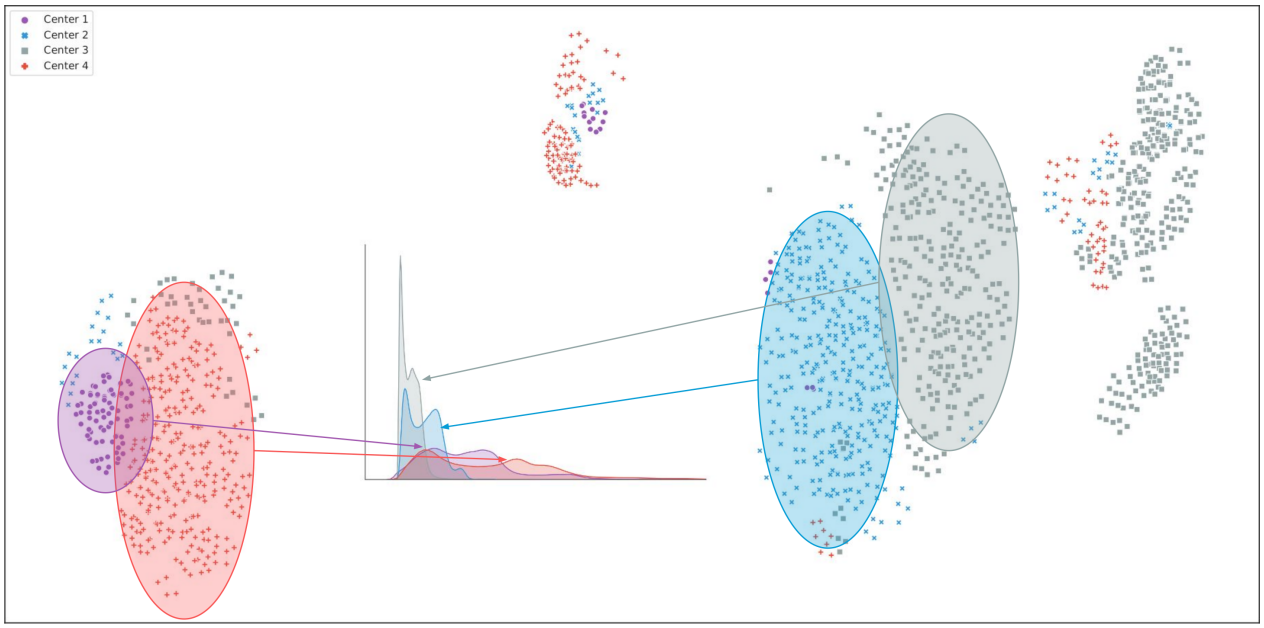


Figure 6.9 Extended visualization based on t-SNE embedding from the domain adaptation scenario in Figure 6.8b. The chart in the middle represents the pixel distribution from each center. Note how similar distributions tend to form clusters on the prediction space. Best viewed in color.

CHAPTER 7 GENERAL DISCUSSION

In this section, a general discussion is presented regarding the relationship between the articles and the proposed research questions.

In Chapter 4, where a method was developed and evaluated to segment the spinal cord gray matter, we answer the question of how modern Deep Learning methods can improve on the current state-of-the-art results. The developed method achieved better results in 8 out of 10 metrics when evaluated against other 6 independently developed methods, showing that modern Deep Learning architectures can indeed improve on the previous results even when compared to the U-Net [64], one of the most used architectures for segmentation tasks.

In Chapter 5, a semi-supervised method was developed not only the supervision signal from labeled data samples but also from the unlabeled data that are more commonly available in medical imaging. This work answers the research question on how can segmentation methods be extended to take leverage of unlabeled data. By extending a previous work on semi-supervised learning, we developed a semi-supervised training procedure that uses unlabeled data while still using all traditionally data augmentation techniques. This work demonstrated that by introducing unlabeled data during the training procedure, it was possible to achieve better results than only using the supervision signal from annotated data.

In Chapter 6, an unsupervised domain adaptation technique was developed to address the generalization gap present when a Deep Learning model that was trained on a source domain, is applied to another slightly different domain. This work answers the question regarding how this generalization can be mitigated using only unlabeled data from the target domain. This work builds upon the previous semi-supervised segmentation work and shows that when evaluated with unseen domains, this technique can achieve significantly better results when compared to the same model using only labeled data from the source domain.

It is important to note that the three mentioned themes used a publicly available and multi-center dataset that contained a realistic dataset size. For domain adaptation, this dataset was split on a non-standard way in order to avoid contamination from all centers in the test set and provide a careful realistic evaluation of a practical scenario.

These methods were also open-sourced on public GitHub repositories or as a tool inside the Spinal Cord Toolbox (SCT) framework that is also open-source and freely available.

CHAPTER 8 CONCLUSION, LIMITATIONS AND RECOMMENDATIONS

Given the evaluations presented in this work, it is clear that Deep Learning methods are very promising. It was shown in this research that in the pure supervised context, Deep Learning showed significant improvements when compared to previously developed methods. When comparing the Dice score, that measures the overlapping between predicted segmentations and gold standards, the developed method achieved 0.85 while the best previous approach achieved 0.80. These are, to the best of our knowledge, the state-of-the-art single model results up to the moment of publication of this thesis. A limitation of this work is that only healthy subjects were evaluated, however, we believe that an extensive evaluation with diseased patients with multiple sclerosis, or ALS is paramount to assess its performance out of the healthy control groups.

It also clear that by taking unlabeled data, which is often unused for model training in medical imaging, semi-supervised learning approaches can indeed improve the segmentation results without requiring any additional labeled sample. When comparing the Dice score, our method using unlabeled data achieved a value of 0.70, while the supervised only result was 0.67. A limitation of this work on semi-supervised learning is a single dataset evaluation with multiple methods, however, this is a difficult task because there is a lack of standard datasets and as mentioned in [44], realistic evaluation of these methods requires an equal budget hyper-parameter optimization procedures and same network architectures, that takes a lot of effort and might still be biased due to the pre-defined range of hyper-parameter priors.

We have also shown that improvements can be achieved to reduce the generalization gap often present on models trained on the source domain data and then applied to a different target domain, as is common in multi-center studies. Our results showed that by using our technique, a model trained on source domain and adapted to an unseen domain, achieved 0.74 Dice score when evaluated in the same unseen domain, while a purely supervised technique achieved only a 0.69 Dice score. A limitation of this work on domain adaptation is the lack of hyper-parameter optimization for the target domain, which is a difficult matter that was left for further research avenues given that if there is no labeled data in the target domain, the choice of hyper-parameters might be suboptimal.

In this work, we showed 3 main methods to improve the segmentation performance of the spinal cord gray matter, although not limited to spinal cord gray matter only. Deep Learning, however, is not a panacea for medical imaging. It still has many open challenges, such as the generalization gaps found when these models are applied to data coming from different

distributions, a quite common scenario with medical images. In practice, medical imaging often breaks the assumptions where machine learning models rely on, such as the identical distribution between training and test sets. This is an active area of research, however, since it's an issue that touches a fundamental assumption of these models, it is definitely not trivial to deal with.

While more and more data is being made available, the lack of annotations is still a problem in the medical imaging community, not to mention the usual scenario of using private datasets that makes studies almost unfeasible to compare. It is clear that privacy-preserving techniques [148] can mitigate the problem, however, these techniques are still in the early stages of development.

Further developments are also ongoing, such as the development of techniques that can be used to inform machine learning models with the MRI parametrization, therefore improving the generalization of a single model to different contrasts. We believe that these techniques can provide important improvements when used to a wide variety of contrasts.

Other potential research opportunities are related to the introduction of better/different inductive biases. Examples of inductive biases can include rotational invariance, symmetry, or anatomical structure priors. Nowadays, CNNs only provide a certain level of rotation invariance, for example, not to mention that it is very cumbersome to incorporate anatomical priors [149] into these models. However, it is clear that by adding more inductive biases, these models will be able to generalize better and potentially have a lower sample complexity.

Another important line of study is regarding the transition from research to clinical practice. Many other techniques can improve the results of this work, such as ensemble of models, test-time data augmentation, better pre-processing and post-processing techniques, to name a few. However, without large and labeled datasets, it becomes really difficult to evaluate the performance of these models for practical applications. Evaluations are often over-optimistic due to the reasons mentioned in the Chapter 6, where proper evaluation splits are often ignored in many challenges and studies of the field, therefore we would like to emphasize the importance of proper evaluation techniques that can reflect a realistic practical scenario.

It is also important to focus on fundamental aspects of machine learning models, such as the aforementioned issues related to the strong assumptions made by employed learning principles. Applied research is important, however, without a long term agenda on these fundamental issues, the field can enter into a stage of diminishing returns where only very small improvements will be able to be achieved without rethinking the underlying learning principles.

BIBLIOGRAPHY

[1] D. Weishaupt, V. D. Kochli, and B. Marincek, *How does MRI work? : an introduction to the physics and function of magnetic resonance imaging.* Springer, 2006.

[2] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. D. Leener, S. M. Dupont, P. Freund, C. A. G. Wheeler-Kingshott, F. Grussu, R. Henry, B. A. Landman, E. Ljungberg, B. Lyttle, S. Ourselin, N. Papinutto, S. Saporito, R. Schlaeger, S. A. Smith, P. Summers, R. Tam, M. C. Yiannakas, A. Zhu, and J. Cohen-Adad, "Spinal cord grey matter segmentation challenge," *NeuroImage*, vol. 152, pp. 312–329, 2017.

[3] H.-H. Chang, A. H. Zhuang, D. J. Valentino, and W.-C. Chu, "Performance measure characterization for evaluating neuroimage segmentation algorithms," *NeuroImage*, vol. 47, no. 1, pp. 122–135, aug 2009.

[4] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *ArXiv e-prints*, mar 2016.

[5] L. R. Squire, D. Berg, F. E. Bloom, S. Du Lac, A. Ghosh, and N. C. Spitzer, *Fundamental Neuroscience: Fourth Edition*, 2012.

[6] S. A. Amukotuwa and M. J. Cook, "Spinal Disease: Neoplastic, Degenerative, and Infective Spinal Cord Diseases and Spinal Cord Compression," *Clinical Gate*, pp. 511–538, 2007.

[7] F. Prados, M. J. Cardoso, M. C. Yiannakas, L. R. Hoy, E. Tebaldi, H. Kearney, M. D. Liechti, D. H. Miller, O. Ciccarelli, C. A. M. G. Wheeler-Kingshott, and S. Ourselin, "Fully automated grey and white matter spinal cord segmentation," *Scientific Reports*, vol. 6, no. June, p. 36151, 2016.

[8] R. Schlaeger, N. Papinutto, V. Panara, C. Bevan, I. V. Lobach, M. Bucci, E. Caverzasi, J. M. Gelfand, A. J. Green, K. M. Jordan, W. A. Stern, H. C. Von B??dingen, E. Waubant, A. H. Zhu, D. S. Goodin, B. A. C. Cree, S. L. Hauser, and R. G. Henry, "Spinal cord gray matter atrophy correlates with multiple sclerosis disability," *Annals of Neurology*, vol. 76, no. 4, pp. 568–580, 2014.

[9] B. De Leener, M. Taso, J. Cohen-Adad, and V. Callot, "Segmentation of the human spinal cord," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, no. 2, pp. 125–153, 2016.

[10] C. Blaiotta, P. Freund, A. Curt, J. Cardoso, and J. Ashburner, "A probabilistic framework to learn average shaped tissue templates and its application to spinal cord image segmentation," *Proceedings of the 24th Annual Meeting of ISMRM, Singapore*, vol. 1449, 2016.

[11] S. M. Dupont, B. De Leener, M. Taso, A. Le Troter, S. Nadeau, N. Stikov, V. Callot, and J. Cohen-Adad, "Fully-integrated framework for the segmentation and registration of the spinal cord white and gray matter," *NeuroImage*, vol. 150, pp. 358–372, 2017.

[12] A. J. Asman, F. W. Bryan, S. A. Smith, D. S. Reich, and B. A. Landman, "Groupwise multi-atlas segmentation of the spinal cord's internal structure," *Medical Image Analysis*, vol. 18, no. 3, pp. 460–471, 2014.

[13] R. Giraud, V. T. Ta, N. Papadakis, J. V. Manjón, D. L. Collins, and P. Coupé, "An Optimized PatchMatch for multi-scale and multi-feature label fusion," *NeuroImage*, vol. 124, pp. 770–782, 2016.

[14] M. Jorge Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin, "STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis*, vol. 17, no. 6, pp. 671–684, 2013.

[15] E. Datta, N. Papinutto, R. Schlaeger, A. Zhu, J. Carballido-Gamio, and R. G. Henry, "Gray matter segmentation of the spinal cord with active contours in MR images," *NeuroImage*, vol. 147, pp. 788–799, 2017.

[16] A. Porisky, T. Brosch, E. Ljungberg, L. Y. W. Tang, Y. Yoo, B. De Leener, A. Traboulsee, J. Cohen-Adad, and R. Tam, *Grey Matter Segmentation in Spinal Cord MRIs via 3D Convolutional Encoder Networks with Shortcut Connections*. Cham: Springer International Publishing, 2017, pp. 330–337.

[17] Y. LeCun, Y. Bengio, G. Hinton, L. Y., B. Y., and H. G., "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv preprint*, 2017.

[21] D. Amodei, R. Anubhai, E. Battenberg, C. Carl, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep-speech 2: End-to-end speech recognition in English and Mandarin," *Jmlr W&Cp*, vol. 48, p. 28, 2015.

[22] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *NAACL*, 2018.

[23] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[24] S. Grillner, "Interaction between Sensory Signals and the Central Networks Controlling Locomotion in Lamprey, Dogfish and Cat," in *Neurobiology of Vertebrate Locomotion*. London: Palgrave Macmillan UK, 1986, pp. 505–512.

[25] V. Dietz, "Spinal cord pattern generators for locomotion," *Clinical Neurophysiology*, vol. 114, no. 8, pp. 1379–1389, aug 2003.

[26] C. Y. Saab, *The spinal cord*. Chelsea House Publishers, 2006.

[27] M. Filippi and M. A. Rocca, "Linking disability and spinal cord imaging outcomes in MS," *Nature Reviews Neurology*, vol. 9, no. 4, pp. 189–190, apr 2013.

[28] M.-Ê. Paquin, M. El Mendili, C. Gros, S. Dupont, J. Cohen-Adad, and P.-F. Pradat, "Spinal Cord Gray Matter Atrophy in Amyotrophic Lateral Sclerosis," *American Journal of Neuroradiology*, 2017.

[29] P. Freund, A. Curt, K. Friston, and A. Thompson, "Tracking Changes following Spinal Cord Injury," *The Neuroscientist*, vol. 19, no. 2, pp. 116–128, apr 2013.

[30] C. Gros, B. De Leener, A. Badji, J. Maranzano, D. Eden, S. M. Dupont, J. Talbott, R. Zhuoquiong, Y. Liu, T. Granberg, R. Ouellette, Y. Tachibana, M. Hori, K. Kamiya, L. Chougar, L. Stawiarz, J. Hillert, E. Bannier, A. Kerbrat, G. Edan, P. Labauge, V. Callot, J. Pelletier, B. Audoin, H. Rasoanandrianina, J.-C. Brisset, P. Valsasina, M. A. Rocca, M. Filippi, R. Bakshi, S. Tauhid, F. Prados, M. Yiannakas, H. Kearney, O. Ciccarelli, S. Smith, C. A. Treaba, C. Mainero, J. Lefeuvre, D. S. Reich, G. Nair, V. Auclair, D. G. McLaren, A. R. Martin, M. G. Fehlings, S. Vahdat, A. Khatibi, J. Doyon, T. Shepherd, E. Charlson, S. Narayanan, and J. Cohen-Adad, "Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks," *NeuroImage*, vol. 184, pp. 901–915, jan 2019.

[31] B. De Leener, S. Kadoury, and J. Cohen-Adad, "Robust, accurate and fast automatic segmentation of the spinal cord," *NeuroImage*, vol. 98, pp. 528–536, 2014.

[32] J. Jo and Y. Bengio, "Measuring the tendency of CNNs to Learn Surface Statistical Regularities," 2017.

[33] P. W. Stroman, C. Wheeler-Kingshott, M. Bacon, J. M. Schwab, R. Bosma, J. Brooks, D. Cadotte, T. Carlstedt, O. Ciccarelli, J. Cohen-Adad, A. Curt, N. Evangelou, M. G. Fehlings, M. Filippi, B. J. Kelley, S. Kollias, A. Mackay, C. A. Porro, S. Smith, S. M. Strittmatter, P. Summers, and I. Tracey, "The current state-of-the-art of spinal cord imaging: Methods," *NeuroImage*, vol. 84, pp. 1070–1081, jan 2014.

[34] M. Yiannakas, H. Kearney, R. Samson, D. Chard, O. Ciccarelli, D. Miller, and C. Wheeler-Kingshott, "Feasibility of grey matter and white matter segmentation of the upper cervical cord in vivo: A pilot study with application to magnetisation transfer measurements," *NeuroImage*, vol. 63, no. 3, pp. 1054–1059, nov 2012.

[35] P. Held, U. Dorenbeck, J. Seitz, R. Frund, and H. Albrich, "MRI of the abnormal cervical spinal cord using 2D spoiled gradient echo multiecho sequence (MEDIC) with magnetization transfer saturation pulse. A T2* weighted feasibility study." *Journal of neuroradiology. Journal de neuroradiologie*, vol. 30, no. 2, pp. 83–90, mar 2003.

[36] T. M. Mitchell, *Machine Learning*, ser. McGraw-Hill International Editions. McGraw-Hill, 1997.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[38] C. Olivier, B. Schölkopf, and A. Zien, "Semi-Supervised Learning," *Interdisciplinary sciences computational life sciences*, vol. 1, no. 2, p. 524, 2006.

[39] X. Zhu, "Semi-supervised learning literature survey," *Computer Science, University of Wisconsin-Madison*, vol. 2, no. 3, p. 4, 2006.

[40] A. Rasmus, H. Valpola, and M. Berglund, "Semi-Supervised Learning with Ladder Network," *arXiv*, pp. 1–17, 2015.

[41] S. Laine and T. Aila, "Temporal Ensembling for Semi-Supervised Learning."

[42] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017.

[43] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning," 2017.

[44] A. G. B. Oliver, A. G. B. Odena, C. G. B. Raffel, E. G. B. Cubuk, and I. J. G. B. Goodfellow, "Realistic Evaluation of semi-supervised learning algortihms," *International conference on Learning Representations*, pp. 1–15, 2018.

[45] C. Baur, S. Albarqouni, and N. Navab, "Semi-supervised Deep Learning for Fully Convolutional Networks," *Miccai-2017*, pp. 311–319, 2017.

[46] N. Souly, C. Spampinato, and M. Shah, "Semi Supervised Semantic Segmentation Using Generative Adversarial Network," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017, pp. 5689–5697.

[47] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning Bounds for Domain Adaptation," Tech. Rep.

[48] W. M. Kouw, "An introduction to domain adaptation and transfer learning," Tech. Rep., 2018.

[49] C. S. Perone and J. Cohen-Adad, "Promises and limitations of deep learning for medical image segmentation," *Journal of Medical Artificial Intelligence*, vol. 2, no. 0, 2019.

[50] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual Domain Adaptation: A survey of recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 53–69, may 2015.

[51] G. Csurka, "Domain Adaptation for Visual Applications: A Comprehensive Survey," 2017.

[52] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10265 LNCS, pp. 597–609, 2017.

[53] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2015.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," oct 2018.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, pp. 630–645, 2016.

[56] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," dec 2015.

[57] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. . A. Ranzato, "Phrase-Based & Neural Unsupervised Machine Translation," Tech. Rep.

[58] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proceedings of the 27th International Conference on Machine Learning*, no. 3, pp. 807–814, 2010.

[59] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, pp. 541–551, 1989.

[60] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in neural {...}*, pp. 1–9, 2012.

[61] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[62] Y. Ganin and V. Lempitsky, "N4-Fields: Neural Network Nearest Neighbor Fields for Image Transforms," jun 2014.

[63] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, 2015, pp. 3431–3440.

[64] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[65] J. Cohen-Adad and L. L. Wald, "Array Coils," in *Quantitative MRI of the Spinal Cord*. Elsevier, 2014, pp. 59–67.

[66] M. C. Yiannakas, H. Kearney, R. S. Samson, D. T. Chard, O. Ciccarelli, D. H. Miller, and C. A. M. Wheeler-Kingshott, "Feasibility of grey matter and white matter segmentation of the upper cervical cord in vivo: A pilot study with application to magnetisation transfer measurements," *NeuroImage*, vol. 63, no. 3, pp. 1054–1059, nov 2012.

[67] N. Papinutto, R. Schlaeger, V. Panara, E. Caverzasi, S. Ahn, K. J. Johnson, A. H. Zhu, W. A. Stern, G. Laub, S. L. Hauser, and R. G. Henry, "2D phase-sensitive inversion recovery imaging to measure in vivo spinal cord gray and white matter areas in clinically feasible acquisition times," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 3, pp. 698–708, sep 2015.

[68] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 11-18-Dece, pp. 1026–1034, 2016.

[70] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks," *arXiv preprint*, 2017.

[71] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," *Iclr*, pp. 1–14, 2014.

[72] B. De Leener, S. Lévy, S. M. Dupont, V. S. Fonov, N. Stikov, D. Louis Collins, V. Callot, and J. Cohen-Adad, "SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data," *NeuroImage*, vol. 145, pp. 24–43, 2017.

[73] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2018–2025, 2011.

[74] B. H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 95–103.

[75] L. Zhang, Y. Ji, and X. Lin, "Style Transfer for Anime Sketches with Enhanced Residual U-net and Auxiliary Classifier GAN," *CoRR*, 2017.

[76] J. Yosinski, C. Jeff, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding Neural Networks Through Deep Visualization," in *Deep Learning Workshop, International Conference on Machine Learning (ICML)*, 2015.

[77] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *ICLR*, pp. 1–9, 2016.

[78] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 4898–4906.

[79] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. JMLR.org, 2015, pp. 448–456.

[80] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[81] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*, 2016, pp. 565–571.

[82] M. Drozdzal, G. Chartrand, E. Vorontsov, L. Di Jorio, A. Tang, A. Romero, Y. Bengio, C. Pal, and S. Kadoury, "Learning Normalized Inputs for Iterative Estimation in Medical Image Segmentation," *arXiv preprint*, 2017.

[83] P. Simard, D. Steinkraus, and J. Platt, "Best practices for convolutional neural networks applied to visual document analysis," *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, vol. 1, no. Icdar, pp. 958–963, 2003.

[84] M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, "Medical Image Processing, Analysis and Visualization in clinical research," in *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems*, no. February. IEEE Comput. Soc, 2001, pp. 381–386.

[85] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, and P. M. Matthews, "Advances in functional and structural MR image analysis and implementation as FSL," in *NeuroImage*, vol. 23, no. SUPPL. 1. Academic Press, jan 2004, pp. S208–S219.

[86] D. P. Kingma and J. L. Ba, "Adam: a Method for Stochastic Optimization," *International Conference on Learning Representations 2015*, pp. 1–15, 2015.

[87] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[88] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing Neural Networks by Penalizing Confident Output Distributions," *arXiv preprint*, 2017.

[89] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of statistics*, pp. 416–431, 1983.

[90] P. D. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.

[91] M. Hardt and O. Vinyals, "Understanding Deep Learning Requires Re- Thinking Generalization," *International Conference on Learning Representations 2017*, pp. 1–15, 2017.

[92] S. Kornblith, J. Shlens, and Q. V. Le Google Brain, "Do Better ImageNet Models Transfer Better?" Tech. Rep.

[93] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems 27 (Proceedings of NIPS)*, vol. 27, pp. 1–9, 2014.

[94] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, U. Dogan, M. Kloft, F. Orabona, and T. Tommasi, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2016.

[95] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop: Challenges in Representation Learning*, pp. 1–6, 2013.

[96] Y. Grandvalet and Y. Bengio, "Semi-supervised Learning by Entropy Minimization," *Advances in Neural Information Processing Systems - NIPS'04*, vol. 17, pp. 529–536, 2004.

[97] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," 2016.

[98] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017.

[99] B. T. Polyak and A. B. Juditsky, "Acceleration of Stochastic Approximation by Averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[100] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," mar 2015.

[101] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for domain adaptation," pp. 1–15, 2017.

[102] C. S. Perone, E. Calabrese, and J. Cohen-Adad, "Spinal cord gray matter segmentation using deep dilated convolutions," oct 2017.

[103] H. Xiao, Y. Wei, Y. Liu, M. Zhang, and J. Feng, "Transferable Semi-supervised Semantic Segmentation," 2017.

[104] C. Gros, B. De Leener, A. Badji, J. Maranzano, D. Eden, S. M. Dupont, J. Talbott, R. Zhuoquiong, Y. Liu, T. Granberg, R. Ouellette, Y. Tachibana, M. Hori, K. Kamiya, L. Chougar, L. Stawiarz, J. Hillert, E. Bannier, T. Shepherd, E. Charlson, S. Narayanan, and J. Cohen-Adad, "Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks," may 2018.

[105] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 5947–5956.

[106] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[107] C. S. Perone, E. Calabrese, and J. Cohen-Adad, "Spinal cord gray matter segmentation using deep dilated convolutions," *Nature Scientific Reports*, vol. 8, no. 1, 2018.

[108] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Confounding variables can degrade generalization performance of radiological deep learning models," 2018.

[109] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," *arXiv preprint arXiv:1706.05208*, 2017.

[110] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[111] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011.

[112] M. Lai, "Deep learning for medical image segmentation," *arXiv preprint arXiv:1505.02000*, 2015.

[113] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.

[114] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[115] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, 2018.

[116] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[117] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[118] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.

[119] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate To Adapt: Aligning Domains using Generative Adversarial Networks," *Proceedings - 31th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 2018.

[120] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. F. Wang, "Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation," *Proceedings - 31th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 2018.

[121] J. Cao, O. Katzir, P. Jiang, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "Dida: Disentangled synthesis for domain adaptation," *arXiv preprint arXiv:1805.08019*, 2018.

[122] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," *arXiv preprint arXiv:1603.04779*, 2016.

[123] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 443–450.

[124] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[125] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[126] E. A. AlBadawy, A. Saha, and M. A. Mazurowski, "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing," *Medical physics*, vol. 45, no. 3, pp. 1150–1158, 2018.

[127] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 597–609.

[128] C. Chen, Q. Dou, H. Chen, and P.-A. Heng, "Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-ray Segmentation," 2018.

[129] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," 2018.

[130] M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, "Domain-adversarial neural networks to address the appearance variability of histopathology images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10553 LNCS, pp. 83–91, 2017.

[131] M. Javanmardi and T. Tasdizen, "DOMAIN ADAPTATION FOR BIOMEDICAL IMAGE SEGMENTATION USING ADVERSARIAL TRAINING Scientific Computing and Imaging Institute , University of Utah," no. Isbi, pp. 554–558, 2018.

[132] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised Cross-Modality Domain Adaptation of ConvNets for Biomedical Image Segmentations with Adversarial Loss," Tech. Rep., 2018.

[133] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training," *IEEE Transactions on Medical Imaging*, vol. PP, no. c, p. 1, 2018.

[134] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation," *IEEE 15th Symposium on Biomedical Imaging*, no. Isbi, pp. 1038–1042, 2018.

[135] A. Odena, A. Oliver, C. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of semi-supervised learning algorithms," 2018.

[136] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[137] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[138] D. Ruppert, "Efficient estimations from a slowly convergent robbins-monro process," Cornell University Operations Research and Industrial Engineering, Tech. Rep., 1988.

[139] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[140] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.

[141] C. S. Perone and J. Cohen-Adad, "Deep semi-supervised segmentation with weight-averaged consistency targets," *DLMIA MICCAI*, pp. 1–8, sep 2018.

[142] Y. Wu and K. He, "Group Normalization," 2018.

[143] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[144] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 379–387.

[145] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How Does Batch Normalization Help Optimization? (No, It Is Not About Internal Covariate Shift)," 2018.

[146] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren, "Cross validation framework to choose amongst models and datasets for transfer learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6323 LNAI, no. PART 3, 2010, pp. 547–562.

[147] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[148] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015*. IEEE, sep 2016, pp. 909–910.

[149] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. Cook, A. D. Marvao, D. O. Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically Constrained Neural Networks (ACNN): Application to Cardiac Image Enhancement and Segmentation," vol. XX, no. X, pp. 1–10, 2017.