Istituto Italiano
di Tecnologia

Università Degli Studi
di Genova

Department of Pattern analysis and Computer Vision (PAVIS)

Department of Electrical, Electronic and Telecommunications
Engineering and Naval Architecture (DITEN)

PhD in Science and Technology for Electronic and Telecommunication Engineering

Curriculum: Computational Vision, Automatic Recognition and Learning

Deep Learning Approaches for The Segmentation of Multiple Sclerosis Lesions on Brain MRI

Shahab Aslani

*A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy*

Supervisor: Prof. Vittorio Murino
Co-Supervisor: Dr. Diego Sona

Coordinator of the PhD Course: Prof. Mario Marchese

February 2020 - XXXII Cycle

Acknowledgments

Firstly, I would like to thank and express my sincere gratitude to my supervisors, Professor Vittorio Murino and Doctor Diego Sona for their valuable support, educational moral and encouraging words throughout my doctoral journey. I am greatly indebted to them for accepting me as their student and allowing me to grow under their patient guidance.

My very special gratitude goes to Doctor Michael Dayan who offered me professional guidance in medical image analysis, owing to his rich experience and excellent expertise.

I am especially thankful to Professor Ghassan Hamarneh and Professor Roger Tam, who provided me an opportunity to join their team. Without their precious support, this research would have not reached so far.

Living abroad and far from the family and home was a great challenge and finding new friends is what made it to be a sweet one. My special thank goes to all my friends and colleagues in PAVIS, VGM, and MIAL research groups.

I am also grateful to my best friends and brothers, Doctor Amir Ghorbani and Doctor Amir Ghiami who have given me extraordinary friendship and unconditional moral and emotional support. They have been there whenever I need.

Finally, and most importantly, I would like to extend special thanks to my parents, Behzad, Homeira and Aliye for their sacrifices, inspiration, and patients throughout my studies. Without their full support and love, this thesis would have not been possible. I am also thankful to my lovely sisters, Saba, Fereshte and my brother-in-law Ali who supported me with their love and trust all along the way. To them, I dedicate this thesis. Yashasin!!

February 2020, Genova

Abstract

Multiple Sclerosis (MS) is a demyelinating disease of the central nervous system which causes lesions in brain tissues, especially visible in white matter with magnetic resonance imaging (MRI). The diagnosis of MS lesions, which is often performed visually with MRI, is an important task as it can help characterizing the progression of the disease and monitoring the efficacy of a candidate treatment. automatic detection and segmentation of MS lesions from MRI images offer the potential for a faster and more cost-effective performance which could also be immune to expert bias segmentation.

In this thesis, we study automated approaches to segment MS lesions from MRI images. The thesis begins with a review of the existing literature on MS lesion segmentation and discusses their general limitations. We then propose three novel approaches that rely on Convolutional Neural Networks (CNNs) to segment MS lesions.

The first approach demonstrates that the parameters of a CNN learned from natural images, transfer well to the tasks of MS lesion segmentation. In the second approach, we describe a novel multi-branch CNN architecture with end-to-end training that can take advantage of each MRI modalities individually. In that work, we also investigated the combination of MRI modalities leading to the best segmentation performance. In the third approach, we show an effective and novel generalization method for MS lesion segmentation when data are collected from multiple MRI scanning sites and as suffer from (site-)domain shifts. Finally, this thesis concludes with open questions that may benefit from future work. This thesis demonstrates the potential role of CNNs as a common methodological building block to address clinical problems in MS segmentation.

Keywords: *Multiple Sclerosis, Lesions, Brain, Multiple Image Modality, Segmentation, Domain Generalization, Convolutional Neural Network.*

Contents

List of Figures

List of Tables

Common Abbreviations

CNN	Convolutional Neural Network
FCN	Fully Convolutional Neural Network
MS	Multiple Sclerosis
MRI	Magnetic Resonance Images
T1w	T1-weighted MRI
T2w	T2-weighted MRI
PDw	Proton Density-weighted
FLAIR	Fluid Attenuated Inversion Recovery
WM	White Matter
CSF	Cerebrospinal Fluid
GM	Gray Matter
ML	Machine Learning
DSC	Dice Similarity Coefficient
LTPR	Lesion-wise True Positive Rate
LFPR	Lesion-wise False Positive Rate
SD	Average Symmetric Surface Distance
LFPR	Lesion-wise False Positive Rate
HD	Hausdorff Distance
PPV	Positive Prediction Value
VD	Absolute Volume Difference
SC	Overall Evaluation Score

Chapter 1

Introduction

1.1 Background and Motivations

Image segmentation is one of the most important tasks in image processing and computer vision when information needs to be extracted from image components. It is a method to partition an image into several coherent regions/objects, without any attempt however at understanding what these regions/objects represent. Therefore, the definition of the regions/objects is completely ambiguous [28, 85]. Instead, semantic image segmentation which is one of the most crucial tasks for visual scene understanding attempts to partition the image into semantically meaningful regions/objects. This task aims at automatically extracting meaningful information from images. It does so by assigning a unique label/category to every single pixel in the image. This computer vision problem can also be addressed as a dense classification problem. There is a large amount of research on semantic image segmentation, most of them based however on hand-crafted features which are designed be forehand by human experts to extract a given set of chosen characteristics [34, 50, 69, 72, 87].

Recently, deep convolutional neural networks (CNNs) [55] have achieved great success in a large variety of artificial intelligence tasks, including image classification [39, 42, 53, 82, 97] and object detection [33, 38, 77, 78]. In contrast to traditional methods, where handcrafted features are used, CNNs automatically learn representative complex features directly from the data itself.

Based on the successful and excellent performance of CNNs on image recognition, semantic segmentation has also been adapted using CNNs by extending them to fully convolutional networks (FCNs) for pixel-wise classification [63]. Afterward, many semantic segmentation networks have been proposed based on FCNs [6, 15, 16, 17, 58, 73, 110].

Medical image segmentation is one of the most challenging tasks in medical image analysis. In this task, the pixels of organs or lesions need to be separated from the background of medical images. This identification delivers critical information about the shapes and volumes of the mentioned organs or lesions which are necessary for diagnosis, monitoring and treatment [40, 112]. In the last few years, medical image segmentation based on deep neural network, particularly FCNs, has received vast attention [84]. Important improvement has been demonstrated in different problems such as neuronal structures segmentation in microscopy images [80], retinal blood vessel segmentation in fundus images [59], multiple organ segmentation in computed tomography (CT) images [113] and skin lesion segmentation in dermoscopy images [106].

Medical imaging devices to study the brain such as positron emission tomography (PET), CT, magnetic resonance spectroscopy (MRS) and magnetic resonance imaging (MRI) are all used to provide valuable information about shape, size, location, and metabolism of brain tissues, tumors and lesions assisting in diagnosis. While all these imaging techniques are used in combination to provide the highest detailed information about the brain, MRI is considered as the standard and more convenient technique for brain imaging due to its good soft-tissue contrast and wide availability [84]. Recent performances of deep learning methods, specifically FCNs, in several brain MRI segmentation challenges increased their popularity resulting in multiple applications such as brain tumor [37], tissue [66] and lesion segmentation [4].

The focus of the current study is multiple sclerosis (MS) lesion segmentation, one of the most important areas of brain lesion segmentation. MS is one of the most common demyelination diseases and the effects of demyelination is especially visible in white matter with MRI which is typically required to diagnose the disease [93]. During the last years, several methods have been proposed for MS lesion segmentation [10, 83, 86, 96, 104]. However, there is a limited number of methods based on deep learning [8, 9, 32, 36, 81, 101, 102].

In this thesis, we will investigate most of the aforementioned deep learning-based methods for MS lesion segmentation and we will propose new solutions with improvements compared to the state-of-the-art methods.

1.2 Challenges

MS is a chronic, autoimmune and demyelinating disease of the central nervous system causing lesions in brain tissues. This disease is a persistent inflammatory-demyelinating and degenerative disease that is characterised pathologically by areas of inflammation, axonal loss, and gliosis scattered throughout the central nervous system, often causing motor, sensorial, vision, coordination, deambulation, and cognitive impairment. MS is the most non-traumatic disease causing disability, especially in young people. This disease is very common in Europe, the United States, Canada, New Zealand, and Australia. In countries or regions with template climate, the incidence, and prevalence of MS increase with latitude. MS is more common in women than in men. The incidence of MS in children is low. However, it increases drastically after the age of 18, having a peak between 25 and 35 and then slowly decreases, so becoming very rare after 50 age. It is reported that there exist between 1.3 and 2.5 million MS cases in the world and according to the latest studies, the prevalence and incidence of MS have been increasing during the last years [93].

Nowadays, MRI scans are the most common solution to visualize the alterations owing to their sensitivity to detect WM damage especially common in MS [21]. Four standard MRI modalities are traditionally used for visualizing and diagnosing MS lesions including T1-weighted MRI (T1w), T2-weighted MRI (T2w), proton density-weighted (PDw), and fluid-attenuated inversion recovery (FLAIR).

Clinically, there exists two main important phenomena of MS: relapses and progression. Following the changes in lesion volume over time via the precise segmentation of lesions is an important step to understand and characterize the progression of the disease [79]. To this aim, both manual and automated methods are used to compute the total number of lesions and total lesion volume.

The most basic form of assessment includes manually tracing the outline of each MS lesion on each MRI brain slice to compute the total area and volume of lesions [27]. Although manual segmentation of MS lesions is considered the gold standard [88], this approach is affected by various difficulties such as lesions' deformable shapes, locations, intensities and texture characteristics which can be significantly different across patients. Moreover, delineation of 3-dimensional (3D) information from MRI modalities is time-consuming, tedious and prone to intra- and inter-observer variability [96]. This motivates machine learning (ML) experts to develop automated lesion segmentation techniques, which can be orders of magnitude faster and immune to variability in expert bias.

1.3 Datasets

Automated MS lesion segmentation methods are still in the early stage of development and are not fully applicable to real clinical applications. One of the most important problems in this field is the lack of sufficient publicly available dataset. Moreover, most of the existing datasets consist of a very limited number of subjects. In this section, we provide a list of datasets including publicly available and private clinical datasets which were used for this thesis.

1.3.1 ISBI 2015 Longitudinal MS Lesion Segmentation Challenge:

The ISBI dataset [11] includes 19 subjects divided into two sets, 5 subjects in the train set and 14 subjects in the test set. All subjects were scanned using a Philips 3 Tesla (T) MRI scanner. Each subject was scanned more than once at different time-points, ranging from 4 to 6. For each time-point, T1w, T2w, PDw, and FLAIR image modalities were provided. The volumes were composed of 182 slices with FOV=182mm×256mm and 1-millimeter cubic voxel resolution. All images available were already segmented manually by two different raters, therefore representing two ground truth lesion masks. For all 5 training images, lesion masks were made publicly available. For the remaining 14 subjects in the test set, there is no publicly available ground truth. The performance evaluation of the proposed method over the test dataset is done through an online service by submitting the binary masks to the challenge[1] website. The preprocessed version of the images were available online at the challenge website. All images were already skull-stripped using brain extraction tool (BET) [91], rigidly registered to $1mm^3$ MNI-ICBM152 template [74] and N3 intensity normalized [90]. An example of the preprosseced version of a subject related to ISBI dataset can be seen in Figure 1.1.

1.3.2 Clinical Private Dataset 1:

The Neuroimaging Research Unit (NRU) dataset [4] is a private clinical dataset collected by a research team from Ospedale San Raffaele, Milan, Italy. The dataset consists of 37 MS patients acquired on a 3T Philips Ingenia CX scanner (Philips Medical Systems). The following sequences were collected: Sagittal 3D FLAIR sequence, FOV=256mm×256mm, pixel size=1mm×1mm, 192 slices, $1mm$ thick; Sagittal 3D T2w turbo spin-echo (TSE) sequence, FOV=256mm×256mm, pixel size=1mm×1mm, 192 slices, $1mm$ thick; Sagittal 3D high resolution T1w, FOV=256mm×256mm, pixel size=1mm×1mm, 204 slices, $1mm$ thick. For the creation of the ground truth lesion masks, two different readers performed the lesion delineation blinded to each other's results. They estimated the agreement between the two expert raters by using the Dice similarity coefficient as a measure of the degree of overlap between the segmented lesion masks (they found a mean Dice of 0.87). Following the common clinical practice of considering a single consensus mask between raters, the two masks created by the two expert raters were used to generate a high quality, gold standard, mask by intersecting the two binary masks of each rater. An example of a subject from the NRU dataset can be seen in Figure 1.2.

[1] http://iacl.ece.jhu.edu/index.php/MSChallenge

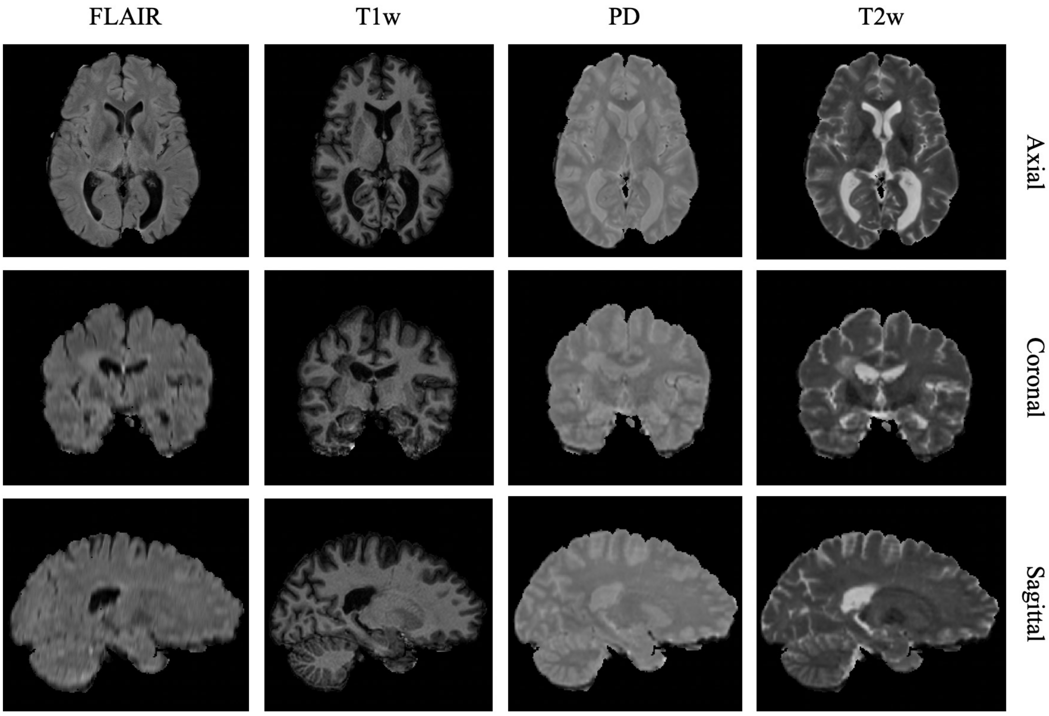


Figure 1.1: An example of a subject from the ISBI dataset with four MRI modalities (FLAIR, T1w, T2w and PD) visualized through orthogonal views of the brain (axial, coronal and sagittal planes).

1.3.3 Clinical Private Dataset 2:

This dataset [9] was collected by an hospital in University of British Columbia (UBC) from 56 different centers (5 sites with 3T scanner and 41 sites with 1.5T scanner). Each site has a different number of patients, ranging from 1 to 11 which resulted in totally 117 patients. Each patient had several scanning sessions, with each session including 4 MRI modalities: T1w, T2w, PDw, and FLAIR. Each volume is composed of 60 slices with FOV=256mm×256mm and 1mm×1mm×3mm voxel resolution. All volumes were already segmented manually by several technicians, using a semi-automated method that was assisted by a clustering method. The gold standard mask was obtained by the intersection of the binary masks. An example of a subject from the UBC dataset can be seen in Figure 1.3.

Table 1.1: Publicly available and private clinical datasets related to MS lesion segmentation.

Dataset	**Date**	**# patients**	**# modalities**	**Publicly available**
ISBI [11]	2015	19	4	✓
Clinical 1 [4]	2018	37	4	
Clinical 2 [9]	2015	117	4	

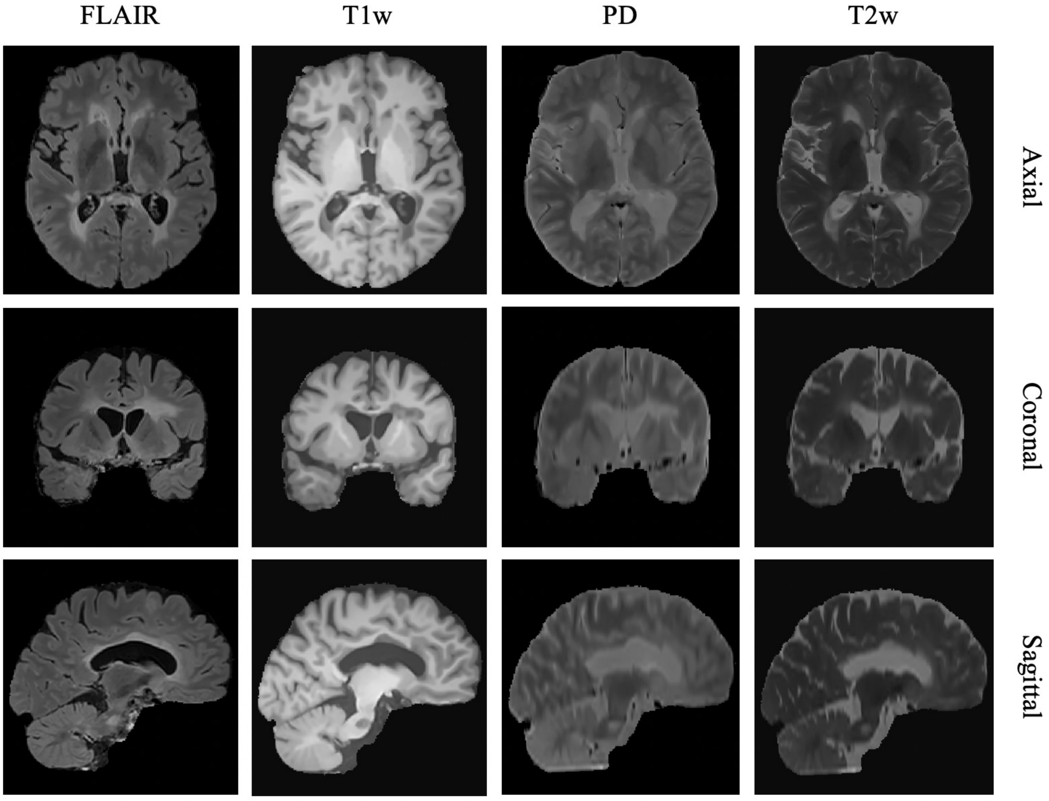


Figure 1.2: An example of a subject related to NRU dataset with four MRI modalities (FLAIR, T1w, T2w and PD) visualized through orthogonal views of the brain (axial, coronal and sagittal planes)

1.4 Evaluation

Different metrics have been used to evaluate the performance of the proposed methods in MS lesion segmentation. In this section, we introduce them briefly.

1.4.1 Dice Similarity Coefficient (DSC):

The *DSC* is a statistic used for measuring the similarity (intersection) of two samples (segmentation results and ground truth labels).

$$DSC = \frac{2TP}{FN + FP + 2TP} \tag{1.1}$$

where *TP*, *FN* and *FP* indicate the true positive, false negative and false positive voxels,

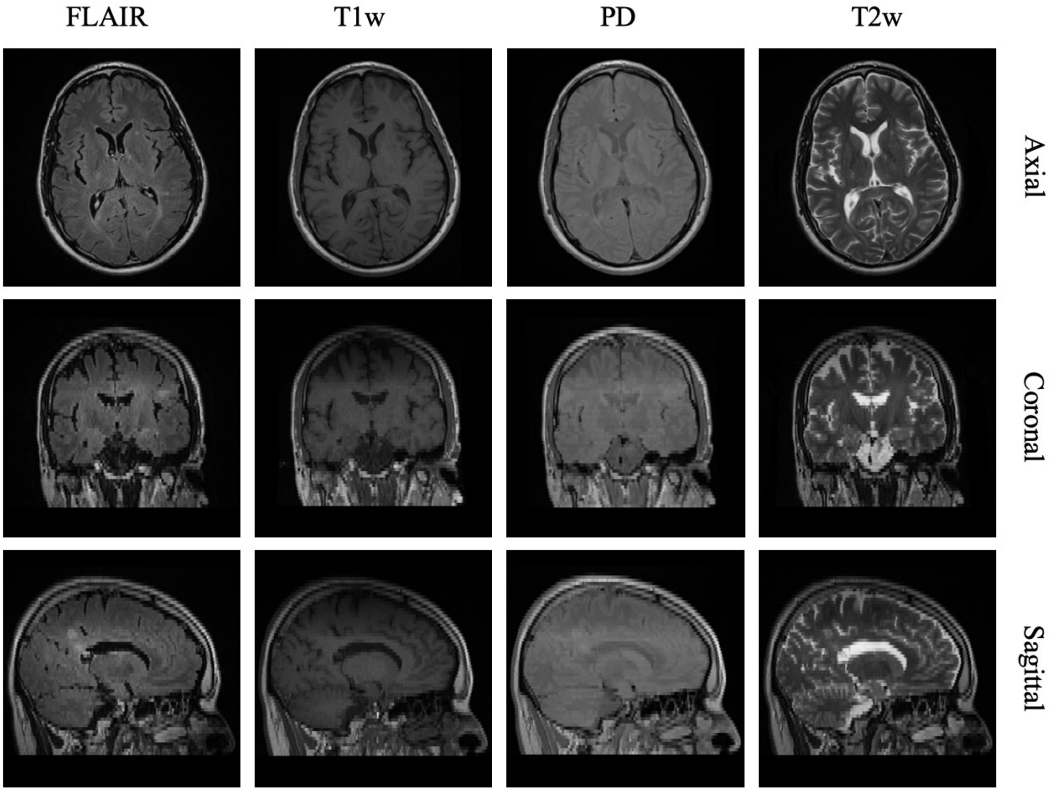


Figure 1.3: An example of a subject from the UBC dataset with four MRI modalities (FLAIR, T1w, T2w and PD) visualized through orthogonal views of the brain (axial, coronal and sagittal planes)

respectively. An illustration is depicted in Figure 1.4.

1.4.2 Lesion-wise True Positive Rate (LTPR):

The *LTPR* is the lesion-wise ratio of true positives to the sum of true positives and false negatives.

$$LTPR = \frac{LTP}{RL} \tag{1.2}$$

where *LTP* denotes the number of lesions in the reference segmentation that overlap with a lesion in the output segmentation (at least one voxel overlap), and *RL* is the total number of lesions in the reference segmentation.

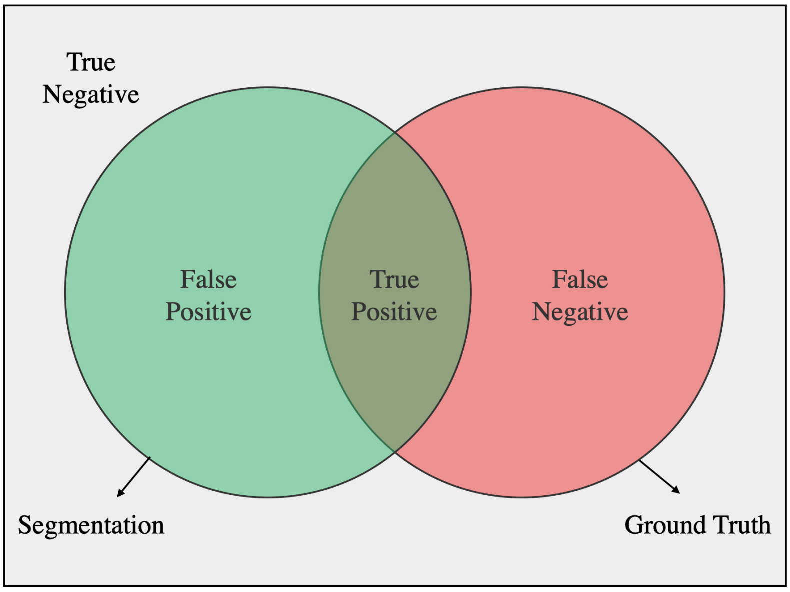


Figure 1.4: Schematic illustration of the measuring the similarity between segmentation result and ground truth label.

1.4.3 Lesion-wise False Positive Rate (LFPR):

The *LFPR* is the lesion-wise ratio of false positives to the sum of false positives and true negatives.

$$LFPR = \frac{LFP}{PL} \tag{1.3}$$

where *LFP* denotes the number of lesions in the output segmentation that do not overlap with a lesion in the reference segmentation and *PL* is the total number of lesions in the produced segmentation.

1.4.4 Average Symmetric Surface Distance (SD):

The *SD* is the average of the distance from the lesions in ground truth mask to the nearest lesion identified in segmentation output plus the distance from the lesions in segmentation output to the nearest lesion identified in ground truth mask.

$$SD = \frac{1}{|N_{gt}| + |N_s|} \cdot \left(\sum_{x \in N_{gt}} \min_{y \in N_s} d(x,y) + \sum_{x \in N_s} \min_{y \in N_{gt}} d(x,y) \right) \tag{1.4}$$

where N_s and N_{gt} are the set of voxels in the contour of the automatic and manual annotation masks, respectively. $d(x, y)$ is the Euclidean distance (quantified in millimetres) between voxel x and y.

1.4.5 Hausdorff Distance (HD):

The *HD* measures the maximum distance between the nearest contours for all pairs of segmented lesion and ground truth lesion.

$$\mathit{HD} = \max\left\{\max_{x\in N_{gt}}\min_{y\in N_s} d(x,y), \max_{x\in N_s}\min_{y\in N_{gt}} d(x,y)\right\} \tag{1.5}$$

1.4.6 Positive Prediction Value (PPV):

The *PPV* is the voxel-wise ratio of the true positives to the sum of the true and false positives.

$$\mathit{PPV} = \frac{\mathit{TP}}{\mathit{TP} + \mathit{FP}} \tag{1.6}$$

1.4.7 Absolute Volume Difference (VD):

The *VD* is the absolute difference in volumes divided by the true volume,

$$\mathit{VD} = \frac{|\mathit{TP}_s - \mathit{TP}_{gt}|}{\mathit{TP}_{gt}} \tag{1.7}$$

where TP_s and TP_{gt} reveal the total number of the segmented lesion voxels in the output and manual annotations masks, respectively.

1.4.8 Overall Evaluation Score (SC):

As described in [11], the ISBI challenge website provides the overall score based on most of the previous metrics described.

$$\mathit{SC} = \frac{1}{|R|\cdot|S|}\cdot\sum_{R,S}\left(\frac{\mathit{DSC}}{8} + \frac{\mathit{PPV}}{8} + \frac{1-\mathit{LFPR}}{4} + \frac{\mathit{LTPR}}{4} + \frac{\mathit{Cor}}{4}\right) \tag{1.8}$$

where S is the set of all subjects, R is the set of all raters and Cor is the Pearson's correlation coefficient of the volumes.

1.5 Thesis Contributions

The following sections summarize the chapters within the thesis. Each section briefly describes a chapter highlighting the contributions and concludes with a reference to the associated published or submitted paper.

1.5.1 Deep 2D Encoder-Decoder CNN for MS Lesion Segmentation

The number of annotated brain MRI images for MS lesion segmentation is always much less than the number of labeled natural images (i.e., non-brain specific images). This makes applying supervised machine learning approaches challenging in MS lesion segmentation. Natural images are completely different in appearance from brain MRI images (Figure 1.5). However, we know that a CNN trained on one domain may generalize to another domain. In chapter 3, we show that the parameters of a CNN trained on natural images in the classification task can generalize well to brain MRI images in MS lesion segmentation task. To this end, we demonstrate that a simple modification on a CNN trained on natural images can produce encouraging MS lesion segmentation results from a clinical perspective. Further, the proposed model is the first whole-brain slice-based (2D) approach allowing to exploit the overall structural information and multi-plane strategy to take advantage of full contextual information for MS lesion segmentation. We evaluate the proposed model on the ISBI dataset (section 1.3.1). Comparing with other state-of-the-art methods, our experiments have shown that the proposed architecture performed better proving evidence that it has a higher capability to effectively identify unhealthy regions while having an overall good overlap with the ground truth in terms of global lesion volume. This can be particularly important in clinical settings where detecting all potential lesions is prioritized over discarding easily identifiable false negatives.

Contributions

- First work to use the parameters from a CNN trained on natural images and fine-tuned to perform MS lesion segmentation of MRI images.
- First whole-brain slice-based FCN for MS lesion segmentation.
- providing evidence that the performance of the proposed method is comparable to the

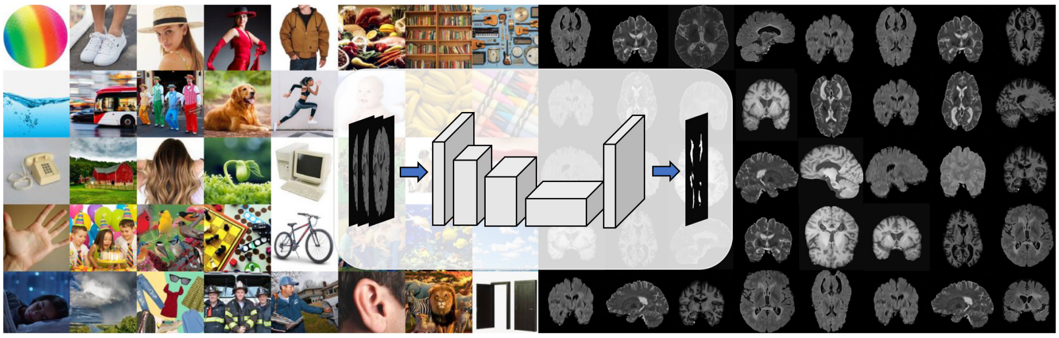


Figure 1.5: The learned parameters of a CNN trained to classify natural images of a large dataset (left side) will generalize to segment MS lesion in brain MRI images (right side).

inter-rater score.

The content of this chapter was published in the MICCAI Workshop on Brain Lesion.

[3] **S. Aslani**, M. Dayan, V. Murino, D. Sona, "Deep 2D Encoder-Decoder Convolutional Neural Network for Multiple Sclerosis Lesion Segmentation in Brain MRI", In International MICCAI BrainLesion Workshop, pages 132-141, Springer, 2018.

1.5.2 Multi-branch CNN for MS Lesion Segmentation

We presented the usefulness of pre-trained parameters to segment MS lesions in the previous section, and the method proposed in this chapter relies on pre-trained parameters. Moreover, we take advantages of a whole-brain slice-based approach and a multi-plane strategy. However, rather than using a single branch CNN, in this chapter, we propose a multi-branch CNN which enables the network to encode information from multiple modalities separately. This feature enables the network to take advantage of each modality individually and allows the network to abstract higher-level features at different granularities specific to each modality. Further, we evaluate different versions of the proposed multi-branch model to find the most performant combination of MRI modalities (T1w, T2w and FLAIR) for MS lesion segmentation. The proposed CNN is evaluated on two different datasets, the ISBI dataset (section 1.3.1) and the NRU dataset (section 1.3.2) showing top performance in comparison to the state-of-the-art.

Contributions

- Multi-modal approach for MS lesion segmentation based on multi-branch CNN.
- Analysis of MRI modalities combination leading to best segmentation performance.

- Providing evidence of top performance on two different datasets.

The content of this chapter was published in NeuroImage.

[4] **S. Aslani**, M. Dayan, L. Storelli, M. Filippi, V. Murino, M.A. Rocca, D. Sona, "Multi-branch Convolutional Neural Network for Multiple Sclerosis Lesion Segmentation", NeuroImage, 169:1-15, 2019.

1.5.3 Scanner Invariant MS lesion Segmentation

Medical data acquisition can vary strongly between centers. Specifically, MRI is often subject to variations due to scanner properties and MRI sequence characteristics. These specificities cause high domain differences between datasets from different centers, which eventually can result in poor generalization. In this chapter, a simple and effective solution is proposed to generalize well our backbone model in the presence of high domain differences. To this aim, an auxiliary loss function has been added to a standard encoder-decoder network to deal with the generalization problem. We test the proposed method showing that using auxiliary loss helps the network to generalize better when using data from multiple centers. The proposed method is evaluated on the UBC dataset (section 1.3.3).

Contributions

- First approach to scanner invariant model for MS lesion segmentation.

- Outperforming base-line methods.

The content of this chapter was published in the ISBI.

[5] **S. Aslani**, V. Murino, M. Dayan, R. Tam, D. Sona, G. Hamarneh, "Scanner Invariant Multiple Sclerosis Lesion Segmentation from MRI", ISBI, 2020.

1.6 Thesis Outline

The remaining of this thesis is organized as follows:

Chapter 2 presents a general overview of basic models for semantic image segmentation using CNNs. It starts with traditional models for natural image segmentation using CNNs. Further, it introduces different semantic segmentation approaches in medical images. Finally, it surveys MS lesion segmentation approaches using algorithms based on deep learning. Chapter 3 through Chapter 5 present the proposed approaches to automated MS lesion segmentation. Finally, Chapter 6 summarizes the proposed works and contributions and suggests future directions.

Chapter 2

Previous Works

During the last decade, there have been significant improvements in semantic image segmentation using CNNs, which have been applied on both natural and medical images [29, 60]. This Chapter will explore literature of semantic image segmentation based on CNNs for natural and medical images, which is mostly attributed to both exploring new architectures by modifying depths, widths, connectivity and proposing new types of components or layers.

2.1 CNN-based Semantic Segmentation of Natural Images

One of the first attempts in CNN-based semantic image segmentation is based on Fully Convolutional Network (FCN), which was proposed by Long et al. [63]. FCN is considered as a stem of most of the successful state-of-the-art methods for semantic image segmentation based on deep learning. The general idea of this approach is to take advantage of existing CNNs as powerful visual models that can learn hierarchies of features [39, 42, 53, 60, 89, 97] and to successfully transfer them into the corresponding FCN versions. This approach is done by replacing the fully connected layers in CNNs with convolution layers to keep the spatial information of the low-resolution attributes which is useful for semantic segmentation. Then those low-resolution attributes are up-sampled using deconvolutional layers to produce a dense pixel-wise classification. One of the most considerable advancements in this approach is that any CNN can be effectively trained end-to-end for semantic image segmentation with inputs of arbitrary sizes. Moreover, FCNs showed state-of-the-art performance over other traditional methods in many datasets like PASCAL VOC [26]. The overall architecture and conventionalizing procedure of a CNN are visualized in Figure 2.1.

Although FCNs have good segmentation performance, they still have a couple of critical limitations. The most important limitation of FCNs is that small objects are often ignored and classified as background. The reason is that the detailed structures of the small objects are often lost or smoothed since the low-level attributes at the end of the network are too

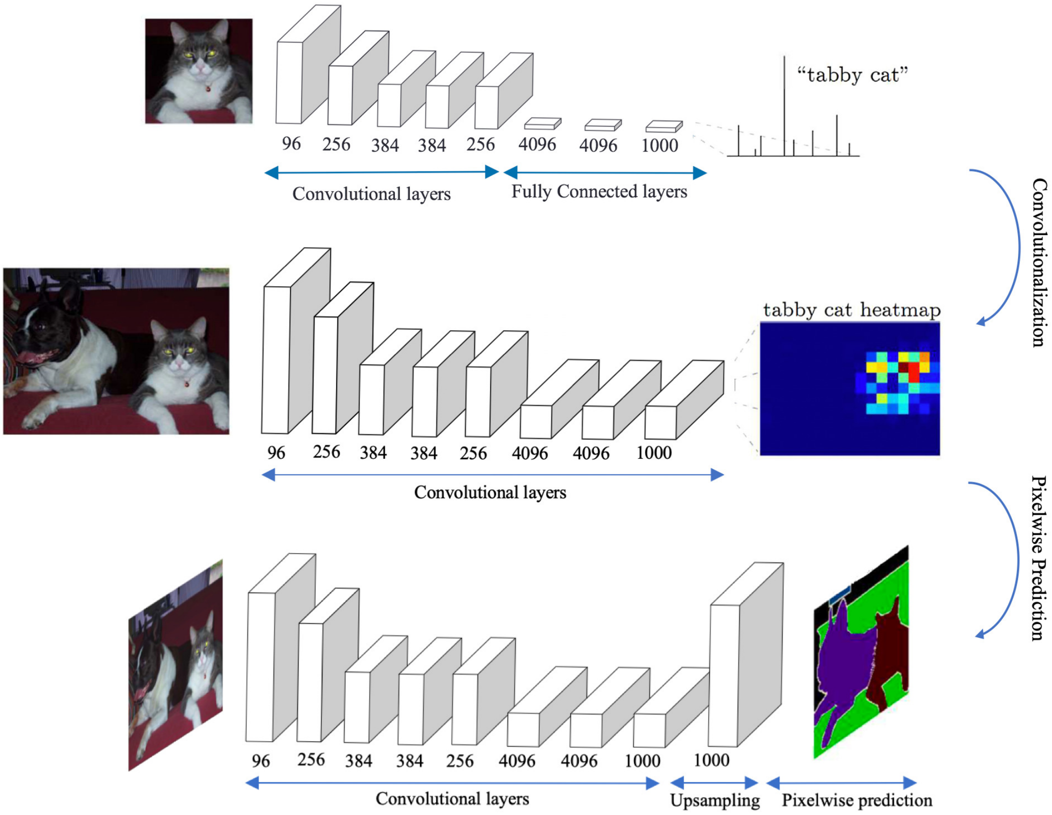


Figure 2.1: overall architecture of FCN [63]. The first row shows a simple CNN for image classification. The second row transfers the mentioned CNN to produce attributes including spatial information of the object by replacing the fully connected layers with convolutional layers. The last row includes an up-sampling stage using deconvolutional layer which allows dense image classification (per-pixel labeling).

coarse and a single deconvolutional procedure for up-sampling is overly simple. To overcome such a limitation, Noh et al. [73] proposed an encoder-decoder network known as Deconv-Net including multi-stage upsampling layers to capture a different level of shape details; lower layers for overall shape and higher layers for class-specific fine details. They proposed two layers in the decoder part of the network including deconvolution and unpooling layers. Generally speaking, pooling operation is used in the encoders to filter noisy activations and keep the robust activations. However, this operation removes spatial information of the objects which is not good for semantic segmentation. To address this problem, an un-pooling layer was proposed to reconstruct the original size of activation in the decoder network by recording the place of maximum activation during the pooling operation and putting it back to its original place during the decoding stage. Although output activations of the un-pooling are enlarged correctly based on information coming from pooling layers, the resulted activations are sparse activations that are useful to find the location of the objects.

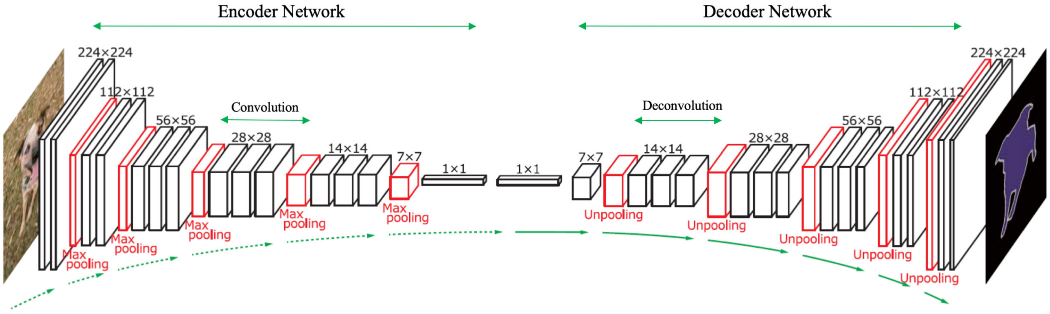


Figure 2.2: An illustration of the Deconv-Net [73] architecture.

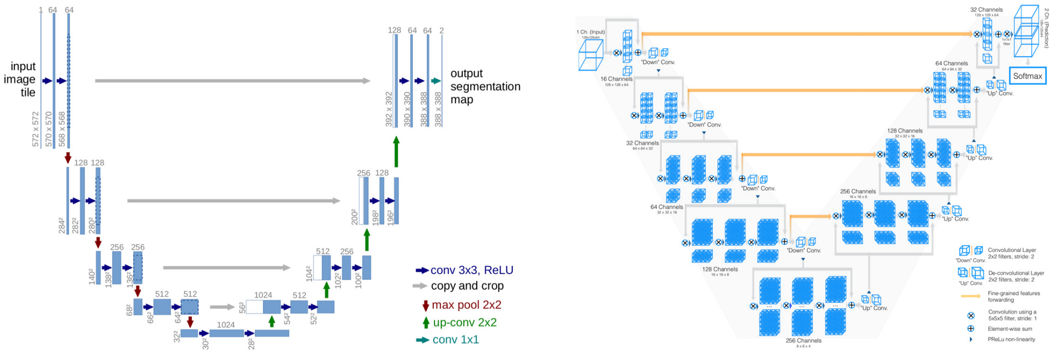


Figure 2.3: An illustration of the U-Net [80] (left) and V-Net [65] (right) architectures.

The deconvolution layer has been used after un-pooling layers for densifying the sparse activations which is useful for capturing class-specific information. Therefore, un-pooling operation attempts to find the location of the objects and the deconvolution layer is useful for class-specific information and also using these layers at multi-stage of the decoder helps to reconstruct fine-detailed information of the objects. Figure 2.2 shows a general overview of the proposed network.

Following the same idea, Ronneberger et al. [80] proposed a network called U-Net including an encoder to capture the context and a symmetric decoder which enables precise localization. The most important contribution in this work is the skip-connections between encoder and decoder which improved the segmentation performance drastically and addressed the problem of vanishing gradients. The general framework of the network can be seen in Figure 2.3. Milletari et al. [65] proposed a similar network known as V-Net adding residual connections and replacing 2D operations with their 3D versions to process volumetric datasets. Moreover, they also proposed a new loss function based on a widely used segmentation metric, Dice. The general framework of the network can be seen in Figure 2.3.

Pyramid scene parsing network (PSP-Net) proposed by Zhao et at. [110] is another CNN-

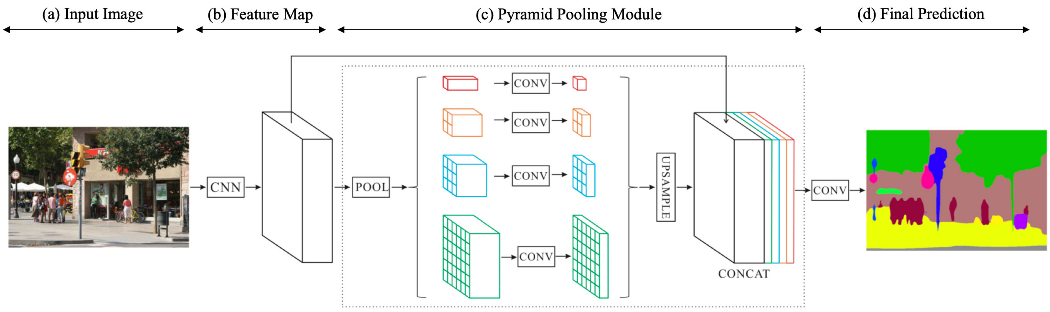


Figure 2.4: An illustration of the PSP-Net [113] architectures. (a) input image, (b) using a CNN to extract the initial feature maps, (c) pyramid parsing module: extracting different sub-region representations followed by up-sampling and concatenation layers to get the final feature maps, (d) final feature maps feeding to a convolution layer to get the final prediction.

based method giving a promising direction for a pixel-level classification task. This network exploits local and global context information, which is the most important problem of FCNs. The aforementioned approach is done by utilizing different region-based context aggregation via a pyramid pooling module. Figure 2.4 shows the general architecture of the PSP-Net.

According to the literature, there exist several networks which are the modified versions (changing the depth of the network by adding/removing blocks) of the described architectures [6, 15, 16, 58, 76].

Recently, Chen et al. [17] proposed a network based on DeepLabV3 [16] which takes the advantage of dilated convolutions [16] and the pyramid parsing module [113]. The proposed network outperformed many state-of-the-art methods on several datasets like PASCAL VOC [26] and Cityscapes [111]. Specifically, DeepLabv3+ [17] uses a simple yet effective decoder module to refine the segmentation results, especially along object boundaries using dilated convolutions and pyramid features. The general framework of the network can be seen in Figure 2.5.

2.2 CNN-based Segmentation of Medical Images

During the last years, deep learning methods, especially CNNs [55] have demonstrated outstanding performance in medical image segmentation. They could provide state-of-the-art results in different problems such as segmentation of neuronal structures [80], retinal blood vessel extraction [59] and brain extraction [52].

In particular, CNN-based medical image segmentation methods can be categorized into two different groups: patch-based (region-based) and image-based (FCN-based) methods.

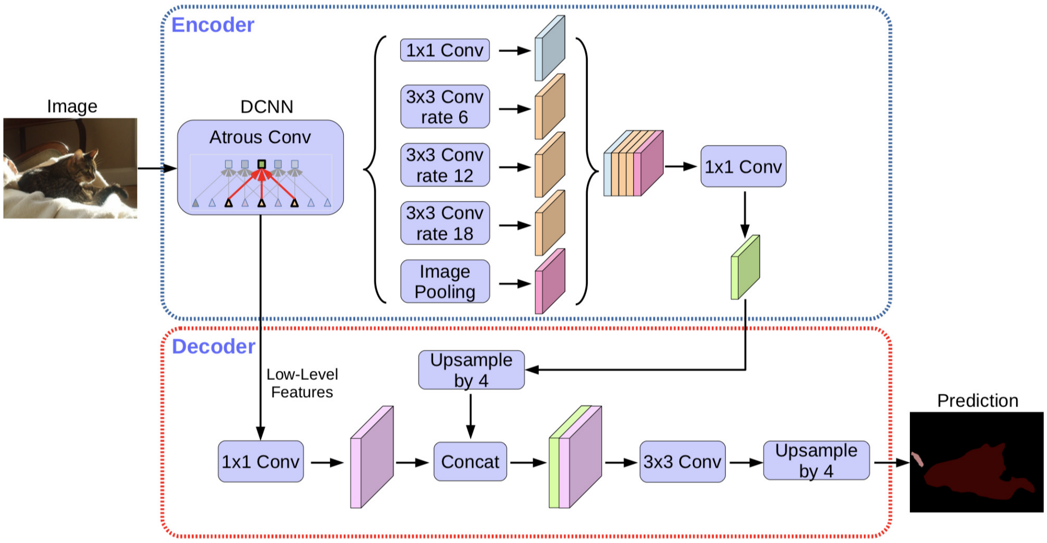


Figure 2.5: An illustration of the DeepLabV3+ [17] architecture. The network includes the encoder part which extracts multi-scale contextual information by applying dilated convolution at multiple scales using pyramid parsing module, while the simple yet effective decoder part refines the segmentation results along object boundaries.

In patch-based methods, a moving window scans the image generating a local representation for each pixel/voxel. Then, a CNN is trained using all extracted patches, classifying the central pixel/voxel of each patch. These methods are frequently used in medical image analysis since they considerably increase the number of training samples. However, they suffer from an increased training time due to repeated computations of the over-lapping features associated with the sliding window. Moreover, they neglect the information on the global structure because of the small size of patches [100]. Figure 2.6 shows different examples of patch-based methods for white matter hyper-intensities segmentation proposed by Ghafoorian et al. [32].

On the contrary, image-based approaches process the entire image exploiting the information on the global structure [9, 100]. These methods can be further categorized into two groups according to the processing of the data: slice-based segmentation of 3D data [100] and 3D-based segmentation [9].

In slice-based segmentation methods, each 3D image is converted to a set of 2D slices, which are then processed individually. Subsequently, the segmented slices are concatenated together to reconstruct the 3D volume. In most of the proposed pipelines based on this approach, the segmentation is not accurate, most likely because the method ignores part of the contextual information. Figure 2.7 shows an example of the slice-based segmentation which is proposed by [100].

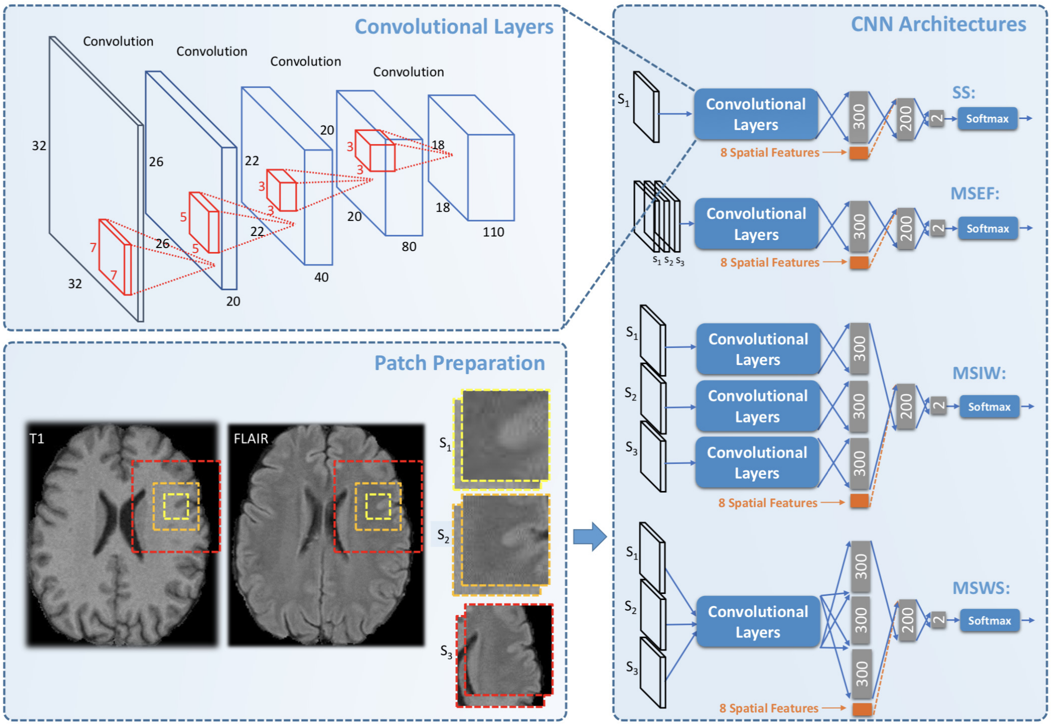


Figure 2.6: Patch preparation process (left bottom) and different CNN architectures (right) proposed by Ghafoorian et al. [32]. As a first stage, they extract patches with three different sizes in patch preparation step. Then a different version of CNNs has been trained based on the early fusion (second row right) and late fusion (third and fourth rows right) of the extracted patched. They also added a set of auxiliary handcrafted spatial features to the networks to increase the segmentation performance.

In 3D-based segmentation, a CNN with 3D kernels is used for extracting meaningful information directly from the original 3D image. The main significant disadvantage of these methods is related to the training procedure, which usually fits a large number of parameters with a high risk of over-fitting in the presence of small datasets. Unfortunately, this is a quite common situation in medical applications [9]. To overcome this problem, 3D cross-hair convolution has been proposed [61, 98], where three 2D filters are defined for each of the three orientations around a voxel (each one is a plane orthogonal to X, Y, or Z axis). Then, the sum of the result of the three convolutions is assigned to the central voxel. The most important advantage of the proposed idea is the reduced number of parameters, which makes training faster than a standard 3D convolution. However, compared to standard 2D convolution (slice-based), still, there are three times more parameters for each layer, which increases the chance of over-fitting in small datasets. An example of 3D-based segmentation method proposed by Milletari et al. [65] can be seen in Figure 2.3 (right).

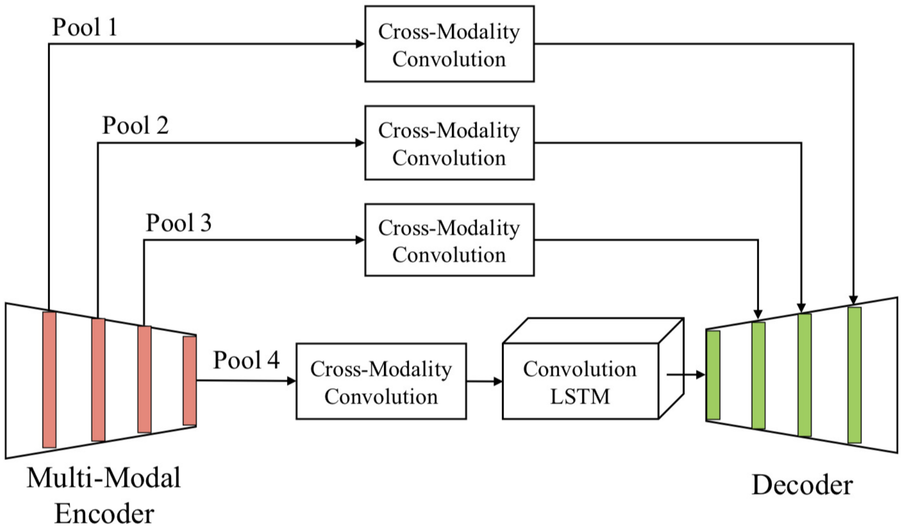

Pool 1
Pool 2
Pool 3
Pool 4
Cross-Modality Convolution
Cross-Modality Convolution
Cross-Modality Convolution
Cross-Modality Convolution
Convolution LSTM
Multi-Modal Encoder
Decoder

Figure 2.7: An illustration of the slice-based segmentation method architecture proposed by [100]. Multi-modal encoder extracts feature maps from different modalities. Then the cross-modality convolution blocks aggregate the extracted features maps after pooling layers. The final feature maps are fed into convolutional LSTM and decoder to generate the final prediction.

2.3 CNN-based Segmentation of Multiple Sclerosis

The literature offers some methods based on CNNs for MS lesion segmentation.

Vaidya et al. [101] proposed a shallow 3D patch-based CNN including two convolution layers, one multi-layer perceptron and a softmax layer as can be seen in Figure 2.8. Sparse convolution idea [57] has been used for effective and fast training. Moreover, they added a post-processing stage, which increased the segmentation performance by applying a WM mask to the output predictions. ISBI dataset [11] including 4 MRI modalities has been used to evaluate the performance of the proposed method.

Ghafoorian et al. [32] developed a deep CNN based on 2D patches to increase the number of training samples and avoid the over-fitting problems of 3D-based approaches. The proposed method was 5 layers patch-based CNN taking 32×32 patches in four channels (ISBI dataset with four MRI modalities [11]) as its input samples. There were also four convolution layers with 15 filters of size 13×13, 25 filters of size 9×94, 60 filters of size 7×7 and eventually 130 filters of size 3×3. A final softmax layer classified the resulting responses to the filters in the last convolution layer.

Similarly, in [8], multiple 2D patch-based CNNs have been designed to take advantage of the

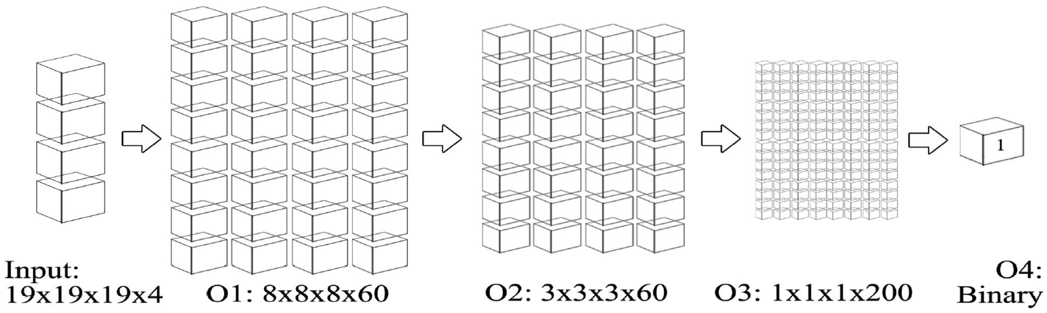
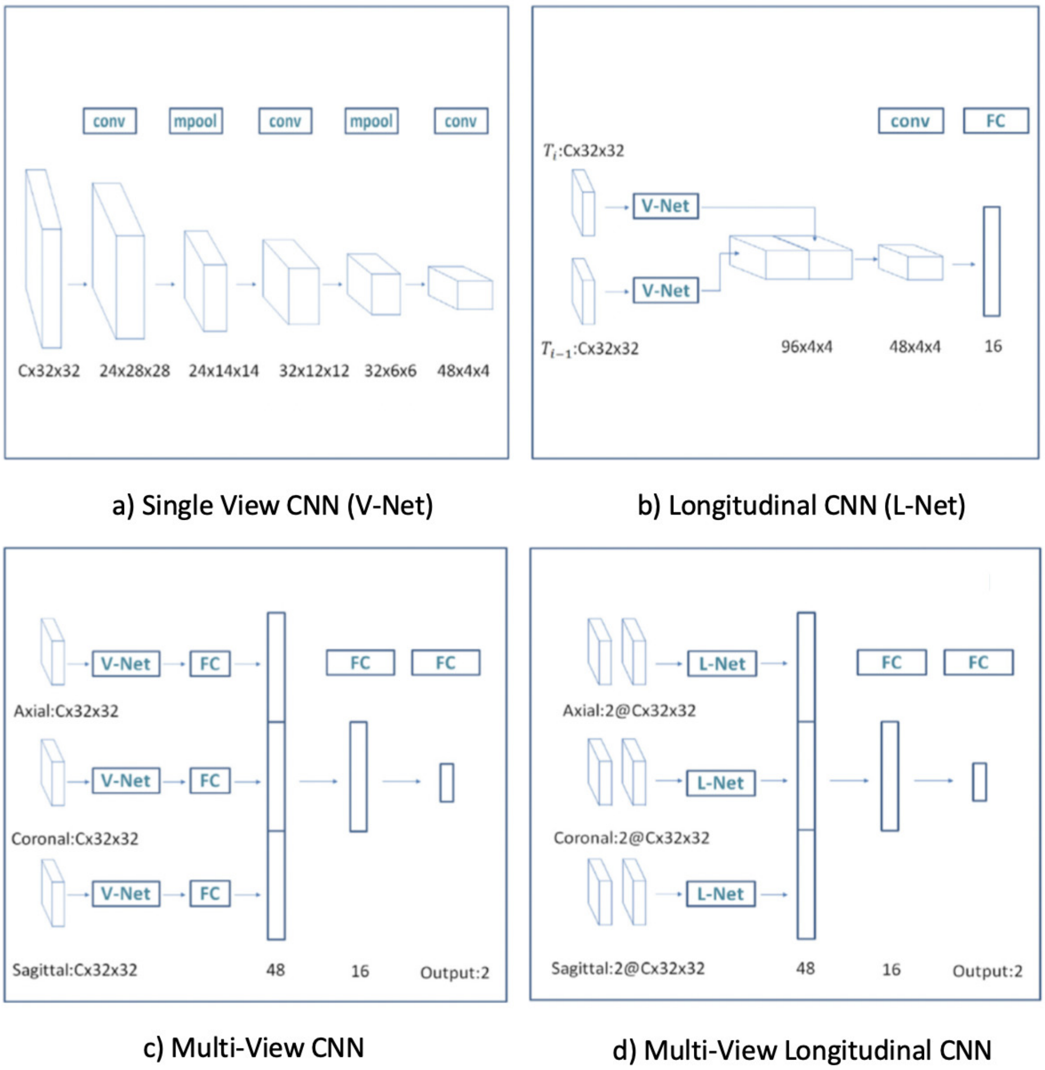


Figure 2.8: An illustration of the patch-based segmentation method proposed by [57]. This 3D patch-based CNN has 4 layers: 1) convolution layer with 60 filters of $4 \times 4 \times 4$ with average pooling of $2 \times 2 \times 2$, 2) convolution layer with 60 filters of $3 \times 3 \times 3$ with average pooling of $2 \times 2 \times 2$, 3) a Multi-layer Perceptron, 4) a softmax layer for final prediction. O1, O2, and O3 show sizes of the respective outputs of the corresponding layers.

common information within longitudinal data. The main difference between the proposed CNNs is the type and number of input patches to the network, for instance: single modality with single time point (SMST), multiple modalities with single time point (MMST), single modality with multiple timepoints (SMMT) and multiple modalities with multiple time points (MMMT). The main network which can be considered as a common block for all other networks is V-Net shown in Figure 2.9(a). This block has the flexibility to be fed with a single input (SMST) or multiple-input (MMST) by modifying the parameter C. The network proposed to process the longitudinal data (SMMT or MMMT) is L-Net shown in Figure 2.9(b). This network includes two V-Nets which process the current and previous time points separately. The two separate representations are then concatenated and processed by other layers to get the final prediction. To take advantage of the full representation of the input patch, they proposed another version of CNN including three different views of a single voxel (axial, coronal and sagittal) as input patches to the network as can be seen in Figure 2.9(c). Each view is processed by separate V-Net and the resulted representations are concatenated for final prediction. Moreover, to take advantage of multiple time points, they replaced the V-Nets blocks with L-Nets shown in Figure 2.9(d). The proposed models have been evaluated using ISBI dataset with four MRI modalities [11].

Valverde et al. [102] proposed a pipeline relying on a cascade of two 3D patch-based CNNs. As a first step, they created a set including all available patches from each single MRI modalities. Then, patches with central voxel intensities less than 0.5 in FLAIR modality have been removed from the set. Moreover, to deal with the data imbalance problem, they randomly removed negative samples (healthy) in the set. They trained the first network using selected patches in the mentioned set, and the second network was used to refine the training procedure utilizing misclassified samples from the first network. The general processing pipeline



a) Single View CNN (V-Net)

b) Longitudinal CNN (L-Net)

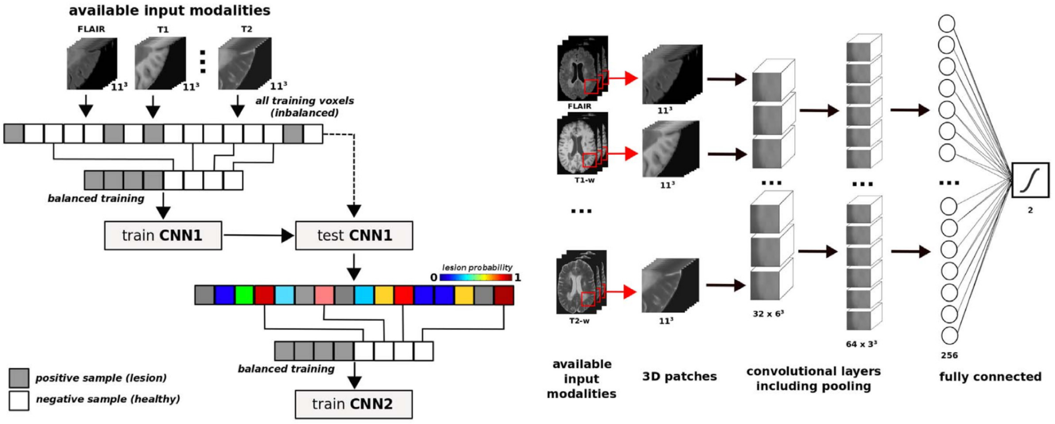
c) Multi-View CNN

d) Multi-View Longitudinal CNN

Figure 2.9: The network architectures proposed by Birenbaum et al. [8]. (a) V-Net: the main block of the proposed CNNs (b) L-Net: the network to process longitudinal dataset (c) Multi-view CNN to take advantage of different views of the each voxel (SMST with c=1 and MMST with c=4 (d) Multi-view longitudinal CNN take advantage of both different views of each voxel and longitudinal dataset (SMMT with c=1 and MMMT with c=4).

of the cascade-based training can be seen in Figure 2.10(a) and also the overall architecture of CNN (7 layers) which was same for first and second training can be seen in Figure 2.10(b). The proposed network has been evaluated using two different datasets: The MICCAI 2008 dataset [94] and a clinical private dataset, both composed by three MRI modalities.

Roy et al. [81] proposed a method based on FCNs (slice-based) including two pathways as can be shown in Figure 2.11. They used different MRI modalities as input for each pathway and the outputs were concatenated and processed with another shallow network to create



a) Cascade-based Pipeline

b) CNN architecture

Figure 2.10: An illustration of 3D patch-based segmentation method proposed by Valverde et al. [102]. (a) Cascade-based pipeline: the output of the first network was used to refine the training procedure of the second network by selecting the misclassified samples from first training (b) The proposed 7-layer CNN model trained using 3D patches from different MRI modalities.

a membership function for healthy or healthy regions. Only a single view of each voxel (axial side) was selected to extracted 2D slices. Then, from each extracted axial slice, small patches with the size of 35 × 35 generated for each MRI modality as input to the network. Note that the sizes of the input and outputs of all layers are kept identical to the original input patch size by zero paddings. They evaluated the proposed method using two datasets: ISBI dataset [11] and a clinical private dataset, using only two MRI modalities: FLAIR and T1w.

Recently, Hashemi et al. [36] proposed a method relying on FCNs (3D-based) using the idea of densely connected blocks. The general architecture of the proposed method can be seen in Figure 2.12. They generated a set including 3D patches related to different MRI modalities with the size of 64 × 64 × 64 and 50% overlap area as the input to the proposed network. The overall architecture is similar to the U-Net [80] with contacting path, expanding path and also shortcut connections between them. However, instead of using simple convolution layers, they used densely connected blocks [42] with the idea of skip connection between layers. They also developed an asymmetric loss function for dealing with highly unbalanced data. They evaluated the proposed method on two publicly available datasets: ISBI dataset [11] and MSSEG dataset [20].

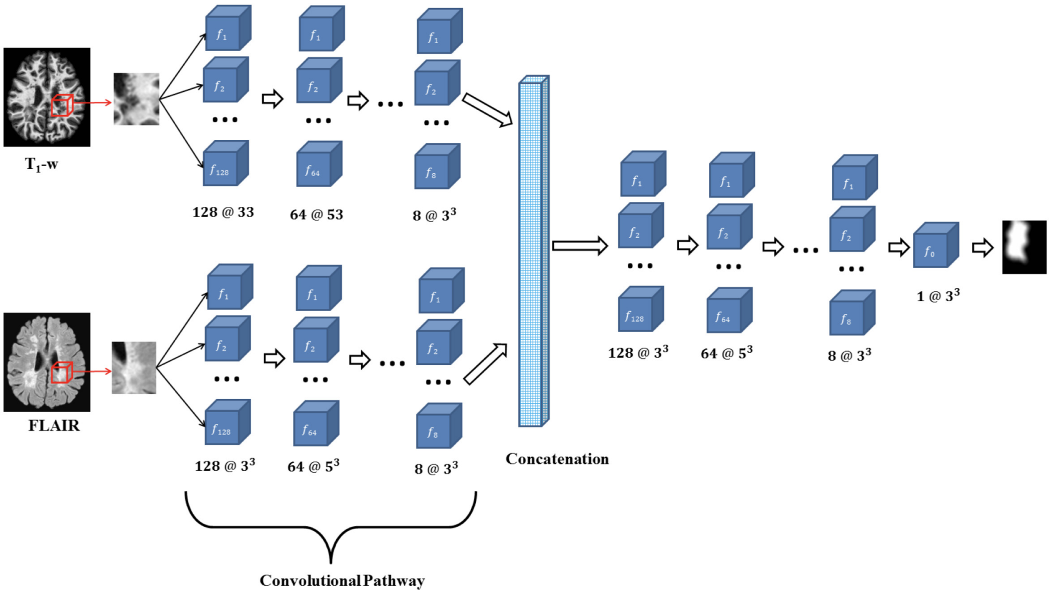


Figure 2.11: The general overview of the network architecture proposed by Roy et al. [81]. 2D patches were extracted from axial slices of two different MRI modalities and were fed into the network in parallel pathways. The feature maps related to each pathway were concatenated and supplied to another shallow pathway to predict the final membership function of the input patch.

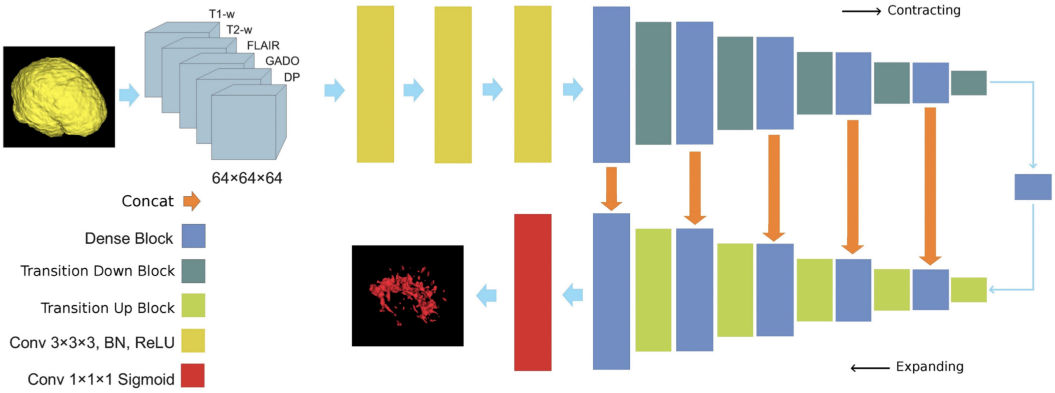


Figure 2.12: The general framework of the proposed network by Hashemi et al. [36]. The proposed network is fed using the patches with a size of 64 × 64 × 64 and five channels corresponding to the five different MRI modalities. It includes eleven densely connected blocks, five transitions down blocks, five transition up blocks and four convolution layers in both contracting and expanding paths. A sigmoid layer is used as the last layer of the network to get the final prediction.

Even though all the proposed patch-based techniques have good segmentation performance, they suffer from lacking global structural information. This means that the global structure of the brain and the absolute location of lesions are not exploited during the segmentation.

In contrast to the above-mentioned methods, Brosch et al. [9] developed a whole-brain segmentation method using a 3D CNN. They used a single shortcut connection between the coarsest and the finest layers of the network, which enables the network to concatenate the features from the deepest layer to the shallowest layer to learn information about the structure and organization of MS lesions. However, they did not exploit middle-level features, which have been shown to have a considerable impact on the segmentation performance.

2.4 Summary and Conclusion

In this Chapter, we presented a general overview of the basic models for semantic image segmentation based on CNNs for natural and medical images.

Given the presented literature related to MS lesion segmentation using CNNs, most of the proposed methods show good segmentation performance. However, as it was mentioned previously, there are some limitations that can be addressed to improve the segmentation performance. For instance, the 2D/3D patch-based methods which are commonly used in medical image segmentation suffer from lacking global structural information. Regarding the slice-based approaches, the segmentation is not accurate, because the methods ignore part of the contextual information due to considering a single view of each voxel (usually axial view). Regarding the 3D-based approaches, the main significant disadvantage of these methods is related to the training procedure, which usually have high risk of overfitting in the presence of small datasets. Moreover, almost all of the proposed methods suffer from poor generalization since they are optimized to produce segmentation performance on a single domain (datasets from a single center).

In the following two Chapters, we propose two deep models to address the lack of global structural and contextual information. Moreover, to avoid overfitting problem, we used a pre-trained network. Then, in the last chapter, a regularization method is proposed to generalize the backbone segmentation model in the presence of multi-center dataset.

Chapter 3

Deep 2D Encoder-Decoder CNN for MS Lesion Segmentation

3.1 Introduction

MS is known as one of the most important diseases of the central nervous system of the brain. The detection, segmentation, and quantification of the MS lesions is an important task as it can help to characterize the progression of the disease and monitor the efficacy of a candidate treatment [62].

Recently, CNNs have shown excellent performance in image classification task and are consistently used in many competitions such as ImageNet challenge in which competitors try to propose solutions to classify hundreds of different natural objects [82]. CNNs not only show state-of-the-art performance when trained for a specific task with millions of images, but experiments have shown that a pri-trained CNN on a dataset can generate a set of useful representations that are generic for different image tasks that the CNN was not originally trained for [25]. Among all tasks, image segmentation is one of the most important and common tasks in which researchers manipulated and adapted pre-trained CNNs to obtain state-of-the-art performance in several datasets [35].

In this Chapter, we propose an automated segmentation approach based on a two-dimensional (2D) CNN pre-trained on ImageNet dataset to segment brain multiple sclerosis lesions from multi-modal magnetic resonance images. The proposed model is made as a combination of two deep sub-networks. An encoding network extracts different feature maps at various resolutions and a decoding part upconvolves the feature maps combining them through shortcut connections during an upsampling procedure. We concentrated on whole-brain slice-based segmentation to prevent both the overfitting present in 3D-based segmentation [9] and the lack of global structure information in patch-based methods [31, 81, 102]. The robustness of the method is improved by exploiting the volumetric slicing in all three possible imaging planes (axial, coronal and sagittal). Indeed, we used the three different imaging axes of each 3D input MRI in an ensemble framework to exploit the contextual information in all three anatomical planes. Moreover, this model has been used as a multi-modal framework to make use of all of the information available within each available MRI modality, typically FLAIR, T1w, and T2w.

3.2 Method

We propose a 2D end-to-end CNN based on the residual network (ResNet) [39]. The core idea of ResNet is the use of identity shortcut connections, which allows for preventing gradient vanishing. Thanks to this benefit, ResNets have shown outstanding performance in computer vision problems, specifically in the image recognition task.

We modified ResNet50 (version with 50 layers) to work as a pixel-level segmentation network. This has been obtained by changing the last prediction layer with a dense pixel-level prediction layer inspired by the idea of the fully convolutional network [63]. Since the output of the last convolutional layer of ResNet is very coarse compared with the input image resolution (32 times smaller than the original image), upsampling such high-level feature maps with a simple operation like bilinear interpolation as described in FCNs [63] is not an effective solution. Therefore, inspired by [80], we propose a multi-pass upsampling network using the advantages of multi-level feature maps with skip connections.

In the following sections, we first describe how the input features were generated by decomposing 3D data into 2D images. Then, we describe the proposed network architecture in detail and the training procedure.

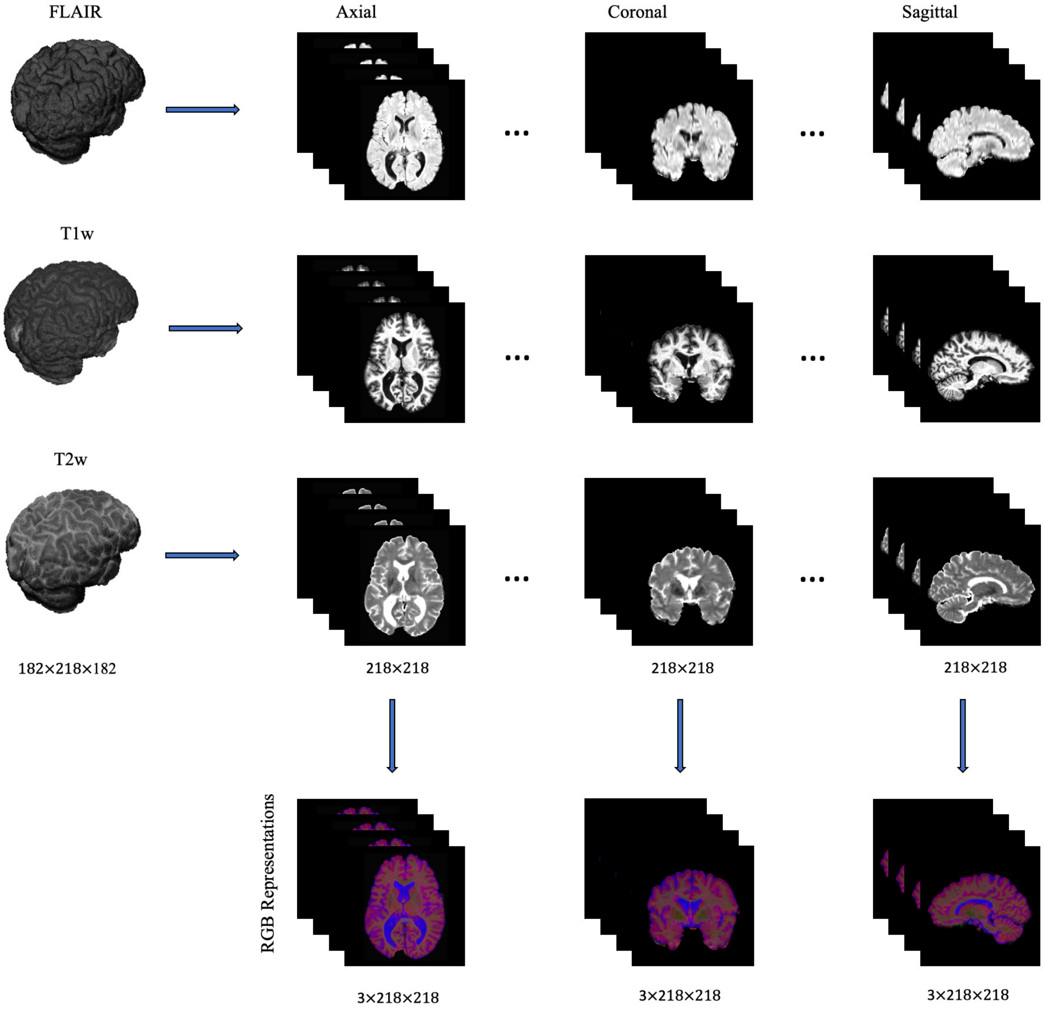


Figure 3.1: Input feature extraction pipeline. From each original 3D MRI image, axial, coronal and sagittal planes were extracted for each modality. Since the size of extracted slices was different with respect to the plane orientations (axial= 182×218, coronal=182×182, sagittal=218×182), all slices were zero-padded while centering the brain so to obtain all slices with the same size (218×218), no matter their orientation. In our specific application, 3 modalities were used (FLAIR, T1w, T2w), hence, multi-channel slices (represented here as RGB images) were created by grouping together the corresponding slices of each modality.

3.2.1 Input Features Preparation

From each original volumetric MRI modality, axial, coronal and sagittal planes are considered by extracting 2D slices along the x, y, z axes of the 3D image. Since the size of the imaging planes differed according to the imaging axes (axial= 182×218, coronal=182×182, sagittal=218×182), we zero-padded each slice (while centering the brain), so to obtain the same consistent size irrespective of the imaging plane. Further, the same operation was applied to all modalities. Then, for all slices belonging to each plane orientation and all modalities were stacked together to create a single multi-channel input stack. Since three modalities were used in our experiments, the obtained multi-channel slices included three channels which can be represented as RGB images. Figure 3.1 illustrates the described procedure using three modalities, FLAIR, T1w, and T2w.

3.2.2 Network Architecture Details

In deep networks, features from deep layers include high-level semantic information. On the contrary, features from the early layers contain low-level spatial information. It was shown that features from the middle layers also provide information which can be effective to increase the performance of the segmentation [80]. Therefore, combining multi-level features from different stages of the network makes the feature map richer than just using single scale feature maps. The intuition behind our architecture is to use these multi-level feature maps by adding multiple upsampling layers with skip connections [80] to the ResNet output of all intermediate layers. The diagram of the proposed network for segmentation can be seen in Figure 3.2.

We divided the ResNet50 into 5 blocks in the downsampling part according to the resolution of feature maps. In the original ResNet50 architecture, the first layer is composed of a 7×7 convolution layer with stride 2 to downsample the input by an order of 2. Then, a 3×3 max pooling layer with stride 2 is applied to further downsample the input followed by a bottleneck block without downsampling. Subsequently, three other bottleneck blocks are applied, each one followed by a downsampling convolution layer with stride 2. Therefore, ResNet50 can be organized into five blocks with different resolutions ($109 \times 109, 54 \times 54, 27 \times 27, 14 \times 14, 7 \times 7$). In the upsampling sub-network, the encoded features from different scales are decoded step by step using upsampling fused features (UFF) blocks. Each UFF block includes one upconvolution layer with kernel size 2×2 and stride 2, one concatenation/fusion layer and two convolution layers with kernel sizes 3×3. After each layer, a rectifier linear activation function (ReLU) is applied [71]. The upconvolution layer is used to transform low-resolution feature maps into the higher resolution maps. Then a simple concatenation layer is used for combining the two sets of input feature maps. Two convolution layers are further used for adaptation as described in [80], and the output goes to the next block. The number of feature maps after each UFF block is halved. At the end of the network, a soft-max layer of size 2 is used to get output probability maps, identifying pixel-wise positive (lesion) or negative (non-lesion) classes.

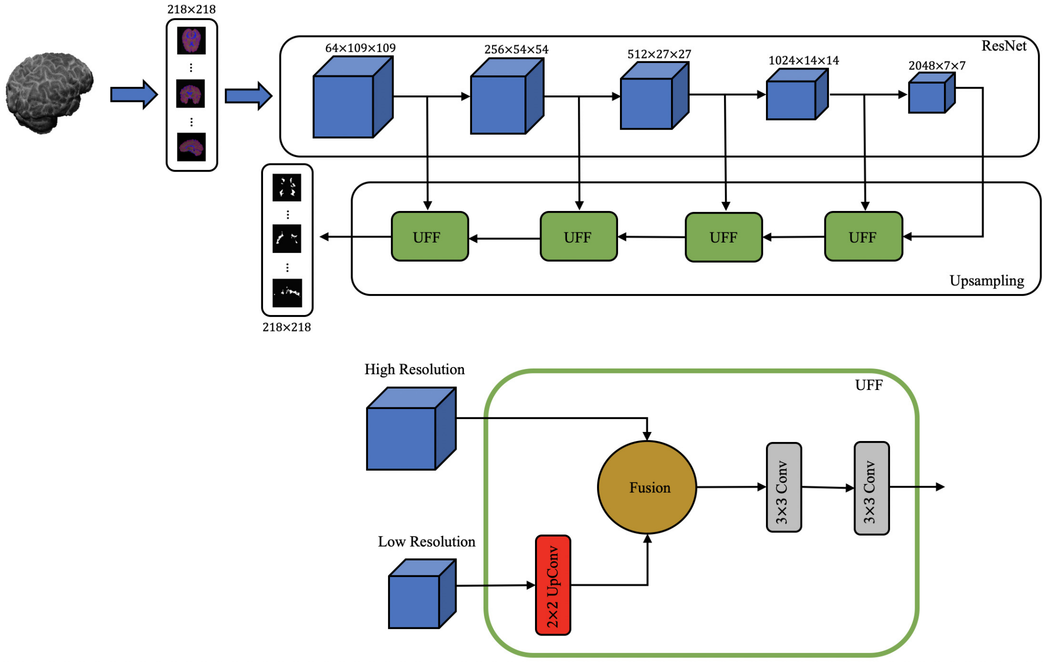


Figure 3.2: General framework of the proposed network for MS lesion segmentation. The first sub-network (ResNet50) encodes the input 2D slices (with the size of 218 × 218) into different feature sets at various resolutions. This sub-network was organized into 5 blocks according to the resolution of the representations during the encoding. For example, the first block denotes 64 representations with resolution 109 × 109. The second sub-network (Upsampling) decodes the representations provided by the encoder network. This sub-network gradually converts low-resolution representations back to the original resolution of the input image using UFF blocks. Each UFF block has two sets of input representations with different resolutions. This block is responsible for upsampling the low-resolution representations and combining them with higher-resolution representations.

3.2.3 Implementation Details

To train the proposed CNN, a training set was created using the pipeline mentioned in section 3.2.1. To remove uninformative samples and limit extremely unbalanced data from the whole training set, a subset was determined by selecting only slices with at least one lesion pixel. This meant that 2D slices without lesions were omitted from the training set.

As suggested in [37], simple off-line data augmentation was applied to the training set to increase training samples. Increasing training samples has been shown to increase the performance of the network. Therefore, we increased the number of the samples by a factor of 5 simply by either rotating each extracted slice by 4 possible angles (5°, 10°, -5°, -10°) and flipping (right to left) the images with their original rotation (no combination of flipping and rotation were included in the data augmentation procedure).

To optimize network weights with early stopping criteria, we split the training set into different training and validation sets depending on the experiments as described in the following section. According to the network initialization, in the first sub-network, the pre-trained ResNet50 on ImageNet was used and the weights from the second sub-network (Upsampling) were randomly initialized. The adaptive learning rate method (ADADELTA) [107] was used to tune the learning rate with an initial learning rate of 0.001. Binary cross-entropy was used as a loss function to train the proposed network. The maximum number of training epochs was fixed to 500, and the best model was selected according to the validation set.

We implemented our proposed model in Python language [1] using Keras[2] [18] with Tensorflow[3] [1] backend. We used a Nvidia GTX Titan X GPU for all experiments.

3.3 Experiments

We evaluate the performance of the proposed method on ISBI dataset (refer to section 1.3.1 for details). From the ISBI dataset, we selected the preprocessed version of the images available online at the challenge website. All images were already skull-stripped using Brain Extraction Tool (BET) [91], rigidly registered to the $1mm^3$MNI-ICBM152 template [74] using FMRIB's Linear Image Registration tool (FLIRT) [46, 47] and N3 intensity normalized [90].

For evaluation purposes, two different experiments were implemented according to the availability of ground truth. In the first experiment, we ignored the official ISBI test set to only considering data with the available ground truth. To get a fair result, we did a leave-one-out cross-validation training (at subject level: 3 subjects for training, 1 subject for validation and 1 subject for testing). In this experiment, *DSC*, *LTPR*, and *LFPR* measures were used to make our results comparable to those obtained in [9, 48, 64].

For the second experiment, the official ISBI test set was used as our test set so the ground truth was not available. We trained the network using leave-one-out cross-validation over all 5 subjects in the training set (4 subjects for training and 1 subject for validation). We did majority voting over all classifiers evaluated the ensemble of 5 models on the test set. The 3D output binary lesion maps were submitted to the website of ISBI for evaluation purposes. In this experiment, a score is measured online (using the challenge website). As described in section 1.4.8, the mentioned score is a weighted average of different metrics including *DSC*, *LTPR*, *LFPR*, *PPV*, and *VD*.

[1] https://www.python.org
[2] https://keras.io
[3] https://www.tensorflow.org

Table 3.1: Comparison of our method with the other state-of-the-art methods. GT1 and GT2 show that the corresponding model was trained using annotation provided by rater 1 and rater 2 as the ground truth respectively.

Method	Rater 1			Rater 2		
	DSC	*LTPR*	*LFPR*	*DSC*	*LTPR*	*LFPR*
Rater 1	-	-	-	0.7320	0.6450	0.1740
Rater 2	0.7320	0.8260	0.3550	-	-	-
Jesson et al. [48]	**0.7040**	0.6111	**0.1355**	**0.6810**	0.5010	**0.1270**
Maier et al. [64] (GT1)	*0.7000*	0.5333	0.4888	*0.6555*	0.3777	*0.4444*
Maier et al. [64] (GT2)	*0.7000*	0.5555	*0.4888*	0.6555	0.3888	*0.4333*
Brosch et al. [9] (GT1)	0.6844	*0.7455*	0.5455	0.6444	*0.6333*	0.5288
Brosch et al. [9] (GT2)	0.6833	*0.7833*	0.6455	0.6588	*0.6933*	0.6199
Ours (GT1)	0.6980	**0.7460**	*0.4820*	0.6510	**0.6410**	0.4506
Ours (GT2)	0.6940	**0.7840**	0.4970	*0.6640*	**0.6950**	0.4420

Note that for each test subject, we first extracted all the slices, following the approach described in the previous section 3.2.1. Feeding each 2D slice to the network, we got as output the associated 2D binary lesion classification map. Since the original data was duplicated three times in the input, once for each slice orientation (coronal, axial, sagittal), concatenating the binary lesion maps belonging to the same orientation resulted in three 3D lesion classification maps. These three lesion maps were combined via majority voting (the most frequent lesion classification was selected).

3.4 Results

In the first experiment, as described previously, we evaluate the performance of our network on the training set. Table 3.1 shows the performance of our method in comparison with other previously proposed methods. As can be seen, our method has the highest performance regarding *LTPR* metric while having a high *DSC* which means that the proposed method can identify lesions with higher precision than other methods, also having a good overlap in terms of lesion volume overall. Figure 3.3 shows an example of the output of our network in comparison to the corresponding ground truth.

In the second experiment, the performance of the proposed method was also evaluated on the official ISBI test set using the challenge web service. At the time we submitted the results, we obtained a score of 89.85 which is comparable to the ISBI inter-rater score scaled to 90. The detailed result for each subject is available online on the ISBI MS lesion segmentation challenge website[4].

[4] http://iacl.ece.jhu.edu/index.php/MSChallenge

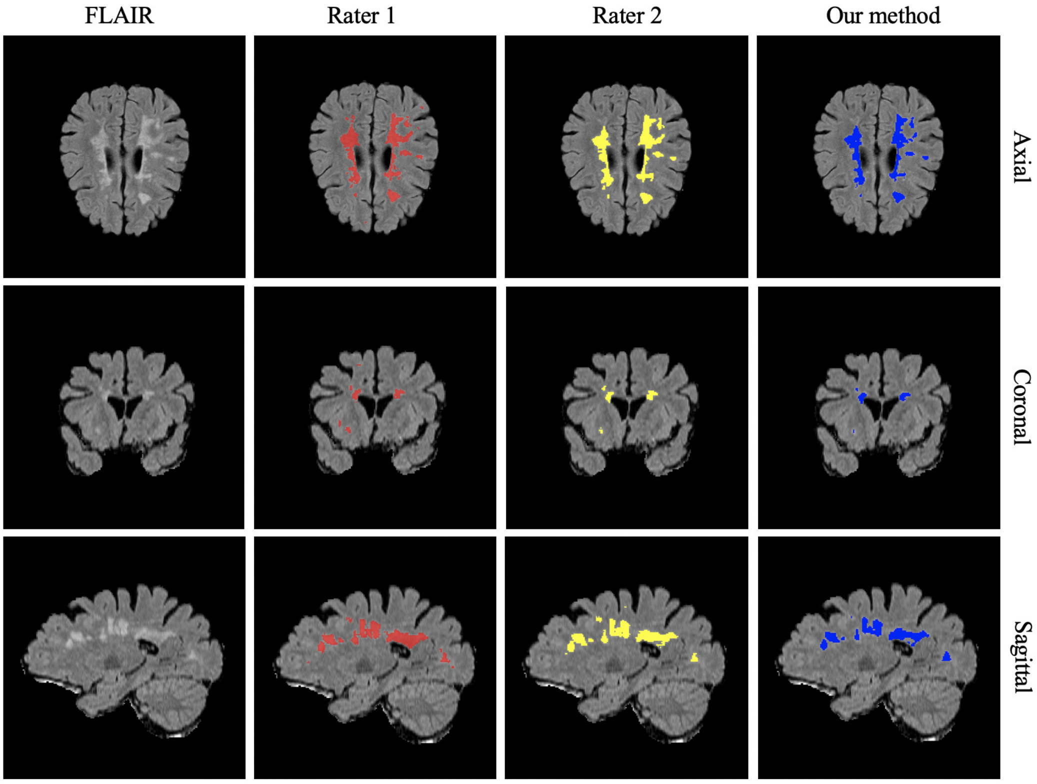


Figure 3.3: An example of our network results in the axial, coronal and sagittal planes. First column: original FLAIR modality from different views, second column: ground truth related to the rater 1, third column: ground truth related to the rater 2, last column: segmentation output from the proposed method

3.5 Discussion and Conclusion

We have proposed an automated method for the brain MS lesion segmentation based on a pre-trained 2D CNN. The presented approach is a deep end-to-end CNN including two pathways, a contracting path which extracts multi-resolution representations by encoding the input image (ResNet) and an expanding path which decodes the provided representations gradually by upsampling and fusing them. Our CNN has been trained using whole-brain slices as inputs to take advantage of the spatial information about the location and shape of MS lesions. Moreover, it has been designed for multi-modality (FLAIR, T1w, T2w) and multi-planes (axial, coronal and sagittal) analysis of MRI images. Transfer learning has showed to be a good solution in deep learning based approaches when inadequate amount of data is available for training which is very common problem in medical domain [13, 14, 41]. Therefore, we have used a CNN pre-trained on ImageNet. This approach not only helps boosting the performance of the network but also significantly reduces overfitting.

The proposed method has been evaluated using the publicly available dataset (ISBI 2015 challenge). Comparing with other state-of-the-art methods, our experiments have shown that the proposed architecture performed better with high capability to effectively identify unhealthy regions (*LTPR*=0.7840) while having overall a good overlap with the ground truth in terms of overall lesion volume (*DSC*=0.6980). This can be particularly important in clinical settings where detecting all potential lesions is prioritized over discarding easily identifiable false negatives.

Unlike previously proposed 3D-based CNN approach by Brosch et al. [9] which used a single short-cut connection between the deepest and the shallowest layers, our proposed architecture includes multiple short-cut connections between several layers of the network combining multi-level features from different stages of the network. In our opinion, the obtained results suggest that the combination of multi-level features during the upsampling procedure helps network to exploit more contextual information of the shape of the lesions. This could explain why the segmentation performance of our proposed network (*DSC*=0.6980) improved compared with the method proposed by Brosch et al. [9] (*DSC*=0.6844).

To avoid overfitting problem, unlike Brosch et al. [9] that proposed a two steps approach for training based on restricted Boltzman machines [56], we used a CNN pre-trained on ImageNet for contracting path of our model. Moreover, they used four modalities in their approach. However, the our approach is based on three modalities. The results in Table 3.1 shows that the proposed model outperforms their model with respect to all available measures.

The proposed method also has some limitations. Our network cannot use 4-dimensional (4D) modalities such as functional MRI or diffusion MRI. Moreover, the maximum number of MRI modalities that can be used in our architecture is three. This results from the fact that we used pre-trained ResNet as the encoder part in our network, which can only handle an input with three channels. Therefore in the case of more modalities, we have to train a model from scratch. Another limitation is that CNN based approaches in MS segmentation highly depend on the training which is costly to acquire due to the time consuming manual segmentation by experts it requires.

Chapter 4

Multi-branch CNN for MS Lesion Segmentation

4.1 Introduction

In the previous chapter, we showed that a pre-trained CNN for classification on natural images can generalize well for other tasks like MS lesion segmentation. It was observed that the proposed model can produce better segmentation performances compared with other CNN based methods, thanks to shortcuts at different resolutions. Moreover, it was shown that the proposed results are very close to the expert level segmentation performances.

As mentioned before, using single pre-trained ResNet as an encoder has some limitations such as it's input with three channels which resulted, in our case, the use of maximum three MRI modalities and in the case of more modalities available, they would be omitted to choose three amongst all. Modalities have different information regarding MS lesions. For instance, sagittal T1w MRI depicts multiple hypointense lesions in the corpus callosum which is characteristic of multiple sclerosis, coronal FLAIR MRI in a patient with MS demonstrates periventricular high signal intensity lesions, which exhibit a typical distribution for multiple sclerosis, Axial T2w MRI in a patient with MS demonstrates numerous white matter plaques in a callosal and pericallosal white matter distribution, etc. Combining all the available modalities with the mentioned different information into a single input to the model can not be the optimal solution for extracting the accurate location of the lesion. Moreover, it is vague that which modality caries informative and effective knowledge regarding the MS lesion segmentation when using deep CNN models.

Unlike our previous single-branch model, in this work, we propose a novel deep architecture consisting of a multi-branch 2D convolutional encoder-decoder network to address the above-mentioned problems. We designed an end-to-end encoder-decoder network including a multi-branch downsampling path as the encoder, a multi-scale feature fusion and a multi-scale upsampling block as the decoder. In the encoder, each branch is assigned to a specific MRI modality to take advantage of each modality individually. During the decoding stage of the network, different scales of the encoded attributes related to each modality, from the coarsest to the finest, including the middle-level attributes, were combined and upconvolved gradually to get fine details (more contextual information) of the lesion shape. Moreover, we evaluate different versions of the proposed model to find the most performant combination of MRI modalities for MS lesion segmentation.

4.2 Method

Following the idea in the previous chapter, we propose a model based on pre-trained 2D CNN. Similar to the approach described in section 3.2, ResNet50 [39] was used and converted to FCN [63]. However, to exploit the MRI multi-modality analysis, we built a pipeline of parallel ResNets without weights sharing. A multi-modal feature fusion block (MMFF) and a multi-scale feature upsampling block (MSFU) were proposed to combine and upsample the features from different modalities and different resolutions, respectively. We also concentrated on whole-brain slice-based segmentation and used three different(orthogonal) planes for each 3D modality as an input to the network. Moreover, we study a multi-plane reconstruction block, which defines and shows the suitable combination of the 2D binary slices of the network output to match the original 3D data.

In the following sub-sections, we first describe how the input features were generated by decomposing 3D data into 2D images. Then, we describe the proposed network architecture in detail and the training procedure. Finally, we introduce the multi-plane reconstruction block, which defines how we combined the 2D binary slices of the network output to match the original 3D data.

4.2.1 Input Features Preparation

We followed the same approach described in section 3.2.1 which for each MRI volume (and each modality), three different plane orientations (axial, coronal and sagittal) were considered (centering the brain by zero-padding each slice) to generate 2D slices along x, y, and z axes as input to the network. The aforementioned procedure was applied to all three modalities (FLAIR, T1w, and T2w). However, unlike our previous approach, we kept separating the extracted slices from each MRI modality (without stacking them together).

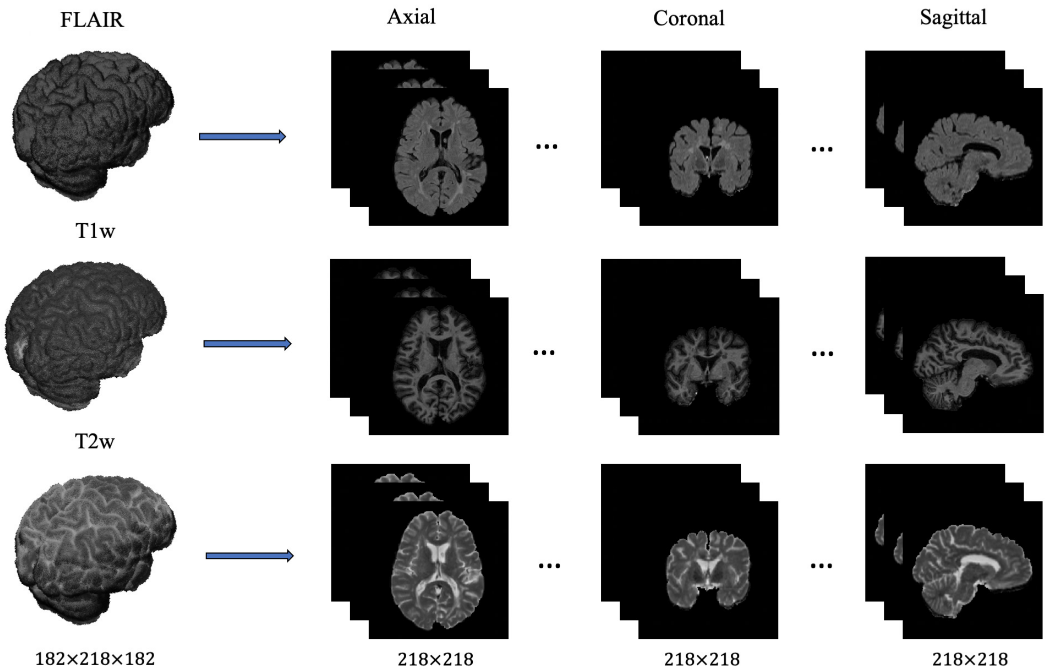


Figure 4.1: Input features preparation. For each subject, three MRI modalities (FLAIR, T1w, and T2w) were considered. 2D slices related to the orthogonal views of the brain (axial, coronal and sagittal planes) were extracted from each modality. Since the size of extracted slices was different with respect to the plane orientations, all slices were zero-padded while centering the brain so to obtain all slices with the same size, no matter their orientation.

Figure 4.1 illustrates the described procedure using FLAIR, T1w, and T2w modalities.

4.2.2 Network Architecture Details

The proposed model essentially integrates multiple ResNets with other blocks to handle multi-modality and multi-resolution approaches, respectively. As can be seen in Figure 4.2, the proposed network includes three main parts: downsampling networks, multi-modal feature fusion using MMFF blocks, and multi-scale upsampling using MSFU blocks.

In the downsampling stage, multiple parallel ResNets (without weights sharing) are used for extracting multi-resolution features, with each ResNet associated to one specific modality (in our experiments, we used FLAIR, T1w, and T2w). As mentioned in section 3.2.2, ResNet can be organized into five blocks according to the resolution of the generated feature maps. Thanks to this organization, we can take advantage of the multi-resolution. Features with the same resolution from different modalities are combined using MMFF blocks as illustrated in Figure 4.3. Each MMFF block includes 1×1 convolutions to reduce the number of feature maps (halving them), followed by 3×3 convolutions for adaptation. A simple concatenation layer is then used to combine the features from different modalities.

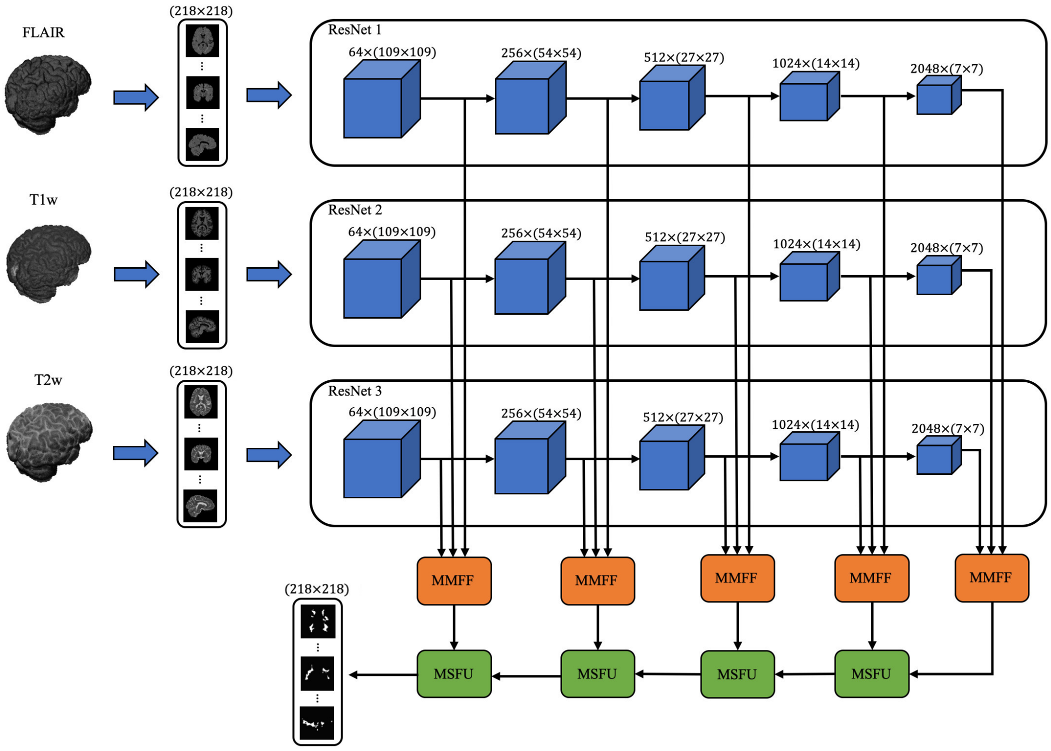


Figure 4.2: General overview of the proposed method. Input data is prepared as described in section 4.2.1, where volumes for each modality (FLAIR, T1w, and T2w) are described by slices (with the size of 218 × 218). Data is presented in input by slices, and the model generates the corresponding segmented slices. The downsampling part of the network (blue blocks) includes three parallel ResNets without weight sharing, each branch for one modality (we used three modalities: FLAIR, T1w, and T2w). Each ResNet can be considered composed by 5 blocks according to the resolution of the representations. For example, the first block denotes 64 representations with resolution 109 × 109. Then, MMFF blocks are used to fuse the representations with the same resolution from different modalities. Finally, the output of MMFF blocks is presented as input to MSFU blocks, which are responsible for upsampling the low-resolution representations and for combining them with high-resolution representations.

In the upsampling stage, MSFU blocks fuse the multi-resolution representations and gradually upsize them back to the original resolution of the input image. Figure 4.3 illustrates the proposed MSFU block consisting of a 1×1 convolution layer to reduce the number of feature maps (halving them) and an upconvolution layer with 2×2 kernel size and a stride of 2, transforming low-resolution feature maps to higher resolution maps. Then, a concatenation layer is used to combine the two sets of feature maps, followed by a 1×1 convolution layer to reduce the number of feature maps (halving them) and a 3×3 convolution layer for adaptation.

After the last MSFU block, a soft-max layer of size 2 is used to generate the output probability maps of the lesions. In our experiments the probabilistic maps were thresholded at 0.5 to generate binary classification for each pixel (lesion vs. non-lesion). It is important to mention that in all proposed blocks before each convolution and upconvolution layer, we use a batch normalization layer [43] followed by a rectifier linear unit activation function [71]. Size and number of feature maps in the input and output of all convolution layers are kept the same.

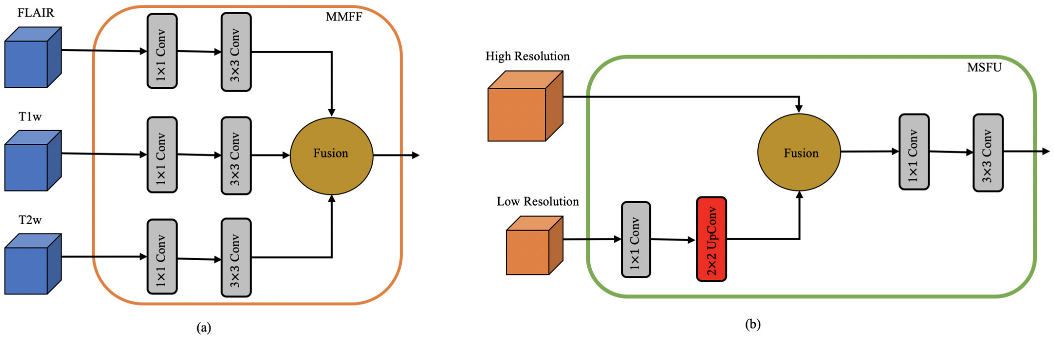


Figure 4.3: Building blocks of the proposed network. a) MMFF block is used to combine representations from different modalities (FLAIR, T1w, and T2w) at the same resolution. b) MSFU block is used to upsample low-resolution features and combine them with higher-resolution features.

4.2.3 Implementation Details

The proposed model was implemented in Python language [1] using Keras[2] [18] with Tensorflow[3] [1] backend. All experiments were done on a Nvidia GTX Titan X GPU. Our multi-branch slice-based network was trained end-to-end. In order to train the proposed CNN, we created a training set including the 2D slices from all three orthogonal views of the brain, as described in Section 4.2.1. Then, to limit extremely unbalanced data and omit uninformative samples, a training subset was determined by selecting only slices containing at least one pixel labeled as lesion (the number of slices ranged approximately from 150 to 300 per subject).

To optimize the network weights and early stopping criterion, the created training set was divided into training, and validation subsets, depending on the experiments as described

[1] https://www.python.org

[2] https://keras.io

[3] https://www.tensorflow.org

in the following section. In all experiments, the split was performed on the subject base, to simulate a real clinical condition and all the hyperparameters were selected through grid search. We trained our network using the Adam optimizer [51] with an initial learning rate of 0.0001 multiplied by 0.95 every 400 steps. The size of mini-batches was fixed at 15 and each mini-batch included random slices from different orthogonal views. The maximum number of training epochs was fixed to 1000 for all experiments, well beyond the average converging rate. The best model was then selected according to the validation set. Experimentally, we found that the best performance on validation set was systematically reached before 1000 training epochs.

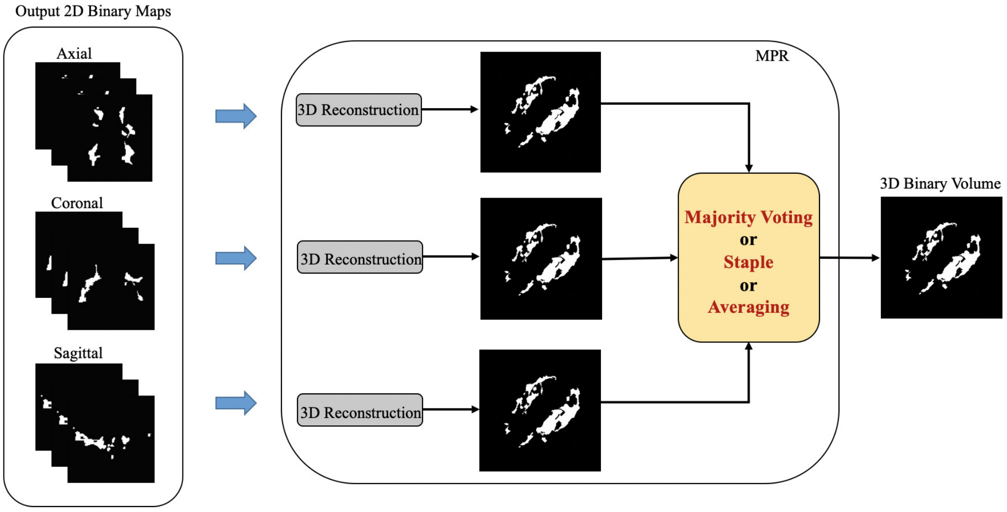


Figure 4.4: The MPR block produces a 3D volumetric binary map by combining the 2D output binary maps of the network. First, the output 2D binary maps associated to each plane orientation (axial, coronal, and sagittal) are concatenated to create three 3D binary maps. Then, a majority vote for each voxel is applied to obtain a single lesion segmentation volume.

Regarding the network initialization, in the downsampling branches, we used ResNet50 pre-trained on ImageNet and all other blocks (MMFFs and MSFUs) were randomly initialized from a Gaussian distribution with zero mean and standard deviation equal to $\sqrt{2/(a+b)}$ where a and b are respectively the number of input and output units in the weight tensor. It is worth noticing that we did not use parameter sharing in parallel ResNets. The soft Dice Loss function (DL) was used to train the proposed network:

$$DL = 1 - \frac{2\sum_i^N g_i p_i}{\sum_i^N {g_i}^2 + \sum_i^N {p_i}^2} \tag{4.1}$$

where $p_i \in [0, 1]$ is the predicted value of the soft-max layer and g_i is the ground truth binary value for each pixel i. We slightly modified the original soft dice loss [65] by replacing (-Dice) with (1-Dice) for visualization purposes. Indeed, the new equation returns positive values in the range $[0, 1]$. This change does not impact the optimization.

4.2.4 3D binary image reconstruction

Output binary slices of the network are concatenated to form a 3D volume matching the original data. To reconstruct the 3D image from the output binary 2D slices, we proposed a multi-planes reconstruction (MPR) block. Feeding each 2D slice to the network, we get as output the associated 2D binary lesion classification map. Since each original modality is duplicated three times in the input, once for each slice orientation (coronal, axial, sagittal), concatenating the binary lesion maps belonging to the same orientation results in three 3D lesion classification maps. To obtain a single lesion segmentation volume, these three lesion maps are combined via majority voting for each voxel (the most frequent classification is selected) as illustrated in Figure 4.4. To justify the choice of majority voting instead of other label fusion methods, we tested alternative well known label fusion methods (refer to the section 4.3.2).

4.3 Experiments

In order to evaluate the performance of the proposed method for MS lesion segmentation, two different datasets were used: the publicly available ISBI 2015 Longitudinal MS Lesion Segmentation Challenge dataset [11] (refer to section 1.3.1 for more details), and an in-house dataset NRU (refer to section 1.3.2).

4.3.1 Experiments on the ISBI dataset

From the ISBI dataset, we selected the preprocessed version of the images available online at the challenge website. All images were already skull-stripped using Brain Extraction Tool (BET) [91], rigidly registered to the $1mm^3$ MNI-ICBM152 template [74] using FMRIB's Linear Image Registration tool (FLIRT) [46, 47] and N3 intensity normalized [90].

To evaluate the performance of the proposed method on the ISBI dataset, two different experiments were performed according to the availability of the ground truth.

Since the ground truth was available only for the training set, in the first experiment we ignored the official ISBI test set for which no ground truth was provided. We only considered data with available ground truth (training set with 5 subjects) as mentioned in [3, 9]. To obtain a fair result, we tested our approach with a nested leave-one-subject-out cross-validation (3 subjects for training, 1 subject for validation and 1 subject for testing - refer to Table 4.1 for more details). To evaluate the stability of the model, this experiment was performed evaluating separately our method on the two sets of labels provided by the two raters.

Table 4.1: This table shows the implementation of first experiment in Section 4.3.1. In this experiment, we evaluated our model using the ISBI dataset with available ground truth (training set with 5 subjects only). We implemented a nested leave-one-subject-out cross-validation (3 subjects for training, 1 subject for validation, and 1 subject for testing). The numbers indicate the subject identifier.

Training	Validation	Testing
1,2,3	4	5
1,2,4	3	5
1,3,4	2	5
2,3,4	1	5
1,2,3	5	4
1,2,5	3	4
1,3,5	2	4
2,3,5	1	4
1,2,4	5	3
1,2,5	4	3
1,4,5	2	3
2,4,5	1	3
1,3,4	5	2
1,3,5	4	2
1,4,5	3	2
3,4,5	1	2
2,3,4	5	1
2,3,5	4	1
2,4,5	3	1
3,4,5	2	1

In the second experiment, the performance of the proposed method was evaluated on the official ISBI test set (with 14 subjects), for which the ground truth was not available, using the challenge web service. We trained our model doing a leave-one-subject-out cross-validation on the whole training set with 5 subjects (4 subjects for training and 1 subject for validation - refer to Table 4.2 for more details). We executed the ensemble of 5 trained models on the official ISBI test set and the final prediction was generated with a majority voting over the ensemble. The 3D output binary lesion maps were then submitted to the challenge website[4] for evaluation.

[4] http://iacl.ece.jhu.edu/index.php/MSChallenge

Table 4.2: This table shows the implementation of the second experiment in Section 4.3.1. In this experiment, our model was evaluated using official ISBI test set including 14 subjects without publicly available ground truth. We trained our model doing a leave-one-subject-out cross-validation on whole training set (4 subject for training, 1 subject for validation, and 14 subject for testing). The numbers indicate the subject identifier.

Training	Validation	Testing
1,2,3,4	5	ISBI test set
1,2,3,5	4	ISBI test set
1,2,4,5	3	ISBI test set
1,3,4,5	2	ISBI test set
2,3,4,5	1	ISBI test set

4.3.2 Experiment on the NRU dataset

In the NRU dataset, all sagittal acquisitions were reoriented in the axial plane and the exceeding portion of the neck was removed. T1w and T2w sequences were realigned to the FLAIR MRI using FLIRT and brain tissues were separated from non-brain tissues using BET [91] on FLAIR volumes. The resulting brain mask was then used on both registered T1w and T2w images to extract brain tissues. Finally, all images were rigidly registered to a $1mm^3$ MNI-ICBM152 template using FLIRT [46, 47] to obtain volumes of size $182 \times 218 \times 182$ and then N3 intensity normalized [90].

To test the robustness of the proposed model, we performed two experiments using the NRU dataset including 37 subjects. In the first experiment, we implemented a nested 4-fold cross-validation over the whole dataset (21 subjects for training, 7 subjects for validation and 9 subjects for testing - refer to Table 4.3 for more details). Since for each test fold we had an ensemble of four nested trained models, the prediction on each test fold was obtained as a majority vote of the corresponding ensemble.

To aggregate the outputs of the ensembles, beyond majority voting, we tested alternative well-known label fusion methods. Specifically, we repeated the aforementioned experiment on the NRU dataset substituting the majority vote framework with averaging and STAPLE (Simultaneous Truth and Performance Level) [105] methods, used to aggregate both the output volumes of the three plane orientations and the output volumes of the different models during cross-validation.

For comparison, we tested three different publicly available MS lesion segmentation software: OASIS (Automated Statistic Inference for Segmentation) [96], TOADS (Topology reserving Anatomy Driven Segmentation) [86], and LST (Lesion Segmentation Toolbox)[83]. OASIS generates the segmentation exploiting information from FLAIR, T1w, and T2w modalities, and it only requires a single thresholding parameter, which was optimized to ob-

Table 4.3: This table gives detailed information regarding the training procedure for the first experiment in Section 4.3.2. In this experiment, we implemented a nested 4-fold cross-validation over the whole NRU dataset including 37 subjects. [A-B @ C-D] denotes subjects A to B and C to D.

Training	Validation	Testing
[17-37]	[10-16]	[1-9]
[10-16 @ 24-37]	[17-23]	[1-9]
[10-23 @ 31-37]	[24-30]	[1-9]
[10-30 @ 31-37]	[31-37]	[1-9]
[8-9 @ 19-37]	[1-7]	[10-18]
[1-7 @ 24-37]	[8-9 @ 19-23]	[10-18]
[1-9 @ 19-23 @ 31-37]	[24-30]	[10-18]
[1-9 @ 19-30]	[31-37]	[10-18]
[8-18 @ 28-37]	[1-7]	[19-27]
[1-7 @ 15-18 @ 27-37]	[8-14]	[19-27]
[1-14 @ 31-37]	[15-18 @ 28-30]	[19-27]
[1-18 @ 28-30]	[31-37]	[19-27]
[8-37]	[1-7]	[28-37]
[1-7 @ 15-27]	[8-14]	[28-37]
[1-14 @ 22-27]	[15-21]	[28-37]
[1-21]	[22-27]	[28-37]

tain the best DSC. TOADS does not need parameter tuning and it only requires FLAIR and T1w modalities for segmentation. Similarly, LST works with FLAIR and T1w modalities only. However, it needs a single thresholding parameter that initializes the lesion segmentation. This parameter was optimized to get the best DSC in this experiment.

We also tested the standard 2D U-Net [80] that was developed for biomedical image segmentation, repeating the training protocol described in Table 4.3. Indeed, we used the same training set as described in Sections 4.2.1 and 4.2.2, with the difference that 2D slices from all modalities were aggregated in multiple channels. This network was trained using the Adam optimizer [51] with an initial learning rate of 0.0001 multiplied by 0.9 every 800 steps. For the sake of comparison, optimization was performed on the soft Dice Loss function (Equation 4.1) [65]. To get the 3D volume from output binary slices of the network, we used the proposed MPR block as described in Section 4.2.4.

Differences in performance metrics between our method and each of the 4 other methods were statistically evaluated with resampling. For a given method M and metric C, resampling was performed by randomly assigning for each subject the sign of the difference in C between method M and our method in 10 million samples. The test was two-sided and corrected for multiple comparisons with Holm's method (28 comparisons in total with 7 metrics assessed for the 4 methods to compare ours with). The alpha significance threshold level was set to 0.05.

While for the ISBI dataset, we evaluated our method on two separate sets of masks, one for each rater, in the NRU dataset, we considered the manual consensus segmentation as a more robust gold standard against which to validate the proposed method. Nevertheless, to evaluate the stability of the model trained with the gold standard labeling, we also tested it separately on the two sets of masks.

In the second experiment, to investigate the importance of each single modality in MS lesion segmentation, we evaluated our model with various combinations of modalities. This means that the model was adapted in the number of parallel branches in the downsampling network. In this experiment, we randomly split the corresponding dataset into fixed training (21 subjects), validation (7 subjects) and test (9 subjects) sets.

Single-branch (SB): In a single-branch version of the proposed model, we used a single ResNet as the downsampling part of the network. Attributes from different levels of the single-branch were supplied to the MMFF blocks. In this version of our model, each MMFF block had single input since there was only one downsampling branch. Therefore, MMFF blocks included a 1×1 convolution layer followed by a 3×3 convolution layer. We trained and tested the single-branch version of our proposed network with each modality separately and also with a combination of all modalities as a multi-channel input.

Multi-branch (MB): The multi-branch version of the proposed model used multiple parallel ResNets in the downsampling network without weights sharing. In this experiment, we used two-branch and three-branch versions, which were trained and tested using two modalities and three modalities, respectively. We trained and tested the mentioned models with all possible combination of modalities (two-branches:[FLAIR, T1w], [FLAIR, T2w], [T1w, T2w] and three-branches: [FLAIR,T1w, T2w]).

4.4 Results

4.4.1 ISBI dataset

In the first experiment, we evaluated our model using three measures: *DSC*, *LTPR*, and *LFPR* to make our results comparable to those obtained in [3, 9, 48, 64]. Table 4.4 summarizes the results of the first experiment when comparing our model with previously proposed methods. The table shows the mean *DSC*, *LTPR*, and *LFPR*. As can be seen in that table, our method outperformed other methods in terms of *DSC* and *LFPR*, while the highest *LTPR* was achieved by our proposed method in privious chapter. Figure 4.5 shows the segmentation outputs of the proposed method for subject 2 (with high lesion load) and subject 3 (with low

Table 4.4: Comparison of our method with other state-of-the-art methods in the first ISBI dataset experiment (in this experiment, only images with available ground truth were considered). GT1 and GT2 denote the corresponding model was trained using annotation provided by rater 1 and rater 2 as ground truth, respectively (the model was trained using GT1 and tested using both GT1 and GT2 and vice versa). Mean values of *DSC*, *LTPR*, and *LFPR* for different methods are shown. Values in bold and italic refer to the first-best and second-best values of the corresponding metrics, respectively.

Method	Rater 1			Rater 2		
	DSC	*LTPR*	*LFPR*	*DSC*	*LTPR*	*LFPR*
Rater 1	-	-	-	0.7320	0.6450	0.1740
Rater 2	0.7320	0.8260	0.3550	-	-	-
Jesson et al. [48]	*0.7040*	0.6111	*0.1355*	*0.6810*	0.5010	*0.1270*
Maier et al. [64] (GT1)	0.7000	0.5333	0.4888	0.6555	0.3777	0.4444
Maier et al. [64] (GT2)	0.7000	0.5555	0.4888	0.6555	0.3888	0.4333
Brosch et al. [9] (GT1)	0.6844	*0.7455*	0.5455	0.6444	*0.6333*	0.5288
Brosch et al. [9] (GT2)	0.6833	*0.7833*	0.6455	0.6588	*0.6933*	0.6199
Aslani et al. [3] (GT1)	0.6980	**0.7460**	0.4820	0.6510	**0.6410**	0.4506
Aslani et al. [3] (GT2)	0.6940	**0.7840**	0.4970	0.6640	**0.6950**	0.4420
Ours (GT1)	**0.7649**	0.6697	**0.1202**	**0.6989**	0.5356	**0.1227**
Ours (GT2)	**0.7646**	0.7002	**0.2022**	**0.7128**	0.5723	**0.1896**

lesion load) compared to both ground truth annotations (rater 1 and rater 2). Confirming the above-mentioned point, Figure 4.5 shows that the proposed method is robust to identify both small and big lesions having overall a good overlap with the ground truth (*DSC*>0.7).

In the second experiment, the official ISBI test set was used. Indeed, all 3D binary output masks computed on the test set were submitted to the ISBI website. Several measures were calculated online by the challenge website. Table 4.5 shows the results on all measures reported as a mean across raters. At the time of the submission, our method had an overall evaluation score of 92.12 on the official ISBI challenge web service[5], making it amongst the top-ranked methods with a published paper or a technical report.

4.4.2 NRU dataset

Table 4.6 reports the results of the first experiment on NRU dataset showing the mean values of *DSC*, *LFPR*, *LTPR*, *PPV*, *VD*, *SD* and *HD*. It summarizes how our method performed compared to others. As shown in the table, our method achieved the best results with respect to *DSC*, *PPV*, *LFPR*, *VD*, *SD* and *HD* measures while showing a good trade-off between *LTPR* and *LFPR*, comparable to the best results of the other methods.

Figure 4.8 shows boxplots of the *DSC*, *LFPR*, *LTPR*, *PPV*, *VD*, *SD* and *HD* evaluation

[5] http://iacl.ece.jhu.edu/index.php/MSChallenge

Table 4.5: Results related to the top-ranked methods (with published papers or technical reports) evaluated on the official ISBI test set and reported on the ISBI challenge website. *SC*, *DSC*, *PPV*, *LTPR*, *LFPR*, and *VD* are mean values across the raters. For detailed information about the metrics, refer to Section 1.4. Values in bold and italic refer to the metrics with the first-best and second-best performances, respectively.

Method	*SC*	*DSC*	*PPV*	*LTPR*	*LFPR*	*VD*
Hashemi et al. [36]	**92.48**	0.5841	**0.9207**	0.4135	**0.0866**	0.4972
Ours	*92.12*	0.6114	*0.8992*	0.4103	*0.1393*	0.4537
Andermatt et al. [2]	92.07	*0.6298*	0.8446	0.4870	0.2013	0.4045
Valverde et al. [102]	91.33	**0.6304**	0.7866	0.3669	0.1529	**0.3384**
Maier et al. [64]	90.28	0.6050	0.7746	0.3672	0.2657	0.3653
Birenbaum et al. [8]	90.07	0.6271	0.7889	**0.5678**	0.4975	*0.3522*
Aslani et al. [3]	89.85	0.4864	0.7402	0.3034	0.1708	0.4768
Deshpande et al. [24]	89.81	0.5960	0.7348	0.4083	0.3075	0.3762
Jain et al. [45]	88.74	0.5560	0.7300	0.3225	0.3742	0.3746
Sudre et al. [95]	87.38	0.5226	0.6690	*0.4941*	0.6776	0.3837
Tomas et al. [99]	87.01	0.4317	0.6973	0.2101	0.4115	0.5109
Ghafoorian et al. [30]	86.92	0.5009	0.5491	0.4288	0.5765	0.5707

Table 4.6: Results related to the first NRU dataset experiment. Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for different methods. Values in bold and italic indicate the first-best and second-best results.

Method	*DSC*	*PPV*	*LTPR*	*LFPR*	*VD*	*SD*	*HD*
TOADS [86]	0.5241	0.5965	**0.4608**	0.6277	*0.4659*	5.4392	13.60
LST [83]	0.3022	0.5193	0.1460	0.3844	0.6966	7.0919	14.35
OASIS [96]	0.4193	0.3483	0.3755	0.4143	2.0588	*3.5888*	18.33
U-NET [80]	*0.6316*	*0.7748*	0.3091	*0.2267*	*0.3486*	3.9373	*9.235*
Ours	**0.6655**	**0.8032**	*0.4465*	**0.0842**	**0.3372**	**2.5751**	**6.728**

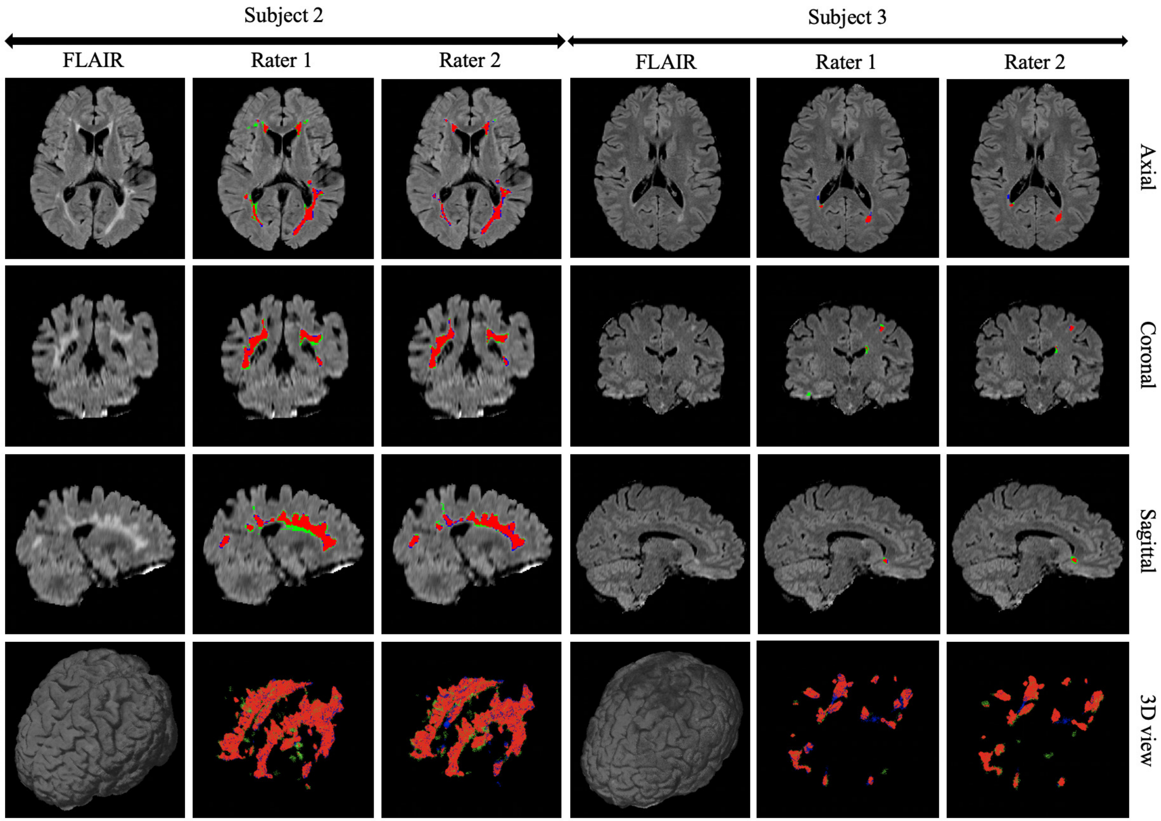


Figure 4.5: Segmentation results of the proposed method on two subjects of the ISBI dataset compared to ground truth annotations provided by rater 1 and rater 2. From left to right, the first three columns are related to subject 2 with high lesion load and reported DSC values of 0.8135 and 0.8555 for rater 1 and rater 2, respectively. Columns 4 to 6 are related to the subject 3 with low lesion load and reported *DSC* values of 0.7739 and 0.7644 for rater 1 and rater 2, respectively. On all images, true positives, false negatives, and false positives are colored in red, green and blue, respectively.

Table 4.7: The proposed model was tested with different combinations of the three modalities in the second NRU dataset experiment. SB and MB denote the single-branch and multi-branch versions of the proposed model, respectively. Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for different methods. Values in bold and italic indicate the first-best and second-best values.

Method	Set of Modalities	DSC	PPV	LTPR	LFPR	VD	SD	HD
SB	FLAIR	0.6531	0.5995	0.6037	0.2090	0.3034	1.892	9.815
	T1w	0.5143	0.5994	0.3769	0.2738	0.3077	4.956	*8.201*
	T2w	0.5672	0.5898	0.4204	0.2735	*0.1598*	4.733	9.389
	FLAIR, T1w, T2w	*0.6712*	0.6029	0.6095	0.2080	0.2944	*1.602*	9.989
MB	FLAIR, T1w	0.6624	*0.6109*	*0.6235*	0.2102	0.2740	1.727	9.526
	FLAIR, T2w	0.6630	0.6021	**0.6511**	*0.2073*	0.3093	1.705	9.622
	T1w, T2w	0.5929	0.6102	0.4623	0.2309	0.1960	4.408	9.004
	FLAIR, T1w, T2w	**0.7067**	**0.6844**	0.6136	**0.1284**	**0.1488**	**1.577**	**8.368**

metrics obtained from the different methods and summarized in Table 4.6. This Figure shows statistically significant differences between model performances for most metrics and methods when compared to ours, after multiple comparison correction with the conservative Holm's method. The output segmentation of all methods applied to a random subject (with medium lesion load) can be seen with different plane orientations in Figure 4.6.

Figure 4.7 depicts the relationship between the volumes of all ground truth lesions and the corresponding estimated size for each evaluated method (one datapoint per lesion). With a qualitative evaluation, it can be seen that TOADS and OASIS methods tend to overestimate lesion volumes as many lesions are above the dashed black line, i.e., many lesions are estimated larger than they really are. On the contrary, LST method tends to underestimate the lesion sizes. U-Net and our method, on the contrary, produced lesions with size more comparable to the ground truth. However, with a quantitative analysis, our model produced the slope closest to unity (0.9027) together with the highest Pearson correlation coefficient (0.75), meaning our model provided the stronger global agreement between estimated and ground truth lesion volumes. Note that a better agreement between lesion volumes does not mean the segmented and ground truth lesions better overlap – the amount of overlap was measured with the *DSC*.

Table 4.7 shows the performance of the proposed model with respect to different combinations of modalities in the second experiment. The SB version of the proposed model used with one modality had noticeably better performance in almost all measures when using FLAIR modality. However, all modalities carry relevant information as better performance in most metrics was obtained when using a combination of modalities. In MB versions of the model, all possible two-branch and three-branch versions were considered. As shown in Table 4.7, two-branch versions including FLAIR modality showed a general better performance than the single-branch version using single modality. This emphasizes the importance

Table 4.8: This table shows the results of the first experiment on the NRU dataset using our model as described in Section 4.3.2. We implemented the same experiment using different methods for fusing output volumes (when merging the outputs from each plane orientation, and also when merging the outputs of models from different cross-validation folds). Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for each method. Values in bold indicate the first-best results.

Method	*DSC*	*PPV*	*LTPR*	*LFPR*	*VD*	*SD*	*HD*
Majority Voting	**0.6655**	*0.8032*	**0.4465**	*0.0842*	**0.3372**	2.575	**6.728**
Averaging	0.5883	**0.8391**	0.3220	**0.0788**	0.4625	3.216	*8.503*
STAPLE [105]	*0.6632*	0.7184	*0.3989*	0.0802	*0.3883*	**2.330**	8.629

Table 4.9: This table indicates the performance of our trained model in first experiment of NRU dataset when using different ground truth masks as testing. Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for each method. Values in bold indicate the first-best results.

Method	*DSC*	*PPV*	*LTPR*	*LFPR*	*VD*	*SD*	*HD*
Rater1	**0.6827**	0.8010	**0.5039**	0.0977	0.3727	**2.085**	**6.704**
Rater2	0.6607	0.7784	0.4458	0.0860	0.3638	2.511	7.009
Gold Standard (Consensus Mask)	0.6655	**0.8032**	0.4465	**0.0842**	**0.3372**	2.575	6.728

of using FLAIR modality together with others (T1w and T2w). However, overall, a combination of all modalities in the three-branch version of the model showed the best general performance compared to the other versions of the network.

In order to aggregate the outputs of the ensembles, beyond majority voting, we tested alternative well known label fusion methods. Specifically, we repeated the first experiment on NRU dataset as described in Section 4.3.2, substituting the majority vote framework with averaging and STAPLE methods, used to aggregate both the output volumes of the three plane orientations and the output volumes of the different models during cross-validation. Table 4.8 indicates the performance of each method. Overall, majority voting had better performance than other methods. Therefore, we selected this method for all experiments.

In the first experiment on NRU dataset, beyond verifying the quality of the proposed model on the ground truth generated from the consensus of two experts, we also compared the performance with the ground truth from each individual experts. The rationale behind the experiment was to assess the consistency of the system across raters. Table 4.9 shows the corresponding results. As expected from the high consensus between the masks provided by the two raters (as mentioned in Section 1.3.2), our trained model using the gold standard mask (derived from the two raters' masks) showed comparable results when evaluated with either raters' masks or the consensus mask as ground truth.

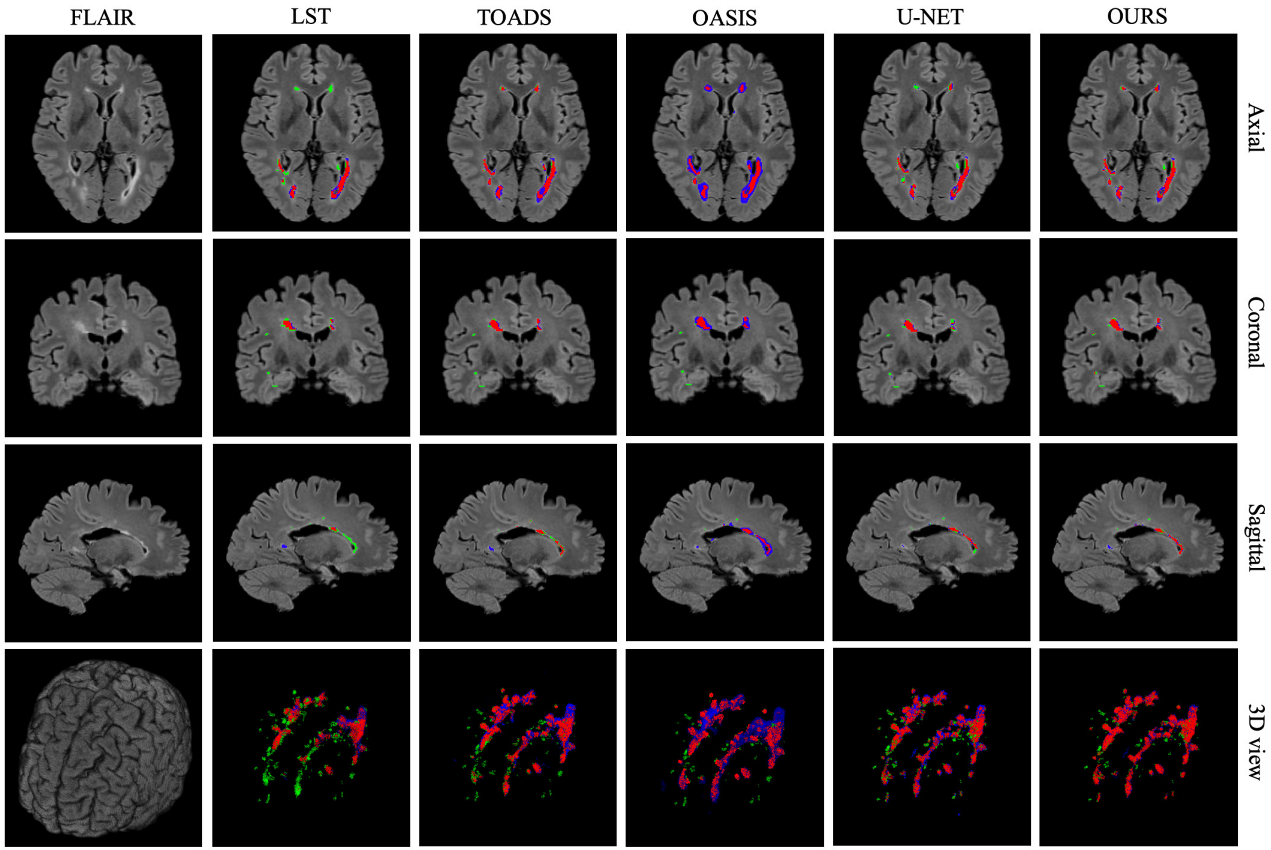


Figure 4.6: Output segmentation results of the different methods for one subject with medium lesion load from the NRU dataset compared with ground truth annotation. Reported *DSC* values for TOADS, OASIS, LST, U-Net and our proposed method for this subject are 0.7110, 0.4266, 0.6505, 0.7290 and 0.7759, respectively. On all images, true positives, false negatives, and false positives are colored in red, green and blue, respectively.

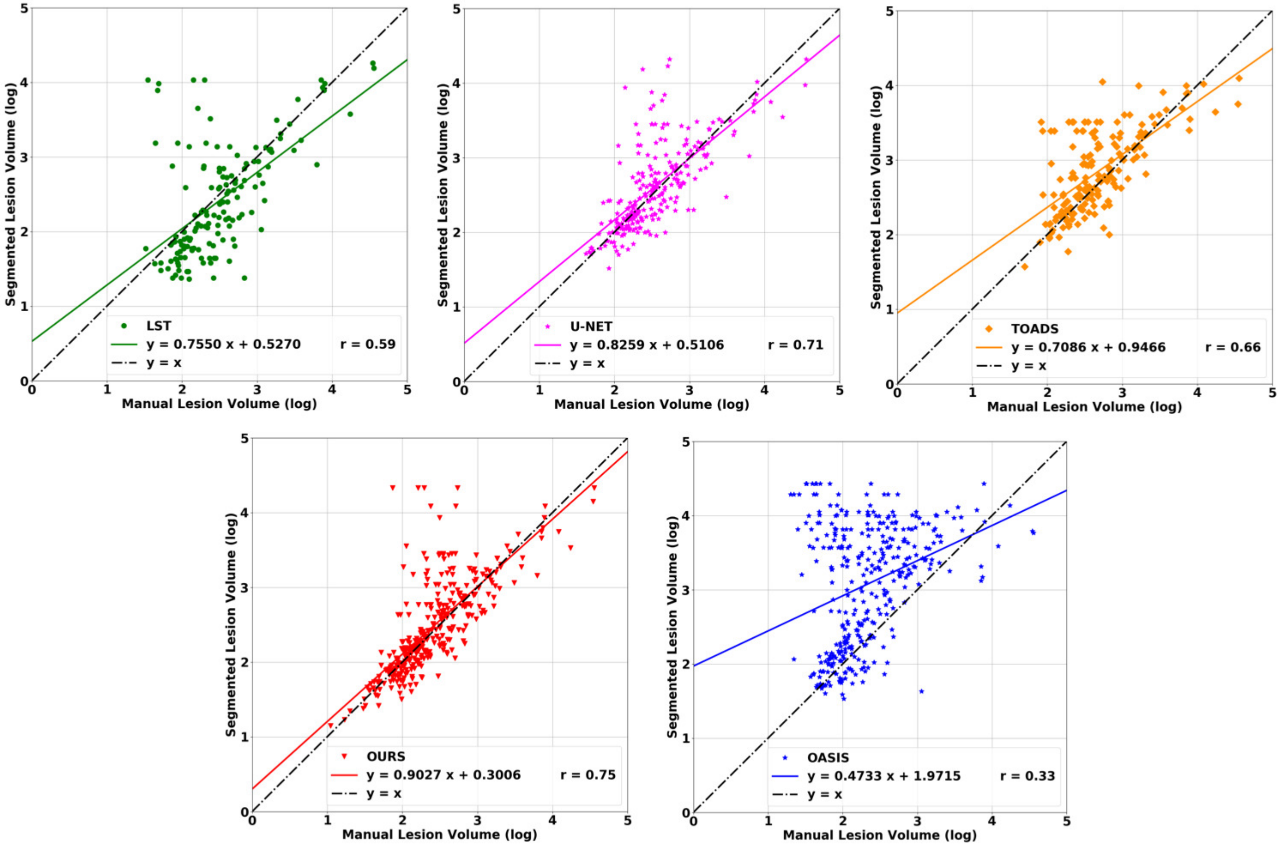


Figure 4.7: Comparison of the lesion volumes produced by manual and automatic segmentation on the NRU dataset with different methods. Each point is associated with a single lesion. Colored (solid) lines indicate the correlation between manual and segmented lesion volumes. Black (dotted) lines indicate the ideal regression line. Slope, intercept, and Pearson's linear correlation (all with $p << 0.001$) between manual and estimated masks can also be seen for different methods.

4.5 Discussion and Conclusions

In this chapter, we have designed an automated pipeline for MS lesion segmentation from multi-modal MRI data. The proposed model is a deep end-to-end 2D CNN consisting of a multi-branch downsampling network, MMFF blocks fusing the features from different modalities at different stages of the network, and MSFU blocks combining and upsampling multi- scale features.

When having insufficient training data in deep learning based approaches, which is very common in the medical domain, transfer learning has demonstrated to be an adequate solution [13, 14, 41]. As we showed in first chapter, not only it helps boosting the performance of the network but also it significantly reduces overfitting. Therefore, in this chapter, we also used the parallel ResNet50s pre-trained on ImageNet as a multi-branch downsampling network while the other layers in MMFF and MSFU blocks were randomly initialized from a Gaussian distribution. We then fine-tuned the whole network on the given MS lesion segmentation task.

In brain image segmentation, a combination of MRI modalities overcomes the limitations of single modality approaches, allowing the models to provide more accurate segmentations [3, 52, 66]. Unlike previously proposed deep networks [3, 9], which stacked all modalities together as a single input, we designed a network with several downsampling branches, one branch for each individual modality. We believe that stacking all modalities together as a single input to a network is not an optimal solution since during the downsampling procedure, the details specific to the the most informative modalities can vanish when mixed with less informative modalities. On the contrary, the multi-branch approach allows the network to abstract higher-level features at different granularities specific to each modality. Independently of the ground truth used for training and testing the model, results in Table 4.4 confirm our claim showing that a network with separate branches generated more accurate segmentations (e.g., *DSC*=0.7649) than single-branch networks with all modalities stacked, as proposed by Brosch et al. [9] (e.g., *DSC*=0.6844) and our model in first chapter [3] (e.g., *DSC*=0.6980). Indeed, the mentioned methods (single-branch) generally obtained higher *LTPR* values (e.g., 0.7455 and 0.7460) than multi-branch (e.g., 0.6697). However, they also obtained very high *LFPR* values showing a significant overestimation of lesion volumes. The proposed method, instead, showed the best trade-off between *LTPR* and *LFPR*.

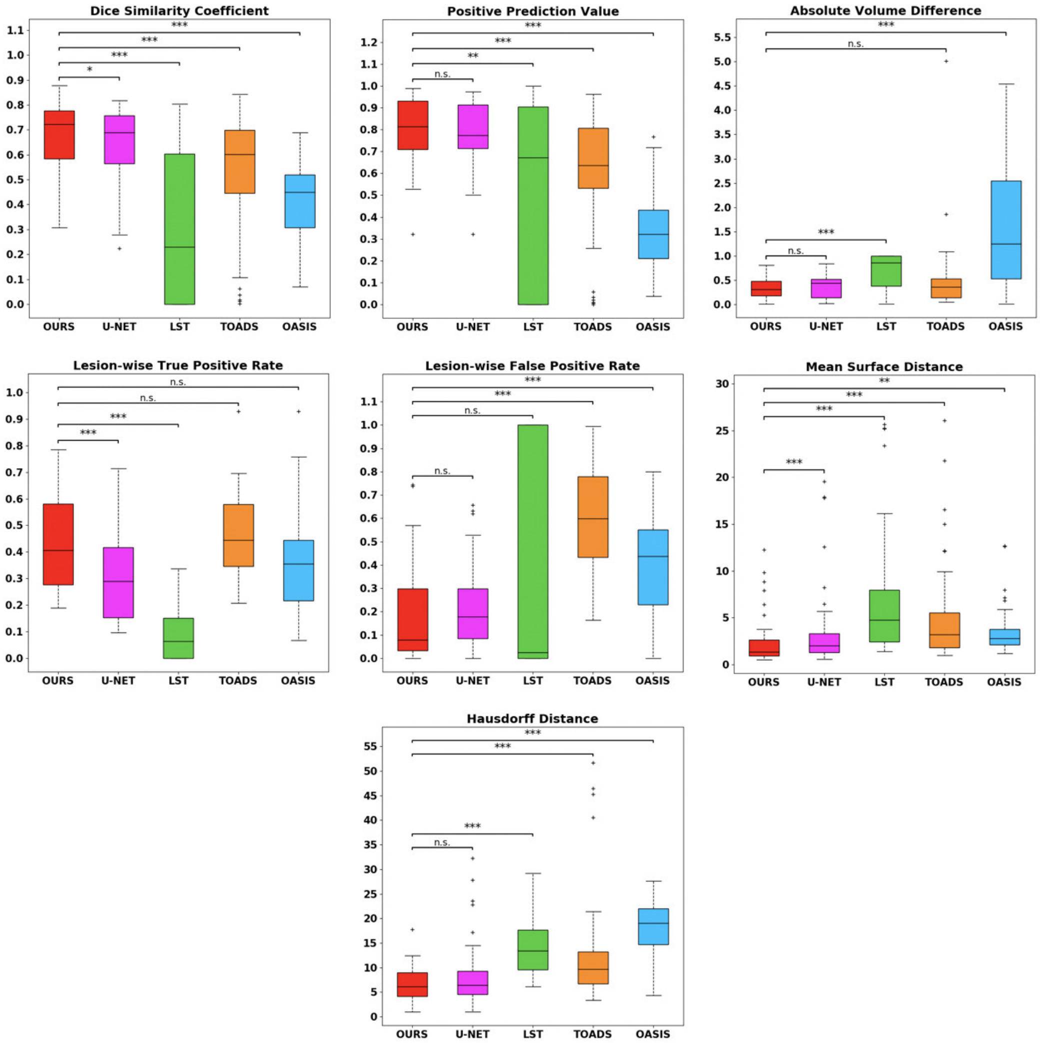


Figure 4.8: Boxplots showing the performance of tested models with all measures on NRU dataset. Among all methods, the proposed one had the best trade-off between the lesion-wise true positive rate and lesion-wise false positive rate, while having the best mean value for dice similarity coefficient, positive prediction value, absolute volume differences, mean surface distance and hausdorff distance. Statistically significant differences between our method and the others were assessed using resampling statistics with multiple comparison correction. The significance threshold was set as $\alpha = 0.05$. p-values were annotated as follows: '*' for $p < 0.05$, '**' for $p < 0.005$, '***' for $p < 0.0005$, and 'n.s.' for non-significant values.

When examining the influence of different modalities, results in Table 4.7 demonstrated that the most important modality for MS lesion segmentation was FLAIR (*DSC*>0.65). This is likely due to the fact that FLAIR sequences benefit from CSF signal suppression and hence provide a higher image contrast between MS lesions and the surrounding normal appearing

WM. Using all modalities together in a SB network (by concatenating them as single multi-channel input) and in a MB network (each modality as single input to each branch) showed good segmentation performance. This could be due to the combination of modalities helping the algorithm identifying additional information regarding the location of lesions. However, supporting our claim that stacking all modalities together as a single input to the network is not an optimal solution, top performance, indeed, was obtained in most measures with the MB network when using all available modalities, as can be seen in Table 4.7.

In deep CNNs, attributes from different layers include different information. Coarse layers are related to high-level semantic information (category specific), and shallow layers are related to low-level spatial information (appearance specific) [63], while middle layer attributes have shown a significant impact on segmentation performance [80]. Combining these multi-level attributes from the different stages of the network makes the representation richer than using single-level attributes, like in the CNN based method proposed in [9], where a single shortcut connection between the deepest and the shallowest layers was used. Following the same idea in the first chapter [3], our model, instead, includes several shortcut connections between all layers of the network, in order to combine multi-scale features from different stages of the network as inspired by U-Net architecture [80]. The results shown in Table 4.4 suggest that the combination of multi-level features during the upsampling procedure helps the network exploiting more contextual information associated to the lesions. This could explain why the performance of our proposed model (*DSC*=0.7649) is higher than the method proposed in [9] (*DSC*=0.6844).

Patch-based CNNs suffer from lacking globalspatial information about the lesions because of the patch size limitation. To deal with this problem, we proposed a whole-brain slice-based approach. Compared with patch-based methods [30, 102], we have shown that our model has better performance for most measures, as seen in Table 4.5. Although the CNN proposed in [102] had the highest *DSC* value among all, our method showed better performance regarding the *LTPR* and *LFPR*, which indicates that our model is robust in identifying the correct location of lesions. The method proposed in [8] has been optimized to have the highest *LTPR*. However, their method showed significantly lower performance in *LFPR*. Compared with this method, our method has better trade-off between *LTPR* and *LFPR*.

As mentioned in [11], manual delineation of MS lesions from MRI modalities is prone to intra- and inter-observer variability, which explains the relatively low *DSC* between two experts delineating the same lesions (∼0.73 for ISBI data as shown in Table 4.4). Automated methods are therefore expected to have a maximum performance in the same order of magnitude when comparing their generated segmentation with the rater's one. Accordingly, it is important to notice that, our model obtained a performance (*DSC*) close to the experts agreement, as can be seen in Table 4.4.

The proposed method also has some limitations. We observed that the proposed pipeline is slightly slow in segmenting a 3D image since segmenting whole-brain slices takes a longer time compared to other CNN-based approaches [81]. The time required to segment a 3D image is proportional to the size of the image and is based on the computational cost of three sequential steps: input features preparation 4.2.1, slice-level segmentation 4.2.2, and 3D image reconstruction 4.2.4. In both the ISBI and NRU datasets, the average time for segmenting an input image with our model, including all 3 steps, was approximately 90 seconds.

Chapter 5

Scanner Invariant MS Lesion Segmentation

5.1 Introduction

Deep learning models, in particular CNNs [55] have shown excellent performance in a large variety of computer vision tasks, including image classification [53], object detection [33], semantic segmentation [63], etc. It is, therefore, common to expect that successful deep models can obtain good performances. However, it has been shown that, in practice, these approaches easily fail to generalize well [12, 108].

According to the literature, the most important reasons for this failure are (i) the small size of the training data which causes overfitting and (ii) the large difference between training and test data which is typically addressed as domain shift. Therefore, one of the most important problems that arises is how to improve the quality of models so that they generalize well to unseen data from a different domain. During the last years, several algorithms have been proposed to tackle the mentioned problem, improving the models generalization through heuristic techniques such as dropout [92], early stopping [67], weight decay [54], data augmentation [44], randomization methods [108], and other theoretical generalization methods [49, 7].

Thanks to these regularization methods, deep models are reaching expert-level accuracy in medical image segmentation. However, they still have a limited clinical application due to the aforementioned challenge (i), which is also considered as one of the most relevant and common problem in medical image analysis tasks [11, 19, 68]. To tackle this challenge, several strategies have been proposed, such as increasing the dataset size using 2.5D representations (slices) rather than full-size 3D images, initializing parameters of the proposed model with pre-trained weights on natural images, and adopting special data augmentation techniques [109]. An effective solution, however, would be to merge datasets collected from different centers. This, however, introduces another important challenge. The medical data acquisition can vary significantly between different centers. For instance, in magnetic resonance imaging (MRI), this procedure is often subject to the variation of several specific properties such as scanner, magnet strength, and acquisition protocol. This causes high domain variability between datasets which eventually can result in poor generalization. In order to tackle this problem, several methods have been proposed such as scanner invariant representations for medical image harmonization [70], one-shot domain adaptation [103], and unsupervised methods [75] for medical image segmentation.

In this chapter, we propose a novel simple domain generalization method to enhance baseline models and diminish the effect of domain differences in the data. To this end, a regularization network, equipped with an auxiliary loss function, is proposed to incorporate regularization into a standard encoder-decoder segmentation network. We tested the model with a standard cross-validation procedure using an in-house dataset on MS lesion segmentation. Results show that the proposed regularization network has a significant impact on the generalization of the standard segmentation network when data from multiple centers are used.

5.2 Method

Generally, the performance of models suffers when they are applied to domains other than the ones they were trained upon. In this work, the goal is to improve the generalizability of a backbone segmentation model by training it on a collection of datasets presenting high domain variability. The selected backbone encoder-decoder model has a traditional loss function used for image segmentation. However, to handle the domain shift problem, we propose a regularization network including an auxiliary loss function that is designed to encourage the model to ignore domain-specific information. This property emerges from optimizing cross entropy or correlation coefficient as detailed below. Training the backbone segmentation network incorporated with regularization network reduces the domain differences problem across the datasets.

5.2.1 Network Architecture Details

The overall architecture of the proposed model includes three main components (Figure 5.1). The first component is a feature extractor network consisting of an encoder ϕ_E which is fed by an input image $x_i \in \mathcal{X}$. The output of the encoder ϕ_E is a p-dimensional vector $r_i = \{r_{ij} \in \mathbb{R}\}_{j=1}^{p} \in \mathcal{R}$ representing the latent features.

The second component is a segmentation network consisting of a decoder ϕ_D which reconstruct from the latent features $r_i \in \mathcal{R}$ a feature representation with the resolution of the input image $x_i \in \mathcal{X}$. The output layer then produces a dense pixel-wise prediction output $s_i \in \mathcal{S}$ using a softmax activation. This network includes a traditional loss term used to update the weights to improve the segmentation performance (see section 5.2.2). Note that the described architecture composed of the mentioned encoder ϕ_E and decoder ϕ_D is very similar to the model presented in [80], using 3D operations rather than 2D ones and removing the regularization layers (dropout). For simplicity, we removed skip-connections in Figure 5.1.

The third component of the presented model is a regularization network ϕ_R including three perceptron layers and a softmax layer. The network receives the latent features $r_i \in \mathcal{R}$ to produce category-wise prediction $c_i \in \mathcal{C}$, which in our case corresponds to the prediction of the input's domain.

Our observation is that during training, the model without ϕ_R learns how to segment the input images also encoding their source domain. This results in overfitting of the model with a domain bias. As a result, the segmentation performance of the network on data coming from unseen source domains is very poor. The goal of the regularization network is therefore to steer the whole model to reduce the domain bias, to obtain a better generalization and, hence, a fairer segmentation performance on seen and unseen domains. To this aim, we introduce an auxiliary loss term whose aim is to confuse the model about the dataset domains, thus forcing the model to learn how to segment the image while maximally reducing the domain bias.

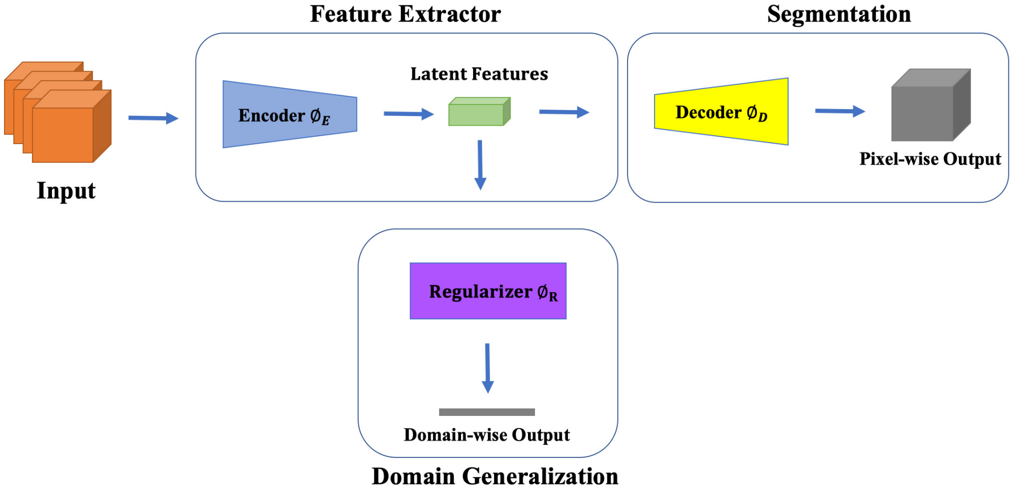


Figure 5.1: Overall architecture of the proposed method including three main components: An encoder network ϕ_E for extracting latent features, a decoder network ϕ_D for segmentation and a regularization network ϕ_R for domain generalization.

5.2.2 The Loss Functions

Our method has been tested using multiple loss terms to enable the network to precisely segment the input image while generalizing over the domains. Specifically, the proposed model was optimized according to the loss function formulated as:

$$\mathcal{L}(\mathcal{X}, \mathcal{S}, \mathcal{C}, \mathcal{H}, \mathcal{G}) = \mathcal{L}_{\text{seg}}(\mathcal{X}, \mathcal{S}, \mathcal{G}) + \lambda \mathcal{L}_{\text{reg}}(\mathcal{X}, \mathcal{C}, \mathcal{H}) \tag{5.1}$$

where $\lambda \in [0, 1]$ is a hyperparameter controlling the trade-off between segmentation and regularization losses. $\mathcal{H}$ and $\mathcal{G}$ are the domain-wise and pixel-wise ground truth, respectively. In this work, we propose three different regularization approaches using loss functions $\mathcal{L}_{\text{reg}}$ based on two well-established measures, namely cross-entropy and the Pearson correlation coefficients. Hence, $\mathcal{L}_{\text{reg}}$ can take on either of these three options:

$$\mathcal{L}_{\text{reg}}(\mathcal{X}, \mathcal{C}, \mathcal{H}) = \{\mathcal{L}_{\text{pc}}(\mathcal{X}, \mathcal{C}, \mathcal{H}), \mathcal{L}_{\text{rand}}(\mathcal{X}, \mathcal{C}, \mathcal{H}), \mathcal{L}_{\text{du}}(\mathcal{X}, \mathcal{C}, \mathcal{H})\} \tag{5.2}$$

Pearson Correlation Loss $\mathcal{L}_{\text{pc}}$: Given an input image $x_i \in \mathcal{X}$, the regularization network ϕ_R generates the corresponding output vector $c_i = \{c_{ij} \in [0, 1]\}_{j=1}^{n} \in \mathcal{C}$ which shows the probabilities for x_i to be in one of the n domains. For each input image x_i the corresponding one-hot encoded vector as ground truth domain labeling is also given by $h_i = (0, 0, ..., 1, 0, ..., 0) \in \mathcal{H}$.

The Pearson correlation coefficient measures the strength of linear correlation or similarity between two variables, where higher values correspond to higher similarity. Hence, to remove the domain bias, the model can be trained to minimize the Pearson correlation between $\mathcal{C}$ and $\mathcal{H}$

$$\mathcal{L}_{\text{pc}}(x_i, c_i, h_i) = \frac{\sum_j (c_{ij} - \overline{c_i})(h_{ij} - \overline{h_i})}{\sqrt{\sum_j (c_{ij} - \overline{c_i})^2 \sum_j (h_{ij} - \overline{h_i})^2}} \tag{5.3}$$

where $\overline{h_i}$ and $\overline{c_i}$ denote the mean values of elements in the vectors h_i and c_i, respectively.

Randomized Cross-Entropy Loss $\mathcal{L}_{\text{rand}}$: The most commonly used loss function for image classification is the cross-entropy:

$$\mathcal{L}_{\text{rand}}(x_i, c_i, h_i) = -\sum_j h_{ij} \log c_{ij} \tag{5.4}$$

which allows comparing the class predictions vector $c_i \in \mathcal{C}$ and the ground truth one-hot encoded vector $h_i \in \mathcal{H}$, penalizing the correct classes having a probability diverging from the expected value. Our ultimate goal is to push the network to filter unnecessary domain information during training. This can be easily obtained by shuffling the ground truth $h_i \in \mathcal{H}$ at each training iteration and for every single input. This encourages the model to avoid learning the correct classes as they are always changing.

Discrete Uniform Loss $\mathcal{L}_{\text{du}}$: Analyzing the problem from a different prospective, to remove the domain bias, the encoder network ϕ_E should generate a representation $r_i \in \mathcal{R}$ from which the domain classifier in ϕ_R cannot extract information. This should correspond to a classifier in ϕ_R that classify any class with equal probability, independently from the input. We can obtain this result training the model with the cross-entropy loss (Equation 5.4), forcing the domain ground truth to be a uniform distribution $h_i : \{h_{ij} = \frac{1}{n}\}_{j=1}^{n} \in \mathcal{H}$

Segmentation Loss $\mathcal{L}_{\text{seg}}$: To fit the model according to the segmentation task, we used a well-known loss function for image segmentation, namely the soft-Dice loss function [4]:

$$\mathcal{L}_{seg}(x_i, s_i, g_i) = 1 - \frac{2\sum_i s_i g_i}{\sum_i s_i^2 + \sum_i g_i^2} \tag{5.5}$$

where $s_i \in \mathcal{S}$ is the dense pixel-wise prediction and $g_i \in \mathcal{G}$ is the corresponding ground truth segmentation. This function penalizes ϕ_E and ϕ_D based on the overlap between the prediction and the ground truth segmentation.

5.2.3 Implementation Details

We designed a backbone model based on the U-Net [80], built by concatenating a down-sampling encoder ϕ_E made by 4 stages with an up-sampling decoder ϕ_D made by 4 stages. However, differently from the U-Net, the regularization layers (dropout) were removed and all 2D operations were replaced by their 3D counterparts. During the training procedure, the backbone model was combined with the proposed regularization network. On the contrary, during the testing phase, the regularization network was removed addressing only the segmentation task. The model was executed on 3D patches with size $64 \times 64 \times 64$ cropped from each volume (with 50% overlap [36]). The model was trained only using patches containing at least one voxel labeled as a lesion. During the test, on the contrary, all patches were used.

The evaluation of the model was performed on the reconstructed full-size volumes, fusing the predictions for all patches.

The proposed model was implemented in Python language [1] using Keras[2] [18] with Tensorflow[3] [1] backend. We trained our model using Adam optimizer with an initial learning rate of 0.0001. The size of batch and the maximum number of training epochs were fixed respectively at 15 and 500 (with 300 steps per epoch). Regarding the model initialization, all blocks were randomly initialized from a Gaussian distribution. The hyperparameter λ in Equation 5.1 was selected through grid search with values equal to 0.2, 0.3 and 0.1 for $\mathcal{L}_{\text{pc}}$, $\mathcal{L}_{\text{du}}$ and $\mathcal{L}_{\text{rand}}$, respectively.

5.3 Experiments

We evaluate the performance of the proposed method on two different datasets: an in-house clinical dataset from UBC (refer to section 1.3.3 for more details) and the publicly available ISBI 2015 Longitudinal MS Lesion Segmentation Challenge dataset [11] (refer to section 1.3.1 for more details).

5.3.1 Experiments on the UBC dataset

In the UBC dataset, all images were skull-stripped using BET [91], and rigidly registered to the $1mm^3$ MNI-ICBM152 template [74] using FLIRT [46].

We considered each site as a separate domain and to keep data balanced over all available sites, a single subject including one time point with four MRI modalities (T1w, T2w, PDw, and FLAIR) was selected from each site. We implemented 5-fold cross-validation over the whole data (60%:20%:20% for training, validation, and test, respectively).

For comparison purpose, we repeated the above-mentioned experiment (using exactly the same folds) for the backbone model without any regularization (denoted as the BM) and the same backbone model with additional dropout layers (denoted as the BDM). Note that BDM model is equivalent the U-net model [80] with replacing the 2D operations by their 3D counterparts.

[1] https://www.python.org

[2] https://keras.io

[3] https://www.tensorflow.org

5.3.2 Experiments on the ISBI dataset

From the ISBI dataset, we selected the preprocessed version of the images available online at the challenge website. All images were already skull-stripped using BET [91], rigidly registered to the $1mm^3$ MNI-ICBM152 template [74] using FLIRT [46, 47] and N3 intensity normalized [90].

In this experiment, we consider the whole UBC dataset as a training set with the same assumption described in aforementioned section 5.3.1. However, the performance of the trained model was evaluated on the official ISBI test set (with 14 subjects), for which the ground truth was not available. The extracted 3D output binary lesion maps were submitted to the challenge website[4] for evaluation. Moreover, for comparison purpose, we also repeated the above-mentioned experiment using the BM model.

5.4 Results

5.4.1 UBC dataset

We evaluated our model using four measures: DSC, LTPR, LFPR, and PPV (refer to section 1.4 for more details).

Table 5.1 summarizes the results of our experiment on the test set comparing our model with other baseline methods. The Table shows the mean value of DSC, LTPR, LFPR, and PPV. As can be seen, our proposed methods outperformed baseline methods on the DSC measure. Moreover, in terms of LTPR and LFPR measures, our model with the randomized and discrete uniform auxiliary loss functions showed more balanced performance compared with the other models. Figure 5.2 shows an example of segmentation of all methods for a random subject. Figure 5.3 compares the DSC performance of the proposed methods with other models on the validation set. Confirming the results reported in the test set, as shown in Figure 5.3, our model with all three possible auxiliary loss terms depicts better DSC performance than the baseline methods.

5.4.2 ISBI dataset

All 3D binary output masks computed on the test set were submitted to the ISBI website. Therefore, the performance of the methods was evaluated by the challenge web service. At

[4] http://iacl.ece.jhu.edu/index.php/MSChallenge

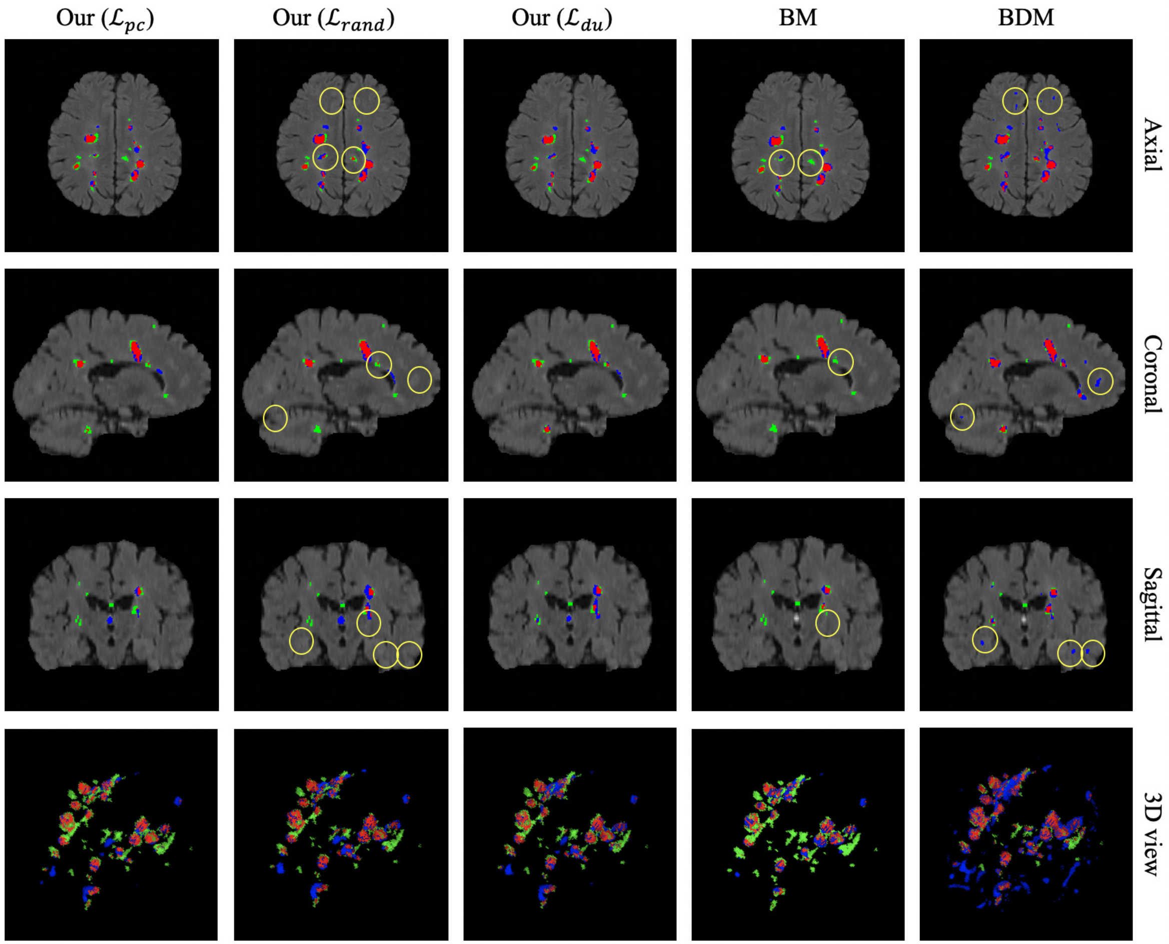


Figure 5.2: Segmentation of a random subject obtained by different methods against ground truth annotation. On all images, true positives, false negatives, and false positives are marked in red, green and blue, respectively (refer to the section for yellow circles).

Table 5.1: Results related to our experiment on UBC dataset. Mean values of DSC, LTPR, LFPR, and PPV were measured for different methods. Values in bold and italic indicate the first-best and second-best results.

Method	DSC	LTPR	LFPR	PPV
Our ($\mathcal{L}_{\mathrm{pc}}$)	0.4638	0.4267	0.3954	0.4865
Our ($\mathcal{L}_{\mathrm{rand}}$)	**0.5001**	0.4618	**0.3348**	**0.5193**
Our ($\mathcal{L}_{\mathrm{du}}$)	*0.4893*	*0.4670*	0.3525	*0.5182*
BM	0.4540	0.4318	*0.3383*	0.5088
BDM	0.4598	**0.5821**	0.5151	0.4577

the time we submitted the results, we obtained a score of 86.65 for our proposed method with randomized cross-entropy loss. Regading the BM model, the overall evaluation score was 85.14. The detailed result for each subject is available online on the ISBI MS lesion segmentation challenge website[5].

5.5 Discussion and Conclusions

In this chapter, we have introduced a generalization method implemented via an auxiliary loss with three variants. We tested this method on medical image segmentation, particularly MS lesion segmentation from MRI modalities in the presence of domain shift originating from multi-center datasets. The proposed model is the combination of a traditional encoder-decoder network for segmentation and an additional regularization network including an auxiliary loss term for domain generalization.

Investigating the impact of the proposed method summarized in Table 5.1, our model always outperformed the baseline models when considering the DSC measures (regardless of which of the adopted auxiliary loss variant was used). However, the best performance in terms of DSC, LFPR, and PPV measures among all tested models is provided by our model with the randomized auxiliary loss function. The BDM model showed the best LTPR measure together with the worst LFPR measure showing that this model has very poor trade-off between LTPR and LFPR.

Confirming the above-mentioned point, Figure 5.2 shows that BDM model tends to over segmented lesion regions (referring to the yellow circles in the last column). Moreover, it can be seen that BM model did not identify some lesions (referring to the yellow circles in third column). However, the proposed method with randomized loss term shows a considerable good performance by not only identifying the mentioned small lesions but also ignoring the false positives (referring to the yellow circles in the second column).

[5] http://iacl.ece.jhu.edu/index.php/MSChallenge

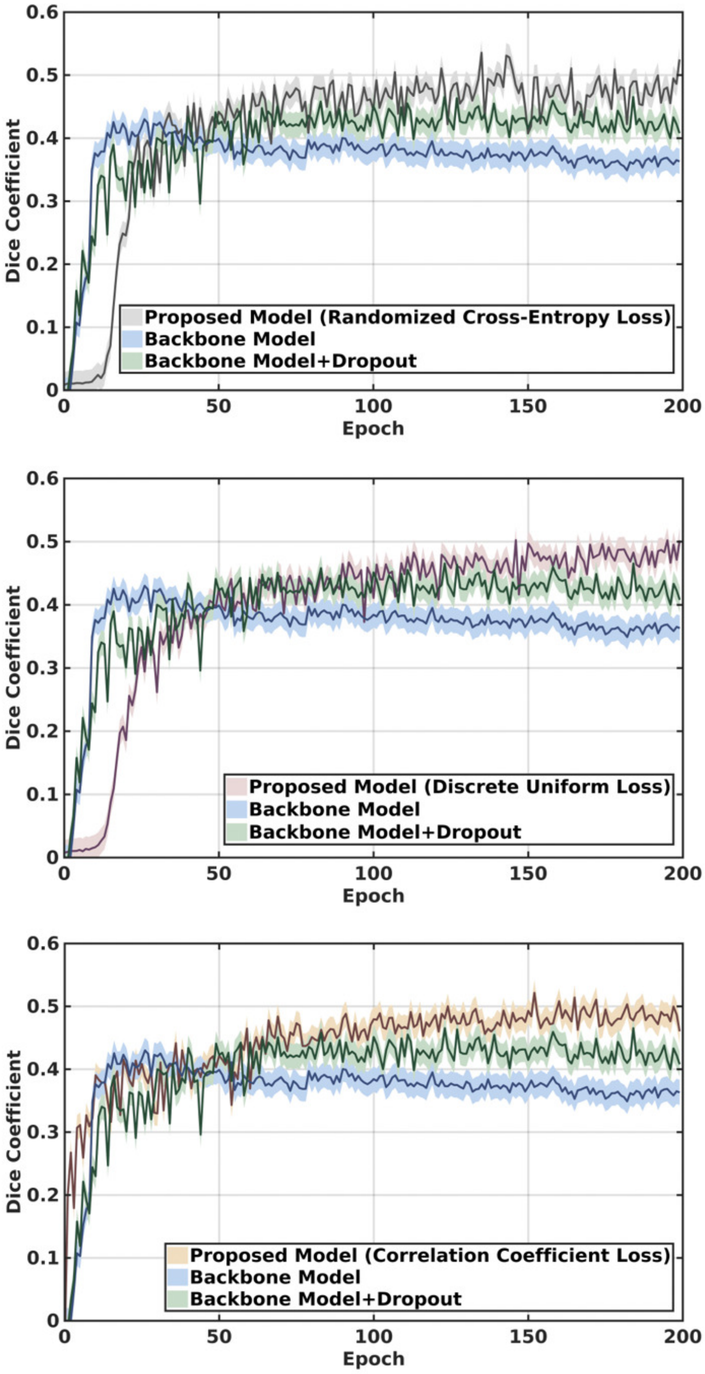


Figure 5.3: Comparison of the DSC measure performance of the proposed methods with other baseline models on validation set during training.

The reported performance related to the proposed method evaluated on the official ISBI test ($SC_{our} = 86.65$) is comparable to the ISBI inter-rater score scaled to 90. Althouth the repoeted score is lower than the other scores related to the state-of-the-art methods (refer to the Table 4.5), it is important to notice that our model has never been trained on the ISBI dataset. Moreover, it showed higher score than the BM model ($SC_{BM} = 85.14$) with highlights the impact of the proposed generalization method.

Chapter 6

Conclusions and Future Work

6.1 Summary and Conclusions

This thesis provided several approaches to advance automated image-based machine diagnosis of MS.

Chapter 2 provided a comprehensive review on state-of-the-art deep learning-based methods, particularly CNNs commonly used as top performing machine-based approaches for natural and medical image segmentation. It started by a review of the literature on CNN-based approaches for semantic segmentation of natural images. Then, it provided a brief review on CNN-based methods for segmentation of medical images. Finally, this chapter discussed several studies for MS lesion segmentation based on CNNs.

Chapter 3 presented our first published work examining whether the parameters of a CNN learned from natural images transfer well to MS lesion segmentation from MRI images. The results show that the parameters of a CNN trained on natural images in the classification task can generalize to brain MRI images in MS lesion segmentation task. Moreover, multiple short-cut connections between several layers of the network combining multi-level features from different stages of the network helps the network exploiting more contextual information about the shape of the lesions.

As the pre-trained CNN worked well for MS lesion segmentation, Chapter 4 discussed our second published work based on a novel multi-branch CNN architecture with end-to-end training that can take advantage of each MRI modality individually for MS lesion segmentation. Further, in the mentioned model, MRI modalities combination was analyzed to identify the best MS lesion segmentation performance. Combination of modalities helps the algorithm leveraging additional information regarding the location of lesions. However, supporting our claim that stacking all modalities together as a single input to the network is not an optimal solution, top performance, indeed, was obtained in most metrics with the proposed multi-branch network when using all available modalities. Moreover, examining the influence of different modalities, results show that the most important modality for MS lesion segmentation is FLAIR.

Going beyond the segmentation of MS lesions in data collected from a single center, Chapter 5, our third published work, presented an effective and novel generalization method for MS lesion segmentation when data are collected from multiple MRI scanning sites and are consequently affected by (site-)domain shifts. The proposed network includes an auxiliary loss function that is designed to encourage the model to ignore domain-specific information. Considering the impact of the proposed method, our model always outperformed the baseline models regardless of which of the adopted auxiliary loss variant was used.

CNNs played a key component in all these approaches, where each proposed work modified the traditional CNN architecture to suit a particular aspect of our tasks. The thesis progression mirrored the increase in sophistication and generalizability of our CNN implementations, starting with a pre-trained CNN (Chapter 3), then relying on a multi-branch pre-trained CNN benefiting from modality specific information (Chapter 4), finally utilizing a generalized CNN in the presence of domain shift originating from multi-center datasets (Chapter 5).

6.2 Future Directions

An open problem in MS lesion segmentation is the identification of cortical and sub-cortical lesions. To this aim, we plan to use other MRI modalities such as double inversion recovery (DIR) sequences, which benefit from signal suppression from both cerebrospinal fluid and WM.

Moreover, we believe that introducing information from tissue class could help improve the network identifying cortical, sub-cortical and WM lesions. Therefore, it could be promising to design a multi-task network for segmenting different parts of the brain including different tissue classes (WM, gray matter, cerebrospinal fluid) and different types of MS lesions (including cortical lesions).

Since the assessment of the disease burden of MS patients from MRI requires the quantification of the volume of hyper intense lesions on T2-weighted images, the final goal of this proposed thesis is to obtain a fully automated and robust MS lesion segmentation tool. This will be particularly useful to facilitate scaling advanced MS analysis based on myelin imaging [22] or multi-modal characterization of white matter tracts [23] to large datasets.

The long term goal, more generally, is the translation of the proposed automated methods into a clinical tool. However, to be fully ready for clinical applications, the proposed methods should be also validated on healthy subjects and in a longitudinal framework. Testing on healthy subjects needs to be done to evaluate the amount of false positives generated by our approach on healthy brain scans. The experiments in a longitudinal framework would be useful to assess the model reliability and capability to identify new, enlarged and stable lesions. Achieving both these aims would bring our methodological approach closer to our ultimate objective of deploying it in clinical settings. Note that automated methods can also help study the effect of treatment in clinical trials involving multiple centers without the bias typically introduced by specific raters. While they cannot help for screening yet (would require research including other diseases to differentiate from) and will not replace clinicians to make the diagnosis, they can help refine the diagnosis by characterizing MS lesions, notably the lesion load.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

[2] S. Andermatt, S. Pezold, and P. C. Cattin. Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. In *International MICCAI Brainlesion Workshop*, pages 31–42. Springer, 2017.

[3] S. Aslani, M. Dayan, V. Murino, and D. Sona. Deep 2d encoder-decoder convolutional neural network for multiple sclerosis lesion segmentation in brain MRI. In *International MICCAI Brainlesion Workshop*, pages 132–141. Springer, 2018.

[4] S. Aslani, M. Dayan, L. Storelli, M. Filippi, V. Murino, M. A. Rocca, and D. Sona. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage*, 196:1–15, 2019.

[5] S. Aslani, V. Murino, M. Dayan, R. Tam, D. Sona, and G. Hamarneh. Scanner invariant multiple sclerosis lesion segmentation from MRI, 2019.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[7] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[8] A. Birenbaum and H. Greenspan. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 58–67. Springer, 2016.

[9] 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*, 35(5):1229–1239, 2016.

[10] M. Cabezas, A. Oliver, S. Valverde, B. Beltran, J. Freixenet, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó. Boost: A supervised approach for multiple sclerosis lesion segmentation. *Journal of Neuroscience Methods*, 237:108–117, 2014.

[11] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148:77–102, 2017.

[12] R. Caruana, S. Lawrence, and C. L. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, pages 402–408, 2001.

[13] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–514. Springer, 2015.

[14] H. Chen, X. Qi, L. Yu, and P.-A. Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[18] F. Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[19] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[20] O. Commowick, F. Cervenansky, and R. Ameli. Msseg challenge proceedings: Multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure, 2016.

[21] A. Compston and A. Coles. Multiple sclerosis. *The Lancet*, 372(9648):1502–1517, 2008.

[22] M. Dayan, S. M. Hurtado Rúa, E. Monohan, K. Fujimoto, S. Pandya, E. M. LoCastro, T. Vartanian, T. D. Nguyen, A. Raj, and S. A. Gauthier. MRI analysis of white

matter myelin water content in multiple sclerosis: a novel approach applied to finding correlates of cortical thinning. *Frontiers in Neuroscience*, 11:284, 2017.

[23] M. Dayan, E. Monohan, S. Pandya, A. Kuceyeski, T. D. Nguyen, A. Raj, and S. A. Gauthier. Profilometry: a new statistical framework for the characterization of white matter pathways, with application to multiple sclerosis. *Human Brain Mapping*, 37(3):989–1004, 2016.

[24] H. Deshpande, P. Maurel, and C. Barillot. Adaptive dictionary learning for competitive classification of multiple sclerosis lesions. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 136–139. IEEE, 2015.

[25] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.

[26] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[27] M. Filippi, M. Horsfield, H. Ader, F. Barkhof, P. Bruzzi, A. Evans, J. Frank, R. Grossman, H. McFarland, P. Molyneux, et al. Guidelines for using quantitative measures of brain magnetic resonance imaging abnormalities in monitoring the treatment of multiple sclerosis. *Annals of Neurology*, 43(4):499–506, 1998.

[28] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. In *International Workshop on Object Representation in Computer Vision*, pages 335–360. Springer, 1996.

[29] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018.

[30] M. Ghafoorian, N. Karssemeijer, T. Heskes, M. Bergkamp, J. Wissink, J. Obels, K. Keizer, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, et al. Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, 14:391–399, 2017.

[31] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. van Uden, C. I. Sanchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):5110, 2017.

[32] M. Ghafoorian and B. Platel. Convolutional neural networks for ms lesion segmentation, method description of diag team. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate

object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[34] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013.

[35] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

[36] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2018.

[37] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[39] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[40] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, pages 1–15, 2019.

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

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[43] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[44] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2017–2025, Cambridge, MA, USA, 2015. MIT Press.

[45] S. Jain, D. M. Sima, A. Ribbens, M. Cambron, A. Maertens, W. Van Hecke, J. De Mey, F. Barkhof, M. D. Steenwijk, M. Daams, et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from mr images. *NeuroImage:*

Clinical, 8:367–375, 2015.

[46] M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[47] M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[48] A. Jesson and T. Arbel. Hierarchical mrf and random forest segmentation of ms lesions and healthy tissues in brain mri. *The Longitudinal MS Lesion Segmentation Challenge*, 2015.

[49] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

[50] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Geometry driven semantic labeling of indoor scenes. In *European Conference on Computer Vision*, pages 679–694. Springer, 2014.

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

[52] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller. Deep mri brain extraction: a 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[54] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957, 1992.

[55] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[56] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[57] H. Li, R. Zhao, and X. Wang. Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification. *arXiv preprint arXiv:1412.4526*, 2014.

[58] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.

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

[60] 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*, 42:60–88, 2017.

[61] S. Liu, D. Zhang, Y. Song, H. Peng, and W. Cai. Triple-crossing 2.5 d convolutional neural network for detecting neuronal arbours in 3d microscopic images. In *International Workshop on Machine Learning in Medical Imaging*, pages 185–193. Springer, 2017.

[62] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira. Segmentation of multiple sclerosis lesions in brain mri: a review of automated approaches. *Information Sciences*, 186(1):164–185, 2012.

[63] 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*, pages 3431–3440, 2015.

[64] O. Maier and H. Handels. Ms lesion segmentation in mri with random forests. *Proc. 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

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

[66] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum. Automatic segmentation of mr brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261, 2016.

[67] G. Montavon, G. Orr, and K.-R. Müller. *Neural networks: tricks of the trade*, volume 7700. Springer, 2012.

[68] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012.

[69] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[70] D. Moyer, G. V. Steeg, C. M. Tax, and P. M. Thompson. Scanner invariant representations for diffusion mri harmonization. *arXiv preprint arXiv:1904.05375*, 2019.

[71] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[72] L. Nanni, S. Ghidoni, and S. Brahnam. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71:158–172, 2017.

[73] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmen-

tation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[74] K. Oishi, K. Zilles, K. Amunts, A. Faria, H. Jiang, X. Li, K. Akhter, K. Hua, R. Woods, A. W. Toga, et al. Human brain white matter atlas: identification and assignment of common anatomical structures in superficial white matter. *Neuroimage*, 43(3):447–457, 2008.

[75] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

[76] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4151–4160, 2017.

[77] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[78] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[79] L. A. Rolak. Multiple sclerosis: it is not the disease you thought it was. *Clinical Medicine and Research*, 1(1):57–60, 2003.

[80] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[81] S. Roy, J. A. Butman, D. S. Reich, P. A. Calabresi, and D. L. Pham. Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks. *arXiv preprint arXiv:1803.09172*, 2018.

[82] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[83] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, et al. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage*, 59(4):3774–3783, 2012.

[84] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

[85] J. Shi and J. Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.

[86] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535, 2010.

[87] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV workshops)*, pages 601–608. IEEE, 2011.

[88] J. Simon, D. Li, A. Traboulsee, P. Coyle, D. Arnold, F. Barkhof, J. Frank, R. Grossman, D. Paty, E. Radue, et al. Standardized mr imaging protocol for multiple sclerosis: Consortium of ms centers consensus guidelines. *American Journal of Neuroradiology*, 27(2):455–461, 2006.

[89] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[90] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998.

[91] S. M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[92] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[93] L. Steinman. Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system. *Cell*, 85(3):299–302, 1996.

[94] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells, and S. Warfield. 3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation. *Midas Journal*, 2008:1–6, 2008.

[95] C. H. Sudre, M. J. Cardoso, W. H. Bouvy, G. J. Biessels, J. Barnes, and S. Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Transactions on Medical Imaging*, 34(10):2079–2102, 2015.

[96] E. M. Sweeney, R. T. Shinohara, N. Shiee, F. J. Mateen, A. A. Chudgar, J. L. Cuzzocreo, P. A. Calabresi, D. L. Pham, D. S. Reich, and C. M. Crainiceanu. Oasis is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri. *NeuroImage: Clinical*, 2:402–413, 2013.

[97] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[98] G. Tetteh, V. Efremov, N. D. Forkert, M. Schneider, J. Kirschke, B. Weber, C. Zimmer, M. Piraud, and B. H. Menze. Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. *arXiv preprint arXiv:1803.09340*, 2018.

[99] X. Tomas-Fernandez and S. K. Warfield. A model of population and subject (mops) intensities with application to multiple sclerosis lesion segmentation. *IEEE Transac-*

tions on Medical Imaging, 34(6):1349–1361, 2015.

[100] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6393–6400, 2017.

[101] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi. Longitudinal multiple sclerosis lesion segmentation using 3d convolutional neural networks. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[102] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

[103] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, and X. Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21:101638, 2019.

[104] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, P. Suetens, et al. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Transactions on Medical Imaging*, 20(8):677–688, 2001.

[105] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903, 2004.

[106] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2016.

[107] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[108] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[109] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, H. Roth, A. Myronenko, D. Xu, and Z. Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

[110] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[111] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017.

[112] T. Zhou, S. Ruan, and S. Canu. A review: Deep learning for medical image segmen-

tation using multi-modality fusion. *Array*, page 100004, 2019.

[113] X. Zhou, R. Takayama, S. Wang, T. Hara, and H. Fujita. Deep learning of the sectional appearances of 3d ct images for anatomical structure segmentation based on an fcn voting method. *Medical Physics*, 44(10):5221–5233, 2017.

List of Publications Related to the Thesis

- **S. Aslani**, M. Dayan, V. Murino, D. Sona, "Deep 2D Encoder-Decoder Convolutional Neural Network for Multiple Sclerosis Lesion Segmentation in Brain MRI", In International MICCAI BrainLesion Workshop, pages 132-141, Springer, 2018. (Related to chapter 3)

- **S. Aslani**, M. Dayan, L. Storelli, M. Filippi, V. Murino, M.A. Rocca, D. Sona, "Multi-branch Convolutional Neural Network for Multiple Sclerosis Lesion Segmentation", NeuroImage, 169:1-15, 2019. (Related to chapter 4)

- **S. Aslani**, V. Murino, M. Dayan, R. Tam, D. Sona, G. Hamarneh, "Scanner Invariant Multiple Sclerosis Lesion Segmentation from MRI", ISBI, 2020. (Related to chapter 5)