Learning brain alterations in multiple sclerosis from multimodal neuroimaging data

Wen Wei

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THESE DE DOCTORAT

Apprentissage Automatique des Altérations Cérébrales Causées par la Sclérose en Plaques en Neuro-Imagerie Multimodale

Wen Wei

INRIA, Équipes Epione et ARAMIS

Dirigée par : Nicholas Ayache, Olivier Colliot
Soutenue le : 5 Juin 2020

Présentée en vue de l'obtention du grade de docteur en **AUTOMATIQUE, TRAITEMENT DU SIGNAL ET DES IMAGES** d'**UNIVERSITE COTE D'AZUR**

Devant le jury, composé de :

Isabelle Bloch	Télécom Paris, France	Rapporteur, Examinateur
Pierrick Coupé	Université de Bordeaux, France	Rapporteur, Examinateur
Hervé Delingette	INRIA Sophia Antipolis, France	Président, Examinateur
Bruno Stankoff	ICM - Hôpital Pitié Salpêtrière, France	Examinateur
Nicholas Ayache	INRIA Sophia Antipolis, France	Directeur de thèse, Examinateur
Olivier Colliot	Sorbonne Université & INRIA, France	Co-directeur de thèse, Invité
Stanley Durrleman	INRIA & ICM, France	Invité

Inria

ICM

Apprentissage Automatique des Altérations Cérébrales Causées par la Sclérose en Plaques en Neuro-Imagerie Multimodale

Learning Brain Alterations in Multiple Sclerosis from Multimodal Neuroimaging Data

Jury :

Président du jury :
Hervé Delingette - INRIA Sophia Antipolis, France

Rapporteurs :
Isabelle Bloch - Télécom Paris, France
Pierrick Coupé - Université de Bordeaux, France

Examinateurs :
Isabelle Bloch - Télécom Paris, France
Pierrick Coupé - Université de Bordeaux, France
Hervé Delingette - INRIA Sophia Antipolis, France
Bruno Stankoff - ICM - Hôpital Pitié Salpêtrière, France
Nicholas Ayache - INRIA Sophia Antipolis, France

Invités :
Olivier Colliot - Sorbonne Université & INRIA, France
Stanley Durrleman - INRIA & ICM, France

Abstract

Multiple Sclerosis (MS) is the most common progressive neurological disease of young adults worldwide and thus represents a major public health issue with about 90,000 patients in France and more than 500,000 people affected with MS in Europe. In order to optimize treatments, it is essential to be able to measure and track brain alterations in MS patients. In fact, MS is a multi-faceted diseases which involves different types of alterations, such as myelin damage and repair. Under this observation, multimodal neuroimaging are needed to fully characterize the disease. Magnetic resonance imaging (MRI) has emerged as a fundamental imaging biomarker for multiple sclerosis because of its high sensitivity to reveal macroscopic tissue abnormalities in patients with MS. Conventional MR scanning provides a direct way to detect MS lesions and their changes, and plays a dominant role in the diagnostic criteria of MS. Moreover, positron emission tomography (PET) imaging, an alternative imaging modality, can provide functional information and detect target tissue changes at the cellular and molecular level by using various radiotracers. For example, by using the radiotracer [^{11}C]PIB, PET allows a direct pathological measure of myelin alteration. However, in clinical settings, not all the modalities are available because of various reasons. In this thesis, we therefore focus on learning and predicting missing-modality-derived brain alterations in MS from multimodal neuroimaging data.

Keywords: Multiple Sclerosis, PET Imaging, MR Imaging, Brain Alterations, Deep Learning, Convolutional Neural Networks (CNN), Generative Adversarial Network (GAN), Image Synthesis, Missing MRI Sequences, Missing Modalities

Résumé

La sclérose en plaques (SEP) est la maladie neurologique évolutive la plus courante chez les jeunes adultes dans le monde et représente donc un problème de santé publique majeur avec environ 90 000 patients en France et plus de 500 000 personnes atteintes de SEP en Europe. Afin d'optimiser les traitements, il est essentiel de pouvoir mesurer et suivre les altérations cérébrales chez les patients atteints de SEP. En fait, la SEP est une maladie aux multiples facettes qui implique différents types d'altérations, telles que les dommages et la réparation de la myéline. Selon cette observation, la neuroimagerie multimodale est nécessaire pour caractériser pleinement la maladie. L'imagerie par résonance magnétique (IRM) est devenue un biomarqueur d'imagerie fondamental pour la sclérose en plaques en raison de sa haute sensibilité à révéler des anomalies tissulaires macroscopiques chez les patients atteints de SEP. L'IRM conventionnelle fournit un moyen direct de détecter les lésions de SEP et leurs changements, et joue un rôle dominant dans les critères diagnostiques de la SEP. De plus, l'imagerie par tomographie par émission de positons (TEP), une autre modalité d'imagerie, peut fournir des informations fonctionnelles et détecter les changements tissulaires cibles au niveau cellulaire et moléculaire en utilisant divers radiotraceurs. Par exemple, en utilisant le radiotraceur [^{11}C]PIB, la TEP permet une mesure pathologique directe de l'altération de la myéline. Cependant, en milieu clinique, toutes les modalités ne sont pas disponibles pour diverses raisons. Dans cette thèse, nous nous concentrons donc sur l'apprentissage et la prédiction des altérations cérébrales dérivées des modalités manquantes dans la SEP à partir de données de neuroimagerie multimodale.

Mots clés: Sclérose en Plaques, TEP, IRM, Altérations Cérébrales, Apprentissage en Profondeur, Réseau de Neurones Convolutifs (CNN), Réseaux Antagonistes Génératifs (GAN), Synthèse d'images, Séquences IRM Manquantes, Modalités Manquantes

Acknowledgements

The period of my PhD has been a truly priceless life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

First and foremost I would like to express my sincere gratitude to my supervisors, Nicholas and Olivier, for their continuous support of my PhD study in two teams, for their patience, motivation, and immense knowledge. I appreciate all their contributions of time, ideas, and funding to make my PhD experience productive and stimulating. Many thanks to Nicholas, who not only encouraged my research, but also expertly guided and taught me to grow as a research scientist. Many thanks also to Olivier, whose advices on both research as well as on my career have been priceless. I could not have imagined having better advisors and mentors for my PhD study.

Apart from my supervisors, I won't forget to express the gratitude to my clinical collaborators: Emilie Poirion, Bruno Stankoff, Benedetta Bodini and Matteo Tonietto, for giving the encouragement and sharing insightful suggestions. I would like also to thank Bruno for being one of the members of the thesis committee. I am especially grateful to Emilie Poirion who helped me a lot on brain image processing and answered lots of my questions about multiple sclerosis.

Then, I want to thank the reviewers of my thesis, Isabelle Bloch, and Pierrick Coupé, for reading and reviewing my manuscript, as well as for attending my thesis defense especially during this difficult time because of COVID-19.

My sincere thanks goes to all the members in the team Epione(Asclepios), including those who have already left. Thanks to Hervé, Xavier, Maxime, and Marco for answering my questions and sharing their ideas. Special thanks to Hervé for being the jury president. Many thanks to Thomas, Marc-Michel, Sophie, Pawel, Raphaël, Roch, Rocio, Sofia, Julian, Shuman, Zihao, Qiao, Nina, Manon, Yann T., Nicholas G., Nicholas C., Clement, Jaume, Luigi, Sara, Buntheng, Benoit, Florent, Santiago, Yingyu, Fanny, Alan, Loic and many

others for the great time we have shared. Special thanks to Isabelle, for supporting and helping me a lot with my travels between Paris and Nice, houses, holidays, VISA and so on. I would like also to thank the members in the team ARAMIS including Ninon, Stanley, Junhao, Fabrizio, Marie-Constance, Igor, Raphael, Tiziana, Simona, Johann, Giulia, Elina, Alexandre R, Alexandre B, Manon, Pascal, Alexis, Jorge, and many others. Moreover, I acknowledge the financial support from INRIA, 3IA Côte d'Azur and ICM.

Lastly, thanks to Danny and Alexandra who treat us like their own children and we have shared unforgettable memories in Villa Plantagenet in Antibes. Many thanks to my parents in law who always encouraged me. I would like to say a heartfelt thank you to my family for all their love and encouragement. A special thanks to my lovely parents who raised me and supported me in all my pursuits. Finally I would like to express appreciation to my lovely, patient, encouraging wife Pengchi, who has been extremely supportive and always been my side throughout my PhD. Without her, my life wouldn't complete so. And also to my darling Yanxi for choosing us as her parents and making my PhD more memorable.

Wen Wei
Antibes
During the COVID-19 lockdown

Contents

1 Introduction

Contents

1.1 Context

1.1.1 Multiple Sclerosis

Multiple Sclerosis is the most common progressive neurological disease of young adults worldwide and thus represents a major public health issue with about 90,000 patients in France and more than 500,000 people affected with MS in Europe [1]. This disease is an autoimmune disease in which the immune system attacks myelinated axons in the central nervous system (CNS), damaging or destroying the myelin (demyelination). This damage disrupts the ability of CNS to transmit signals, leading to various symptoms, including paralysis, sensory disturbances, lack of coordination and visual impairment [Compston, 2008]. Clinically, MS can present as different dynamic phenotypes [Lublin, 2014]:

1) Relapsing-remitting MS (RRMS), the most common disease course, is characterized by clearly defined attacks (also called relapses) followed by periods of partial or complete recovery (remissions). During remissions, all symptoms may disappear, or some symptoms may continue and become permanent;

[1] MS Barameter 2015: http://www.emsp.org/projects/ms-barometer/

2) Secondary progressive MS (SPMS), develops from RRMS for many people. Patients with SPMS generally have fewer relapses and a progressive worsening of neurological function, because nerves have begun to be damaged or lost at this stage;

3) Primary progressive MS (PPMS), is characterized by worsening neurological function and gradual accumulation of disability from the onset of symptoms, without early relapses or remissions.

The cause of MS is still unknown. Scientists believe that a combination of environmental and genetic factors contribute to the risk of developing MS.

1.1.2 Multimodal Neuroimaging in Multiple Sclerosis

Neuroimaging is increasingly used to help clinicians in understanding MS physiopathological mechanisms, such as myelin damage and repair, monitoring disease progression, and improving the accuracy of MS diagnosis and prognosis. In the last decade, magnetic resonance imaging (MRI) has emerged as a fundamental imaging biomarker for multiple sclerosis because of its high sensitivity to reveal macroscopic tissue abnormalities in patients with MS. Conventional MR scanning provides a direct way to detect MS lesions and their changes, and plays a dominant role in the diagnostic criteria of MS [Thompson, 2018]. In particular, T2-weighted image is highly sensitive in detection of hyperintense lesions in the white matter (WM) so that the quantification of T2 lesion load is often used to assess the disease burden. As periventricular lesions are often indistinguishable from the adjacent cerebrospinal fluid (CSF) which is also of high signal on the T2-w, fluid-attenuated inversion recovery (FLAIR) is especially helpful in the evaluation of these lesions due to its ability to suppress the ventricular signal. In addition, double inversion recovery (DIR) has direct application in MS for evaluating cortical pathology. Unlike conventional MR imaging, magnetization transfer ratio (MTR) offers greater pathologic specificity for macromolecules and is utilized to measure myelin content and tissue damage. However, the pathological specificity of MTR is limited since the signal can be influenced by water content and inflammation.

MRI has been regarded as the golden standard in MS research and diagnosis. However, positron emission tomography (PET) imaging, an alternative imaging modality, can provide functional information and detect target tissue changes at the cellular and molecular level by using various radiotracers. In recent years, the researchers have been successful in developing novel tracers

for multiple different aspects of MS to enhance understanding the pathophysiology of the disease [Poutiainen, 2016]. For example, the radiotracer [^{11}C]PIB is used as a myelin tracer in MS clinical settings because of its ability to selectively bind to myelinated white matter regions [Stankoff, 2011]. As mentioned above, all of these multimodal neuroimages play different roles in MS diagnosis and clinical research. However, in clinical settings, not all the modalities are available because of various reasons, such as patients' interruptions resulting in the missing of some MRI pulse sequences. In this work, we therefore focus on learning and predicting missing-modality-derived brain alterations in MS from multimodal neuroimaging data.

1.2 Deep Learning for Medical Image Prediction

In the recent years, deep learning has achieved state-of-the-art results in various areas including computer vision and medical image analysis. In addition, benefit from modern hardware and software resource, deep learning models can be trained very fast and applied for huge high-dimensional datasets. In the particular case of medical image prediction, many researchers are trying to explore how to use deep learning methods to deal with various challenges in this field. Among them, convolutional neural networks (CNNs) and generative adversarial networks (GANs) are two mainly used models.

1.2.1 Convolutional Neural Networks (CNNs)

The architecture of a CNN is analogous to that of the connectivity pattern of neurons in the human brain and was inspired by the organization of the visual cortex. Standard convolutional neural networks include an input, an output layer, as well as multiple hidden layers which are typically a series of convolutional layers followed by additional layers such as pooling layers, fully connected layers and normalization layers. Various methods based on CNNs have been proposed for medical image prediction, for instance, reconstruction of 7T-like images from 3T MRI [Bahrami, 2016b], synthesis of CT images from MRI [Nie, 2016], prediction of positron emission tomography (PET) images with MRI [Li, 2014], and generation of FLAIR from T1-w MRI [Sevetlidis, 2016].

Among the CNN architectures, the most commonly used framework is U-Net [Ronneberger, 2015] which has achieved competitive performance in

both computer vision [Ma, 2018; Zhang, 2018b] and medical imaging fields [Rohé, 2017; Zheng, 2018]. Benefiting from the introduced skip connections in U-Net, the network is able to retrieve the spatial information lost during the down-sampling operations. In addition, the gradient vanishing problem which is a typical issue during the training process is mitigated, since the gradients from the deeper layers can be directly back-propagated to the shallower layers through the skip connections. Improved results have been shown for image prediction by using U-Net model [Han, 2017; Sikka, 2018].

1.2.2 Generative Adversarial Networks (GANs)

The original GAN was proposed by Goodfellow et al. [Goodfellow, 2014] for nature image synthesis. Different from the CNN-based models, the GAN consists of two components: a generator G and a discriminator D. The generator G is trained to generate samples which are as realistic as possible, while the discriminator D is trained to maximize the probability of assigning the correct label both to training examples from the real dataset and samples from G. This adversarial training strategy can make the synthesized image to be indistinguishable from the real ones. In order to constrain the outputs of the generator G, conditional GAN (cGAN) [Mirza, 2014] was proposed in which the generator and the discriminator both receive a conditional variable.

More recently, a lot of works using GAN-based methods have further improved the medical image prediction results, such as PET-to-MRI prediction for the quantification of cortical amyloid load [Choi, 2018] and CT-to-PET synthesis [Bi, 2017]. Several studies also achieved state-of-the-art results via GANs on other modality synthesis, for instance retinal images [Costa, 2018; Zhao, 2018], ultrasound images [Hu, 2017] and endoscopy images [Mahmood, 2018]. Unlike optimizing a single loss function used in standard convolutional neural networks, both the generator and the discriminator in GANs have cost functions that are defined in terms of both players' parameters. Because each player's cost depends on the other player's parameters, but each player cannot control the other player's parameters, this scenario is most straightforward to describe as a game rather than as an optimization problem. Both the generator and the discriminator are trained simultaneously until their losses converge to certain constant numbers, indicating that the GANs model finally finds a Nash equilibrium between the generator and discriminator networks.

1.3 Thesis overview

In this thesis, we aim to propose efficient methods to learn and predict brain alterations in MS from multimodal neuroimaging data. The following chapters correspond to published or submitted articles during the preparation of the thesis.

In **chapter 2**, we propose 3D fully convolutional neural networks to predict FLAIR pulse sequence from some other MRI pulse sequences, such as T1-w, T2-w, and so on. Our approach is tested on a real multiple sclerosis image dataset and evaluated by comparing our approach to other methods. As the FLAIR pulse sequence is used clinically and in research for the detection of WM lesions, we also assess the lesion contrast in the ground truth and the synthesized FLAIR pulse sequences from our method and other methods. This chapter is based on the publication [Wei, 2019a].

In **chapter 3**, we aim to learn and predict myelin content which is quantified by PET imaging and is essential to understand the MS physiopathology, track progression and assess treatment efficacy. For this purpose, we propose Sketcher-Refiner GANs with specifically designed adversarial loss functions to predict the PET-derived myelin content map from multisequence MRI. A visual attention saliency map is also proposed to interpret the attention of neural networks. We compared our method with state-of-the-art methods. Particularly, it is evaluated at both global and voxel-wise levels for myelin content prediction. The work presented in this chapter is published in [Wei, 2019b]. A preliminary version of this work was presented orally at MICCAI 2018 and published in the proceedings of the conference [Wei, 2018b].

In **chapter 4**, our goal is to further learn and predict myelin changes (i.e. demyelination-remyelination cycles) for MS individual *longitudinal* analysis. The method is based on conditional flexible self-attention GAN (cFSAGAN) which is specifically adjusted for high-dimensional medical images and able to capture the relationships between the spatially separated lesional regions during the image synthesis process. Jointly applying the sketch-refinement process described in chapter 3, the result is further improved and the method is shown to outperform the state-of-the-art methods qualitatively and quantitatively. Importantly, the clinical evaluations of our method for the prediction of myelin content for MS individual longitudinal analysis show similar results to the PET-derived gold standard. This study has been submitted to the conference MICCAI 2020 [Wei, 2020a] and the journal NeuroImage [Wei, 2020b].

In **chapter 5**, we finally summarize the main contributions of the thesis and discuss the perspectives for future research work.

2

FLAIR MR Image synthesis from Multisequence MRI using 3D Fully Convolutional Networks for Multiple Sclerosis

Contents

This chapter corresponds to the following scientific articles:

- [Wei, 2019a] *Fluid-attenuated Inversion Recovery MRI Synthesis from Multisequence MRI using Three-dimensional Fully Convolutional Networks for Multiple Sclerosis*
  W.Wei, E.Poirion, B.Bodini, S.Durrleman, O.Colliot, B.Stankoff, N.Ayache
  *Journal of Medical Imaging (**JMI**), 6(01):27, February 2019*

- [Wei, 2018a] *FLAIR MR Image Synthesis by Using 3D Fully Convolutional Networks for Multiple Sclerosis*
  W.Wei, E.Poirion, B.Bodini, S.Durrleman, O.Colliot, B.Stankoff, N.Ayache
  ***ISMRM-ESMRMB 2018** - Joint Annual Meeting, Paris, France*

2.1 Introduction

Multiple sclerosis (MS) is a demyelinating and inflammatory disease of the central nervous system and a major cause of disability in young adults [Compston, 2008]. MS has been characterized as a white matter (WM) disease with the formation of WM lesions, which can be visualized by magnetic resonance imaging (MRI) [Paty, 1988; Barkhof, 1997]. The fluid-attenuated inversion recovery (FLAIR) MRI pulse sequence is commonly used clinically and in research for the detection of WM lesions which appear hyperintense compared to the normal appearing WM tissue (NAWM). Moreover, the suppression of the ventricular signal, characteristic of the FLAIR images, allows an improved visualization of the periventricular MS lesions [Woo, 2006], and can also suppress any artifacts created by CSF. In addition, the decrease of the dynamic range of the image can make the subtle changes easier to see. Typical MRI pulse sequences used in a clinical setting are shown in Fig. 2.1. WM lesions (red rectangles) characteristic of MS are clearly best seen on FLAIR pulse sequences. However, in a clinical setting, some MRI pulse sequences can be missing because of limited scanning time or patients' interruptions in case of anxiety, confusion or severe pain. Hence, there is a need for predicting the missing FLAIR when it has not been acquired during patients' visits. FLAIR may also be absent in some legacy research datasets, that are still of major interest due to their number of subjects and long follow-up periods, such as ADNI [Mueller, 2005]. Furthermore, the automatically synthesized MR images may also improve brain tissue classification and segmentation results as suggested in the works of Iglesias et al. [Iglesias, 2013] and Van Tulder and Bruijne [Van Tulder, 2015], which is an additional motivation for this work.

In the work of Roy et al. [Roy, 2010], the authors proposed an atlas-based patch matching method to predict FLAIR from T1-w and T2-w. In this approach, given a set of atlas images $(I_{\mathrm{T1}}, I_{\mathrm{T2}}, I_{\mathrm{FLAIR}})$ and a subject S with $(S_{\mathrm{T1}}, S_{\mathrm{T2}})$, the corresponding FLAIR $\hat{S}_{\mathrm{FLAIR}}$ is formed patch by patch. A pair of patches in $(S_{\mathrm{T1}}, S_{\mathrm{T2}})$ is extracted and used to find the most similar one in the set of patches extracted from the atlas $(I_{\mathrm{T1}}, I_{\mathrm{T2}})$. Then the corresponding patch in I_{FLAIR} is picked and used to form $\hat{S}_{\mathrm{FLAIR}}$.

In the work of Jog et al. [Jog, 2014], random forests (RF) are used to predict FLAIR given T1-w, T2-w, and PD. In this approach, a patch at position i is extracted from each of these three input pulse sequences. All these three patches are then rearranged and concatenated to form a column vector X_i. The vector X_i and the corresponding intensity y_i in FLAIR at position of i are

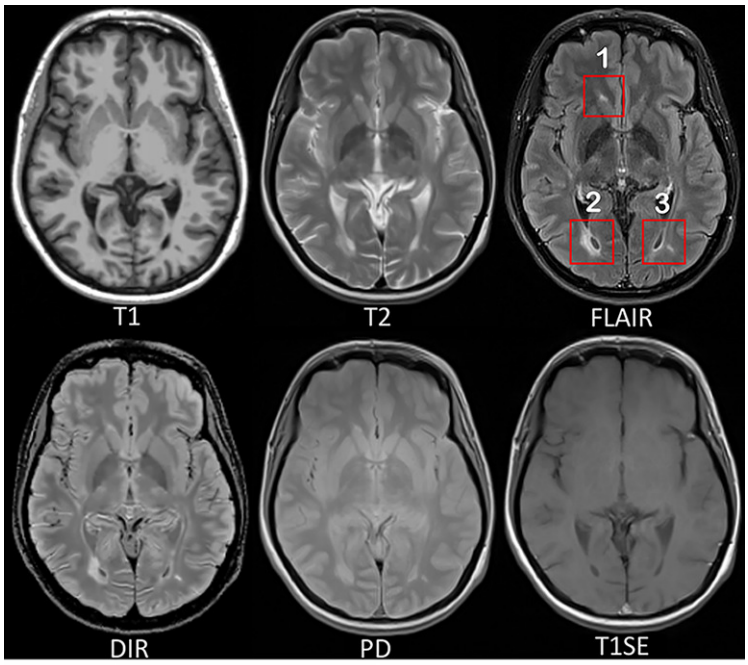


Fig. 2.1: MRI pulse sequences usually used in a clinical setting.
T1-w provides an anatomical reference and T2-w is used for WM lesions visualization. However, on the T2-w, periventricular lesions are often indistinguishable from the adjacent cerebrospinal fluid (CSF) which is also of high signal. WM lesions (red rectangles) characteristic of MS are best seen on FLAIR pulse sequence because of the suppression of the ventricular signal. Double inversion recovery (DIR) has direct application in MS for evaluating cortical pathology. Proton density (PD) and T1 spin-echo (T1SE) are also used clinically.

used to train the RF. There are also some other close research fields doing subject-specific image synthesis of a target modality from another modality. For example, in the works of Huynh et al. [Huynh, 2016] and Burgos et al. [Burgos, 2014], computed tomography (CT) imaging is predicted from MRI pulse sequences.

Recently, deep learning has achieved state-of-the-art results in several computer vision domains, such as image classification [He, 2016], object detection [Chen, 2017], segmentation [Shelhamer, 2017] and also in the fields of medical image analysis [Zhou, 2017]. Various methods of image enhancement and reconstruction using a deep architecture have been proposed, for instance, reconstruction of 7T-like images from 3T MRI [Bahrami, 2016b] and of CT images from MRI [Nie, 2016], and prediction of positron emission tomography (PET) images with MRI [Li, 2014]. The research work most similar to ours is Sevetlidis et al. [Sevetlidis, 2016]. In this method, FLAIR is generated from T1-w MRI by a five-layer 2D deep neural network (DNN) which treats the input image slice-by-slice.

However, these FLAIR synthesis methods have their own shortcomings. The method in the work of Roy et al. [Roy, 2010] breaks the input images into patches. During inference process, the extracted patch is then used to find the most similar patch in the atlas. But this process is often computationally expensive. Moreover, the result heavily depends on the similarity between the source image and the images in the atlas. This makes the method fail in the presence of abnormal tissue anatomy since the images in the atlas do not have the same pathology. The learning based methods in Refs. Jog et al. [Jog, 2014] and Sevetlidis et al. [Sevetlidis, 2016] are less computationally intensive, because they store only the mapping function. However, they do not take into account the spatial nature of 3D images and can cause discontinuous predictions between adjacent slices. Moreover, many works used multiple MRI pulse sequences as the inputs [Roy, 2010; Jog, 2014], but none of them evaluated how each pulse sequence influences the prediction results.

In order to overcome the disadvantages mentioned above, we propose 3D Fully Convolutional Neural Networks (3D FCNs) to predict FLAIR. The proposed method can learn an end-to-end and voxel-to-voxel mapping between other MRI pulse sequences and the corresponding FLAIR. Our networks have three convolutional layers and the performance is evaluated qualitatively and quantitatively. Moreover, we propose a pulse-sequence-specific saliency map (P3S map) to visually measure the impact of each input pulse sequence on the prediction result.

2.2 Method

Standard convolutional neural networks are defined for instance in Refs. LeCun et al. [LeCun, 1989] and Krizhevsky et al. [Krizhevsky, 2012]. Their architectures basically contain three components: convolutional layers, pooling layers, and fully-connected layers. A convolutional layer is used for feature learning. A feature at some locations in the image can be calculated by convolving the learned feature detector and the patches at those locations. A pooling layer is used to progressively reduce the spatial size of feature maps in order to reduce the computational cost and the number of parameters. However, the use of a pooling layer can cause the loss of spatial information, which is important for image prediction, especially the lesion regions. Moreover, a fully-connected layer has all the hidden units connected to all the previous units, so it contains majority of the total parameters and an additional fully-connected layer makes it easy to reach the hardware

limits both in memory and computation power. Therefore, we propose fully convolutional neural networks composed of only three convolutional layers.

2.2.1 3D Fully Convolutional Neural Networks

Our goal is to predict FLAIR pulse sequences by finding a non-linear function s, which maps multi-pulse-sequence source images $\boldsymbol{I}_{\mathbf{source}} = (I_{\mathrm{T1}}, I_{\mathrm{T2}}, I_{\mathrm{PD}}, I_{\mathrm{T1SE}}, I_{\mathrm{DIR}})$, to the corresponding target pulse sequence $\boldsymbol{I}_{\mathbf{target}}$. Given a set of source images $\boldsymbol{I}_{\mathbf{source}}$, and the corresponding target pulse sequence $\boldsymbol{I}_{\mathbf{target}}$, our method finds the non-linear function by solving the following optimization problem:

$$\hat{s} = \arg\min_{s \in S} \frac{\sum_{i=1}^{N} \| (\boldsymbol{I}^{\boldsymbol{i}}_{\mathbf{target}}, s(\boldsymbol{I}^{\boldsymbol{i}}_{\mathbf{source}})) \|_2}{N} \tag{2.1}$$

where S denotes a group of potential mapping functions, N is the number of subjects and mean-square-error (MSE) is used as our loss function which calculates a discrepancy between the predicted images and the ground truth.

In order to learn the non-linear function, we propose the architecture of our 3D fully convolutional neural networks shown in Fig. 2.2. The input layer is composed of the multi-pulse-sequence source images $\boldsymbol{I}_{\mathbf{source}}$ which are arranged as channels and then sent altogether to the network. Our network architecture consists of three convolutional layers ($L = 3$) followed by rectified linear functions ($relu(x) = max(x, 0)$). If we denote the m_{th} feature map at a given layer as $\boldsymbol{h}^{\boldsymbol{m}}$, whose filters are determined by the weights $\boldsymbol{k}^{\boldsymbol{m}}$ and bias b^m, then the feature map $\boldsymbol{h}^{\boldsymbol{m}}$ is obtained as follows:

$$\boldsymbol{h}^{\boldsymbol{m}} = max(\boldsymbol{k}^{\boldsymbol{m}} * \boldsymbol{x} + b^m, 0) \tag{2.2}$$

where the size of input $\boldsymbol{x}$ is $H \times W \times D \times M$. Here, H, W, D indicate the height, width and depth of each pulse sequence or feature map and M is the number of the pulse sequences or feature maps. To form a richer representation of the data, each layer is composed of multiple feature maps $\{\boldsymbol{h}^{\boldsymbol{m}} : 1, ..., F\}$, also referred as channels. Note that the kernel $\boldsymbol{k}$ has a dimension $H_k \times W_k \times D_k \times M \times F$ where H_k, W_k, D_k are the height, width and depth of the kernel respectively. The kernel $\boldsymbol{k}$ operates on $\boldsymbol{x}$ with M channels, generating $\boldsymbol{h}$ with F channels. The parameters $\boldsymbol{k}, b$ in our model

can be efficiently learned by minimizing the function 2.1 using stochastic gradient descent (SGD).

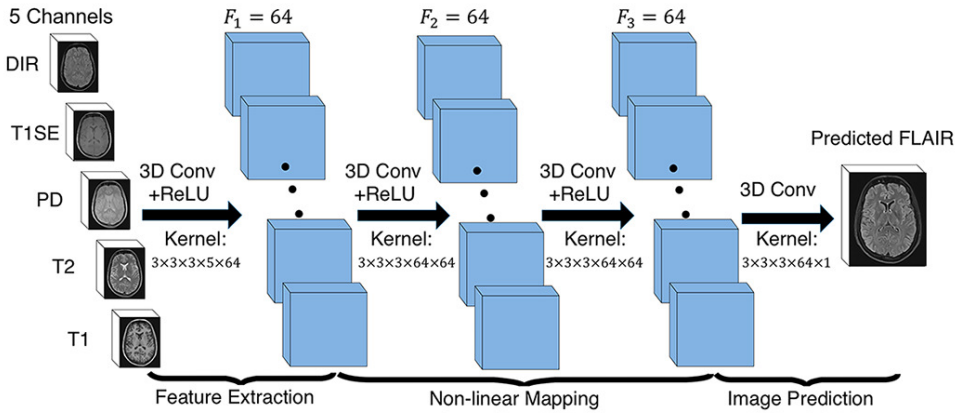


Fig. 2.2: The proposed 3D fully convolutional neural networks. Our network architecture consists of three convolutional layers. The input layer is composed of 5 pulse sequences arranged as channels. The first layer extracts a 64-dimensional feature from input images through convolution process with a $3 \times 3 \times 3 \times 5 \times 64$ kernel. The second and third layers apply the same convolution process to find a non-linear mapping for image prediction.

2.2.2 Pulse-sequence-specific Saliency Map (P3S Map)

Multiple MRI pulse sequences are used as inputs to predict FLAIR. Given a set of input pulse sequences and a target pulse sequence, we would like to assess the contribution of each pulse sequence on the prediction result. One method is class saliency visualization proposed in the work of Simonyan et al. [Simonyan, 2013], which is used for image classification to see which pixels influence the most the class score. Such pixels can be used to locate the object in the image. We call the method presented in this paper *pulse-sequence-specific saliency map* to visually measure the impact of each pulse sequence on the prediction result. Our P3S map is the absolute partial derivative of the difference between the predicted image and the ground truth with respect to the input pulse sequence of subject i. It is calculated by standard backpropagation.

$$\boldsymbol{M_i} = \left| \frac{\partial \| \boldsymbol{I}^i_{\mathbf{target}} - \hat{\boldsymbol{I}}^i_{\mathbf{target}} \|_2}{\partial \boldsymbol{I}^i_{source}} \right| \tag{2.3}$$

where i denotes the subject, $\boldsymbol{I}_{\mathbf{target}}$ and $\hat{\boldsymbol{I}}_{\mathbf{target}}$ are the ground truth and the predicted image, respectively.

2.2.3 Materials and Implementation Details

Our dataset contains 24 subjects including 20 MS patients (8 women, mean age 35.1, sd 7.7) and 4 age- and gender-matched healthy volunteers (2 women, mean age 33, sd 5.6). Each subject underwent the following pulse sequences:

a) T1-w ($1 \times 1 \times 1.1\text{mm}^3$)
b) T2-w and Proton Density (PD) ($0.9 \times 0.9 \times 3\text{mm}^3$)
c) FLAIR ($0.9 \times 0.9 \times 3\text{mm}^3$)
d) T1 spin-echo (T1SE, $1 \times 1 \times 3\text{mm}^3$)
e) Double Inversion Recovery (DIR, $1 \times 1 \times 1\text{mm}^3$)

All have signed written informed consent to participate in a clinical imaging protocol approved by the local ethics committee. The preprocessing steps include intensity inhomogeneity correction [Tustison, 2010] and intra-subject affine registration [Greve, 2009] onto FLAIR space. Finally, each preprocessed image has a size of $208 \times 256 \times 40$ and a resolution of $0.9 \times 0.9 \times 3\text{mm}^3$.

Our networks have three convolutional layers ($L = 3$). The filter size is $3 \times 3 \times 3$ and for every layer the number of the filters is 64 which is designed with empirical knowledge from the widely-used FCN architectures, such as ResNet [He, 2016]. We used Theano [Theano, 2016] and Keras [Chollet, 2015] libraries for both training and testing. The whole data is first normalized by using $\bar{\boldsymbol{x}} = (\boldsymbol{x} - mean)/std$, where $mean$ and std are calculated over all the voxels of all the images in each sequence. We do not use any data augmentation. Our networks were then trained using standard SGD optimizer with 0.0005 as the learning rate and 1 as the batch size. The stopping criteria used in our work is early stopping. We stopped the training when the generalization error increased in p successive q-length-strips:

- $STOP_p$: stop after epoch t iff $STOP_{p-1}$ stops after epoch $t - q$ and $E_{ge}(t) > E_{ge}(t - q)$
- $STOP_1$: stop after first end-of-strip epoch t and $E_{ge}(t) > E_{ge}(t - q)$

where $q = 5, p = 3$ and $E_{ge}(t)$ is the generalization error at epoch t. It takes 1.5 days for training and less than 2 seconds for predicting one image on a NVIDIA GeForce GTX TITAN X.

Our method is validated through a 5-fold cross validation in which the dataset is partitioned into 5 folds (4 folds have 5 subjects with 1 healthy subject in each fold and the last fold has 4 subjects). Subsequently 5 iterations of training and validation are performed such that within each iteration one different fold is held-out for validation and remaining four folds are used for training. The validation error is used as an estimate of the generalization error. And then we compared it qualitatively and quantitatively with four state-of-the-art approaches : modality propagation [Ye, 2013], random forests (RF) with 60 trees [Jog, 2014], U-Net [Ronneberger, 2015], and voxel-wise multilayer perceptron (MLP) which consists of 2 hidden layers and 100 hidden neurons for each layer, trained to minimize the mean squared error. The patch size used in modality propagation and RF is $3\times3\times3$ as suggested in their works [Ye, 2013; Jog, 2014]. The U-Net architecture is separated in 3 parts: downsampling, bottleneck and upsampling. The downsampling path contains 2 blocks. Each block is composed of two $3 \times 3 \times 3$ convolution layers and a max-pooling layer. Note that the number of feature maps doubles at each pooling, starting with 16 feature maps for the first block. The bottleneck is built from simply two 64-width convolutional layers. And the upsampling path also contains 2 blocks. Each block includes a deconvolution layer with stride 2, a skip connection from the downsampling path and two $3 \times 3 \times 3$ convolution layers. Lastly, we use our pulse-sequence-specific saliency map to visually measure the contribution of each input pulse sequence.

2.3 Experiments and Results

2.3.1 Model Parameters and Performance Trade-offs

Number of Filters

Generally, the wider the network is, the more features can be learned so that the better performance can be obtained. Based on this, besides our default setting ($F_1 = F_2 = F_3 = 64$), we also did two experiments for comparison: (1) a wider architecture ($F_1 = F_2 = F_3 = 96$) and (2) a thinner architecture ($F_1 = F_2 = F_3 = 32$). The training process is the same as described in the section 2.2.3. The results are shown in Table 2.1. We can observe that increasing the width of network from 32 to 64 leads to a clear improvement. However, increasing the filter numbers from 64 to 96 only slightly improved the performance. However, if less computational cost is needed, a thinner network which can also achieve a good performance is more suitable.

Tab. 2.1: Comparison of Different Number of Filters

	MSE (SD)	Number of Parameters	Inference Time (sec)
$F_1 = F_2 = F_3 = 32$	1094.52 (49.46)	60.6 K	0.72
$F_1 = F_2 = F_3 = 64$	918.07 (41.70)	213.7 K	1.34
$F_1 = F_2 = F_3 = 96$	909.84 (38.68)	513.5 K	2.58

Number of Layers

It is indicated in Ref. He et al. [He, 2016] that neural networks could benefit from increasing the depth of the networks. We thus tested two different number of layers by adding or removing a 64-width layer based on our default setting ($L = 3$), i.e. (1) $L = 2$ and (2) $L = 4$. The comparison result is shown in Fig. 2.3. It can be found that when $L = 2$, the result is worse than our default setting ($L = 3$). However, when we increased the number of layers to $L = 4$, it converges slower and finally to the same level as the 3-layer network. In addition, we also designed a much deeper network ($L = 6$) by adding three more 64-width layers on our default setting ($L = 3$). It is shown in the Fig. 2.3 that the performance even dropped and failed to surpass the 3-layer network. The cause of this could be the complexity is increasing while the networks are going deeper. During the training process, it is thus more difficult to converge or falls into a bad minimum.

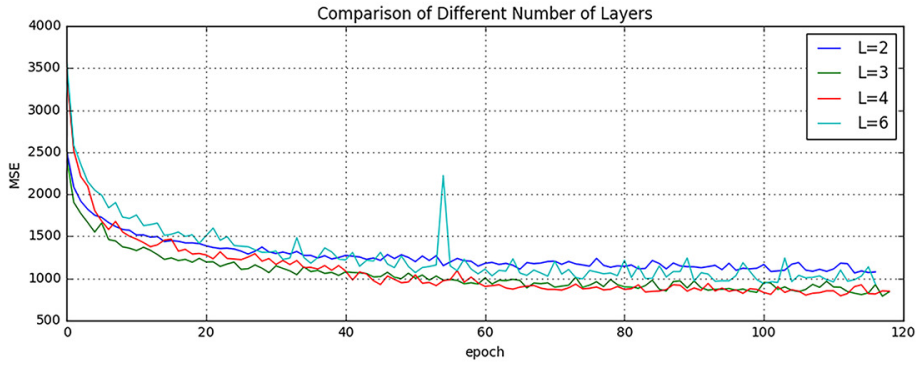


Fig. 2.3: Comparison of Different Number of Layers.
Shown are learning curves for different number of layers ($L = 2, 3, 4, 6$). As the network goes deeper, the result can be increased. However, deeper structure cannot always lead to better results, sometimes even worse.

2.3.2 Evaluation of Predicted Images

Image quality is evaluated by mean square error (MSE) and structural similarity (SSIM). Table 2.2 shows the result of MSE and SSIM on 5-fold cross validation. Our method is statistically significantly better than the rest of the methods ($p < 0.05$) except for U-Net which got the best result on two

folds for MSE and three folds for SSIM. However, the difference with our method is very small and we outperformed at the average level. Furthermore, the number of the parameters in U-Net is 375.6K which is much more than ours (213.7K). If less computational cost is needed, our method is preferred. To further evaluate the quality of our method, in particular on the MS lesions detection, we have chosen to evaluate the MS lesion contrast with the NAWM ($Ratio\,1$) and the surrounding NAWM ($Ratio\,2$), defined by a dilatation of 5 voxels around the lesions. Given the mean intensity of each region $\mathrm{I}_i(R)$ of subject i, $Ratio\,1$ and $Ratio\,2$ are defined as:

$$Ratio\,1 = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{I}_i(Lesions)}{\mathrm{I}_i(NAWM)}, Ratio\,2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{I}_i(Lesions)}{\mathrm{I}_i(SNAWM)} \tag{2.4}$$

As seen from Table 2.3, our method achieves statistically significantly better performances ($p < 0.01$) than other methods on both $Ratio\,1$ and $Ratio\,2$ which reflects a better contrast for MS lesions. The evaluation results can be visualized in Fig. 2.4 with the absolute difference maps on the 2nd and 4th rows. It can be observed that RF and U-Net can generate the good global anatomical information but the MS lesion contrast is poor. This can be truly reflected by a good MSE and SSIM (See in Table 2.2), but a low $Ratio\,1/\,2$ (See in Table 2.3). On the contrary, our method can well keep the anatomical information and also yield the best contrast for WM lesions.

Moreover, we input the synthetic FLAIR and the ground truth to a brain segmentation pipeline [Coupé, 2018] to generate automatic segmentations of WM lesions. A similar segmentation should be obtained if the FLAIR synthesis is good enough and the DICE score is used to compare the overlap of the segmentations previously obtained from both the synthetic FLAIR and the ground truth. We got a very good WM lesion segmentation agreement with a mean (SD) DICE of 0.73(0.12). Some examples are shown in Fig. 2.5.

2.3.3 Pulse-Sequence-Specific Saliency Map (P3S Map)

It can often happen that not all the subjects have the five complete protocols (T1-w, T2-w, T1SE, PD, and DIR). Therefore, it might be useful to measure the impact of each input pulse sequence. Our proposed P3S map is to visually measure the contribution of each input pulse sequence. It can be

Tab. 2.2: Quantitative comparison between our method and other methods

(a) Mean Square Error (Standard Deviation)

	Random Forest 60	Modality Propagation	Multilayer Perceptron	U-Net	Our Method
Fold 1	993.68 (67.21)	2194.79 (118.73)	1532.89 (135.82)	921.69 (38.51)	**905.05 (26.06)**
Fold 2	1056.76 (125.51)	2037.69 (151.23)	1236.53 (100.95)	**912.03 (38.58)**	913.34 (39.95)
Fold 3	945.38 (59.42)	1987.32 (156.11)	1169.78 (142.43)	916.16 (38.97)	**898.76 (46.90)**
Fold 4	**932.67 (74.48)**	2273.58 (217.85)	1023.35 (97.93)	938.34 (52.54)	945.33 (63.80)
Fold 5	987.63 (78.34)	1934.25 (140.06)	1403.57 (146.35)	**908.11 (36.13)**	927.88 (31.80)
Average	983.22 (80.99)	2085.53 (156.80)	1273.22 (124.70)	919.26 (40.95)	**918.07 (41.70)**

(b) Structural Similarity (Standard Deviation)

	Random Forest 60	Modality Propagation	Multilayer Perceptron	U-Net	Our Method
Fold 1	0.814 (0.044)	0.727 (0.044)	0.770 (0.052)	0.847 (0.038)	**0.868 (0.036)**
Fold 2	0.822 (0.038)	0.718 (0.045)	0.773 (0.045)	**0.856 (0.025)**	0.854 (0.028)
Fold 3	0.832 (0.040)	0.713 (0.047)	0.790 (0.044)	0.854 (0.036)	**0.880 (0.031)**
Fold 4	0.850 (0.032)	0.708 (0.049)	0.786 (0.044)	**0.853 (0.031)**	0.846 (0.035)
Fold 5	0.830 (0.041)	0.723 (0.039)	0.781 (0.047)	**0.861 (0.034)**	0.850 (0.027)
Average	0.830 (0.039)	0.718 (0.045)	0.780 (0.046)	0.854 (0.033)	**0.860 (0.031)**

Tab. 2.3: Evaluation of MS lesion contrast (Standard Deviation)

	Random Forest 60	Modality Propagation	Multilayer Perceptron	U-Net	Our Method	*Ground Truth*
Ratio 1	1.33 (0.07)	1.31 (0.06)	1.39 (0.11)	1.34 (0.09)	**1.47 (0.13)**	*1.66 (0.12)*
Ratio 2	1.15 (0.04)	1.13 (0.04)	1.20 (0.05)	1.17 (0.04)	**1.22 (0.07)**	*1.33 (0.09)*

observed in Fig. 2.6 that T1-w, DIR, and T2-w contribute more for FLAIR MRI prediction than PD or T1SE. In the P3S map, the intensity reflects the contribution of each input pulse sequence. In particular, from the P3S map we can easily find which sequence affects more the generation of which

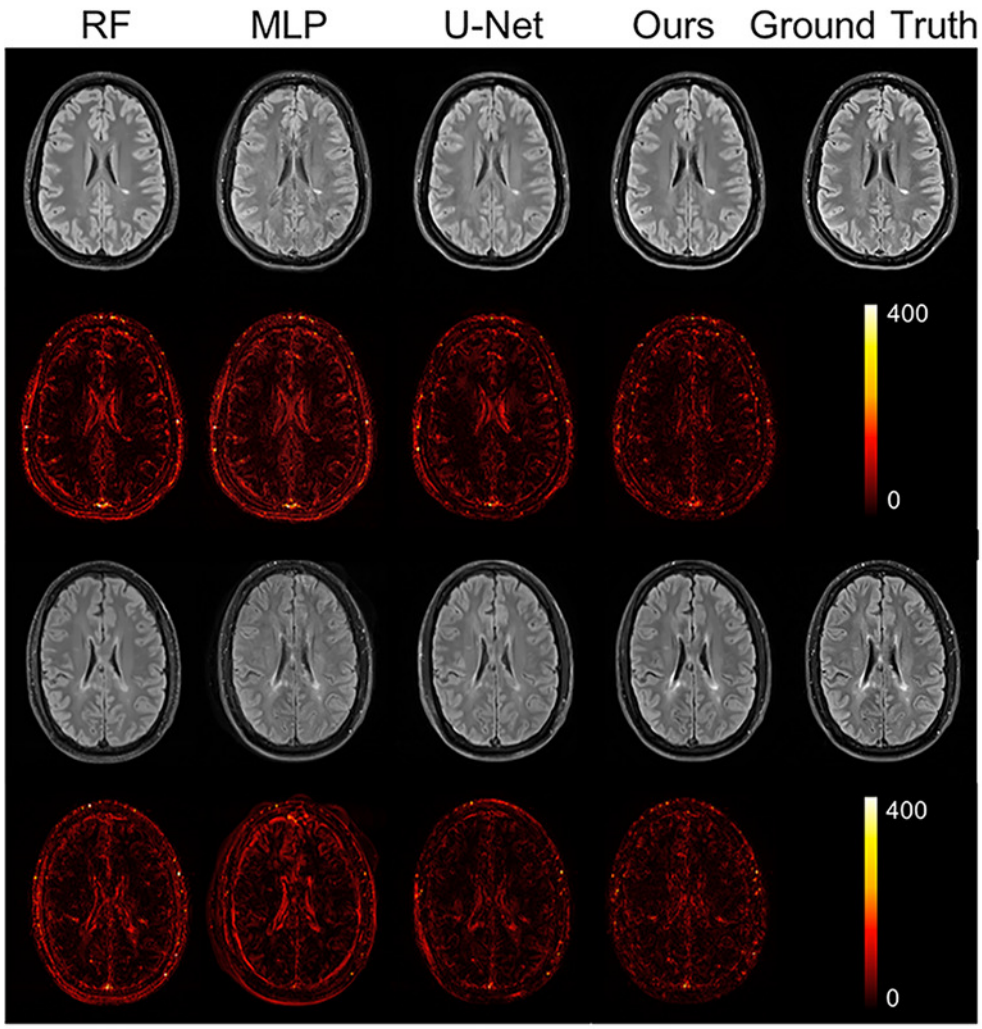


Fig. 2.4: Qualitative comparison of the methods to predict FLAIR sequence. Shown are synthetic FLAIR obtained by RF with 60 trees, MLP, U-Net, and our method followed by the true FLAIR. The 2nd and 4th rows show the absolute difference maps between each synthetic FLAIR and the ground truth.

specific ROIs. For example, as shown in the first row of Fig. 2.6, even though generally DIR is the most important sequence (see Table 2.4(a)), T1-w contributes more for the synthesis of ventricle which can be proved by the high degree of resemblance of ventricle between T1-w and FLAIR (see 2nd row of Fig. 2.6).

In order to test our P3S Map, five experiments have been designed. In each one, we removed one of the five pulse sequences (T1-w, T2-w, T1SE, PD, and DIR) from the input images. Table 2.4(a) shows the testing result on 5-fold cross validation by using MSE as the error metric. As shown in the table, these results are consistent with the observation revealed by our P3S map. The DIR, T1-w and T2-w contribute more than T1SE and PD. In

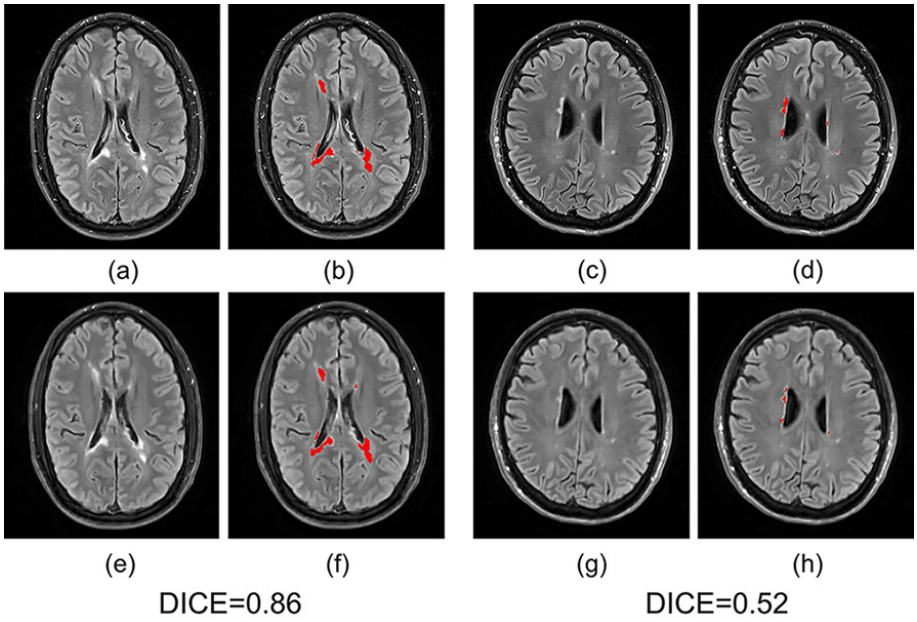


Fig. 2.5: Examples of WM lesion segmentation for a high and a low DICE. The WM lesions are very small and diffuse, so even a slight difference in the overlap can cause a big decrease for the DICE score. (a)(c) True FLAIR. (e)(g) Predicted FLAIR. (b)(d) Segmentation of WM lesions (red) using true FLAIR. (f)(h) Segmentation of WM lesions using predicted FLAIR.

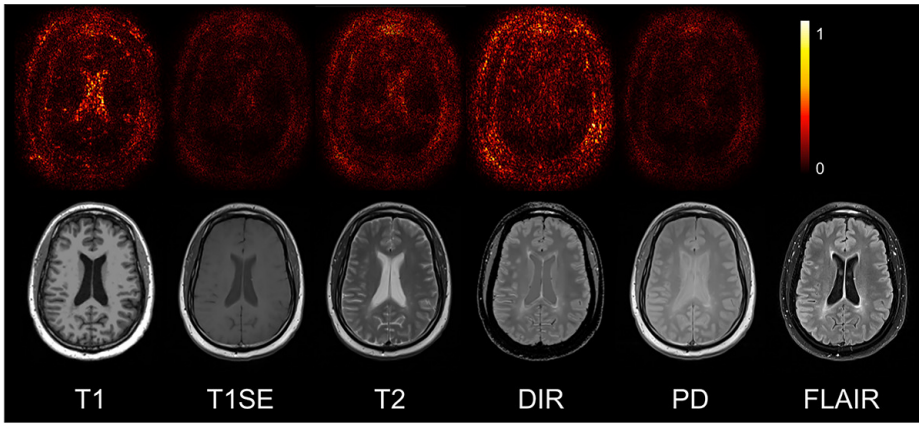


Fig. 2.6: Pulse-Sequence-Specific Saliency Maps for input pulse Sequences. The first row is the saliency maps for T1, T1SE, T2, PD, and DIR, respectively. And the second row is the corresponding multi-sequence MR images. It can be found that T1-w, DIR, and T2-w contribute more for FLAIR MRI prediction than PD or T1SE.

particular, DIR is the most relevant pulse sequences for FLAIR prediction. However, DIR is not commonly used in clinical settings. We thus show a performance comparison between other methods in Table 2.4(b). It can be observed that when DIR is missing, the performance decreases for all the methods suggesting a high similarity between DIR and FLAIR. In addition,

even though DIR is not such common, we still got an acceptable result for FLAIR prediction without DIR.

Besides, some legacy research datasets do not have T1SE or PD, we thus predicted FLAIR from different combinations of T1, T2, DIR and PD (see in Table 2.4(c) and Fig. 2.7). It indicates that our method can be used to get an acceptable predicted FLAIR from the datasets which only contain some sequences. From Table 2.4(c) we can also infer that adding a pulse sequence improves the prediction result.

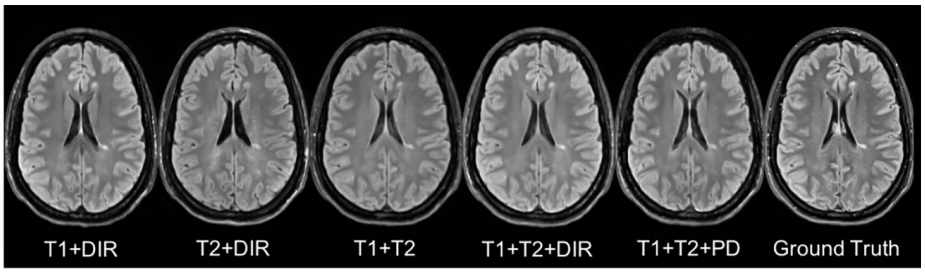


Fig. 2.7: Different Combinations of T1, T2, DIR and PD as input sequences. Shown are synthesized FLAIR with different MRI pulse sequences as inputs from T1+DIR to T1+T2+PD. A better performance can be achieved when both DIR and T1 exist.

2.4 Discussion and Conclusion

We introduced 3D fully convolutional neural networks for FLAIR prediction from multiple MRI pulse sequences, and a sequence-specific saliency map for investigating each pulse sequence contribution. Even though the architecture of our method is simple, the nonlinear relationship between the source images and FLAIR can be well captured by our network. Both the qualitative and quantitative results have shown its competitive performance for FLAIR prediction. Compared to previous methods, representative patches selection is not required so that this speeds up the training process. Additionally, 2D Convolutional Neural Networks (2D CNNs) become popular in computer vision, however they are not suitable to directly use 2D CNNs for volumetric medical image data. Unlike Refs. Sevetlidis et al. [Sevetlidis, 2016] and Jog et al. [Jog, 2014], our method can better keep the spatial information between slices. Moreover, the generated FLAIR has a good contrast for MS lesions. In practice, in some datasets, not all the subjects have all the pulse sequences. Our proposed P3S map can be used to reflect the impact of each input pulse sequence on the prediction result so that the pulse sequences which contribute very little can be removed. Furthermore, DIR is often used for the detection of MS cortical gray matter lesions and if we have DIR, we

can use it to generate FLAIR so that the acquisition time for FLAIR can be saved. Also, our P3S map can be generated by any kinds of neural networks trained by standard backpropagation.

Our 3D FCNs have some limitations. The synthetic images appear slightly more blurred and smoother than the ground truth. This maybe because we use a more traditional loss L2 distance as our objective function. As mentioned in the work of Isola et al. [Isola, 2016], the use of L1 distance can encourage less blurring and generate sharper image. Additionally, the proposed P3S is generated after the data normalization which may affect the gradient. However, the network is changed as the normalization strategy changes. And the saliency map is based on the network. Moreover, the dataset should be ideally partitioned into training-validation-test sets. However, our dataset only has 24 subjects which is quite small to split into training-validation-test set. Instead, we divided it into training-testing set and the testing error is used as an estimate of the generalization error.

In the future, it would be interesting to also assess the utility of the method in the context of other WM lesions (e.g. age-related WM hyperintensities). Specifically, FLAIR is the pulse sequence of choice for studying different types of white matter lesions [Koikkalainen, 2016], including leucoaraiosis (due to small vessel disease) that is commonly found in elderly subjects, that is associated to cognitive decline and is a common co-pathology in neurodegenerative dementias.

Tab. 2.4: FLAIR prediction results by using different input pulse sequences

(a) Mean Square Error (Standard Deviation)

	Removed Pulse Sequence				
	T1	T1SE	T2	PD	DIR
Fold 1	959.75 (60.58)	926.89 (73.25)	981.15 (83.45)	945.79 (67.23)	1097.99 (93.27)
Fold 2	987.13 (91.47)	940.00 (86.34)	994.47 (78.47)	919.09 (69.82)	1097.00 (98.57)
Fold 3	942.76 (59.22)	938.98 (64.27)	940.92 (69.44)	924.59 (61.39)	1065.08 (101.95)
Fold 4	999.64 (100.57)	940.56 (72.98)	939.60 (76.22)	932.46 (59.49)	1151.93 (113.21)
Fold 5	986.55 (71.25)	936.89 (63.23)	953.35 (70.12)	933.12 (65.23)	1068.72 (98.56)
Average	975.16 (76.62)	936.67 (72.00)	961.90 (75.54)	931.01 (64.63)	1096.14 (101.11)

(b) Performance Comparison by removing DIR (Standard Deviation)

	Random Forest 60	Multilayer Perceptron	U-Net	Our Method
Fold 1	**1035.17 (102.37)**	1589.62 (131.32)	1068.59 (100.28)	1097.99 (93.27)
Fold 2	1167.52 (127.67)	1375.28 (121.12)	**998.66 (106.79)**	1097.00 (98.57)
Fold 3	1170.36 (105.37)	1316.53 (128.46)	1135.24 (128.15)	**1065.08 (101.95)**
Fold 4	1218.38 (129.01)	1235.26 (117.26)	1175.68 (107.33)	**1151.93 (113.21)**
Fold 5	1189.64 (108.28)	1537.61 (135.78)	**1003.54 (95.18)**	1068.72 (98.56)
Average	1156.21 (114.54)	1410.86 (126.79)	**1076.34 (107.55)**	1096.14 (101.11)

(c) Mean Square Error (Standard Deviation)

	Input Pulse Sequences				
	T1+DIR	T2+DIR	T1+T2	T1+T2+DIR	T1+T2+PD
Fold 1	966.67 (70.12)	993.25 (99.35)	1375.83 (123.68)	926.88 (83.68)	1281.06 (112.57)
Fold 2	953.87 (68.57)	974.88 (86.32)	1562.46 (132.68)	944.39 (79.23)	1324.17 (121.37)
Fold 3	998.71 (84.90)	1007.69 (103.87)	1158.65 (112.29)	961.19 (71.68)	1261.68 (128.91)
Fold 4	973.24 (77.79)	998.56 (98.23)	1078.67 (103.89)	931.47 (69.31)	1143.58 (98.95)
Fold 5	968.55 (71.59)	986.57 (91.33)	1212.59 (126.79)	958.28 (73.45)	1156.79 (102.67)
Average	972.21 (74.60)	992.19 (95.82)	1277.64 (119.87)	944.44 (75.47)	1233.46 (112.89)

3

Predicting PET-derived Demyelination from Multisequence MRI using Sketcher-Refiner Adversarial Training for Multiple Sclerosis

Contents

This chapter corresponds to the following publications:

- [Wei, 2019b] *Predicting PET-derived Demyelination from Multimodal MRI using Sketcher-Refiner Adversarial Training for Multiple Sclerosis*
  W.Wei, E.Poirion, B.Bodini, S.Durrleman, N.Ayache, B.Stankoff, O.Colliot
  *Medical Image Analysis (**MedIA**), August, 2019*

- [Wei, 2018b] *Learning Myelin Content in Multiple Sclerosis from Multimodal MRI through Adversarial Training*
  W.Wei, E.Poirion, B.Bodini, S.Durrleman, N.Ayache, B.Stankoff, O.Colliot
  *21st International Conference On Medical Image Computing and Computer Assisted Intervention (**MICCAI 2018**)*

3.1 Introduction

Multiple Sclerosis (MS) is the most common cause of chronic neurological disability in young adults, with a clinical onset typically occurring between 20 and 40 years of age [Compston, 2008]. In the central nervous system (CNS), myelin is a biological membrane that enwraps the axon of neurons. Myelin acts as an insulator, enhancing the neural signal conduction velocity as well as balancing the system energy. MS pathophysiology predominately involves autoimmune aggression of central nervous system myelin sheaths. The demyelinating lesions in CNS can cause various symptoms depending on their localizations, such as motor or sensory dysfunction, visual disturbance and cognitive deficit [Compston, 2008]. Therefore, a reliable measure of the tissue myelin content is essential as it would allow to understand key physiopathological mechanisms, such as myelin damage and repair, to track disease progression and to provide an endpoint for clinical trials, for instance assessing neuroprotective and pro-myelinating therapies.

Positron emission tomography (PET) is a nuclear medicine imaging technology based on the injection of a specific radiotracer which will bind to the biological targets within brain tissues. Thus, the imaging procedure offers the potential to investigate neurological diseases at the cellular level. Moreover, another advantage of PET is the absolute quantification of the tracer binding that directly reflects the concentration of the biological target in the tissue of the interest, with excellent sensitivity to changes. [^{11}C]PIB is used as a myelin tracer in MS clinical settings because of its ability to selectively bind to myelinated white matter regions [Stankoff, 2011]. This tracer was initially developed as a marker of beta-amyloid deposition found in the gray matter of patients with Alzheimer's disease (AD) [Rabinovici, 2007]. Nevertheless, note that the signal in myelin is more subtle than for amyloid plaques. However, using PET to quantify myelin content in MS lesions is limited by several drawbacks. First, PET imaging is expensive and not offered in the majority of medical centers in the world. Moreover, it is invasive due to the injection of a radioactive tracer. In addition, the spatial resolution of PET is limited (around 4-5 mm for most cases). As

the myelin content used for MS clinical studies is measured in MS lesions, the quantitative measurements taken from PET images will suffer from the partial volume effect.

On the contrary, MR imaging is a widely available and non-invasive technique. During the past decades, many efforts have been devoted to understand how macroscopic MS lesions visualized on MRI could drive neurological disability over the course of the disease. Even though conventional MRI sequences have a great sensitivity to detect the white matter (WM) lesions in MS, they do not provide a direct and reliable measure of myelin. Specially, they cannot distinguish, within MS lesions, demyelinated voxels from non-demyelinated or remyelinated voxels. Therefore, it would be of considerable interest to be able to predict the PET-derived myelin content map from multimodal MRI. Figure 3.1 illustrates some examples of the ground truth ([^{11}C]PIB PET data) and input multimodal MR images. It can be found that the imaging mechanisms between PET and MRI are very different making our prediction task more difficult.

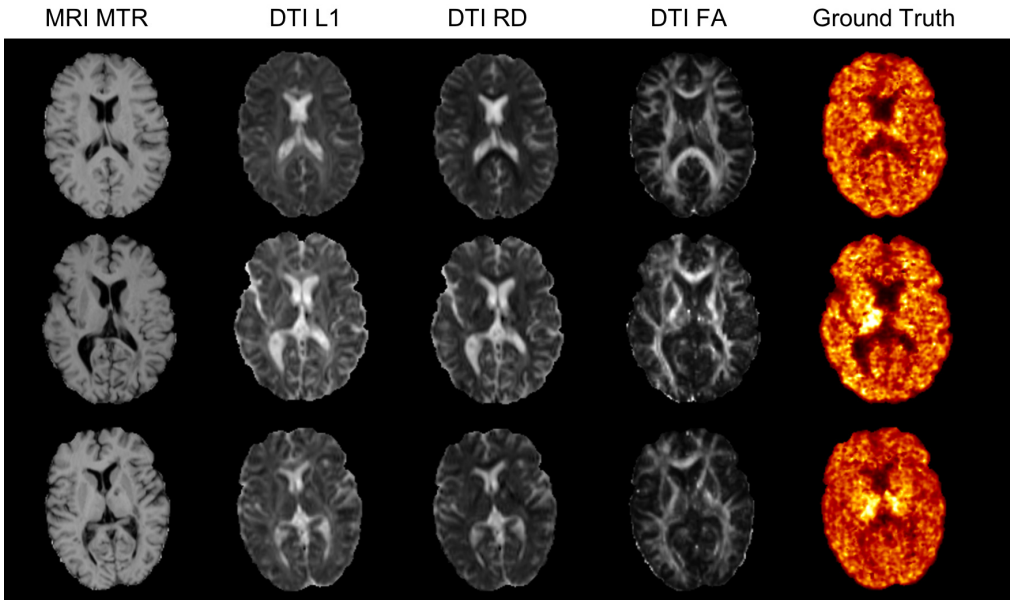


Fig. 3.1: Some examples of the ground truth ([^{11}C]PIB PET data) and input MR images including magnetization transfer ratio (MTR) and three measures derived from diffusion tensor imaging (DTI): fractional anisotropy (FA), radial diffusivity (RD) and axial diffusivity (AD). The relationship between the MR images and the PET data is complex and highly non-linear.

3.1.1 Related Work

To the best of our knowledge, there is currently no method for predicting PET-derived myelin content from MRI. On the other hand, various methods

focusing on estimating one modality image from another modality have been proposed over the last decade. These methods can be mainly classified into the following categories.

(A) **Atlas Registration**. These methods [Hofmann, 2008; Burgos, 2014] usually need an atlas dataset including the pairs of the source and the target modalities. For example, Burgos et al. [Burgos, 2014] proposed to predict a pseudo-CT image from a given MR image. All the MR images in the atlas database are registered to the given MRI. The resulted deformation fields are then applied to register each CT in the atlas database to the given MRI space. The target CT can thus be synthesized through the fusion of the aligned atlas CT images. However, the performance of the atlas-based methods highly depends on the registration accuracy and the quality of the synthesized image may also rely on the priori knowledge for tuning large amounts of parameters in registration step. Moreover, while they seem well adapted to synthesize the overall anatomy (as is typically required in the case of CT synthesis for attenuation correction), they may not be able to accurately predict subtle lesional features, whose location can be highly variable between patients.

(B) **Searching-based methods**. Given a database containing N exemplar pairs of the source image and the target image $\{S_n, T_n\}, n \in N$, the basic idea behind these methods [Ye, 2013; Roy, 2010] is that the local similarity between the new subject source image S_{new} and database source images S_n should indicate the same similarity between the database target images T_n and the image to be synthesized T_{new}. Roy et al. [Roy, 2010] applied this idea to predict FLAIR from T1-w and T2-w. Equally, Ye et al. [Ye, 2013] proposed to generate T2 and DTI-FA from T1 MRI. However, the result heavily depends on the similarity between the source image and the images in the database. This may make the method fail in the presence of abnormal tissue anatomy since the images in the atlas do not have the same pathological features as the patient to predict. Moreover, these methods need to break the image into patches in advance. During inference process, the extracted patch is then used to find the most similar patch in the database. But this process is often computationally expensive.

(C) **Learning-based methods**. Learning-based methods aim to find a non-linear function which maps the source modality to the corresponding target modality. Vemulapalli et al. [Vemulapalli, 2015] proposed an unsupervised approach to generate T1-MRI from T2-MRI and vice versa. The authors aimed to maximize a global mutual information and a local spatial consistency for target image synthesis. In the work

of Jog et al. [Jog, 2014], the authors presented an approach to predict FLAIR given T1-w, T2-w, and PD using random forest. In this approach, a patch at position m is extracted from each of these three input pulse sequences. All these three patches are then rearranged and concatenated to form a column vector X_m. The vector X_m and the corresponding intensity y_m in FLAIR at position of m are used to train the model. Similarly, Huynh et al. [Huynh, 2016] used the structured random forest and auto-context model to predict CT image from MR images. Although these methods have been successful, it appears that the extraction and the fusion of the patches are usually computational expensive. Moreover, the source images are often represented by the extracted features which will influence the final image synthesis quality.

Meanwhile, deep learning techniques [Sevetlidis, 2016; Xiang, 2018; Wang, 2018a] have emerged as a powerful alternative and alleviate the above drawbacks for medical image synthesis. For instance, Sevetlidis et al. [Sevetlidis, 2016] generate FLAIR from T1-w MRI using a deep encoder-decoder network which works on the whole image instead of the image patches. There are also many works trying to generate CT images from MR images using deep learning methods, such as for dose calculation [Han, 2017; Wolterink, 2017; Maspero, 2018] and attenuation correction [Leynes, 2018; Liu, 2018]. In the work of Choi and Lee [Choi, 2018], the authors used GANs to generate the MRI from the PET for the quantification of cortical amyloid load. Bi et al. [Bi, 2017] used multi-channel GANs to synthesis PET images from CT images. Regarding PET synthesis from MRI, several works have already been proposed [Sikka, 2018; Li, 2014; Pan, 2018]. A 3D convolutional neural network (CNN) based on U-Net architecture [Sikka, 2018] and a two-layer CNN [Li, 2014] have been proposed to predict FDG PET from T1-w MRI for AD classification. In recent years, generative adversarial networks (GANs) have been vigorously studied in various image generation tasks, such as conditional GANs for image-to-image translation [Isola, 2016]. The work of Denton et al. [Denton, 2015] also proposed a LAPGAN using a sequence of conditional GANs into the laplacian pyramid framework for the image generation. Regarding the medical image synthesis, Pan et al. [Pan, 2018] proposed a 3D cycle consistent generative adversarial network (3D-cGAN) to generate PET images for AD diagnosis. Note that all these PET synthesis works were devoted to the prediction of the radiotracer FDG. Predicting myelin content (as defined by PIB PET) is a more difficult task because the signal is more subtle and with weaker relationship to anatomical information that could be found in MR images. Moreover, only a single

MRI pulse sequence is used for PET synthesis in these works. However, as suggested in Chartsias et al. [Chartsias, 2018], using multimodal MRI can improve the synthesis performance.

3.1.2 Contributions

In this work, we therefore propose a learning-based method to predict PET-derived demyelination from multiparametric MRI. Consisting of two conditional GANs, our proposed Sketcher-Refiner GANs can better learn the complex relationship between myelin content and multimodal MRI data by decomposing the problem into two steps: 1) sketching anatomy and physiology information and 2) refining and generating images reflecting the myelin content in the human brain. As MS lesions are the areas where demyelination can occur, we thus design an adaptive loss to force the network to pay more attention to MS lesions during the prediction process. Besides, in order to interpret the neural networks, a visual attention saliency map has also been proposed.

A preliminary version of this work was published in the proceedings of the MICCAI 2018 conference [**10.1007/978-3-030-00931-1_59**]. The present paper extends the previous work by: 1) quantitatively comparing our approach to other state-of-the-art techniques; 2) using visual attention saliency maps to better interpret the neural networks; 3) comparing different combinations of MRI modalities and features to assess which is the optimal input; 4) describing the methodology with more details; 5) providing a more extensive account of background and related works.

3.2 Method

3.2.1 Sketcher-Refiner Generative Adversarial Networks

We propose Sketcher-Refiner Generative Adversarial Networks (GANs) with specifically designed adversarial loss functions to generate the [^{11}C]PIB PET distribution volume ratio (DVR) parametric map, which can be used to quantify the demyelination, using multimodal MRI as input. Our method is based on the adversarial learning strategy because of its outstanding performance for generating a perceptually high-quality image. We introduce a sketch-refinement process in which the Sketcher generates the preliminary

anatomical and physiological information and the Refiner refines and generates images reflecting the tissue myelin content in the human brain. We describe the details in the following.

3D Conditional GANs

Generative adversarial networks (GANs) [Goodfellow, 2014] are generative models which consist of two components: a generator G and a discriminator D. Given a database $\boldsymbol{y}$, the generator G defined with parameters θ_g aims to learn the mapping from a random noise vector z to data space denoted as $G(z;\theta_g)$. The discriminator $D(y;\theta_d)$ defined with parameters θ_d represents the probability that y comes from the dataset $\boldsymbol{y}$ rather than $G(z;\theta_g)$. On the whole, the generator G is trained to generate samples which are as realistic as possible, while the discriminator D is trained to maximize the probability of assigning the correct label both to training examples from $\boldsymbol{y}$ and samples from G. In order to constrain the outputs of the generator G, conditional GAN (cGAN) [Mirza, 2014] was proposed in which the generator and the discriminator both receive a conditional variable x. More precisely, D and G play the two-player conditional minimax game with the following cross-entropy loss function:

$$\begin{aligned}\min_G \max_D \mathcal{L}(D,G) = \mathbb{E}_{x,y\sim p_{\text{data}}(x,y)}[\log D(x,y)] - \\ \mathbb{E}_{x\sim p_{\text{data}}(x),z\sim p_z(z)}[\log(1-D(x,G(x,z)))]\end{aligned} \tag{3.1}$$

where p_{data} and p_z are the distributions of real data and the input noise. Both the generator G and the discriminator D are trained simultaneously, with G trying to generate an image as realistic as possible, and D trying to distinguish the generated image from real images.

Sketcher-Refiner GANs

Using multimodal MRI denoted as I_{M}, our goal is to predict the [^{11}C]PIB PET distribution volume ratio (DVR) parametric map I_{P} which can be used to quantify the demyelination. The multiple input modalities I_{M} are arranged as channels with a dimension of $l \times h \times w \times c$, where l, h, w indicate the size of each input modality and c is the number of the modalities. As the signal of the myelin is very subtle, we thus propose a sketch-refinement process. Figure 3.2 shows the architecture of our method consisting of two cGANs named **Sketcher** and **Refiner** with 4 MRI modalities as inputs. Working on the whole images, we decompose the prediction problem into two steps:

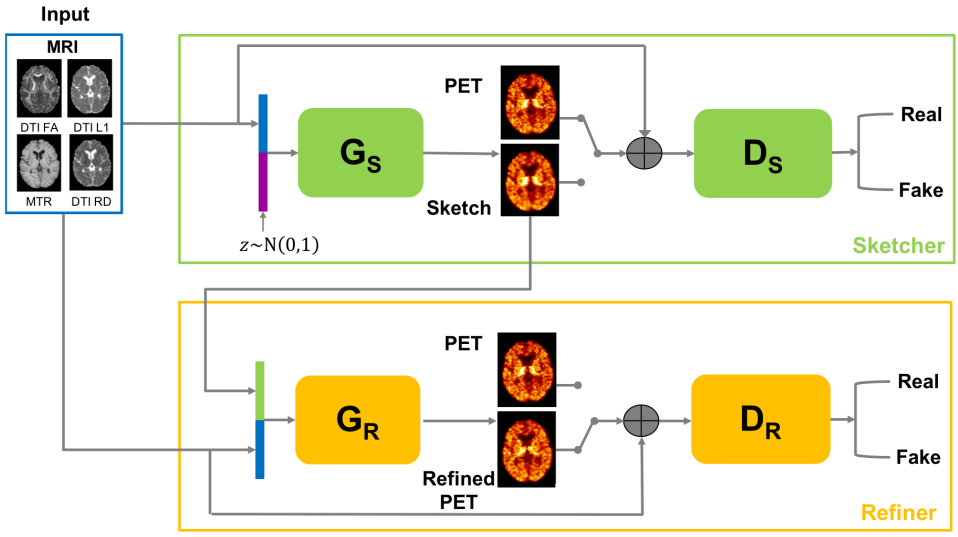


Fig. 3.2: The proposed Sketcher-Refiner GANs. The Sketcher receives MR images and generates the preliminary anatomy and physiology information. The Refiner receives MR images and the output of the Sketcher. Then it refines and generates the synthetic PET images.

1. **Sketcher:** it receives a set of MR image pulse sequences I_{M}. Based on these MR images, it sketches the preliminary anatomy and physiology information.
2. **Refiner:** it receives both the MR image pulse sequences I_{M} and the image generated from previous step I_{S}. Then it refines and generates quantitative images reflecting the tissue myelin content in the human brain. To that purpose, the Refiner pays more attention to lesional areas (where demyelination may occur), using a loss that treats separately lesion, normal appearing white matter (NAWM) defined as the white matter outside visible lesions, and other regions.

Therefore, the Sketcher and the Refiner have the following cross-entropy losses:

$$\begin{aligned}\min_{G_{\mathrm{S}}}\max_{D_{\mathrm{S}}}\mathcal{L}(D_{\mathrm{S}},G_{\mathrm{S}}) = \mathbb{E}_{I_{\mathrm{M}},I_{\mathrm{P}}\sim p_{\mathrm{data}}(I_{\mathrm{M}},I_{\mathrm{P}})}[\log D_{\mathrm{S}}(I_{\mathrm{M}},I_{\mathrm{P}})]- \\ \mathbb{E}_{I_{\mathrm{M}}\sim p_{\mathrm{data}}(I_{\mathrm{M}}),z\sim p_z(z)}[\log(1-D_{\mathrm{S}}(I_{\mathrm{M}},G_{\mathrm{S}}(I_{\mathrm{M}},z)))]\end{aligned} \tag{3.2}$$

$$\begin{aligned}\min_{G_{\mathrm{R}}}\max_{D_{\mathrm{R}}}\mathcal{L}(D_{\mathrm{R}},G_{\mathrm{R}}) = \mathbb{E}_{I_{\mathrm{M}},I_{\mathrm{P}}\sim p_{\mathrm{data}}(I_{\mathrm{M}},I_{\mathrm{P}})}[\log D_{\mathrm{R}}(I_{\mathrm{M}},I_{\mathrm{P}})]- \\ \mathbb{E}_{I_{\mathrm{M}}\sim p_{\mathrm{data}}(I_{\mathrm{M}}),I_{\mathrm{s}}\sim G_{\mathrm{S}}(I_{\mathrm{M}},z)}[\log(1-D_{\mathrm{R}}(I_{\mathrm{M}},G_{\mathrm{R}}(I_{\mathrm{M}},I_{\mathrm{s}})))]\end{aligned} \tag{3.3}$$

where D_{S}, D_{R} and G_{S}, G_{R} represent the discriminators and the generators in the Sketcher and the Refiner respectively. The underlying network architectures for the Sketcher and the Refiner are described in Section 3.2.4.

3.2.2 Adversarial Loss with Adaptive Regularization

Here, we propose specific adversarial losses that produce the desired behaviors for the Sketcher and the Refiner. Previous work of Isola et al. [Isola, 2016] has shown that it can be useful to combine the GAN objective function with a traditional constraint, such as L1 and L2 loss. They further suggested using L1 loss rather than L2 loss to encourage less blurring. We hence mixed the GANs' loss function with the following L1 loss for the Sketcher:

$$\mathcal{L}_{L1}(G_{\mathrm{S}}) = \frac{1}{N}\sum_{i=1}^{N} |I_{\mathrm{P}}^{i} - G_{\mathrm{S}}(I_{\mathrm{M}}^{i}, z^{i})| \tag{3.4}$$

where N is the number of subjects and i denotes the index of a subject.

In CNS, myelin constitutes most of the white matter (WM). Knowing that the demyelinated voxels are mainly found within the MS lesions, we thus want the Refiner network to pay more attention to MS lesions than to the other regions during the prediction process. Most other methods [Roy, 2010; Burgos, 2014; Ye, 2013; Xiang, 2018] tried to synthesize the whole image without any specific focus on some regions of interest. Unlike these methods, to focus the Refiner generator on MS lesions where demyelination happens, the whole image is divided into three regions of interest (ROIs): lesions, NAWM and "other". We thus defined for the Refiner a weighted L1 loss in which the weights are adapted to the number of voxels in each ROI indicated as N_{Les}, N_{NAWM} and N_{other}. Given the masks of the three ROIs: R_{Les}, R_{NAWM} and R_{other}, the weighted L1 loss for the Refiner is defined as follows:

$$\begin{aligned}\mathcal{L}_{L1}(G_{\mathrm{R}}) = \frac{1}{N \times M}\sum_{i=1}^{N}\Big(\frac{1}{N_{\mathrm{Les}}}\sum_{j\in R_{\mathrm{Les}}} |I_{\mathrm{P}}^{i,j} - \hat{I}_{\mathrm{P}}^{i,j}| + \\ \frac{1}{N_{\mathrm{NAWM}}}\sum_{j\in R_{\mathrm{NAWM}}} |I_{\mathrm{P}}^{i,j} - \hat{I}_{\mathrm{P}}^{i,j}| + \frac{1}{N_{\mathrm{other}}}\sum_{j\in R_{\mathrm{other}}} |I_{\mathrm{P}}^{i,j} - \hat{I}_{\mathrm{P}}^{i,j}|\Big)\end{aligned} \tag{3.5}$$

where $\hat{I}_{\mathrm{P}}$ is the prediction output from the Refiner, M is the number of voxels in a PET image, and i, j is the index of a subject and a voxel respectively.

To sum up, our overall objective functions are defined as follows:

$$\begin{aligned}G_{\mathrm{S}}^{*} &= \arg\min_{G_{\mathrm{S}}}\max_{D_{\mathrm{S}}} \mathcal{L}(D_{\mathrm{S}}, G_{\mathrm{S}}) + \lambda_{\mathrm{S}}\mathcal{L}_{L1}(G_{\mathrm{S}}) \\ G_{\mathrm{R}}^{*} &= \arg\min_{G_{\mathrm{R}}}\max_{D_{\mathrm{R}}} \mathcal{L}(D_{\mathrm{R}}, G_{\mathrm{R}}) + \lambda_{\mathrm{R}}\mathcal{L}_{L1}(G_{\mathrm{R}})\end{aligned} \tag{3.6}$$

where λ_{S} and λ_{R} are hyper-parameters which balance the contributions of two terms in the Sketcher and the Refiner respectively.

3.2.3 Visual Attention Saliency Map

Convolutional neural networks and other deep neural networks have achieved breakthrough results in various tasks. However, the lack of interpretability limits the use in clinical applications, because the black-box character of a neural network makes it hard to decompose into understandable components. Broadly speaking, it is necessary to build transparent models which can explain their predictions.

We propose a visual attention saliency map to generate the visual explanations showing the concentration regions of the neural networks for the prediction. Inspired by the work of Simonyan et al. [Simonyan, 2013], our visual attention saliency map is the absolute partial derivative of the prediction loss with respect to the input images $\boldsymbol{I}_{\mathrm{M}}$ defined as follows:

$$\boldsymbol{M} = \left| \frac{\partial Loss}{\partial \boldsymbol{I}_{\mathrm{M}}} \right| \tag{3.7}$$

Given the input images $\boldsymbol{I}_{\mathrm{M}}$, the attention saliency map $\boldsymbol{M}$ is calculated by standard backpropagation. In fact, the saliency maps derived from the generators and the discriminators are different. In GAN, the discriminator is used as a classifier to distinguish if the input is in class “True" or “Fake" . Therefore, the saliency map derived from the discriminator should intuitively highlight salient image regions that most contribute the category classification. In our work, the goal is to interpret the attention of the neural networks for the image synthesis. Therefore, our proposed saliency map is that of the generator.

3.2.4 Network architectures

Both the Sketcher and the Refiner in our method have the same architectures for their generators (respectively for their discriminators). For the generators, we use the 3D U-Net architecture which is widely used and has achieved competitive performance in both computer vision [Ma, 2018; Zhang, 2018b] and medical imaging fields [Rohé, 2017; Zheng, 2018]. The advantage of U-Net [Ronneberger, 2015] is the introduction of skip connections. They help feed the information between the end and the start of the network, allowing a more direct way for the gradient to flow uninterruptedly. In

addition, these skip connections also allow the network to retrieve the spatial information lost during the down-sampling operations. In addition, the spatial information between adjacent slices can be well preserved by the 3D architecture. As shown in Fig. 3.3 (A), the U-Net architecture is symmetric and built with fully convolutional networks with skip connections. It has an Encoder which extracts the spatial features from the input image, and a Decoder which constructs the final output from the encoded features. The Encoder follows the typical architecture of a convolutional network. It includes a sequence of two convolution layers and a convolution with stride 2 for downsampling. This sequence is repeated 3 times and the number of feature maps doubles after each sequence. A progression of two convolutional layers is used to connect the Encoder and the Decoder which inversely involves the 3 repeated sequences of a deconvolution layer with stride 2 and two convolution layers. In all three levels, the output of the convolutional layer (prior to the downsampling operation) in the Encoder is transferred to the output of the upsampling operation in the Decoder by using skip connections. Our 3D U-Net starts with 32 feature maps for the first block (see details in Fig. 3.3 (A)). LeakyReLU is used to allow a stable training of GANs with 0.2 as slope coefficient. The convolution kernel size is $3 \times 3 \times 3$. Batch normalization [Ioffe, 2015] and dropout are applied after each LeakyReLU layer. The rate for dropout layer is 50%.

For the discriminator, a traditional approach in GANs is to use a global discriminator: the discriminator is trained to globally distinguish if the input comes from the true dataset or from the generator. However, the generator may try to over-emphasize certain image features in some regions so that it can make the global discriminator fail to differentiate a real or fake image. In our problem, each region in the PET image has its own myelin content. A key observation is that any local region in a generated image should have a myelin content that is similar to that of the homologous region in the real image. Therefore, instead of using a traditional global network, we define a 3D patch discriminator trained by local patches from input images. As shown in Fig. 3.3 (B), the input image is firstly divided into patches with size $l \times w \times h$ and then the 3D patch discriminator classifies all the patches separately. The final loss of the 3D patch discriminator is the sum of the cross-entropy losses from all the local patches. The PatchGAN was first used in Isola et al. [Isola, 2016] which took the overlapped 2D patches as inputs. Unlike their work, our inputs are 3D patches which need more computational resource. In addition, if we use overlapping patches, the number of patches would be 1.2 million comparing to only 35 thousand in their work. Therefore, considering the computational cost and the GPU memory consumption, we chose to use non-overlapping patches. Its architecture is a traditional CNN

including a series of $3 \times 3 \times 3$ stride 1 convolution layers followed by batch normalization, LeakyReLU and Downsampling. At the end, a fully-connected layer with two nodes and a softmax layer are used to produce the final decision.

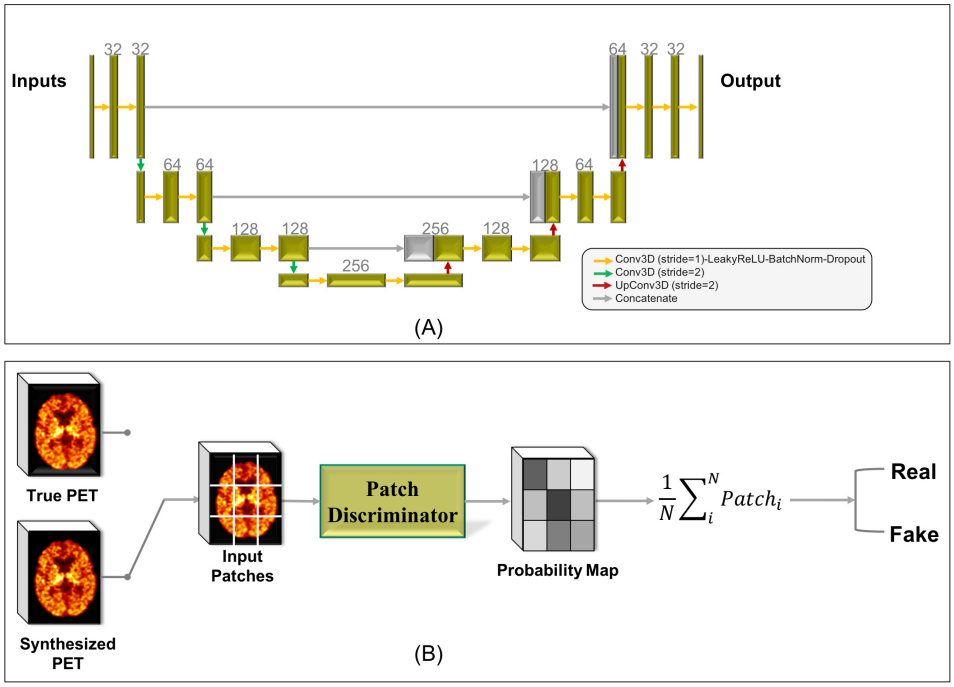


Fig. 3.3: Architectures proposed for the generator (panel A) and for the discriminator (panel B) in our GANs. (A) The 3D U-Net shaped generator with implementation details shown in the image. (B) The proposed 3D patch discriminator which takes all the patches and classifies them separately to output a final loss.

3.3 Experiments and Evaluations

3.3.1 Overview

- **Dataset:** Our dataset includes 18 MS patients (12 women, mean age 31.4 years, sd 5.6) and 10 age-matched healthy volunteers (8 women, mean age 29.4, sd 6.3). The clinical and demographic information is detailed in Bodini et al. [Bodini, 2016]. For each participant, we used the following data:
  a) **MR IMAGES:** MR images were collected using a 3 Tesla Siemens TRIO 32-channel TIM system including Magnetisation Transfer Ratio map (MTR) ($1 \times 1 \times 1.1\text{mm}^3$), and three measures derived from Diffusion Tensor Imaging (DTI): Fractional Anisotropy (FA), Radial Diffusivity (RD) and Axial Diffusivity (AD) ($2 \times 2 \times 2\text{mm}^3$).

The three ROIs (lesions, NAWM and “other") used in Eq. 3.5 were delineated as follows. The hyperintense lesions of MS patients were manually contoured by an expert rater on T2-w scans with reference to FLAIR images. The corresponding lesion masks were generated and aligned to the individual T1-w scan using FLIRT algorithm in the FSL package [Jenkinson, 2012]. After performing a “lesion-filling" procedure in patients only, T1-w scans were segmented using FreeSurfer [Fischl, 2012] to obtain a WM mask. The NAWM is then defined as the WM outside visible lesions on T2-w scans.

b) **PET IMAGES:** PET examinations were performed on a high-resolution research tomograph (HRRT; CPS Innovations, Knoxville, TN) which achieves an intraslice spatial resolution of 2.5mm, with 25-cm axial and 31.2-cm transaxial fields of view. The 90-minute emission scan was initiated with a 1-minute intravenous bolus injection of [^{11}C]PIB (mean = 358 $\pm$ 34 MBq). The Logan graphical reference method [Logan, 1996] was applied at the voxel level on PET scans in native space to obtain [^{11}C]PIB PET distribution volume ratio (DVR) parametric map ($1.22 \times 1.22 \times 1.22\text{mm}^3$).

All participants signed written informed consent to participate in the study, which was approved by the local ethics committee of the Pitié-Salpêtrière hospital. The preprocessing steps mainly consist of brain extraction [Smith, 2002], intensity inhomogeneity correction [Tustison, 2010] and affine intra-subject registration of MR data onto [^{11}C]PIB PET DVR image space using FLIRT algorithm in the FSL package [Jenkinson, 2012]. Finally, we removed part of the background by cropping images to $128 \times 160 \times 128$ with a resolution of $1.22 \times 1.22 \times 1.22\text{mm}^3$. The details of acquisition parameters and PET data quantification are described in Bodini et al. [Bodini, 2016] and Veronese et al. [Veronese, 2015].

- **Training details:** The whole data was first normalized by using $\bar{\boldsymbol{x}} = (\boldsymbol{x} - mean)/std$, where $mean$ and std were calculated over all the voxels of all the images in each sequence. We did not use any data augmentation. During the training process, we first iteratively trained D_S and G_S of the Sketcher for 400 epochs by fixing our Refiner. Then we iteratively trained D_R and G_R of the Refiner from scratch for another 400 epochs by fixing our Sketcher. The optimization was performed with the ADAM solver with 10^{-4}, 5×10^{-5} as initial learning rates for the Sketcher and the Refiner respectively. We used 3-fold cross validation (2 folds have 9 subjects with 3 healthy subjects in each fold and the last fold has 10 subjects with 4 healthy subject). Our Sketcher-Refiner GANs was implemented with the Keras [Chollet,

2015] library with Theano [Theano, 2016] as backend. Two GTX 1080 Ti GPUs were used for training.

In practice, the input noise z is often ignored by the conditional GANs, such as the work of Isola et al. [Isola, 2016]. Actually, in initial experiments, we found that the result was marginally improved by introducing the input noise z which is consistent with Hong et al. [Hong, 2018]. Moreover, the input noise z is used to provide some slight variation in the generated images. If we remove the noise vector, the network can still learn the mapping but it becomes deterministic. Since the output of the Refiner should be deterministic and similar to the true PET image, we kept the noise vector z for the Sketcher and removed it from the Refiner.

3.3.2 Comparisons with state-of-the-art methods

We compared our method with several state-of-the-art methods including a 2-layer DNN [Li, 2014], a 3D U-Net [Sikka, 2018] and a single cGAN [Bi, 2017; Ben-Cohen, 2017] (corresponding to the Sketcher in our approach). The 2-layer DNN consists of two convolutional layers with a filter size of $7 \times 7 \times 7$. To better detect the features, the number of feature maps in each layer is augmented to 64 instead of 10 as mentioned in the paper [Li, 2014]. The architecture of the 3D U-Net is the same as shown in Fig. 3.3 (A). It is similar to 3D U-Net used in the work of [Sikka, 2018], but with a LeakyRelu layer as the last layer instead of sigmoid as our output is not in the range [0,1]. In the works of [Sikka, 2018] and Li et al. [Li, 2014], their proposed methods were aimed to discriminate Alzheimer's disease from normals, the authors thus segmented the images and used gray matter as an input, which is not applicable to our problem. Moreover, unlike the preprocessing step in their paper, we did not downsample our images. In terms of loss function, the L1 loss is optimized for both the 2-layer DNN and the 3D U-Net. In the work of Bi et al. [Bi, 2017], the authors used each patient's lesion label as a separate channel in inputs for CT-to-PET synthesis. As the healthy volunteers in our dataset do not have any lesion, we just took MR images as inputs. To adjust to the 3D image, the 2D cGANs used in Bi et al. [Bi, 2017] and Ben-Cohen et al. [Ben-Cohen, 2017] were extended to 3D architecture which corresponds to the Sketcher (see in Fig. 3.2) in our approach and the loss function was the same as described in Bi et al. [Bi, 2017]. Furthermore, to better compare with our proposed methods, we also provided the information about the location of lesions for the 3D U-Net and the Sketcher by applying the proposed weighted L1 loss. These

state-of-the-art methods were replicated to the maximum extent possible based on details provided in the paper, as their codes are not available.

Figure 3.4 shows the qualitative comparison and the true [^{11}C]PIB PET DVR parametric map. We can find that the 2-layer DNN failed to find the non-linear mapping between the multimodal MRI and the myelin content in PET. Especially, some anatomical or structural traces (that are not present in the ground truth) can still be found in the 2-layer-DNN predicted PET. This highlights that the relationship between myelin content and multimodal MRI data is complex, and only two layers are not powerful enough to encode-decode it. It is also shown that the 3D U-Net and the Sketcher (cGAN) generate blurry outputs with the primitive shape and basic information. On the other hand, after the refinement process by our Refiner, the output is more similar to the ground truth and the myelin content is better predicted. According to this, we can also conclude that the iterative training process can refine and improve the results.

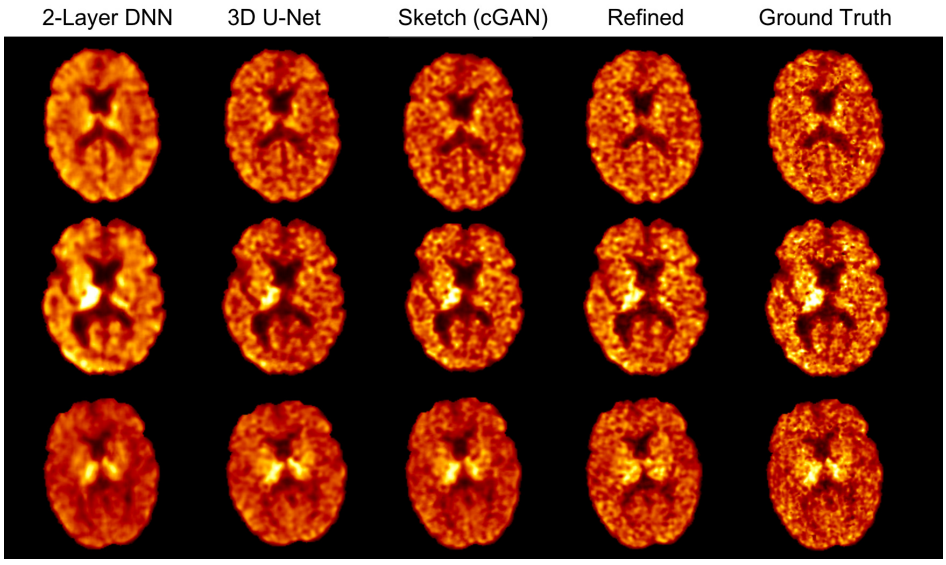


Fig. 3.4: Qualitative comparison of the results of our method (“Refined"), of a 2-layer DNN , of a 3D U-Net and of a single cGAN (corresponding to the Sketcher in our approach and denoted as “Sketch") to the ground truth.

We then performed a quantitative comparison in terms of global image quality (Table 4.1). Image quality is evaluated by mean square error (MSE) and peak signal-to-noise ratio (PSNR) defined as follows:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|(\boldsymbol{I}_{\mathbf{P}}^{i}, \hat{\boldsymbol{I}}_{\mathbf{P}}^{i})\|_2 \tag{3.8}$$

$$PSNR = 20 \cdot \log_{10}(MAX_{I_P}) - 10 \cdot \log_{10}(MSE) \tag{3.9}$$

where MAX_{I_P} is the maximum voxel value of the image.

Our method is shown to outperform all the other methods for both metrics. The difference with the 2-layer-DNN, the 3D U-Net with weighted L1 loss (for both MSE and PSNR), the 3D U-Net (for MSE) and the Sketcher with weighted L1 loss (for PSNR) are statistically significant ($p < 0.05$ by two-sided T-test). We can also find that the performance of the Sketcher is better than 3D U-Net. This can be caused by the use of adversarial training which can make the output image indistinguishable.

Tab. 3.1: Image quality metrics obtained with our method and the other methods. MSE: mean square error; PSNR: peak signal-to-noise ratio. Results are displayed as mean (standard deviation).

	MSE	PSNR
2-Layer DNN	0.0136 (0.0048)*	27.767 (1.214)*
3D U-Net	0.0107 (0.0041)*	29.297 (0.986)
3D U-Net+L1W	0.0113 (0.0043)*	28.606 (1.007)*
Sketcher	0.0094 (0.0038)	29.475 (0.981)
Sketcher+L1W	0.0103 (0.0042)	29.077 (0.995)*
Refiner (Proposed)	**0.0083 (0.0037)**	**30.044 (1.095)**

* indicates our method is significantly better with $p < 0.05$ by two-sided T-test

Then, we quantitatively compared the ability of the different methods to accurately synthesize myelin content in the three ROIs: 1) white matter (WM) in healthy controls (HC); 2) normal-appearing white matter (NAWM) in MS patients; 3) lesions in MS patients. The myelin content prediction discrepancy was defined as the mean absolute difference between the mean myelin content of the ground truth and that of the prediction PET across subjects and ROIs.

Results are shown in Table 3.2. Our method is more accurate than other methods on these three ROIs. Of note, the highest difference between our method and the others is in the MS lesions. This demonstrates that our neural networks indeed payed more attention to MS lesions during the image synthesis process, thanks to the specific loss of the Refiner network.

Furthermore, we also applied the proposed weighted L1 loss to both 3D U-Net and cGANs for comparison. We can find that in terms of global image quality measured by MSE and PSNR shown in Table 4.1, the cGAN and 3D U-Net using the weighted L1 loss performed respectively worse than the ones using the simple L1 loss function. However, the comparison of myelin

Tab. 3.2: Comparison of myelin content prediction discrepancy (defined as mean absolute difference between the ground truth and the predicted PET) in three defined ROIs between our method and other methods. WM in HC: white matter in healthy controls; NAWM: normal appearing white matter in patients. Results are displayed as mean (standard deviation).

	WM in HC	NAWM	MS Lesions
2-Layer DNN	0.059 (0.040)	0.041 (0.036)	0.131 (0.051)*
3D U-Net	0.053 (0.034)	0.039 (0.033)	0.035 (0.027)
3D U-Net+L1W	0.054 (0.034)	0.038 (0.031)	0.032 (0.029)
Sketcher	0.053 (0.041)	0.034 (0.022)	0.030 (0.017)
Sketcher+L1W	0.052 (0.037)	0.035 (0.027)	0.027 (0.022)
Refiner (Proposed)	**0.048 (0.026)**	**0.029 (0.021)**	**0.022 (0.015)**

* indicates our method is significantly better with $p < 0.05$ by two-sided T-test

prediction discrepancy in Table 3.2 suggests that using the weighted L1 loss will result in a better prediction in our regions of interest especially MS lesions. All of the above results demonstrate that the simple L1 loss can drive the network towards the global image generation. On the contrary, the weighted L1 loss specializes in the generation of a specific region.

3.3.3 Refinement Iteration Effect

We have demonstrated that the overall qualitative and quantitative results have been improved after our proposed refinement process. To compare the effect of different refinement iterations, we assess the performance with respect to the number of iterations (from 0 to 3). Note that the iteration 0 is our Sketcher and an additional Refiner is used for each new iteration (so 1 iteration corresponds to the proposed "Sketcher-Refiner method"). We studied the evolution of MSE (Fig. 3.5 (A)) and of the prediction discrepancy in 3 ROIs (Fig. 3.5 (B)). One can see a dramatic improvement when using the Refiner on top of the Sketcher (iteration 1). Iteration 2 also leads to an improvement, but it is much smaller. In the third iteration, the MSE and the prediction discrepancy in WM in HC worsen. Considering the trade-off between the marginally improved performance and the extra training time after first iteration, we suggest to use only one iteration.

3.3.4 Global Evaluation of Myelin Prediction

We compared the myelin content distribution of the ground truth to that of the predicted PET images in three ROIs by all the methods. From Fig. 3.6, we can see that the average PET value in the different regions can be predicted by all the methods except the 2-layer DNN whose prediction in MS lesions is

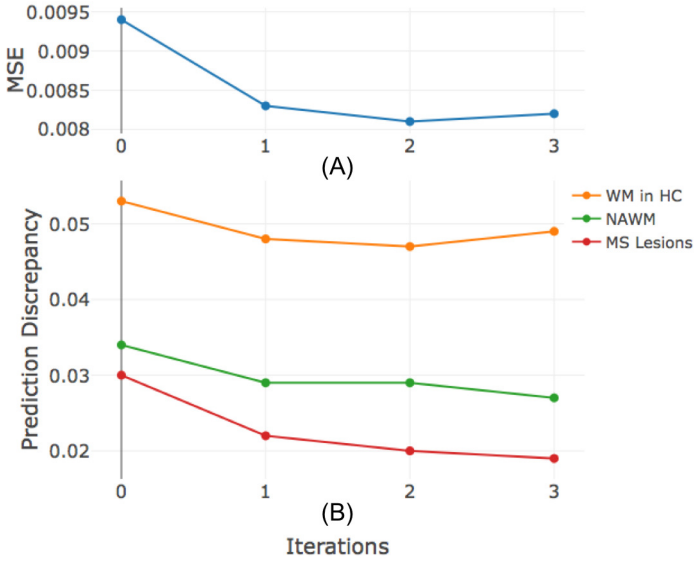


Fig. 3.5: Performance assessment with respect to different number of iterations. Note that the iteration 0 is our Sketcher and an additional Refiner is used for each new iteration.

inconsistent with the gold standard. Specifically, both with the gold standard and our synthetic data, there is no significant difference ($p = 0.88$ by two-sided T-test) between NAWM in patients and WM in HC, while a statistically significant reduction of myelin content in lesions compared to NAWM can be found ($p < 0.0001$ by two-sided T-test).

Further, we presented the Bland-Altman plots for WM/NAWM and MS lesions (Fig. 3.7) for all the methods at the individual level. It can be seen that our method (the Refiner) achieved the best results with 0.0091 and -0.06 as the mean bias for WM/NAWM and the lesions respectively. In particular, the proposed refinement process, passing from the Sketcher to the Refiner, presents a remarkable performance gain especially in the MS lesions. For the Sketcher, it is better than 3D U-Net in WM/NAWM but has similar performance in the lesions. By contrast, the 2-layer CNN achieved the worst performance.

3.3.5 Voxel-wise Evaluation of Myelin Prediction

We also evaluated the ability of our method to predict myelin content at the voxel-wise level in MS lesions. Within each MS lesion of each patient, each voxel was classified as demyelinated or non-demyelinated according to a procedure defined and validated in a previous clinical study [Bodini, 2016]. This method involves the determination of a threshold to separate demyeli-

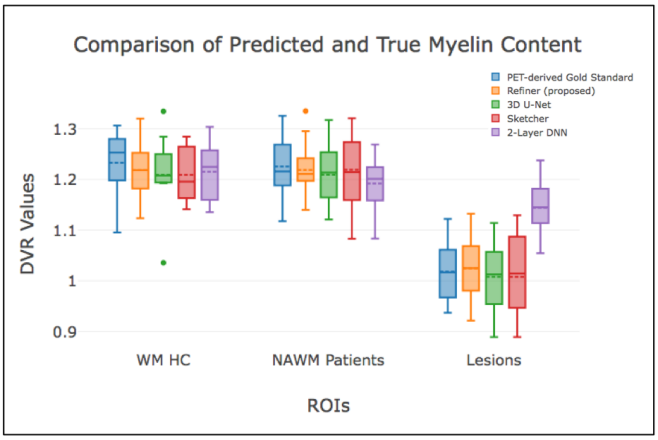


Fig. 3.6: Group level evaluation for all the methods. The box plots show the median (middle solid line), mean (middle dotted line) and min-max (below and above line) DVR for each ROI for PET-derived DVR parametric map used as gold standard (blue) and the prediction results from our method (yellow), 3D U-Net (green), Sketcher (red) and 2-Layer DNN (violet).

nated from non-demyelinated voxels. This threshold being determined at the group-level, the procedure involves a non-linear inter-subject registration onto MNI space performed using FNIRT algorithm in the FSL package [Jenkinson, 2012].

We first measured the percentage of demyelinated voxels over total lesion load of each patient for both the ground truth and the predicted PET as shown in Figure 3.8 (A). Our prediction results approximate the ground truth for most of the patients. We then compared, in each patient, the masks of demyelinated voxels classified from both the true and the predicted PET within MS lesions. The average DICE index between the demyelination maps derived from the ground truth and our predicted PET is 0.83 ± 0.12. This is a strong agreement, demonstrating the ability of our method to predict the demyelination in MS lesions at the voxel-wise level. Examples of demyelinated voxel masks are shown in Figure 3.8 (B).

3.3.6 Attention in Neural Networks

Our proposed *Visual Attention Saliency Map* is used to interpret the attention of neural networks for image prediction. In case of a single modality, the attention saliency map will have the same dimension as the input image. In case of the multimodal images, the size of the map will be 4D (3D+modality channel). We took the maximum value across the modality channels to derive the final attention saliency map.

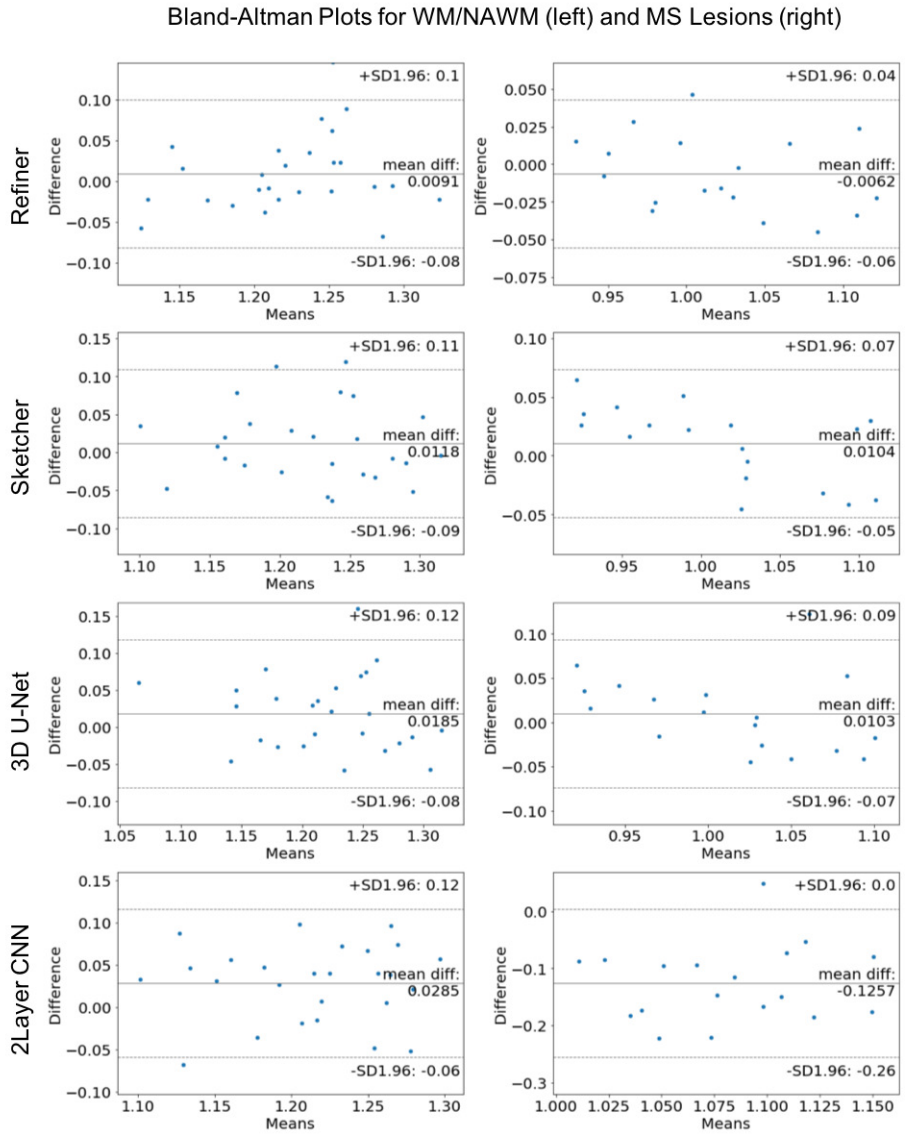


Fig. 3.7: Bland-Altman Plots for WM/NAWM (left) and MS lesions (right) at the individual level for all the methods.

Figure 3.9 displays the attention saliency maps derived from the generators. The maps allow displaying which regions are the most important for the prediction. We can observe that the neural networks using weighted L1 loss pay more attention to voxels located within MS lesions, which are the most important for demyelination quantification. On the other hand, one can see that a neural network using an unweighted L1 loss focuses more on the ventricle regions which have no myelin content and thus no interest for us. We can thus conclude that our designed loss function is able to effectively shift the attention of the neural networks towards the MS lesions.

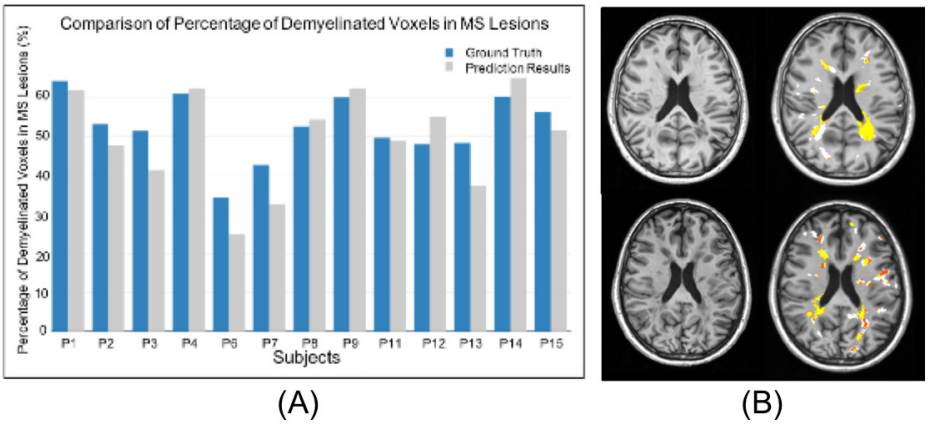


Fig. 3.8: (A) Percentage of demyelinated voxels in white matter MS lesions for each patient computed from the ground truth (blue) and from our method (grey). (B) Demyelinated voxels classified from the ground truth and our predicted PET within MS lesions in two example patients. Agreement between methods is marked in yellow (both true and predicted PET indicated demyelination) and white (both methods did not indicate demyelination). Disagreement is marked in red (demyelination only with the true PET) and orange (only with the predicted PET). The DICE coefficients in these two cases are 0.88 (1st row) and 0.72 (2nd row). The corresponding T1-w MR images are also shown on the left in each row. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

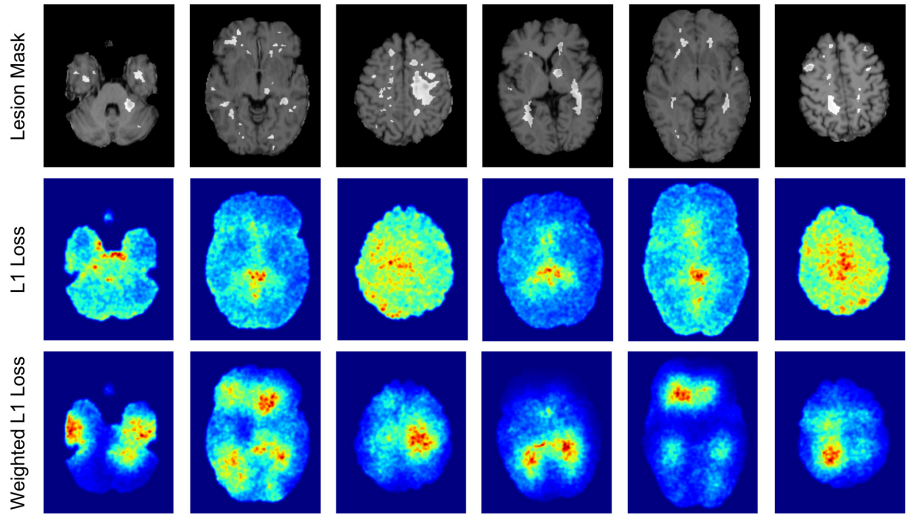


Fig. 3.9: The proposed visual attention saliency map. The white regions shown in first row are MS lesion masks. The second row shows some examples of the attention of neural networks when L1 loss is used as the traditional constraint in the loss function, without the specific weighting scheme that we proposed. The third row shows the corresponding attention of neural networks when our proposed weighted L1 loss is applied. It is clear that our designed loss function is able to effectively shift the attention of neural networks towards MS lesions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3.7 Contribution of Multimodal MRI Images

In this work, we chose to use MTR as well as three measures derived from DTI (FA, RD and AD) as our input images because, among MRI features, they are considered the most indicative of myelin content. Nevertheless, they likely contain redundant information. We thus compared the predictions using: 1) only MTR; 2) MTR+RD; 3) MTR+DTI.

Table 3.3 shows the corresponding image quality metrics (MSE and PSNR as defined in Eq. 4.2(a) and 4.6). It can be found that only using MTR leads to the worst results in terms of MSE and PSNR. Adding DTI RD, the results are slightly better. But these improvements are small. By contrast, when the other two DTI measures (FA and AD) are added, the performances are improved dramatically from 0.0094 to 0.0083 for MSE and from 29.524 to 30.044 for PSNR. This is consistent with the findings in Chartsias et al. [Chartsias, 2018] that adding an additional input modality resulted in a performance improvement and the best performance is achieved when all the input modalities are used.

Tab. 3.3: Image quality metrics for different combinations of MRI features. MTR: magnetization transfer ratio. RD: radial diffusivity. DTI: all three diffusion tensor imaging metrics. MSE: mean square error. PSNR: peak signal-to-noise ratio. Results are displayed as mean (standard deviation).

	MSE	PSNR
MTR	0.0094 (0.0043)	29.524 (1.671)
MTR+RD	0.0092 (0.0043)	29.581 (1.679)
MTR+DTI	0.0083 (0.0037)	30.044 (1.095)

Table 3.4 compares the prediction of myelin content for the different combinations of MRI features. It shows that the prediction discrepancy for all three ROIs decreased markedly when DTI RD is added. The main reason is that RD reflects the diffusion along the radial direction which increases with demyelination. Therefore, DTI RD can provide some extra information and contribute for myelin content prediction. On the other hand, adding other DTI metrics (FA and AD) only slightly improved the performances and this improvement was not significant ($p > 0.5$).

3.4 Discussion

In this work, we proposed a method to predict the PET-derived myelin content from multimodal MR images. Our approach called Sketcher-Refiner GANs, consists of two conditional GANs with specifically designed adversarial

Tab. 3.4: Comparison of myelin content prediction discrepancy (defined as MD) in three defined ROIs by using different combinations of MRI features. MTR: magnetization transfer ratio. RD: radial diffusivity. DTI: all three diffusion tensor imaging metrics. Results are displayed as mean (standard deviation).

	WM in HC	NAWM	MS Lesions
MTR	0.059 (0.040)	0.036 (0.021)	0.037 (0.029)
MTR+RD	0.050 (0.030)	0.031 (0.019)	0.025 (0.017)
MTR+DTI	0.048 (0.026)	0.029 (0.021)	0.022 (0.015)

loss functions. A visual attention saliency map is also proposed to interpret the attention of neural networks. The experimental results demonstrate its superior performance for PET image synthesis and myelin content prediction compared with the state-of-the-art methods.

The demyelination in lesional regions and myelin content in normal-appearing white matter can be well predicted by our method. At the global level, the distribution of the myelin content derived from the ground truth in three ROIs is very similar to that derived from our synthetic PET. Precisely, both with the ground truth and the synthetic PET, no difference can be found between NAWM in patients and WM in HC while a significant reduction is found in MS lesions comparing to NAWM in patients. Using a previously validated clinical research procedure, we showed that our prediction results approximate the percentage of demyelinated voxels derived from the ground truth individually. At the voxel-wise level, there was a high concordance between the demyelination maps derived from the ground truth and from the predicted PET. Even though these results will need to be confirmed in large populations, this demonstrates the potential of method for clinical management of patients with MS.

Furthermore, we compared our approach with the state-of-the-art methods through different aspects. First, by using MSE and PSNR as image quality metrics, we demonstrate a superior performance than the others. Second, we evaluate the myelin prediction at a global level in three relevant ROIs. Although there is no significant difference between the proposed method and almost all other methods, our approach is shown to outperform the others in all three ROIs especially with the highest performance in MS lesions. This demonstrates that our neural networks indeed made more efforts on MS lesions during the image synthesis process, thanks to the specific loss of the Refiner network.

The methods in Sikka et al. [Sikka, 2018] and Li et al. [Li, 2014] and Pan et al. [Pan, 2018] have been proposed to predict FDG-PET using MR

images for AD diagnosis. However, the myelin signal is much more subtle than the metabolic signal found in FDG PET. Moreover, its relationship to the anatomical information found in MRI is weaker. Thus, prediction of myelin content is a more difficult image synthesis problem. We addressed this difficult problem by a sketch-refinement process with two cGANs. The idea of using multiple GANs for image synthesis has already been explored in previous works, such as cascade GANs in Wang et al. [Wang, 2016]. Specifically, the cascade GANs designed in Wang et al. [Wang, 2016] is to address the problem that part of the data distribution might be ignored by the previous GANs. Therefore, the authors proposed to iteratively train multiple GANs until no further improvements are obtained. But unlike the traditional cascade GANs, our two GANs have different specifically designed cost functions (Eq. 3.4 and 3.5) for sketching anatomy and physiology information (Sketcher) and refining myelin content (Refiner). Indeed, the adaptive weights in the Refiner,s loss function force it to shift its attention on MS lesions where demyelination happens. By contrast, without such information, the Refiner would be driven towards generation of normal anatomy, which forms the majority of the image content but is of no interest for our problem. Furthermore, similar to the Dice loss proposed by Milletari et al. [Milletari, 2016], our proposed weighted L1 loss can also mitigate the effect of class imbalance by assigning weights to samples of different class to make the network not ignore the infrequent class.

In addition, in the works of Sikka et al. [Sikka, 2018] and Li et al. [Li, 2014] and Pan et al. [Pan, 2018], only a single MRI pulse sequence is used for prediction, for example Sikka et al. [Sikka, 2018] and Li et al. [Li, 2014] only use T1-w MRI as the input. However, we showed improved performances can be achieved by including more modalities as inputs. Using MTR+RD instead of only MTR can dramatically increase the myelin content prediction results especially in MS lesions. Adding AD and FA only marginally improved the results compared to MTR+RD. However, AD, FA and RD are all computed from a single DTI acquisition. Therefore, adding AD and FA does not require acquisition of more MRI sequences and does not increase the scanning time. We thus recommend using MTR+DTI since this leads to the best results, even though the improvement is small compared to MTR+RD. In fact, using multiple modalities for image synthesis and segmentation has also been studied in Chartsias et al. [Chartsias, 2018] and Havaei et al. [Havaei, 2016]. In their works, multichannel neural networks have been used. During the inference step, each modality is provided independently to convolutional neural networks. After encoding each modality into latent representations, multiple fusion strategies such as the mean-variance fusion [Havaei, 2016] or the max fusion [Chartsias, 2018], have been applied. However, the fusion

strategies maybe unsuitable for image synthesis task which takes multiple modalities as inputs. Some abnormal tissue regions which are important but do not form majority of the image may be ignored after the fusion step. Especially, the location and the shape of subtle lesional features can be highly variable between patients. Furthermore, the use of multichannel neural networks can lead to high computational cost. Because each input modality is treated independently by a neural network, the number of parameters will be dramatically increased. On the contrary, our multiple input modalities are arranged as channels and do not need the fusion strategy, which can alleviate the above problems. Besides, we use 3D operations for all the networks to better model the 3D spatial information and thus could alleviate the discontinuity problem across slices of 2D networks.

In order to interpret the attention of neural networks, we also proposed a visual attention saliency map. The advantage of our saliency map is that it can be generated by any kinds of neural networks and calculated by standard backpropagation. In our work, as it is only used for the visualization of the attention of neural networks, no backpropagation modification is applied. However, according to different applications, different strategies can be used to modify backpropagation, for example: 1) *Guided Backpropagation* [Springenberg, 2014] which only propagates positive gradients for positive activations; 2) *RELU Backpropagation* [Zeiler, 2014] which only propagates positive gradients. Moreover, class activation maps (CAM) [Zhou, 2016] and Grad-CAM [Selvaraju, 2017] are other ways to visualize and understand CNNs. Instead of using gradients with respect to output, these methods use a global average pooling layer and visualize the weighted combination of the feature maps at the penultimate (pre-softmax) layer to obtain class-discriminative visualizations.

There are also some limitations to our work. First, the proposed weighted L1 loss needs the masks of different ROIs so that the generator can pay more attention to the MS lesions. However, in practice, these masks are not always available. In particular, in this work, the MS lesions were manually segmented. It remains to be seen if automatic methods could be used for that process. This is left for future work. Second, in the preprocessing steps, we did the intra-subject registration onto [^{11}C]PIB PET image space which is a common step when using multiple modalities as inputs. However, the quality of the synthesized image can be influenced by the registration accuracy because of image noise and different selections of parameters in the registration step. In the future work, a spatial transformation layer could be integrated in the neural networks in order to avoid the influence from registration or alignment of different modalities. The use of combined

MR-PET systems can also avoid this problem. Third, only a small, single-center, dataset is used in our work to evaluate our proposed method. Further experiments on larger, multi-center, datasets, will thus be needed to assess the generalizability of the approach more in depth. Such further validation is crucial before translation to the clinic can be considered. Last, in our work the input MR data was restricted to MTR and DTI derived metrics. These inputs were selected based on their potential to provide at least indirect information about myelin content (based on the literature and discussion with MS experts). However, it could be that other MR sequences or features (such as for example T1/T2 ratio) provide complementary information. This would need to be assessed in future work.

3.5 Conclusion

We proposed Sketcher-Refiner GANs with specifically designed adversarial loss functions to predict the PET-derived myelin content from multimodal MRI. The prediction problem is solved by a sketch-refinement process in which the Sketcher generates the preliminary anatomy and physiology information and the Refiner refines and generates images reflecting the tissue myelin content in the human brain. Both qualitative and quantitative results demonstrate that our method outperforms the state-of-the-art approaches. Moreover, our method allowed to accurately predict myelin content prediction at both global and voxel-wise levels. The evaluation results show that the demyelination in MS lesions, and myelin content in both patients' NAWM and controls' WM can be well predicted by our method.

4

Predicting PET-derived Myelin Content from Multisequence MRI for Individual Longitudinal Analysis in Multiple Sclerosis

Contents

This chapter corresponds to the following publications:

- [Wei, 2020b] *Predicting PET-derived Myelin Content from Multisequence MRI for Individual Longitudinal Analysis in Multiple Sclerosis*

W.Wei, E.Poirion, B.Bodini, M.Tonietto S.Durrleman,
O.Colliot, B.Stankoff, N.Ayache
Submitted to NeuroImage

- [Wei, 2020a] *Conditional Flexible SAGAN for Predicting PET-derived Myelin Content in Multiple Sclerosis from Multisequence MRI*
W.Wei, E.Poirion, B.Bodini, M.Tonietto S.Durrleman,
O.Colliot, B.Stankoff, N.Ayache
Submitted to MICCAI2020

4.1 Introduction

Multiple sclerosis (MS) is a demyelinating and inflammatory disease of the central nervous system (CNS). The principal hallmark of MS is the presence of focal demyelinating lesions, consisting of a loss of myelin surrounding the axon, leading to the degeneration of the axon in an extent that is variable between patients and between plaques. These lesions appear abundantly in the white matter (WM). However, the demyelination process can be repaired by the generation of a new sheath of myelin around the axon, a process termed remyelination. These pathological features of multiple sclerosis might be highly dynamic over time but largely heterogeneous across patients [Bodini, 2016; Patrikios, 2006]. Therefore, a reliable measure of the tissue myelin changes is essential to push forward our understanding of mechanisms involved in the pathology of MS, and to monitor individual patients in the clinical setting or in the context of clinical trials focused on repair therapies.

Over the years, magnetic resonance imaging (MRI) has been increasingly used in the diagnosis of MS and it is currently the most useful paraclinical tool to assess this diagnosis. Although conventional MRI pulse sequences, such as T2-weighted or fluid attenuated inversion recovery (FLAIR) sequences, are sensitive techniques to detect WM lesions of MS, they lack the specificity for the underlying pathological process, and especially have limitations in differentiating between inflammation, axonal loss, demyelination and remyelination. Semi-quantitative MRI techniques, such as magnetization transfer ratio map (MTR), diffusion weighted imaging or T2 relaxometry, also have potential for the measurement of myelin content, but their ability to do so is only partially characterized, and furthermore MTR is affected not only by myelin, but also by water content and inflammation [Petiet, 2019].

Positron emission tomography (PET) is an alternative imaging modal, which can target specific tissue substrates and detect tissue changes at the cellular and molecular level. In a recent longitudinal study [Bodini, 2016], PET imaging with amyloid radiotracer [^{11}C]PIB could detect a decreased tracer uptake in WM lesions compared to normal-appearing WM (NAWM), which paralleled myelin content. Furthermore, longitudinal data presented by the authors support the ability of [^{11}C]PIB to capture demyelination and remyelination in lesions over time. Note that PET imaging is invasive due to the injection of a radioactive tracer. In addition, it cannot be used in all clinical centers as it is an expensive imaging technique and not available in the majority of medical centers in the world. Therefore, it would be of high interest to predict the individual PET-derived myelin dynamic changes from multisequence MRI.

4.1.1 Related work

To the best of our knowledge, this is, to date, the first work to predict PET-derived demyelination and remyelination for individual longitudinal analysis in MS. On the contrary, there has been amounts of works focusing on image modal prediction and synthesis. We present an overview of these methods by two categories: (1)unimodal synthesis whose input and output are the *same* modal, such as MRI-to-MRI, and (2) cross-modal synthesis whose input and output are *different* modals, such as MRI-to-CT.

Unimodal Synthesis

Unimodal synthesis has shown wide applications, e.g., image denoising and artifact reduction [Wang, 2019; Chen, 2013; Xu, 2012; Zhang, 2017; Liu, 2012; Tian, 2011], image super-resolution [Bahrami, 2016a; Kaplan, 2019; Hagiwara, 2019], and inter-modal conversion [Roy, 2010; Ye, 2013; Sevetlidis, 2016; Chartsias, 2018; Wei, 2019a]. Extensive efforts have been dedicated to decrease the influence from noise and artifacts in low-dose CT. Dictionary learning based approaches were developed for low dose X-ray CT reconstruction [Chen, 2013; Xu, 2012; Zhang, 2017]. The works in [Liu, 2012; Tian, 2011] proposed iterative algorithms by minimizing the total variation to reduce noise and artifacts in CT images. Nevertheless, these algorithms still lost some anatomical details and suffered from remaining artifacts. Searching-based methods have also been used for unimodal synthesis, such as 3T-to-7T MRI super-resolution [Bahrami, 2016a], MRI-to-MRI conversion [Ye, 2013; Roy, 2010]. However, the result heavily depends on the similarity between the source image and the images in the database. This

may make the method fail in the presence of abnormal tissue anatomy since the images in the atlas do not have the same pathological features as the patient to predict. Moreover, above methods often need a high computational cost which would limit their use in practical applications.

To alleviate above issues, deep learning methods have recently shown remarkable ability to automatically learn the underlying features with better descriptive power. Moreover, the end-to-end whole-image-based models are less computationally expensive compared with above small-patch-based methods. For instance, a deep encoder-decoder network was used to generate FLAIR from T1-w MRI which works on the whole image instead of the small image patches [Sevetlidis, 2016]. There are also several works trying to do unimodal synthesis for image super-resolution using deep learning models, such as estimating full-dose PET image from low-dose image [Kaplan, 2019], improving the synthetic FLAIR image quality [Hagiwara, 2019]. Moreover, the works in [Chartsias, 2018; Wei, 2019a] demonstrated that using multisequence MRI can improve the MRI-to-MRI synthesis performance. It is thus suggested using multisequence as inputs when it is possible.

Cross-modal Synthesis

Given a subject's modal (source) x_{source}, the goal is to accurately synthesize another modal (target) of the same subject y_{target}. Over the past decade, lots of methods have been proposed. Two main approaches are atlas-based [Burgos, 2014; Lee, 2017] and learning-based methods. Atlas-based methods need an atlas dataset including the co-registered pairs of the source and target modals defined as X and Y respectively. All the images $x_i \in X$ are first registered to the given source modal x_{source}. The geometric transformations are then applied to each y_i in the atlas database. The target modal y_{target} can thus be synthesized through the fusion of the aligned target modals Y in the atlas dataset. Although these methods [Burgos, 2014; Lee, 2017] demonstrated a good ability for overall anatomy synthesis and can be used for MRI-to-CT synthesis for attenuation correction, they may be unable to accurately predict subtle lesional features, whose location can be highly variable between patients. In addition, these methods highly rely on the registration accuracy and the synthesized image quality may also depend on the prior knowledge for tuning large amounts of parameters in registration step.

An alternative is learning-based methods which aim to find the nonlinear relationship between the source x_{source} and the corresponding target modal

y_{target}. This nonlinear mapping is learned on training data through nonlinear regression models such as random forest [Huynh, 2016] and neural networks [Nie, 2016; Leynes, 2018; Liu, 2018]. A structured random forest and auto-context model has been proposed to synthesize CT image from MRI for attenuation correction [Huynh, 2016]. In this approach, the input MRI is first partitioned into patches and fed into structured random forest to generate corresponding CT patch. An auto-context model is then used to refine the prediction. However, the input images should be presented by the crafted features which will influence the image synthesis result. For the same purpose, 3D convolutional neural networks has been used for attenuation correction through pseudo-CT images [Nie, 2016; Leynes, 2018; Liu, 2018]. In recent years, generative adversarial networks (GANs) have achieved promising results in nature image synthesis [Zhang, 2018a; Isola, 2016; Zhu, 2017] because of the ability to learn the messy and complicated representations of data. As the nonlinear function is more complex in multi-modal synthesis, many works investigated the possibility to use GANs to do so. The work in [Choi, 2018] used GANs to synthesize the MR images by using the PET for the quantification of cortical amyloid load. The authors in [Bi, 2017] used multi-channel GANs to synthesize PET images from CT images. There are also several studies working on other modal synthesis, such as retinal images [Costa, 2018; Zhao, 2018], ultrasound images [Hu, 2017] and endoscopy images [Mahmood, 2018].

Regarding MRI-to-PET synthesis, a U-Net shaped 3D convolutional neural network (CNN) [Sikka, 2018] and a two-layer CNN [Li, 2014] have been proposed to predict FDG PET from T1-w MRI for AD classification. Similarity, the works in [Pan, 2018; Wang, 2018c] used 3D GANs to synthesize FDG PET images for AD diagnosis. Different from these MRI-to-PET works which were devoted to the prediction of the radiotracer FDG, our goal is to predict myelin content as defined by [^{11}C]PIB PET. Predicting myelin content (as defined by [^{11}C]PIB PET) is a more difficult task because the signal is more subtle and with weaker relationship to anatomical information that could be found in MR images. In our recent works [Wei, 2019b; Wei, 2018b], we proposed Sketcher-Refiner GANs to predict the myelin content from multisequence MRI. As this method is based on the conditional GANs, the long-range dependencies between MS lesions are not considered by the networks. Moreover, in our previous works [Wei, 2019b; Wei, 2018b], we only predicted the static demyelination process without the prediction of dynamic demyelination-remyelination process.

4.1.2 Contributions

In this work, by using multisequence MR images, we propose a method through adversarial training to predict the PET-derived dynamic myelin *changes* for MS individual *longitudinal analysis*. The novelties and contributions of our paper are as follows:

1. In order to model the relationships between spatially separated lesional regions during the 3D image synthesis process, we propose a conditional flexible self-attention GAN (CF-SAGAN) to capture these long-range dependencies.
2. Medical images are often high-dimensional which makes the model easily reach the memory constraints when calculating the attention maps used in self-attention mechanism. To address this problem, our CF-SAGAN is improved and specifically adjusted for high-dimensional medical images.
3. Demyelination and remyelination are quantified within MS lesions. An adaptive attention regularization for MS lesions is designed so that the neural networks can pay more attention on the MS lesions during the image generation process.
4. Compared with the state-of-the-art methods, our method is shown to outperform these methods qualitatively and quantitatively.
5. Importantly, our method for the prediction of myelin content changes in patients with MS shows similar clinical correlations to the PET-derived gold standard indicating the potential for clinical management of patients with MS. To the best of our knowledge, this is, to date, the first work to do so.

A preliminary version of this work was sent to the MICCAI2020 conference [Wei, 2020a]. The present paper extends the previous work by: (1) proposing an improved adaptive attention regularization which can not only lead a remarkable local image quality on MS lesions, but can also take consideration of other regions generating a competitive global image quality; (2) studying the contribution of different regularization terms; (3) defining three specific indices to evaluate dynamic myelin changes in more detail; (4) calculating the clinical correlation with the clinical score EDSS; (5) describing the methodology with more details; (6) providing more details of background and related works.

4.2 Method

4.2.1 Overview

GANs have proven very successful in generating images. The basic generative adversarial networks (GANs) [Goodfellow, 2014] consist of two networks: a generator G and a discriminator D. In GANs, the generator G tries to learn the mapping from a latent variable z (typically random noise) to an image in target domain, and the discriminator D aims to distinguish between the true image and the fake image which is generated by the generator G. These two components are learned and compete with each other, with G aiming to generate images as realistic as possible, and D aiming to tell apart generated and real images. In order to constrain the outputs of the generator G for image synthesis, conditional GAN (cGAN) [Mirza, 2014] was proposed in which the generator and the discriminator both receive a conditional input vector.

MS lesions vary in size, location and intensity, but they may display the same basic features of pathology, such as demyelination in the WM. Under this scenario, to better predict the myelin changes reflected by the [^{11}C]PIB PET distribution volume ratio (DVR) parametric map from multimodal MRI, it is necessary to model the relationships between spatially separated lesional regions. The classic GAN represents both generator G and discriminator D as convolutional networks. During the image synthesis process, the convolutional operations can only process a local neighborhood information because of the limited local receptive field. For example, in a convolution operation, it is hard to model the correlation between top-left and bottom-right positions. These long-range dependencies across different regions in the image can be modeled by stacking several convolutional layers to result in large receptive fields. But doing so, the earlier layers can be almost negligible because of the vanishing gradient, and the optimization algorithms may have trouble on huge parameter space and thus make the GANs training more unstable [Kodali, 2017]. Moreover, the medical images are usually high-dimensional. Increasing the depth of the networks and the size of the convolution kernels can dramatically increase the computational cost. To address this situation, inspired by the work in [Zhang, 2018a], we introduce a flexible self-attention layer to capture the long-term dependencies during the 3D image synthesis process. A sketch-refinement process is also applied to improve image quality. We describe the details in the following.

4.2.2 Conditional Flexible Self-Attention GAN (CF-SAGAN)

In this section, we present a conditional flexible self-attention GAN which combines both ideas of conditional GANs [Mirza, 2014] and attention mechanism [Zhang, 2018a], and is adaptively designed for high-dimensional medical images.

A recent self-attention mechanism presented in [Vaswani, 2017] is modeled by a *Transformer* model with three major concepts: *Key, Value and Query*. In our work, the same as [Zhang, 2018a], the C convolutional feature maps $x \in \mathbb{R}^{C \times L}$ is branched out into three copies, corresponding to the three components: *Key* $f(x)$, *Value* $h(x)$ and *Query* $g(x)$ with $f(x) = W_f x, g(x) = W_g x$, and $h(x) = W_h x$. The above weight matrices $W_f \in \mathbb{R}^{\overline{C} \times C}, W_g \in \mathbb{R}^{\overline{C} \times C}, W_h \in \mathbb{R}^{C \times C}$ are part of the model parameters which are implemented by $1 \times 1 \times 1$ convolutions. Each key, value and query is reduced from the high dimensional features to the dimension of $\overline{C} = C/8$ in our implementation for computational efficiency.

Then we transpose the key $f(x)$ and matrix-multiply it by the query $g(x)$ and take the Softmax on all the rows to calculate the attention map:

$$\beta_{j,i} = Softmax(f(x_i)^T g(x_j)) \tag{4.1}$$

Where $\beta_{j,i}$ indicates the extent to which the model attends to the i_{th} location when synthesizing the j_{th} region. After integrating the attention map into the self-attention layer, the output of the attention layer $o \in \mathbb{R}^{C \times L}$ is defined as follows:

$$o_j = \sum_{i=1}^{N} \beta_{j,i} h(x_i) \tag{4.2}$$

Theoretically the self-attention mechanism is able to capture the long-range dependencies across different image regions which are not covered by the convolution kernels. However, as the medical images are generally high-dimensional, the storage of the attention map can easily reach the memory limits. To adapt to our needs, we propose a flexible self-attention layer as shown in Fig. 4.1. The pooling layers are inserted to decrease the size of the input feature maps and the output of the attention layer is reshaped to meet

the size of the input feature maps. By doing so, our flexible self-attention layer reduces the size of the attention map by the cubic of the pool size p. Furthermore, the output of the attention layer o is multiplied by a scale parameter γ and added back to the original input feature maps:

$$y = \gamma o + x \tag{4.3}$$

While the scaling parameter γ is increased gradually from 0 during the training, the network is configured to first rely on the cues in the local regions and then gradually learn to assign more weight to the regions that are further away.

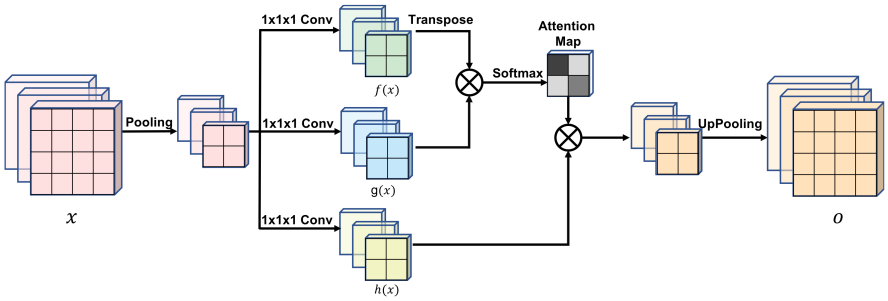


Fig. 4.1: The proposed flexible self-attention layer. The input feature maps x and the output o have the same size. A Pooling and an UpPooling operations have been added to meet the high-dimensional medical image usage.

As the input and the output have the same size, our flexible self-attention layer can be inserted between any two convolutional layers. In our CF-SAGAN, the proposed flexible self-attention layer has been used to train both the generator G and the discriminator D, optimizing the traditional cross-entropy loss function.

4.2.3 Adaptive Attention Regularization for MS Lesions

The underlying image synthesis process for the prediction of PET-derived myelin changes follows that in [Wei, 2018b] which proposed a sketch-refinement process. We extended this approach by using two CF-SAGANs to improve the prediction performance. Our first CF-SAGAN (Sketcher) aims to sketch the anatomy and physiology information from multimodal MR images. The L1 loss is used to regularize the model globally and encourage less blurring. The other CF-SAGAN used as Refiner takes both the output of the sketcher and the input multimodal MR images to refine and generate the final image reflecting the tissue myelin content. Almost all of the state-of-the-art methods, such as [Roy, 2010; Burgos, 2014; Ye, 2013; Xiang, 2018],

aimed to synthesize the whole image with no local attentions. As the myelin changes are mainly quantified within MS lesions, we introduced an adaptive attention regularization to make the Refiner focus more on MS lesions during the image generation process.

Dividing the whole image into three regions of interest (ROIs) from n subjects: lesions R_{Les}, normal appearing white matter (NAWM) R_{NAWM} defined as the white matter outside visible lesions, and "other" R_{other}, with the number of voxels in each region $m_{\text{Les}}, m_{\text{NAWM}}, m_{\text{other}}$ respectively, the proposed adaptive attention regularization is described as follows:

$$\mathcal{L}_{L1}(G_{\text{R}}) = \frac{1}{2nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j|I^{i,j} - \hat{I}^{i,j}|,$$
$$\omega_j = \begin{cases} 1 - \frac{m_{\text{Les}}}{m}, j \in R_{\text{Les}} \\ 1 - \frac{m_{\text{NAWM}}}{m}, j \in R_{\text{NAWM}} \\ 1 - \frac{m_{\text{other}}}{m}, j \in R_{\text{other}} \end{cases} \tag{4.4}$$

where I and $\hat{I}$ are the true image and the prediction output from the Refiner, and i, j is the index of a subject and a voxel respectively.

While prior works [Wei, 2019b; Wei, 2018b; Wei, 2020a] used a weighted L1 loss to change model's attention, our proposed adaptive attention regularization, also regarded as normalized weighted L1 loss, can alleviate the influence from MS lesion size variety across patients and penalize proportionally for different regions.

4.2.4 Clinical Longitudinal Dataset

Our clinical dataset consists of a longitudinal collection of 18 MS patients (12 women, mean age 31.4 years, sd 5.6) which are clinically assessed and scored using the expanded disability status scale (EDSS) [Kurtzke, 1983] and the multiple sclerosis severity score (MSSS) [Roxburgh, 2005], and 10 age-matched healthy volunteers (8 women, mean age 29.4, sd 6.3). All the patients first underwent MRI and PET scan at baseline (t_0). Then they all repeated the whole protocol after either 1-2 months or 3-4 months (t_1) to explore the best time interval for dynamic remyelination and demyelination quantification. The healthy volunteers only underwent one scan. All participants signed written informed consent to participate in the study, which was approved by the local ethics committee of the Pitié-Salpêtrière hospital. The clinical and demographic information is detailed

in [Bodini, 2016]. At t_1, because of missing MR images, 3 patients were excluded from the t_1 dataset. Finally, for each participant, we used the following data:

a) **PET IMAGES:** PET examinations were performed on a high-resolution research tomograph (HRRT; CPS Innovations, Knoxville, TN) which achieves an intraslice spatial resolution of 2.5mm, with 25-cm axial and 31.2-cm transaxial fields of view. The 90-minute emission scan was initiated with a 1-minute intravenous bolus injection of [^{11}C]PIB (mean = 358 $\pm$ 34 MBq). The Logan graphical reference method [Logan, 1996] was applied at the voxel level on PET scans in native space to obtain [^{11}C]PIB PET DVR parametric map ($1.22 \times 1.22 \times 1.22\text{mm}^3$).

b) **MR IMAGES:** MR images were collected using a 3 Tesla Siemens TRIO 32-channel TIM system including Magnetisation Transfer Ratio map (MTR) ($1 \times 1 \times 1.1\text{mm}^3$), and three measures derived from Diffusion Tensor Imaging (DTI): Fractional Anisotropy (FA), Radial Diffusivity (RD) and Axial Diffusivity (AD) ($2 \times 2 \times 2\text{mm}^3$). The three ROIs (lesions, NAWM and “other") used in Eq. 4.4 were delineated as follows. WM lesions of MS patients were manually contoured by an expert rater on T2-w scans with reference to FLAIR images. The corresponding lesion masks were generated and aligned to the individual T1-w scan using FLIRT algorithm in the FSL package [Jenkinson, 2012]. After performing a “lesion-filling" procedure in patients only, T1-w scans were segmented using FreeSurfer [Fischl, 2012] to obtain a WM mask. The NAWM is then defined as the WM outside visible lesions on T2-w scans.

4.2.5 Indices of Myelin Content Change

Following a validated procedure [Bodini, 2016], voxels characterized as demyelinated were identified as those whose DVR value fall below one standard deviation of the mean DVR value of all the voxels in healthy controls that were localized at the same distance from the CSF. This step returned individual maps of demyelinated voxels inside WM lesions in patients, which were generated for each of the 2 time-points (baseline-t_0, follow-up-t_1). In each patient, individual maps of remyelinating and demyelinating voxels inside WM lesions were computed based on the trajectory of each voxel.

From the demyelinated map at both time points (t_0 and t_1), the percentage of demyelinated voxels over the total lesion load measured at baseline (t_0)

was calculated for each patient. To further measure the myelin content changes, we defined the following indices:

- **Global index of myelin content change:** It is the difference between the derived percentage at t_1 and the corresponding percentage at t_0. This index reflects the subject-specific prevalence of either myelin loss or myelin repair over the follow-up interval.
- **Index of dynamic demyelination:** It is defined as the proportion of normally myelinated voxels at baseline t_0 which were then classified as demyelinated voxels at t_1. This index reflects the ongoing myelin loss.
- **Index of dynamic remyelination:** It is defined as the proportion of lesional voxels classified as demyelinated at baseline t_0 which then arrived at a normal myelin level at t_1. This index reflects ongoing myelin repair.

4.2.6 Network Architectures

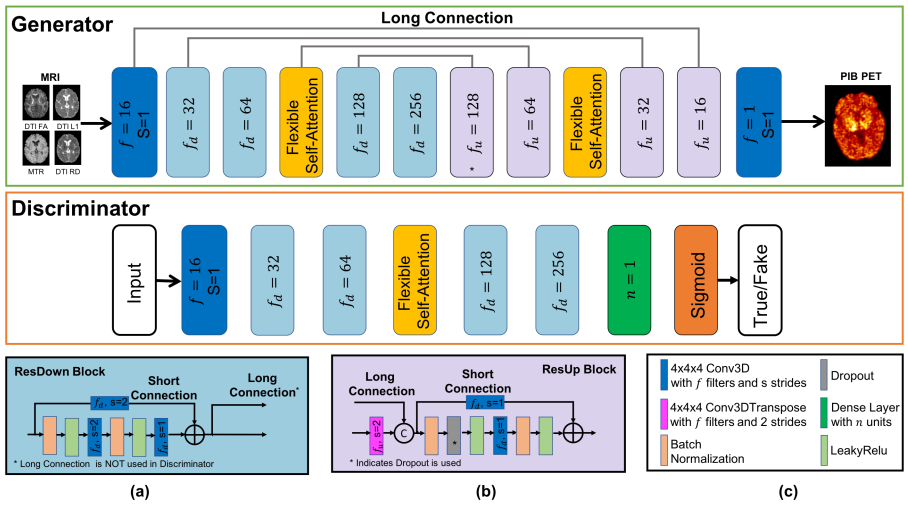


Fig. 4.2: The overall framework of the proposed CF-SAGAN. (a-c) Illustration of the different units and layers used in the Generator and the Discriminator. The values f_d, f_u, f indicate the number of feature maps in each unit and s is the stride number. The parameter n refers to the units number in the Dense layer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In our work, the two CF-SAGAN used as Sketcher and Refiner are both three-dimensional to fully capture the spatial information between slices. Both of them share the same architectures for their generators and discriminators. Instead of using the whole image, we applied large overlapped patches ($64 \times 64 \times 64$) to train the models, which can save computational resource

and also provide sufficient training samples. We detailed the architectures of the generator and the discriminator as follows.

- **Generator architecture:** The generator is critical to the quality of the generated images. We used an encoder-decoder as a backbone architecture. The encoder is used to extract the spatial features from the input image, and the decoder aims to construct the final output from the encoded features. To help the network to retrieve the lost spatial information during the down-sampling operations and boost the information flow between the end and the start of the network, we introduced the long skip connection to the network architecture which can also be regarded as U-Net [Ronneberger, 2015]. In addition, we added the short connection into the generator as shown in *ResDown Block* and *ResUp Block* in Fig. 4.2 (a) and (b) respectively. This improvement, also called residual connection [He, 2016], is known for its ability to mitigate the problem of vanishing gradient by allowing this alternate shortcut path for gradient to flow through, and enhance the feature exchanges across layers. Taking the advantages of our proposed flexible self-attention unit and long/short skip connections, our generator is illustrated in Fig. 4.2. The encoder includes a sequence of a convolutional layer, four ResDown blocks and a flexible self-attention unit between. The number of feature maps starts from 16 and doubles after each convolutional layer or ResDown block. The decoder inversely involves four ResUp blocks with a flexible self-attention in between and a convolutional layer to output the final image. All of the different levels in the encoder are transferred to the corresponding levels in the decoder by using long connections.
- **Discriminator architecture:** The discriminator is used as a classifier to distinguish if the input is in class “True" or “Fake". We adopted the *3D PatchDiscriminator* from the work of [Wei, 2019b]. Each downsampling operation is realized by a ResDown block. In addition, the flexible self-attention unit is added to reinforce the discriminator to capture complicated long-range constraints on the global image structure. The general architecture is demonstrated in Fig. 4.2. Note that the long connections in ResDown block are not used in the discriminators.

4.3 Experiments and Evaluation

4.3.1 Implementation and Training Details

We used the clinical longitudinal dataset described in the Section 4.2.4 for the prediction of dynamic myelin changes. The model is trained, validated and tested on baseline (t_0) dataset by using 3-fold cross validation. For each cross-validation split, the dataset is divided into a training set with 2/3 subjects (including 3 subjects for validation), and a testing set consisting of 1/3 subjects. The fixed model is then directly applied to the images of these 1/3 subjects from t_1 *dataset*. This can also be considered as a way to evaluate model generalization. For both time points (t_0 and t_1), the preprocessing steps mainly consist of brain extraction [Smith, 2002], intensity inhomogeneity correction [Tustison, 2010] and affine intra-subject registration of MR data onto [^{11}C]PIB PET DVR image space using FLIRT algorithm in the FSL package [Jenkinson, 2012]. Finally, we cropped the images into 50% overlapped $64 \times 64 \times 64$ patches with a resolution of $1.22 \times 1.22 \times 1.22\text{mm}^3$. The details of acquisition parameters and PET data quantification are described in [Bodini, 2016] and [Veronese, 2015].

Our model was implemented with Tensorflow [Martin, 2015]. The convolution kernel size is $3 \times 3 \times 3$ and the rate for dropout layer is 50%. To train the model, the whole data underwent the zero mean and unit variance normalization. Data augmentation is also applied including three rotations (90, 180 and 270 degrees), scaling up by 1.25 and scaling down by 0.75. During the training process, we first iteratively trained the Sketcher for 370 epochs by fixing our Refiner. Then we trained the Refiner from scratch for another 370 epochs by fixing our Sketcher. The optimization was performed with the ADAM solver with 10^{-4}, 5×10^{-5} as initial learning rates for the Sketcher and the Refiner respectively.

It is known that training a GAN model can become unstable and even produce an early model collapse. In our work, we used several techniques to improve the stability of training our CF-SAGAN models. First, we used strided convolutions for downsampling as shown in Fig. 4.2. By doing so, the network can learn its own spatial downsampling. Similarly, the upsampling operation is done by strided deconvolutional layer. Second, as suggested in [Salimans, 2016], the labels used by the discriminator for true/fake samples are smoothed to reduce the vulnerability of neural networks to adversarial examples. Instead of setting hard labels (0 and 1), the label was set as a random number between 0 and 0.1 for 0 labels, and a random value between 0.9 and 1.0 to represent 1 labels. In this way, the task for the discriminator becomes more challenging and it can match the difficulty

of the generator task, so that the adversarial training becomes balanced. For the same reason, the third technique is imbalanced learning rate. In our work, the discriminator was updated two times per generator update step during the training process. Last, LeakyReLU is used to allow a stable training of CF-SAGANs with 0.2 as slope coefficient.

4.3.2 Evaluation of Global Image Quality

Our method is compared with several state-of-the-art methods including a 2-layer DNN [Li, 2014], a 3D U-Net [Sikka, 2018], a cGAN [Bi, 2017] and a Sketcher-Refiner framework by using two cGANs (denoted as Refiner cGAN) [Wei, 2019b]. All state-of-the-art methods were kept the same architecture and parameter settings as described in the work of [Wei, 2019b]. The image quality is evaluated by taking into account both the qualitative differences through human perception as well as quantitative aspect measured by mean square error (MSE) and peak signal-to-noise ratio (PSNR) which are defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{N} \|(I^i, \hat{I}^i)\|_2 \tag{4.5}$$

$$PSNR = 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE) \tag{4.6}$$

where MAX_I is the maximum voxel value of the image.

The quantitative evaluation is summarized in Table 4.1. The proposed Refiner CF-SAGAN ranks the best among all the methods in both metrics. Comparing the proposed Refiner CF-SAGAN and the 2-layer-DNN, the Refiner CF-SAGAN outperformed by 42.65% in terms of MSE ($p < 0.05$) and 9.07% in PSNR ($p < 0.05$). Our Refiner CF-SAGAN also achieved better results than 3D U-Net, in all metrics, with improvements of 27.10% in MSE ($p < 0.05$) and 3.37% in PSNR ($p < 0.05$). It can also be found that the CF-SAGAN integrating our proposed flexible self-attention unit demonstrates a superior performance than a simple cGAN. This improvement consistently indicates the ability of the proposed flexible self-attention layer to learn long-range spatial dependencies and thus to generate high quality images. Based on the CF-SAGAN, the image quality is further improved when the sketch-refinement process is applied to output the final image (denoted as Refiner CF-SAGAN in Table 4.1).

Tab. 4.1: Image quality metrics obtained with our method and the other methods. MSE: mean square error; PSNR: peak signal-to-noise ratio. Results are displayed as mean (standard deviation).

	MSE	PSNR
2-Layer DNN	0.0136 (0.0048)*	27.767 (1.214)*
3D U-Net	0.0107 (0.0041)*	29.297 (0.986)*
cGAN	0.0094 (0.0038)	29.475 (0.981)
Refiner cGAN	0.0083 (0.0037)	30.044 (1.095)
CF-SAGAN	0.0085 (0.0042)	29.942 (1.065)
Refiner CF-SAGAN (Proposed)	**0.0078 (0.0038)**	**30.285 (0.0993)**

* indicates our method is significantly better with $p < 0.05$ by two-sided T-test

The similar conclusion can be drawn from the qualitative comparison shown in Fig. 4.3(B) with the corresponding ground truth (the right-most column). As observed, the image quality of the 2-Layer-DNN is obviously the worst. In particular, some anatomical or structural traces which are not present in the ground truth can still be found in the 2-layer-DNN predicted image. It can also be clearly seen that the GAN-based methods generally outperform the 2-layer-DNN and 3D U-Net with more shape information. These enhancements can be attributed to the use of adversarial loss to generate a more realistic output. This also highlights that the relationship between myelin content and multimodal MRI data is complex, and a single simple network is not powerful enough to model it. It is worth noting that the performance is boosted after the refinement process, i.e. from cGan to Refiner cGAN and from CF-SAGAN to Refiner CF-SAGAN. Therefore, taking the advantages of adversarial training, the sketch-refinement process and the proposed flexible self-attention, our Refiner CF-SAGAN can generate objectively sharper, less blurry images which are closest to the ground truth.

4.3.3 Evaluation of Adaptive Attention Regularization

To study the contribution of the proposed adaptive attention regularization term, we conduct comparison experiments for different combinations of each of three models (3D U-Net, cGAN and CF-SAGAN) using one of regularization terms including L1, weighted L1 (denoted as WL1) used in [Wei, 2019b] and the proposed adaptive attention regularization, also regarded as normalized weighted L1, denoted as NWL1. As demyelination and remyelination are quantified within MS lesions, the performance is evaluated by myelin content prediction discrepancy within MS lesions (defined as mean absolute difference between the ground truth and the predicted PET) for local image quality and also by MSE for global image quality.

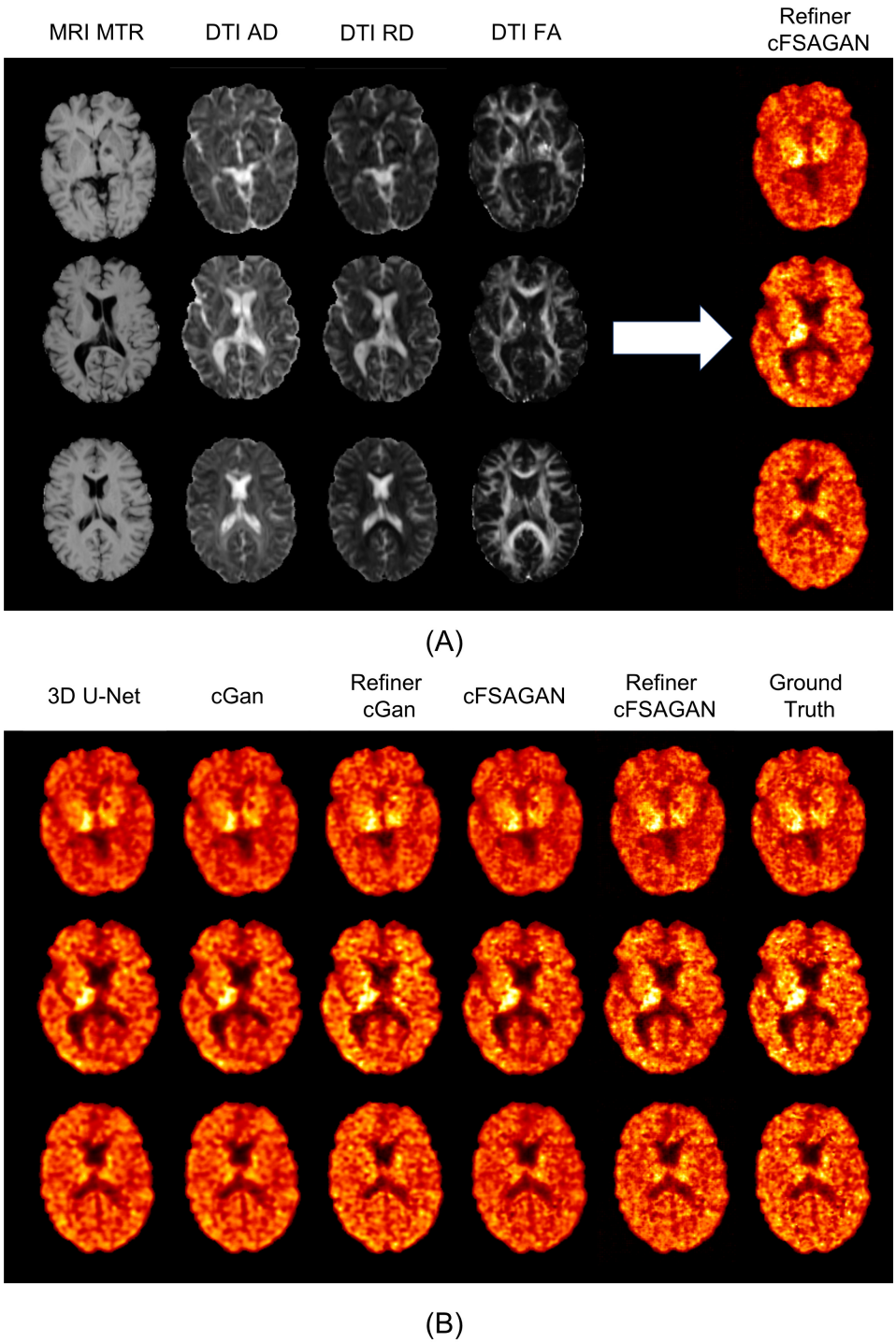


Fig. 4.3: (A) Illustration of multisequence MRI as inputs and our prediction results. (B) Qualitative comparison of the results of our proposed framework (CF-SAGAN), of the refined version of our proposed method ("Refiner CF-SAGAN"), and of the other state-of-the-art methods.

The comparison results are provided in Table 4.2 with Table 4.2(a) showing the performance on global image quality measured by MSE and Table 4.2(b) showing the performance on local image quality measured by myelin content

prediction discrepancy within MS lesions. We can clearly see that the models using L1 loss presented a performance superior to the other combinations on MSE, but they achieved the worst results on myelin content prediction inside MS lesions. Compared with the models using WL1 which got 0.0113, 0.0103 and 0.0097 as MSE for 3D U-Net, cGAN and CF-SAGAN respectively, the methods using NWL1 are shown to outperform by a large margin on MSE with 0.0111, 0.0098 and 0.0089 for 3D U-Net, cGAN and CF-SAGAN respectively. The NWL1-based methods have even shown slightly better performance on myelin content prediction discrepancy inside MS lesions (two of three models using NWL1 got best results and cGAN with NWL1 obtained 0.028 which is very close the best value 0.027). In our work, the two CF-SAGAN are respectively used as Sketcher for the global generation of anatomy and physiology information, and as Refiner for the local refinement for MS lesions. The applications of L1 for the Sketcher and NWL1 for the Refiner exactly take full advantage of their different characteristics. It is also found in experiments that the NWL1 allows faster convergence and stable training.

Both the qualitative and quantitative experimental results demonstrate that our method can synthesize high quality image. In the next section, our proposed method will be evaluated clinically for the prediction of myelin content for MS individual longitudinal analysis.

Tab. 4.2: Comparison of different regularization terms

(a) Comparison of MSE obtained from different methods using different regularization terms. Results are displayed as mean (standard deviation).

	L1	WL1	NWL1
3D U-Net	0.0107 (0.0041)	0.0113 (0.0043)	0.0111 (0.0047)
cGAN	0.0094 (0.0038)	0.0103 (0.0042)	0.0098 (0.0044)
CF-SAGAN	0.0085 (0.0042)	0.0097 (0.0039)	0.0089(0.0041)

(b) Comparison of MS lesion myelin content prediction discrepancy (defined as mean absolute difference between the ground truth and the predicted PET) obtained from different methods using different regularization terms. Results are displayed as mean (standard deviation).

	L1	WL1	NWL1
3D U-Net	0.035 (0.027)	0.032 (0.029)	0.031 (0.032)
cGAN	0.030 (0.017)	0.027 (0.022)	0.028 (0.019)
CF-SAGAN	0.027 (0.020)	0.024 (0.016)	0.022 (0.017)

4.3.4 Evaluation of Static Demyelination Prediction

As mentioned in Section 4.2.5, within MS lesions each voxel can be classified as demyelinated or non-demyelinated. Applying the previous clinical procedure described in [Bodini, 2016] for both time points (t_0 and t_1), we

measured the percentage of demyelinated voxels over total lesion load of each patient for both the ground truth and the predicted PET. It can be seen from the Fig. 4.4 (A) that our prediction results are nearly the same as the ground truth for all of the patients for both time points showing a very good accuracy of our method. Furthermore, the DICE index is used to measure, for each patient, the agreement between the masks of demyelinated voxels classified from both the true and the predicted PET within MS lesions for both time points. As a result, we got 0.91±0.06 and 0.89±0.08 for t_0 and t_1 respectively. Comparing with the prior work [Wei, 2019b] which got 0.83±0.12 for t_0 time point, our method demonstrated a better ability to predict the demyelination in MS lesions at the voxel-wise level. Examples of demyelinated voxel masks are shown in Fig. 4.4 (B).

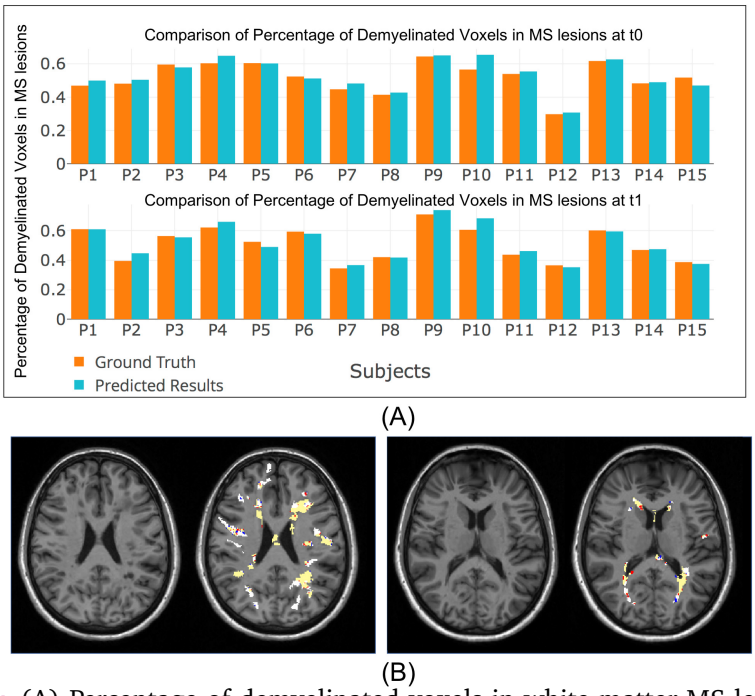


Fig. 4.4: (A) Percentage of demyelinated voxels in white matter MS lesions for each patient computed from the ground truth (orange) and from our method (blue) for t_0 (top) and t_1 (bottom). (B) Demyelinated voxels classified from the ground truth and our predicted PET within MS lesions in two example patients. Agreement between methods is marked in yellow (both true and predicted PET indicated demyelination) and white (both methods did not indicate demyelination). Disagreement is marked in red (demyelination only with the predicted PET) and blue (only with the true PET). The DICE coefficients in these two cases are 0.89 (left) and 0.85 (right). The corresponding T1-w MR images are also shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3.5 Evaluation of Dynamic Demyelination and Remyelination Prediction

To quantify the myelin changes in MS lesions, we calculated the three indices proposed in Section 4.2.5 for both the ground truth and our method. We first calculated the *global index of myelin content change* which is the difference between two time points. This index reflects a balance between a predominant demyelination and remyelination. After sorting by the myelin changes derived from the ground truth, the comparison result is shown in Fig. 4.5 (A). It shows that although there is a high between-patient variability ranging from -0.13 to +0.14, the PET-derived overall myelin changes can be well reproduced by our predicted image. Specifically, most of the patients can be well classified into the right category (demyelination or remyelination). The only false classification is due to the tiny change which is only 0.006 according to the ground truth (see P8 in Fig. 4.5 (A)).

For each patient, we then calculated the *index of dynamic demyelination* and *index of dynamic remyelination* for both the ground truth and our prediction results. These two indices which are defined in Section 4.2.5 are able to reflect the ongoing myelin loss and repair. From the comparison results shown in 4.5 (B)(C), it can be found that our prediction results approximate the ground truth for most of the patients. In particular, comparing with the ground truth, we achieved almost the same level for the majority of patients. We furthermore compared, in each patient, the masks of dynamic demyelination and remyelination voxels classified from both the true and the predicted PET within MS lesions. The average DICE indices are 0.71 ± 0.11 and 0.69 ± 0.12 for the dynamic demyelination and remyelination maps derived from the ground truth and our predicted [^{11}C]PIB PET. Considering the slight myelin changes between two time points and the high variability between patients, this agreement demonstrates a promising ability of our method to predict the dynamic demyelination and remyelination in MS lesions. Three examples are illustrated in Fig. 4.6, showing a small prediction discrepancy from our method and a high between-patient variability.

4.3.6 Clinical Correlation

It is concluded in the recent clinical research work [Bodini, 2016] that there is no significant association between the index of dynamic demyelination and EDSS ($p = 0.72$), but the index of remyelination is a significant explanatory factor for EDSS with -0.67 as beta-coefficient ($p = 0.006$). We thus

calculate the correlation between our synthesized-PET-derived indices of myelin content change and the clinical score EDSS. In our work, because all the mandatory MRI data were not available for the initial patients list, the clinical correlation have been recomputed for both the synthesized-PET-derived and true-PET-derived indices based on a subset of subject to check the consistency of our method. As no significant change was found in EDSS during follow-up, the EDSS measured at baseline was used for clinical correlation which is calculated using multiple linear regression with EDSS as response variable and age, gender and T2-w lesion load as additional covariates.

As shown in Table 4.3(a), no significant effect was found on EDSS for the true-PET-derived index of dynamic demyelination ($p = 0.578$), whereas a significant association was detected between the true-PET-derived index of dynamic remyelination and EDSS ($p = 0.021$; beta-coefficient= -0.703) showing patients with lower disability presenting a higher proportions of remyelinating voxels. A similar finding can be observed between EDSS and the synthesized-PET-derived indices (Table 4.3(b)) with no significant correlation for the index of dynamic demyelination ($p = 0.676$) and a significant inverse correlation for the index of dynamic remyelination ($p = 0.012$; beta-coefficient= -0.734). This observation demonstrates the consistency and the ability of our method for the prediction of PET-derived myelin content in MS.

4.4 Discussion

In this work, we proposed a method to predict PET-derived demyelination and remyelination for individual longitudinal analysis in MS from multimodal MR images. The method is based on our proposed conditional flexible self-attention GAN (CF-SAGAN) which is specifically adjusted for high-dimensional medical images and able to capture the relationships between the spatially separated lesional regions during the image synthesis process. Our result is further improved dramatically by following the sketch-refinement process with the second CF-SAGAN as the Refiner and using our proposed adaptive attention regularization to make the network focus on the MS lesions.

Our method demonstrates superior performance qualitatively and quantitatively compared to the state-of-the-art approaches including a 2-layer DNN [Li, 2014], a 3D U-Net [Sikka, 2018], a cGAN [Bi, 2017] and a Sketcher-

Tab. 4.3: Clinical Correlation

(a) Correlation between **true-PET-derived** index of dynamic demyelination/remyelination and EDSS with age, gender and T2 lesion load as covariates.

Dependent Variable: EDSS Score	Coefficient	Standard Error	t	p	Beta-coefficient
Index of demyelination	0.090	0.157	0.58	0.578	0.216
Age	-0.019	0.090	-0.21	0.840	-0.065
Gender	-0.893	1.126	-0.79	0.446	-0.269
T2 Lesion Load	1.01e-5	1.75e-5	0.58	0.577	0.207
Index of remyelination	-0.233	0.085	-2.75	**0.021***	-0.703
Age	-0.047	0.067	-0.70	0.498	-0.164
Gender	-0.062	0.869	-0.07	0.944	-0.019
T2 Lesion Load	5.22e-6	1.16e-5	0.45	0.661	0.107

*Tests significant at significance level $p = 0.05$.

(b) Correlation between **synthesized-PET-derived** index of dynamic demyelination/remyelination and EDSS with age, gender and T2 lesion load as covariates.

Dependent Variable: EDSS Score	Coefficient	Standard Error	t	p	Beta-coefficient
Index of demyelination	0.071	0.165	0.43	0.676	0.173
Age	-0.014	0.091	-0.15	0.882	-0.048
Gender	-0.980	1.120	-0.87	0.402	-0.295
T2 Lesion Load	9.99e-6	1.92e-5	0.52	0.614	0.205
Index of remyelination	-0.258	0.084	-3.07	**0.012***	-0.734
Age	-0.050	0.064	-0.78	0.452	-0.017
Gender	-0.063	0.816	-0.08	0.940	-0.019
T2 Lesion Load	7.47e-6	1.1e-5	0.68	0.513	0.153

*Tests significant at significance level $p = 0.05$.

Refiner framework by using two cGANs (denoted as Refiner cGAN) [Wei, 2019b]. From human perception, our method can generate objectively more realistic images which are the most similar to the ground truth. By using MSE and PSNR as image quality metrics, our method is shown to outperform the other approaches on all metrics.

In addition of the image quality, it is also important to do clinical evaluation with the previous clinical procedure described in [Bodini, 2016]. From the aspect of static demyelination prediction, our prediction results approximate the true-PET-derived percentage of demyelinated voxels individually for both time points. In particular, a better agreement, between demyelination maps derived from the true and the predicted PET, has been achieved than our prior work [Wei, 2019b]. Regarding dynamic demyelination and remyelination prediction, the three indices of myelin content change derived from our predicted PET images are very similar to those derived from the ground truth. Moreover, the same clinical correlation between the index of dynamic remyelination and EDSS can be found from both the true and our predicted PET images. All these findings indicate the potential of our method for clinical management of patients with MS, although it still needs to be validated and confirmed in large populations.

Several studies have already explored the possibility to synthesize FDG-PET from MR images [Sikka, 2018; Li, 2014; Pan, 2018; Wang, 2018c]. But none of them considered the local attentions and the output images were synthesized directly without any local focus. Our previous work in [Wei, 2019b] applied a weighted L1 loss (WL1) to make the network pay more attention on MS lesions where demyelination and remyelination are quantified. On this basis, we extended and proposed an adaptive attention regularization term in this work (denoted as NWL1 in Section 4.3.3). It can be found in Table 4.2 that L1 loss can yield the best global image quality but with the worst local image quality. Both the WL1 and the NWL1 are good at local image synthesis, but the NWL1 can achieve superior performance on global image quality than WL1. The main reason is that these three regularization terms play different roles during the image generation process. The L1 loss drives the network to output images which are only globally close to the ground truth, but without any specific attention on some regions of interest. The WL1 is designed to transfer more attention on the pre-defined ROIs. However, the larger the weight is assigned to these regions, the less attention on the other regions will be payed during the generation process. Especially, in some patients, when MS lesions are extremely small, the network would only focus on these tiny regions and cannot output an anatomically and structurally plausible image. Nevertheless, our NWL1 can not only make the network pay more attention on these specifically pre-defined ROIs leading to a remarkable local image prediction, but can also find a balance and take consideration of other regions generating a competitive global image quality.

In fact, the tracer [^{11}C]PIB was initially used for β-amyloid plaques in Alzheimer's disease (AD). But, myelin signal quantified by [^{11}C]PIB PET is more subtle than amyloid plaques and with weaker relationship to the anatomical information found in MR images making this synthesis problem more difficult. Multiple GANs can be used to improve synthesis quality as proved in several works [Wang, 2016; Nie, 2018; Wei, 2018b]. Inspired by this idea and given the ability of our CF-SAGAN to capture the relationships between the spatially separated lesional regions, we used two CF-SAGANs to improve the prediction performance. Unlike the traditional cascade GANs used in [Wang, 2016; Nie, 2018], our two CF-SAGANs named as Sketcher and Refiner act as different roles for sketching anatomy and physiology information (Sketcher) and refining myelin content (Refiner) due to the use of NWL1 mentioned above. In practice, the number of GANs can be increased depending on different tasks and the gained performance after each iteration. We found that the performance will not be improved after two iterations by using three CF-SAGANs and the performance is nearly the same between two and three CF-SAGANs. The number of CF-SAGANs is thus fixed to two to save computational resource. The same conclusion is also conducted in [Nie, 2018; Wei, 2018b].

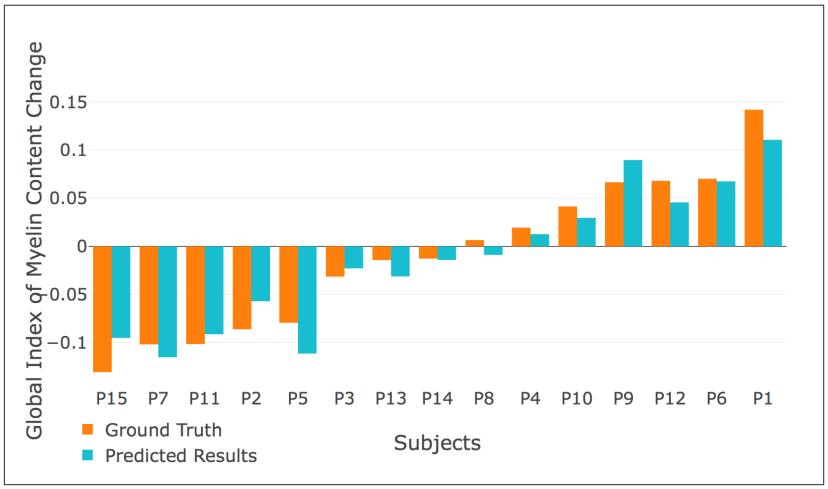
Since the recent idea of self-attention mechanism has been first introduced in [Vaswani, 2017] for machine translation, lots of methods have been developed and improved on this basis for nature language processing [Shen, 2018; Ambartsoumian, 2018] and nature image/video problems in computer vision [Wang, 2018b; Fu, 2019; Tang, 2019; Zhang, 2018a]. However, unlike their low-dimensional datasets, medical images are often high-dimensional which makes the model easily reach the memory constraints when calculating the attention maps used in self-attention mechanism. We addressed this problem by a flexible self-attention layer. The key idea is to insert Pooling layers to decrease the size of the input feature maps and then use UpPooling layers to resize the image to match the shape of the input. The pooling operation reduces the size of the attention map by the cubic of the pooling size p, making it possible to perform attention on high-dimensional data. In our experiments, we used MaxPooing as pooling operations. The other pooling operations can also be explored, such as AveragePooling, GlobalMaxPooling, etc.

The proposed method might be further improved by considering several limitations of our work. First, like the weighted L1 loss, the needs of the masks of different ROIs still remains for our proposed NWL1. In practice, these masks cannot be always available. In the future work, instead of using manually predefined attention, a self-learned attention could be helpful to

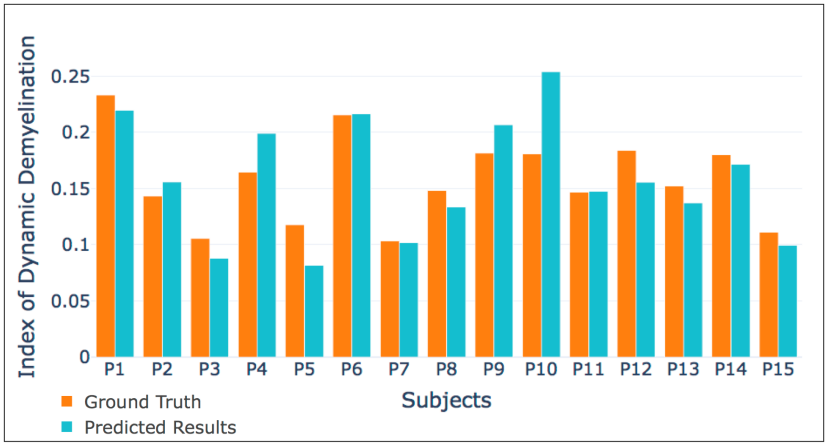
solve this problem. Second, multisequence MR images are used as inputs to provide as much information as possible to the network. However, the subjects with incomplete MR images cannot be used and have to be excluded from the dataset. This is a great loss especially for small medical image datasets. It would be helpful for future work to discover the way to deal with multisequence images independently, so that every incomplete MR series can be used for training. Furthermore, the intra-subject registration used in our work as a preprocessing step for multisequence MR images may also influence the synthesized image quality because of image noise and different selections of parameters in the registration step. The use of combined MR-PET systems can also avoid this preprocessing step. Last, our method is only evaluated on a small, single-center dataset. Further experiments on larger, multi-center, datasets, will thus be needed to assess the generalizability of the approach more in depth. Such further validation is crucial before translation to the clinic can be considered.

4.5 Conclusion

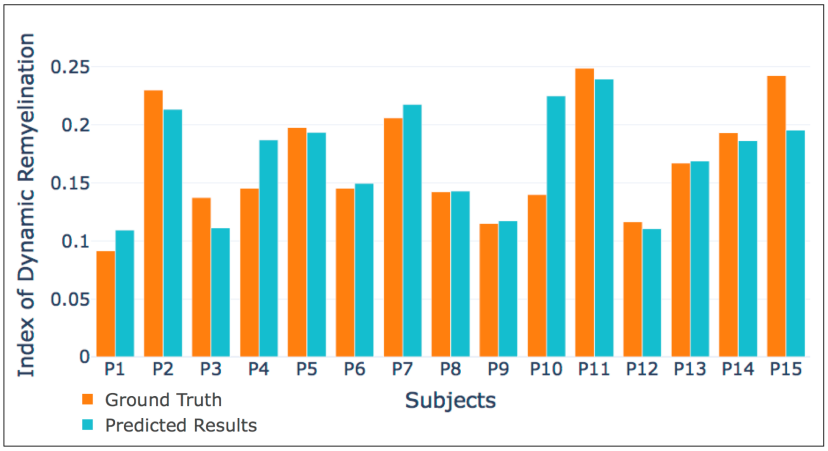
In this paper, we proposed a method to predict PET-derived demyelination and remyelination for individual longitudinal analysis in MS. The method is based on our proposed conditional flexible self-attention GAN (CF-SAGAN) which is specifically adjusted for 3D medical images and able to capture the relationships between the spatially separated lesional regions during the 3D image synthesis process. We also introduced an adaptive attention regularization which can not only lead a remarkable local image quality on MS lesions, but can also take consideration of other regions generating a competitive global image quality. Jointly applying the sketch-refinement process, our approach is shown to outperform the state-of-the-art methods qualitatively and quantitatively. Importantly, the clinical evaluations of our method for the prediction of myelin content from multisequence for MS individual longitudinal analysis show similar results to the PET-derived gold standard.



(A)



(B)



(C)

Fig. 4.5: The patient-level comparison for three myelin change indices computed from the ground truth (orange) and our method (blue). (A) Global index of myelin content change values for each patient. Patients with positive values indicate a predominant demyelination process. Patients with negative values means a predominant process of remyelinating. (B) Index of dynamic demyelination for each patient. (C) Index of dynamic remyelination for each patient.

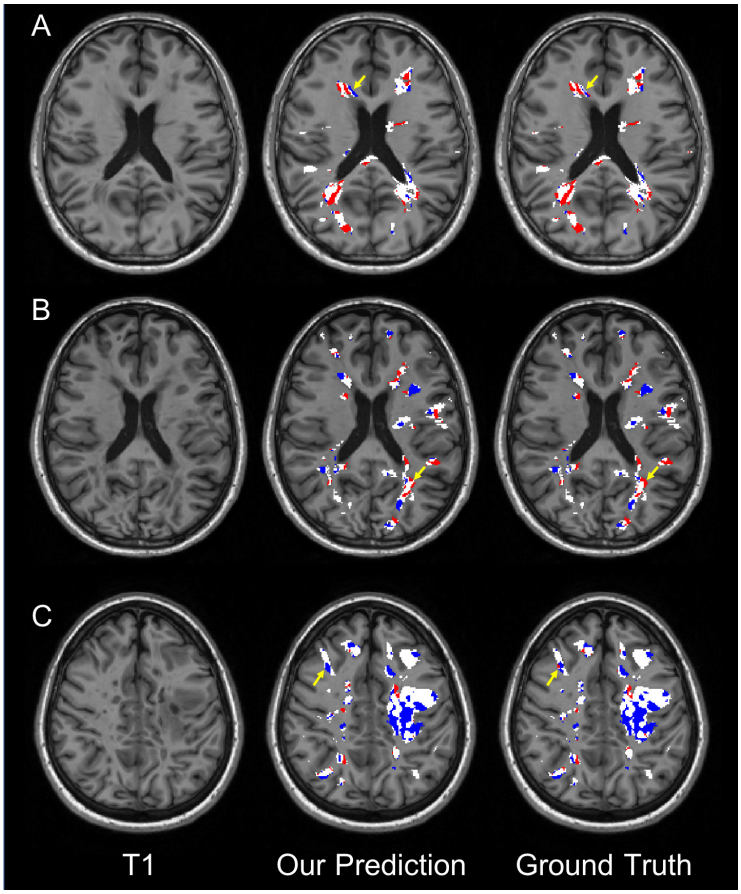


Fig. 4.6: Three examples of lesional myelin content changes showing demyelinating (in red) and remyelinating (in blue) voxels derived from the true longitudinal [^{11}C]PIB PET (right) and our predictions (middle), localized inside white matter (WM) lesions (in white), overlaid onto the corresponding T1-w MR image (left). These three patients respectively show a clear prevalence of demyelination over remyelination (Patient A), an active demyelination together with moderate remyelination in all visible WM lesions (Patient B) and an extensive process of remyelination (Patient C). The yellow arrows indicate some prediction discrepancies. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5 Conclusion and Perspectives

Contents

In this thesis, we proposed several deep learning methods to learn and predict brain alterations in MS from multimodal neuroimaging data. We summarize the main contributions in the following section and discuss some perspectives for future research work.

5.1 Main Contributions

5.1.1 Predicting FLAIR MR Image from Multisequence MRI

In chapter 2, we focused on predicting the missing FLAIR MRI pulse sequence on which WM lesions characteristic of MS can be better seen. To address this concern, we introduced 3D fully convolutional neural networks for FLAIR prediction from other multiple MRI pulse sequences, and a sequence-specific saliency map for investigating each pulse sequence contribution. Although the architecture of our model is simple, the nonlinear relationship between the source images and FLAIR can be well captured by our network. Both the qualitative and quantitative results have shown its competitive performance

for FLAIR prediction. In addition, unlike the 2D CNNs, we extended the model to 3D to better keep the spatial information between slices. Moreover, the FLAIR pulse sequence generated from our model has a better contrast for MS lesions which is important for the detection of MS lesions. Furthermore, we evaluated the contribution of each input pulse sequence to the prediction result so that the pulse sequences which contribute very little can be removed to save computational resource.

5.1.2 Predicting PET-derived Demyelination from Multisequence MRI

In chapter 3, we proposed a method to predict PET-derived demyelination from multisequence MRI. To the best of our knowledge, it is the first work doing this. In order to better learn the complex relationship between myelin content and multisequence MRI data, we propose Sketcher-Refiner generative adversarial networks consisting of two conditional GANs thereby decomposing the problem into two steps: 1) sketching anatomy and physiology information and 2) refining and generating images reflecting the tissue myelin content in the human brain. As the demyelinated voxels are classified within the Ms lesions, we thus designed an adaptive loss to force the network to pay more attention on MS lesions instead of the other regions during the prediction process. Besides, in order to interpret the neural networks, a visual attention saliency map has also been proposed. The experimental results demonstrate its superior performance for PET image synthesis and myelin content prediction compared with the state-of-the-art methods. Importantly, our prediction results show similar results to the PET-derived gold standard at both global and voxel-wise levels. Last but not least, we compared different combinations of MRI pulse sequences and features to evaluate which is the optimal input.

5.1.3 Predicting PET-derived Dynamic Myelin Changes from Multisequence MRI

In chapter 4, we extended the approach described in chapter 3 by introducing a conditional flexible self-attention GAN (cFSAGAN) to model the relationships between spatially separated lesional regions during the 3D image synthesis process. In particular, the proposed cFSAGAN is improved and specifically adjusted for high-dimensional medical images which can make the model reach the memory constraints when calculating the attention maps. In addition, an adaptive attention regularization is proposed for MS

lesions where demyelination and remyelination are quantified. Compared with the state-of-the-art methods and our previous method, this improved method is shown to be able to synthesize [^{11}C]PIB PET images and outperform the other methods qualitatively and quantitatively. More important, clinical evaluation of our method for **individual longitudinal analysis** in MS show similar results to the PET-derived gold standard.

5.2 Publications

- [Wei, 2020b] *Predicting PET-derived Myelin Content from Multisequence MRI for Individual Longitudinal Analysis in Multiple Sclerosis*
  W.Wei, E.Poirion, B.Bodini, M.Tonietto, S.Durrleman, O.Colliot, B.Stankoff, N.Ayache
  Submitted to NeuroImage in March 2020
- [Wei, 2020a] *Conditional Flexible SAGAN for Predicting PET-derived Myelin Content in Multiple Sclerosis from Multisequence MRI*
  W.Wei, E.Poirion, B.Bodini, M.Tonietto, S.Durrleman, O.Colliot, B.Stankoff, N.Ayache
  Submitted to MICCAI2020 in March 2020
- [Wei, 2019b] *Predicting PET-derived Demyelination from Multimodal MRI using Sketcher-Refiner Adversarial Training for Multiple Sclerosis*
  W.Wei, E.Poirion, B.Bodini, S.Durrleman, N.Ayache, B.Stankoff, O.Colliot
  *Medical Image Analysis (**MedIA**), Volume 58, p.101546, December 2019,*
- [Wei, 2019a] *Fluid-attenuated Inversion Recovery MRI Synthesis from Multisequence MRI using Three-dimensional Fully Convolutional Networks for Multiple Sclerosis*
  W.Wei, E.Poirion, B.Bodini, S.Durrleman, O.Colliot, B.Stankoff, N.Ayache
  *Journal of Medical Imaging (**JMI**), 6(01):27, 014005, February 2019*
- [Wei, 2018b] *Learning Myelin Content in Multiple Sclerosis from Multimodal MRI through Adversarial Training (**Oral**)*
  W.Wei, E.Poirion, B.Bodini, S.Durrleman, N.Ayache, B.Stankoff, O.Colliot
  *21st International Conference On Medical Image Computing and Computer Assisted Intervention (**MICCAI 2018**), LNCS, vol 11072. Springer, Cham*
- [Wei, 2018a] *FLAIR MR Image Synthesis by Using 3D Fully Convolutional Networks for Multiple Sclerosis*

W.Wei, E.Poirion, B.Bodini, S.Durrleman, O.Colliot, B.Stankoff, N.Ayache
ISMRM-ESMRMB 2018 *- Joint Annual Meeting, #2785, Paris, France*

5.3 Perspectives

5.3.1 Deep Learning for Medical Imaging Synthesis

In clinical settings, it is very common that certain modalities are useful and expected but infeasible to acquire. In this thesis the proposed deep learning based methods contribute to the field of medical image synthesis and the synthesized MRI/PET images are used for for MS disease. In practice, these kinds of deep learning methods can also be applied for other clinical usages with synthesized data, such as attenuation correction with pseudo-CT images [Nie, 2016; Leynes, 2018; Liu, 2018], AD diagnosis with synthesized PET images [Pan, 2018], cortical amyloid load quantification with synthesized MR images [Choi, 2018], etc. Although we should declare that the synthesized images cannot be used for direct clinical diagnosis, these kinds of data can be of great benefit for indirect clinical usage without any addition cost.

In this thesis, we studied the possibility to combine patient-specific multi-modal neuroimaging data for image synthesis. However, complex diseases are caused by a combination of genetic and environmental factors. In reality, a patient's data is thus inherently complex and may contain multi-site, multi-time, multi-trail, multi-type clinical data and even multi-omics data (such as gene and protein expression data). Deep learning have demonstrated the capability to learn the complex patterns and relationships from data and we have already shown in our work that the more informative data we use, the better result we can achieve. Under this observation, it would be a promising direction to explore how to make good use of these informative but complex data.

5.3.2 Synthesized Data for Deep Learning

In fact, the relationship between synthesized data and deep learning is a virtuous circle and the synthesized data can also contribute to the deep learning methods. In recent years, with increasing breakthroughs in deep learning algorithms and power of computational resources (such as GPU, TPU), the deep learning models become bigger, deeper and more complex

with a huge number of parameters. The amount and quality of training data are thus dominant influencers on deep learning models' performance. However, collecting and labeling medical images from real cases by human experts is very expensive and tedious. Synthesized images are hence great substitutes in case of data shortage. Moreover, medical imaging datasets are often unbalanced as pathologic findings are generally rare. With a great success of image synthesis, abundant abnormal images containing known pathologic characteristics can be generated to provide a big set of information which might not be available from the real image datasets [Shin, 2018; Bailo, 2019; Gupta, 2019]. Therefore, how to make good use of the synthesized images or along with real images to improve deep learning models' performance is an interesting research direction.

5.3.3 Interpretable Deep Learning for Clinical Usage

In recent years, deep learning have rapidly dominated the field of medical image analysis [Duncan, 2020]. With the remarkable performances achieved by deep learning, the complexity of the methods has increased dramatically. Therefore, the system becomes less interpretable and may cause distrust. It is precisely because of the lack of interpretability, we suggested not using synthesized medical images for direct clinical diagnosis. Actually, more interpretable deep learning systems are needed in clinical routine. With such interpretable systems, clinicians can decide whether they should follow/trust the outputs provided by deep learning systems [Rueckert, 2020]. The interpretability can also help clinicians understand and explain the system outputs, or study failures. Moreover, the clinicians can even inspect the models to see if the elements coherent with domain knowledge are well learned by the models. In this thesis, we proposed some ways to interpret neural networks. For example, an approach to identify the most relevant MRI sequence from multi-sequence MRI is introduced in Chapter 2 and a visual attention saliency map to generate the visual explanations showing the concentration regions of the neural networks is proposed in Chapter 3. Meanwhile, in sister fields, researchers have already studied several approaches to dissect deep learning methods [Bau, 2019; Bau, 2017]. Hence, building transparent models and improving interpretability through the development of more efficient ways is necessary and critical for clinical usage.

Bibliography

[Ambartsoumian, 2018] Artaches Ambartsoumian and Fred Popowich. “Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers”. In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Oct. 2018, pp. 130–139 (cit. on p. 72).

[Bahrami, 2016a] K. Bahrami, F. Shi, X. Zong, et al. “Reconstruction of 7T-Like Images From 3T MRI”. In: *IEEE Transactions on Medical Imaging* 35.9 (Sept. 2016), pp. 2085–2097 (cit. on p. 51).

[Bahrami, 2016b] Khosro Bahrami, Feng Shi, Islem Rekik, and Dinggang Shen. “Convolutional Neural Network for Reconstruction of 7T-like Images from 3T MRI Using Appearance and Anatomical Features”. In: *LABELS 2016, DLMIA 2016*. Vol. 10008. LNCS. Springer, 2016 (cit. on pp. 3, 9).

[Bailo, 2019] Oleksandr Bailo, DongShik Ham, and Young Min Shin. “Red Blood Cell Image Generation for Data Augmentation Using Conditional Generative Adversarial Networks”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019 (cit. on p. 81).

[Barkhof, 1997] F. Barkhof, M. Filippi, D. H. Miller, et al. “Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis”. In: *Brain* 120 (Pt 11) (Nov. 1997), pp. 2059–2069 (cit. on p. 8).

[Bau, 2017] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. “Network Dissection: Quantifying Interpretability of Deep Visual Representations”. In: *Computer Vision and Pattern Recognition*. 2017 (cit. on p. 81).

[Bau, 2019] David Bau, Jun-Yan Zhu, Hendrik Strobelt, et al. “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2019 (cit. on p. 81).

[Ben-Cohen, 2017] Avi Ben-Cohen, Eyal Klang, Stephen P. Raskin, Michal Marianne Amitai, and Hayit Greenspan. "Virtual PET Images from CT Data Using Deep Convolutional Networks: Initial Results". In: *Simulation and Synthesis in Medical Imaging*. Ed. by Sotirios A. Tsaftaris, Ali Gooya, Alejandro F. Frangi, and Jerry L. Prince. Cham: Springer International Publishing, 2017, pp. 49–57 (cit. on p. 36).

[Bi, 2017] Lei Bi, Jinman Kim, Ashnil Kumar, Dagan Feng, and Michael Fulham. "Synthesis of Positron Emission Tomography (PET) Images via Multi-channel Generative Adversarial Networks (GANs)". In: *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*. Ed. by M. Jorge Cardoso, Tal Arbel, Fei Gao, et al. Cham: Springer International Publishing, 2017, pp. 43–51 (cit. on pp. 4, 27, 36, 53, 63, 69).

[Bodini, 2016] Benedetta Bodini, Mattia Veronese, Daniel García-Lorenzo, et al. "Dynamic Imaging of Individual Remyelination Profiles in Multiple Sclerosis". In: *Annals of Neurology* 79.5 (2016), pp. 726–738 (cit. on pp. 34, 35, 40, 50, 51, 59, 62, 66, 68, 71).

[Burgos, 2014] Ninon Burgos, M. Jorge Cardoso, Kris Thielemans, et al. "Attenuation Correction Synthesis for Hybrid PET-MR Scanners: Application to Brain Studies". In: *IEEE Transactions on Medical Imaging* 33.12 (Dec. 2014), pp. 2332–2341 (cit. on pp. 9, 26, 31, 52, 57).

[Chartsias, 2018] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris. "Multimodal MR Synthesis via Modality-Invariant Latent Representation". In: *IEEE Transactions on Medical Imaging* 37.3 (Mar. 2018), pp. 803–814 (cit. on pp. 28, 44, 46, 51, 52).

[Chen, 2013] Y. Chen, F. Yu, L. Luo, and C. Toumoulin. "Improving abdomen tumor low-dose CT images using dictionary learning based patch processing and unsharp filtering". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. July 2013, pp. 4014–4017 (cit. on p. 51).

[Chen, 2017] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. "Learning Efficient Object Detection Models with Knowledge Distillation". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al. Curran Associates, Inc., 2017, pp. 742–751 (cit. on p. 9).

[Choi, 2018] Hongyoon Choi and Dong Soo Lee. "Generation of Structural MR Images from Amyloid PET: Application to MR-Less Quantification". In: *Journal of Nuclear Medicine* 59.7 (2018), pp. 1111–1117 (cit. on pp. 4, 27, 53, 80).

[Chollet, 2015] François Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015 (cit. on pp. 13, 35).

[Compston, 2008] A. Compston and A. Coles. "Multiple sclerosis". In: *Lancet* 372.9648 (2008), pp. 1502–1517 (cit. on pp. 1, 8, 24).

[Costa, 2018] P. Costa, A. Galdran, M. I. Meyer, et al. "End-to-End Adversarial Retinal Image Synthesis". In: *IEEE Transactions on Medical Imaging* 37.3 (Mar. 2018), pp. 781–791 (cit. on pp. 4, 53).

[Coupé, 2018] P. Coupé, T. Tourdias, P. Linck, J. E. Romero, and J. V. Manjón. "LesionBrain: An Online Tool for White Matter Lesion Segmentation". In: *International Workshop on Patch-based Techniques in Medical Imaging– Patch-MI 2018*. LNCS. Springer, 2018 (cit. on p. 16).

[Denton, 2015] Emily L Denton, Soumith Chintala, arthur szlam arthur, and Rob Fergus. "Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 1486–1494 (cit. on p. 27).

[Duncan, 2020] J. S. Duncan, M. F. Insana, and N. Ayache. "Biomedical Imaging and Analysis in the Age of Big Data and Deep Learning [Scanning the Issue]". In: *Proceedings of the IEEE* 108.1 (2020), pp. 3–10 (cit. on p. 81).

[Fischl, 2012] Bruce Fischl. "FreeSurfer". In: *Neuroimage* 62.2 (2012), pp. 774–781 (cit. on pp. 35, 59).

[Fu, 2019] Jun Fu, Jing Liu, Haijie Tian, et al. "Dual Attention Network for Scene Segmentation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on p. 72).

[Goodfellow, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680 (cit. on pp. 4, 29, 55).

[Greve, 2009] Douglas N. Greve and Bruce Fischl. "Accurate and robust brain image alignment using boundary-based registration". In: *NeuroImage* 48.1 (2009), pp. 63–72 (cit. on p. 13).

[Gupta, 2019] Anant Gupta, Srivas Venkatesh, Sumit Chopra, and Christian Ledig. “Generative Image Translation for Data Augmentation of Bone Lesion Pathology”. In: *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*. Vol. 102. Proceedings of Machine Learning Research. PMLR, Aug. 2019, pp. 225–235 (cit. on p. 81).

[Hagiwara, 2019] A. Hagiwara, Y. Otsuka, M. Hori, et al. “Improving the Quality of Synthetic FLAIR Images with Deep Learning Using a Conditional Generative Adversarial Network for Pixel-by-Pixel Image Translation”. In: *American Journal of Neuroradiology* (2019) (cit. on pp. 51, 52).

[Han, 2017] Xiao Han. “MR-based synthetic CT generation using a deep convolutional neural network method”. In: *Medical Physics* 44.4 (2017), pp. 1408–1419 (cit. on pp. 4, 27).

[Havaei, 2016] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. “HeMIS: Hetero-Modal Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Ed. by Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells. Cham: Springer International Publishing, 2016, pp. 469–477 (cit. on p. 46).

[He, 2016] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on CVPR*. IEEE Computer Society, 2016, pp. 770–778 (cit. on pp. 9, 13, 15, 61).

[Hofmann, 2008] Matthias Hofmann, Florian Steinke, Verena Scheel, et al. “MRI-Based Attenuation Correction for PET/MRI: A Novel Approach Combining Pattern Recognition and Atlas Registration”. In: *Journal of Nuclear Medicine* 49.11 (2008), pp. 1875–1883 (cit. on p. 26).

[Hong, 2018] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. “Conditional Generative Adversarial Network for Structured Domain Adaptation”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (cit. on p. 36).

[Hu, 2017] Yipeng Hu, Eli Gibson, Li-Lin Lee, et al. “Freehand Ultrasound Image Simulation with Spatially-Conditioned Generative Adversarial Networks”. In: *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*. Ed. by M. Jorge Cardoso, Tal Arbel, Fei Gao, et al. Cham: Springer International Publishing, 2017, pp. 105–115 (cit. on pp. 4, 53).

[Huynh, 2016] T. Huynh, Y. Gao, J. Kang, et al. “Estimating CT Image From MRI Data Using Structured Random Forest and Auto-Context Model”. In: *IEEE Trans Med Imaging* 35.1 (2016), pp. 174–183 (cit. on pp. 9, 27, 53).

[Iglesias, 2013] Juan Eugenio Iglesias, Ender Konukoglu, Darko Zikic, et al. “Is Synthesizing MRI Contrast Useful for Inter-modality Analysis?” In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. Vol. 8149. LNCS. Springer, 2013 (cit. on p. 8).

[Ioffe, 2015] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. ICML’15. Lille, France, 2015, pp. 448–456 (cit. on p. 33).

[Isola, 2016] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. “Image-to-Image Translation with Conditional Adversarial Networks”. In: *arxiv* (2016) (cit. on pp. 21, 27, 31, 33, 36, 53).

[Jenkinson, 2012] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. “FSL”. In: *NeuroImage* 62.2 (2012). 20 YEARS OF fMRI, pp. 782–790 (cit. on pp. 35, 41, 59, 62).

[Jog, 2014] A. Jog, A. Carass, D. L. Pham, and J. L. Prince. “Random Forest FLAIR Reconstruction from T1, T2, and PD -Weighted MRI”. In: *Proc IEEE Int Symp Biomed Imaging* 2014 (2014), pp. 1079–1082 (cit. on pp. 8, 10, 14, 20, 27).

[Kaplan, 2019] Sydney Kaplan and Yang-Ming Zhu. “Full-Dose PET Image Estimation from Low-Dose PET Image Using Deep Learning: a Pilot Study”. In: *Journal of Digital Imaging* 32.5 (Oct. 2019), pp. 773–778 (cit. on pp. 51, 52).

[Kodali, 2017] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. *On Convergence and Stability of GANs*. 2017. arXiv: 1705.07215 (cit. on p. 55).

[Koikkalainen, 2016] Juha Koikkalainen, Hanneke Rhodius-Meester, Antti Tolonen, et al. “Differential diagnosis of neurodegenerative diseases using structural MRI data”. In: *NeuroImage: Clinical* 11 (2016), pp. 435–449 (cit. on p. 21).

[Krizhevsky, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *NIPS 25*. Curran Associates, 2012, pp. 1097–1105 (cit. on p. 10).

[Kurtzke, 1983] John F. Kurtzke. “Rating neurologic impairment in multiple sclerosis”. In: *Neurology* 33.11 (1983), pp. 1444–1444 (cit. on p. 58).

[LeCun, 1989] Y. LeCun, B. Boser, J. S. Denker, et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Comput.* 1.4 (Dec. 1989), pp. 541–551 (cit. on p. 10).

[Lee, 2017] Junghoon Lee, Aaron Carass, Amod Jog, Can Zhao, and Jerry L. Prince. “Multi-atlas-based CT synthesis from conventional MRI with patch-based refinement for MRI-based radiotherapy planning”. In: *Medical Imaging 2017: Image Processing*. Ed. by Martin A. Styner and Elsa D. Angelini. Vol. 10133. International Society for Optics and Photonics. SPIE, 2017, pp. 434–439 (cit. on p. 52).

[Leynes, 2018] Andrew P. Leynes, Jaewon Yang, Florian Wiesinger, et al. “Zero-Echo-Time and Dixon Deep Pseudo-CT (ZeDD CT): Direct Generation of Pseudo-CT Images for Pelvic PET/MRI Attenuation Correction Using Deep Convolutional Neural Networks with Multiparametric MRI”. In: *Journal of Nuclear Medicine* 59.5 (2018), pp. 852–858 (cit. on pp. 27, 53, 80).

[Li, 2014] Rongjian Li, Wenlu Zhang, Heung-Il Suk, et al. “Deep Learning Based Imaging Data Completion for Improved Brain Disease Diagnosis”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*. Vol. 8675. LNCS. Springer International Publishing, 2014, pp. 305–312 (cit. on pp. 3, 9, 27, 36, 45, 46, 53, 63, 69, 71).

[Liu, 2012] Yan Liu, Jianhua Ma, Yi Fan, and Zhengrong Liang. “Adaptive-weighted total variation minimization for sparse data toward low-dose x-ray computed tomography image reconstruction”. In: *Physics in Medicine and Biology* 57.23 (Nov. 2012), pp. 7923–7956 (cit. on p. 51).

[Liu, 2018] Fang Liu, Hyungseok Jang, Richard Kijowski, Tyler Bradshaw, and Alan B. McMillan. “Deep Learning MR Imaging–based Attenuation Correction for PET/MR Imaging”. In: *Radiology* 286.2 (2018), pp. 676–684 (cit. on pp. 27, 53, 80).

[Logan, 1996] Jean Logan, Joanna S. Fowler, Nora D. Volkow, et al. “Distribution Volume Ratios without Blood Sampling from Graphical Analysis of PET Data”. In: *Journal of Cerebral Blood Flow & Metabolism* 16.5 (1996), pp. 834–840 (cit. on pp. 35, 59).

[Lublin, 2014] Fred D. Lublin, Stephen C. Reingold, Jeffrey A. Cohen, et al. “Defining the clinical course of multiple sclerosis”. In: *Neurology* 83.3 (2014), pp. 278–286 (cit. on p. 1).

[Ma, 2018] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. “DocUNet: Document Image Unwarping via a Stacked U-Net”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (cit. on pp. 4, 32).

[Mahmood, 2018] Faisal Mahmood, Richard Chen, and Nicholas J. Durr. "Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training". In: *IEEE Transactions on Medical Imaging* 37.12 (2018) (cit. on pp. 4, 53).

[Martin, 2015] Abadi Martin, Ashish Agarwal, Paul Barham, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015 (cit. on p. 62).

[Maspero, 2018] Matteo Maspero, Mark H F Savenije, Anna M Dinkla, et al. "Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy". In: *Physics in Medicine & Biology* 63.18 (Sept. 2018), p. 185001 (cit. on p. 27).

[Milletari, 2016] F. Milletari, N. Navab, and S. Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*. Oct. 2016, pp. 565–571 (cit. on p. 46).

[Mirza, 2014] Mehdi Mirza and Simon Osindero. "Conditional Generative Adversarial Nets". In: *CoRR* abs/1411.1784 (2014) (cit. on pp. 4, 29, 55, 56).

[Mueller, 2005] Susanne G Mueller, Michael W Weiner, Leon J Thal, et al. "The Alzheimer's Disease Neuroimaging Initiative". In: *Neuroimaging clinics of North America* 15.4 (Nov. 2005), pp. 869–xii (cit. on p. 8).

[Nie, 2016] Dong Nie, Xiaohuan Cao, Yaozong Gao, Li Wang, and Dinggang Shen. "Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks". In: *Deep Learning and Data Labeling for Medical Applications*. Ed. by Gustavo Carneiro, Diana Mateus, Loïc Peter, et al. Cham: Springer International Publishing, 2016, pp. 170–178 (cit. on pp. 3, 9, 53, 80).

[Nie, 2018] D. Nie, R. Trullo, J. Lian, et al. "Medical Image Synthesis with Deep Convolutional Adversarial Networks". In: *IEEE Transactions on Biomedical Engineering* 65.12 (Dec. 2018), pp. 2720–2730 (cit. on p. 72).

[Pan, 2018] Yongsheng Pan, Mingxia Liu, Chunfeng Lian, et al. "Synthesizing Missing PET from MRI with Cycle-consistent Generative Adversarial Networks for Alzheimer's Disease Diagnosis". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Cham: Springer International Publishing, 2018, pp. 455–463 (cit. on pp. 27, 45, 46, 53, 71, 80).

[Patrikios, 2006] Peter Patrikios, Christine Stadelmann, Alexandra Kutzelnigg, et al. “Remyelination is extensive in a subset of multiple sclerosis patients”. In: *Brain* 129.12 (Aug. 2006), pp. 3165–3172 (cit. on p. 50).

[Paty, 1988] D. W. Paty, J. J. Oger, L. F. Kastrukoff, et al. “MRI in the diagnosis of MS: a prospective study with comparison of clinical evaluation, evoked potentials, oligoclonal banding, and CT”. In: *Neurology* 38.2 (Feb. 1988), pp. 180–185 (cit. on p. 8).

[Petiet, 2019] Alexandra Petiet, Isaac Adanyeguh, Marie-Stéphane Aigrot, et al. “Ultrahigh field imaging of myelin disease models: Toward specific markers of myelin integrity?” In: *Journal of Comparative Neurology* 527.13 (2019), pp. 2179–2189 (cit. on p. 50).

[Poutiainen, 2016] Pekka Poutiainen, Merja Jaronen, Francisco J. Quintana, and Anna-Liisa Brownell. “Precision Medicine in Multiple Sclerosis: Future of PET Imaging of Inflammation and Reactive Astrocytes”. In: *Frontiers in Molecular Neuroscience* 9 (2016), p. 85 (cit. on p. 3).

[Rabinovici, 2007] G. D. Rabinovici, A. J. Furst, J. P. O’Neil, et al. “11C-PIB PET imaging in Alzheimer disease and frontotemporal lobar degeneration”. In: *Neurology* 68.15 (2007), pp. 1205–1212 (cit. on p. 24).

[Rohé, 2017] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. “SVF-Net: Learning Deformable Image Registration Using Shape Matching”. In: *MICCAI 2017*. Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Québec, Canada: Springer International Publishing, Sept. 2017, pp. 266–274 (cit. on pp. 4, 32).

[Ronneberger, 2015] O. Ronneberger, P.Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 9351. LNCS. Springer, 2015, pp. 234–241 (cit. on pp. 3, 14, 32, 61).

[Roxburgh, 2005] R. H.S.R. Roxburgh, S. R. Seaman, T. Masterman, et al. “Multiple Sclerosis Severity Score”. In: *Neurology* 64.7 (2005), pp. 1144–1151 (cit. on p. 58).

[Roy, 2010] Snehashis Roy, Aaron Carass, Navid Shiee, Dzung L. Pham, and Jerry L. Prince. “MR contrast synthesis for lesion segmentation”. In: *Proc IEEE Int Symp Biomed Imaging*. 2010, pp. 932–935 (cit. on pp. 8, 10, 26, 31, 51, 57).

[Rueckert, 2020] D. Rueckert and J. A. Schnabel. “Model-Based and Data-Driven Strategies in Medical Image Computing”. In: *Proceedings of the IEEE* 108.1 (2020), pp. 110–124 (cit. on p. 81).

[Salimans, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, et al. “Improved Techniques for Training GANs”. In: *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 2234–2242 (cit. on p. 62).

[Selvaraju, 2017] R. R. Selvaraju, M. Cogswell, A. Das, et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 618–626 (cit. on p. 47).

[Sevetlidis, 2016] Vasileios Sevetlidis, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. “Whole Image Synthesis Using a Deep Encoder-Decoder Network”. In: *Simulation and Synthesis in Medical Imaging, SASHIMI 2016*. Vol. 9968. LNCS. Springer, 2016, pp. 127–137 (cit. on pp. 3, 9, 10, 20, 27, 51, 52).

[Shelhamer, 2017] Evan Shelhamer, Jonathan Long, and Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4 (2017), pp. 640–651 (cit. on p. 9).

[Shen, 2018] Tao Shen, Jing Jiang, Tianyi Zhou, et al. “DiSAN: directional self-attention network for RNN/CNN-free language understanding”. English. In: *The Thirty-Second AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence (AAAI), 2018, pp. 5446–5455 (cit. on p. 72).

[Shin, 2018] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, et al. “Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks”. In: *Simulation and Synthesis in Medical Imaging*. Ed. by Ali Gooya, Orcun Goksel, Ipek Oguz, and Ninon Burgos. Cham: Springer International Publishing, 2018, pp. 1–11 (cit. on p. 81).

[Sikka, 2018] Apoorva Sikka, Skand Vishwanath Peri, and Deepti R. Bathula. “MRI to FDG-PET: Cross-Modal Synthesis Using 3D U-Net for Multi-modal Alzheimer’s Classification”. In: *Simulation and Synthesis in Medical Imaging*. Ed. by Ali Gooya, Orcun Goksel, Ipek Oguz, and Ninon Burgos. Cham: Springer International Publishing, 2018, pp. 80–89 (cit. on pp. 4, 27, 36, 45, 46, 53, 63, 69, 71).

[Simonyan, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *CoRR* abs/1312.6034 (2013) (cit. on pp. 12, 32).

[Smith, 2002] Stephen M. Smith. “Fast robust automated brain extraction”. In: *Human Brain Mapping* 17.3 (2002), pp. 143–155 (cit. on pp. 35, 62).

[Springenberg, 2014] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. “Striving for Simplicity: The All Convolutional Net”. In: *CoRR* abs/1412.6806 (2014) (cit. on p. 47).

[Stankoff, 2011] Bruno Stankoff, Leorah Freeman, Marie-Stéphane Aigrot, et al. “Imaging central nervous system myelin by positron emission tomography in multiple sclerosis using [methyl-11C]-2-(4’-methylaminophenyl)- 6-hydroxybenzothiazole”. In: *Annals of Neurology* 69.4 (2011), pp. 673–680 (cit. on pp. 3, 24).

[Tang, 2019] Hao Tang, Dan Xu, Nicu Sebe, et al. “Multi-Channel Attention Selection GAN With Cascaded Semantic Guidance for Cross-View Image Translation”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on p. 72).

[Theano, 2016] Theano. “Theano: A Python framework for fast computation of mathematical expressions”. In: *arXiv e-prints* abs/1605.02688 (May 2016) (cit. on pp. 13, 36).

[Thompson, 2018] Alan J Thompson, Brenda L Banwell, Frederik Barkhof, et al. “Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria”. In: *The Lancet Neurology* 17.2 (2018), pp. 162–173 (cit. on p. 2).

[Tian, 2011] Zhen Tian, Xun Jia, Kehong Yuan, Tinsu Pan, and Steve B Jiang. “Low-dose CT reconstruction via edge-preserving total variation regularization”. In: *Physics in Medicine and Biology* 56.18 (Aug. 2011), pp. 5949–5967 (cit. on p. 51).

[Tustison, 2010] N. J. Tustison, B. B. Avants, P. A. Cook, et al. “N4ITK: Improved N3 Bias Correction”. In: *IEEE Transactions on Medical Imaging* 29.6 (June 2010), pp. 1310–1320 (cit. on pp. 13, 35, 62).

[Van Tulder, 2015] Gijs Van Tulder and Marleen de Bruijne. “Why Does Synthesized Data Improve Multi-sequence Classification?” In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Vol. 9349. LNCS. Springer, 2015 (cit. on p. 8).

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al. Curran Associates, Inc., 2017, pp. 5998–6008 (cit. on pp. 56, 72).

[Vemulapalli, 2015] R. Vemulapalli, H. V. Nguyen, and S. K. Zhou. “Unsupervised Cross-Modal Synthesis of Subject-Specific Scans”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015, pp. 630–638 (cit. on p. 26).

[Veronese, 2015] Mattia Veronese, Benedetta Bodini, Daniel García-Lorenzo, et al. “Quantification of [11C]PIB PET for Imaging Myelin in the Human Brain: A Test—Retest Reproducibility Study in High-Resolution Research Tomography”. In: *Journal of Cerebral Blood Flow & Metabolism* 35.11 (2015), pp. 1771–1782 (cit. on pp. 35, 62).

[Wang, 2016] Yaxing Wang, Lichao Zhang, and Joost van de Weijer. “Ensembles of Generative Adversarial Networks”. In: *CoRR* abs/1612.00991 (2016). eprint: 1612.00991 (cit. on pp. 46, 72).

[Wang, 2018a] Chengjia Wang, Gillian Macnaught, Giorgos Papanastasiou, Tom MacGillivray, and David Newby. “Unsupervised Learning for Cross-Domain Medical Image Synthesis Using Deformation Invariant Cycle Consistency Networks”. In: *Simulation and Synthesis in Medical Imaging*. Ed. by Ali Gooya, Orcun Goksel, Ipek Oguz, and Ninon Burgos. Cham: Springer International Publishing, 2018, pp. 52–60 (cit. on p. 27).

[Wang, 2018b] X. Wang, R. Girshick, A. Gupta, and K. He. “Non-local Neural Networks”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 7794–7803 (cit. on p. 72).

[Wang, 2018c] Yan Wang, Luping Zhou, Lei Wang, et al. “Locality Adaptive Multi-modality GANs for High-Quality PET Image Synthesis”. In: *MICCAI 2018*. Cham: Springer, 2018, pp. 329–337 (cit. on pp. 53, 71).

[Wang, 2019] Zihao Wang, Clair Vandersteen, Thomas Demarcy, et al. “Deep Learning Based Metal Artifacts Reduction in Post-operative Cochlear Implant CT Imaging”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, et al. Cham: Springer International Publishing, 2019, pp. 121–129 (cit. on p. 51).

[Wei, 2018a] Wen Wei, Emilie Poirion, Benedetta Bodini, et al. “FLAIR MR Image Synthesis by Using 3D Fully Convolutional Networks for Multiple Sclerosis”. In: *Proceedings of the Joint Annual Meeting ISMRM-ESMRMB*. Ed. by ISMRM. Paris, France, 2018 (cit. on pp. 7, 79).

[Wei, 2018b] Wen Wei, Emilie Poirion, Benedetta Bodini, et al. “Learning Myelin Content in Multiple Sclerosis from Multimodal MRI Through Adversarial Training”. In: *MICCAI 2018*. Cham: Springer, 2018, pp. 514–522 (cit. on pp. 5, 24, 53, 57, 58, 72, 79).

[Wei, 2019a] Wen Wei, Emilie Poirion, Benedetta Bodini, et al. “Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis”. In: *Journal of Medical Imaging* 6.1 (2019), pp. 1–9 (cit. on pp. 5, 7, 51, 52, 79).

[Wei, 2019b] Wen Wei, Emilie Poirion, Benedetta Bodini, et al. “Predicting PET-derived demyelination from multimodal MRI using sketcher-refiner adversarial training for multiple sclerosis”. In: *Medical Image Analysis* 58 (2019), p. 101546 (cit. on pp. 5, 23, 53, 58, 61, 63, 64, 67, 70, 71, 79).

[Wei, 2020a] Wen Wei, Emilie Poirion, Benedetta Bodini, et al. “Conditional Flexible SAGAN for Predicting PET-derived Myelin Content in Multiple Sclerosis from Multisequence MRI”. In: Submitted to MICCAI 2020, 2020 (cit. on pp. 5, 50, 54, 58, 79).

[Wei, 2020b] Wen Wei, Emilie Poirion, Benedetta Bodini, et al. “Conditional Flexible SAGAN for Predicting PET-derived Myelin Content in Multiple Sclerosis from Multisequence MRI”. In: *Under Review* (2020) (cit. on pp. 5, 49, 79).

[Wolterink, 2017] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, et al. “Deep MR to CT Synthesis Using Unpaired Data”. In: *Simulation and Synthesis in Medical Imaging*. Ed. by Sotirios A. Tsaftaris, Ali Gooya, Alejandro F. Frangi, and Jerry L. Prince. Cham: Springer International Publishing, 2017, pp. 14–23 (cit. on p. 27).

[Woo, 2006] John H. Woo, Lana P. Henry, Jaroslaw Krejza, and Elias R. Melhem. “Detection of Simulated Multiple Sclerosis Lesions on T2-weighted and FLAIR Images of the Brain: Observer Performance”. In: *Radiology* 241.1 (2006), pp. 206–212 (cit. on p. 8).

[Xiang, 2018] Lei Xiang, Qian Wang, Dong Nie, et al. “Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image”. In: *Medical Image Analysis* 47 (2018), pp. 31–44 (cit. on pp. 27, 31, 57).

[Xu, 2012] Q. Xu, H. Yu, X. Mou, et al. “Low-Dose X-ray CT Reconstruction via Dictionary Learning”. In: *IEEE Transactions on Medical Imaging* 31.9 (Sept. 2012), pp. 1682–1697 (cit. on p. 51).

[Ye, 2013] Dong Hye Ye, Darko Zikic, Ben Glocker, Antonio Criminisi, and Ender Konukoglu. “Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 606–613 (cit. on pp. 14, 26, 31, 51, 57).

[Zeiler, 2014] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, 2014, pp. 818–833 (cit. on p. 47).

[Zhang, 2017] Y. Zhang, X. Mou, G. Wang, and H. Yu. “Tensor-Based Dictionary Learning for Spectral CT Reconstruction”. In: *IEEE Transactions on Medical Imaging* 36.1 (Jan. 2017), pp. 142–154 (cit. on p. 51).

[Zhang, 2018a] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. “Self-Attention Generative Adversarial Networks”. In: *arXiv:1805.08318* (2018) (cit. on pp. 53, 55, 56, 72).

[Zhang, 2018b] Hang Zhang, Kristin Dana, Jianping Shi, et al. “Context Encoding for Semantic Segmentation”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (cit. on pp. 4, 32).

[Zhao, 2018] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. “Synthesizing retinal and neuronal images with generative adversarial nets”. In: *Medical Image Analysis* 49 (2018), pp. 14–26 (cit. on pp. 4, 53).

[Zheng, 2018] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache. “3D Consistent and Robust Segmentation of Cardiac Images by Deep Learning With Spatial Propagation”. In: *IEEE Transactions on Medical Imaging* 37.9 (Sept. 2018), pp. 2137–2148 (cit. on pp. 4, 32).

[Zhou, 2016] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. “Learning Deep Features for Discriminative Localization.” In: *CVPR* (2016) (cit. on p. 47).

[Zhou, 2017] S.K. Zhou, H. Greenspan, and D. Shen. *Deep Learning for Medical Image Analysis*. Elsevier Science, 2017 (cit. on p. 9).

[Zhu, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. 2017 (cit. on p. 53).

List of Figures

List of Tables