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

Wen Wei[a,b,c,*], Emilie Poirion[c], Benedetta Bodini[c,d], Stanley Durrleman[b,c], Nicholas Ayache[a], Bruno Stankoff[c,d], Olivier Colliot[b,c]

[a]*Université Côte d'Azur, Inria, Epione Project-Team, Sophia Antipolis, France*
[b]*Inria, Aramis project-team, Paris, France*
[c]*Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France*
[d]*APHP, Hôpital Saint Antoine, Neurology Department, Paris*

ARTICLE INFO

Keywords: Multimodal MRI, PET Imaging, Adversarial Training, Multiple Sclerosis

ABSTRACT

Multiple sclerosis (MS) is the most common demyelinating disease. In MS, demyelination occurs in the white matter of the brain and in the spinal cord. It is thus essential to measure the tissue myelin content to understand the physiopathology of MS, track progression and assess treatment efficacy. Positron emission tomography (PET) with [^{11}C]PIB is a reliable method to measure myelin content in vivo. However, the availability of PET in clinical centers is limited. Moreover, it is expensive to acquire and invasive due to the injection of a radioactive tracer. By contrast, MR imaging is non-invasive, less expensive and widely available, but conventional MRI sequences cannot provide a direct and reliable measure of myelin. In this work, we therefore propose, to the best of our knowledge for the first time, a method to predict the PET-derived myelin content map from multimodal MRI. To that purpose, we introduce a new approach called Sketcher-Refiner generative adversarial networks (GANs) with specifically designed adversarial loss functions. The first network (Sketcher) generates global anatomical and physiological information. The second network (Refiner) refines and generates the tissue myelin content. A visual attention saliency map is also proposed to interpret the attention of neural networks. Our approach is shown to outperform the state-of-the-art methods in terms of image quality and myelin content prediction. Particularly, our prediction results show similar results to the PET-derived gold standard at both global and voxel-wise levels indicating the potential for clinical management of patients with MS.

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 and Coles, 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 and Coles, 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

*Corresponding author: Email: wen.wei@inria.fr;

myelinated white matter regions (Stankoff et al., 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 et al., 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 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.

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 et al., 2008; Burgos et al., 2014) usually need an atlas dataset including the pairs of the source and the target modalities. For example, Burgos et al. (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 et al., 2013; Roy et al., 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. (2010) applied this idea to predict FLAIR from T1-w and T2-w. Equally, Ye et al. (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. (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. (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. (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 et al., 2016; Xiang et al., 2018; Wang et al., 2018) have emerged as a powerful alternative and alleviate the above drawbacks for medical image synthesis. For instance, Sevetlidis et al. (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 et al., 2017; Maspero et al., 2018) and attenuation correction (Leynes et al., 2018; Liu et al., 2018). In the work of Choi and Lee (2018), the authors used GANs to generate the MRI from the PET for the quantification of cortical amyloid load. Bi et al. (2017) used multi-channel GANs to synthesis PET images from CT images. Regarding PET synthesis from MRI, several

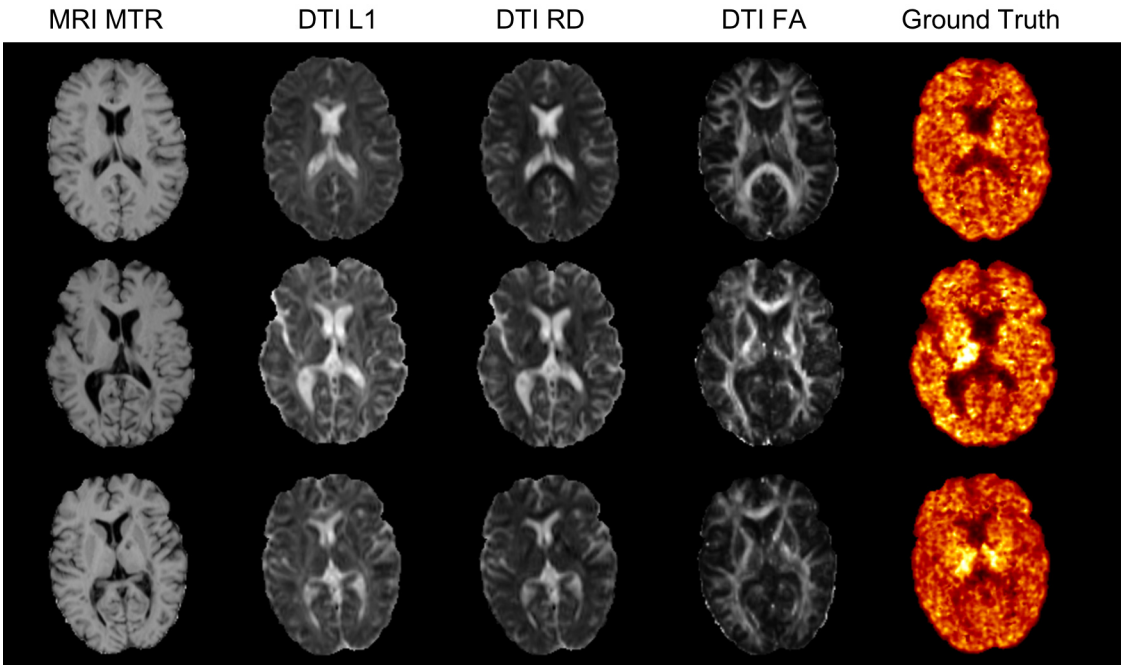


Fig. 1. Some examples of the ground truth ([^{11}C]PIB PET data) and input MR images including magnetization transfer ratio (MTR) and three measures derived from diffusion tensor imaging (DTI): fractional anisotropy (FA), radial diffusivity (RD) and axial diffusivity (AD). The relationship between the MR images and the PET data is complex and highly non-linear.

works have already been proposed (Sikka et al., 2018; Li et al., 2014; Pan et al., 2018). A 3D convolutional neural network (CNN) based on U-Net architecture (Sikka et al., 2018) and a two-layer CNN (Li et al., 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 et al., 2016). The work of Denton et al. (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. (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. (2018), using multimodal MRI can improve the synthesis performance.

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 (Wei et al., 2018). 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.

2. Method

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.

2.1.1. 3D Conditional GANs

Generative adversarial networks (GANs) (Goodfellow et al., 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 and Osindero, 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:

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

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.

2.1.2. 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 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:

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:

$$\min_{G_{\text{S}}} \max_{D_{\text{S}}} \mathcal{L}(D_{\text{S}},G_{\text{S}}) = \mathbb{E}_{I_{\text{M}},I_{\text{P}}\sim p_{\text{data}}(I_{\text{M}},I_{\text{P}})}[\log D_{\text{S}}(I_{\text{M}},I_{\text{P}})] - \mathbb{E}_{I_{\text{M}}\sim p_{\text{data}}(I_{\text{M}}),z\sim p_z(z)}[\log(1-D_{\text{S}}(I_{\text{M}},G_{\text{S}}(I_{\text{M}},z)))] \tag{2}$$

$$\min_{G_{\text{R}}} \max_{D_{\text{R}}} \mathcal{L}(D_{\text{R}},G_{\text{R}}) = \mathbb{E}_{I_{\text{M}},I_{\text{P}}\sim p_{\text{data}}(I_{\text{M}},I_{\text{P}})}[\log D_{\text{R}}(I_{\text{M}},I_{\text{P}})] - \mathbb{E}_{I_{\text{M}}\sim p_{\text{data}}(I_{\text{M}}),I_{\text{s}}\sim G_{\text{S}}(I_{\text{M}},z)}[\log(1-D_{\text{R}}(I_{\text{M}},G_{\text{R}}(I_{\text{M}},I_{\text{s}})))] \tag{3}$$

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 2.4.

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. (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:

$$\mathcal{L}_{L1}(G_{\text{S}}) = \frac{1}{N}\sum_{i=1}^{N} |I_{\text{P}}^i - G_{\text{S}}(I_{\text{M}}^i, z^i)| \tag{4}$$

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 et al., 2010; Burgos et al., 2014; Ye et al., 2013; Xiang et al., 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:

$$\mathcal{L}_{L1}(G_{\text{R}}) = \frac{1}{N\times M}\sum_{i=1}^{N}\Big(\frac{1}{N_{\text{Les}}}\sum_{j\in R_{\text{Les}}}|I_{\text{P}}^{i,j}-\hat{I}_{\text{P}}^{i,j}| + \frac{1}{N_{\text{NAWM}}}\sum_{j\in R_{\text{NAWM}}}|I_{\text{P}}^{i,j}-\hat{I}_{\text{P}}^{i,j}| + \frac{1}{N_{\text{other}}}\sum_{j\in R_{\text{other}}}|I_{\text{P}}^{i,j}-\hat{I}_{\text{P}}^{i,j}|\Big) \tag{5}$$

where $\hat{I}_{\text{P}}$ 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.

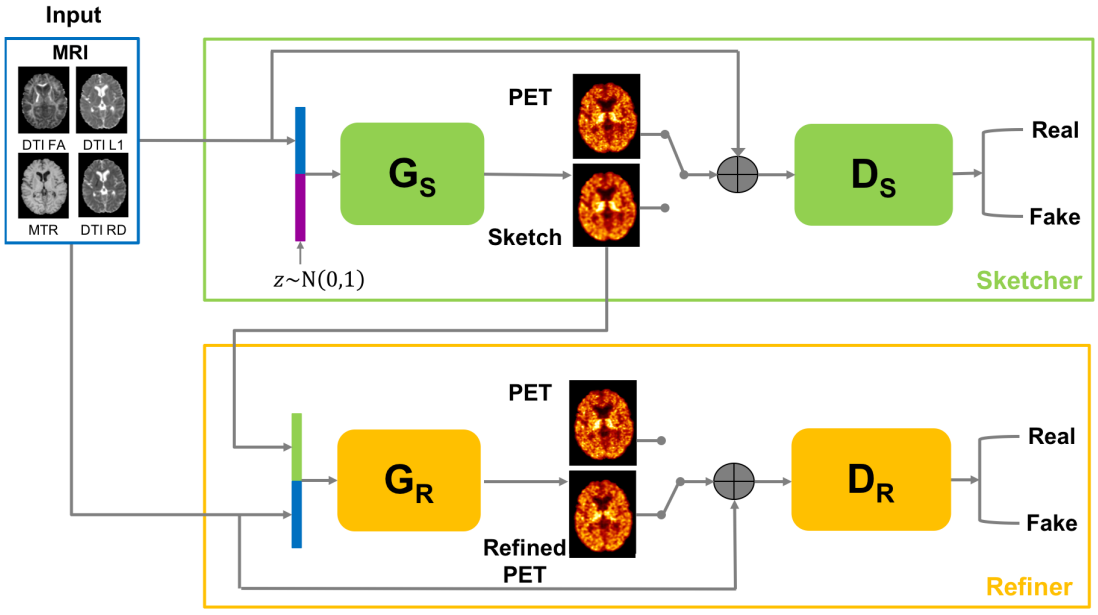


Fig. 2. The proposed Sketcher-Refiner GANs. The Sketcher receives MR images and generates the preliminary anatomy and physiology information. The Refiner receives MR images and the output of the Sketcher. Then it refines and generates the synthetic PET images.

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

$$\begin{aligned} G_{\mathrm{S}}^{*} &= \arg\min_{G_{\mathrm{S}}}\max_{D_{\mathrm{S}}} \mathcal{L}(D_{\mathrm{S}}, G_{\mathrm{S}}) + \lambda_{\mathrm{S}}\mathcal{L}_{L1}(G_{\mathrm{S}}) \\ G_{\mathrm{R}}^{*} &= \arg\min_{G_{\mathrm{R}}}\max_{D_{\mathrm{R}}} \mathcal{L}(D_{\mathrm{R}}, G_{\mathrm{R}}) + \lambda_{\mathrm{R}}\mathcal{L}_{L1}(G_{\mathrm{R}}) \end{aligned} \tag{6}$$

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

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. (2013), our visual attention saliency map is the absolute partial derivative of the prediction loss with respect to the input images $\boldsymbol{I}_{\mathbf{M}}$ defined as follows:

$$\boldsymbol{M} = \left| \frac{\partial Loss}{\partial \boldsymbol{I}_{\mathbf{M}}} \right| \tag{7}$$

Given the input images $\boldsymbol{I}_{\mathbf{M}}$, 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.

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 et al., 2018; Zhang et al., 2018) and medical imaging fields (Rohé et al., 2017; Zheng et al., 2018). The advantage of U-Net (Ronneberger et al., 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 (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 down-sampling. 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 (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 and Szegedy, 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 overemphasize 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 (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. (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.

3. Experiments and Evaluations

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. (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. 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 et al., 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 ± 34 MBq). The Logan graphical reference method (Logan et al., 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 et al., 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 et al., 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. (2016) and Veronese et al. (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 et al., 2015) library with Theano (Theano Development Team, 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. (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. (2018). Moreover, the input noise z is used to provide some slight variation in the generated images. If we remove the noise vector,

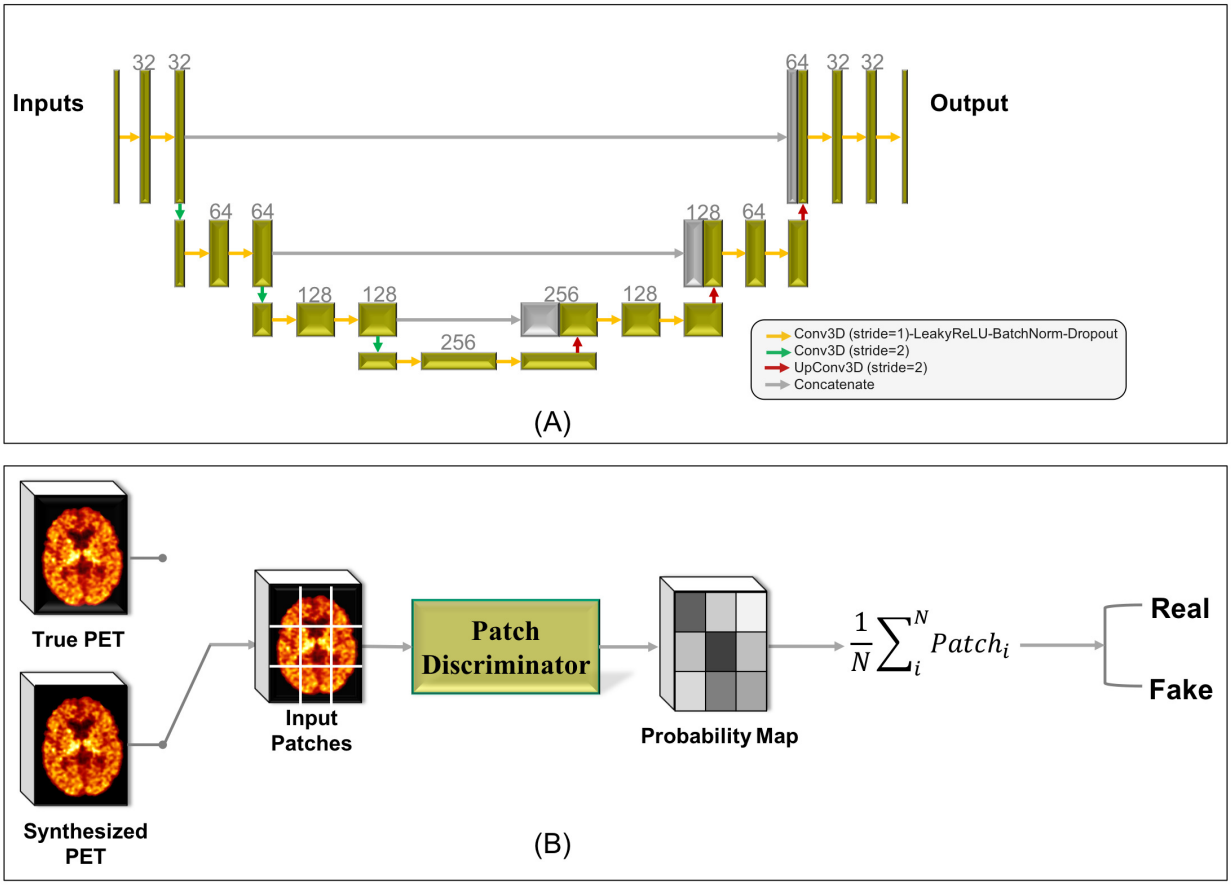


Fig. 3. Architectures proposed for the generator (panel A) and for the discriminator (panel B) in our GANs. (A) The 3D U-Net shaped generator with implementation details shown in the image. (B) The proposed 3D patch discriminator which takes all the patches and classifies them separately to output a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

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.

3.2. Comparisons with state-of-the-art methods

We compared our method with several state-of-the-art methods including a 2-layer DNN (Li et al., 2014), a 3D U-Net (Sikka et al., 2018) and a single cGAN (Bi et al., 2017; Ben-Cohen et al., 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 et al., 2014). The architecture of the 3D U-Net is the same as shown in Fig. 3 (A). It is similar to 3D U-Net used in the work of Sikka et al. (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 et al. (2018) and Li et al. (2014), their proposed methods were aimed to discriminate Alzheimers 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. (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. (2017) and Ben-Cohen et al. (2017) were extended to 3D architecture which corresponds to the Sketcher (see in Fig. 2) in our approach and the loss function was the same as described in Bi et al. (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 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 struc-tural 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.

We then performed a quantitative comparison in terms of global image quality (Table 1). Image quality is evaluated by mean square error (MSE) and peak signal-to-noise ratio (PSNR) defined as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|(\boldsymbol{I}_{\mathbf{P}}^{i}, \hat{\boldsymbol{I}}_{\mathbf{P}}^{i})\|_2 \tag{8}$$

$$PSNR = 20 \cdot \log_{10}(MAX_{I_{\mathrm{P}}}) - 10 \cdot \log_{10}(MSE) \tag{9}$$

where $MAX_{I_{\mathrm{P}}}$ 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.

Table 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 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 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 prediction discrepancy in Table 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. 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. 5 (A)) and of the prediction discrepancy in 3 ROIs (Fig. 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.

3.4. 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. 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 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. 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.

3.5. 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

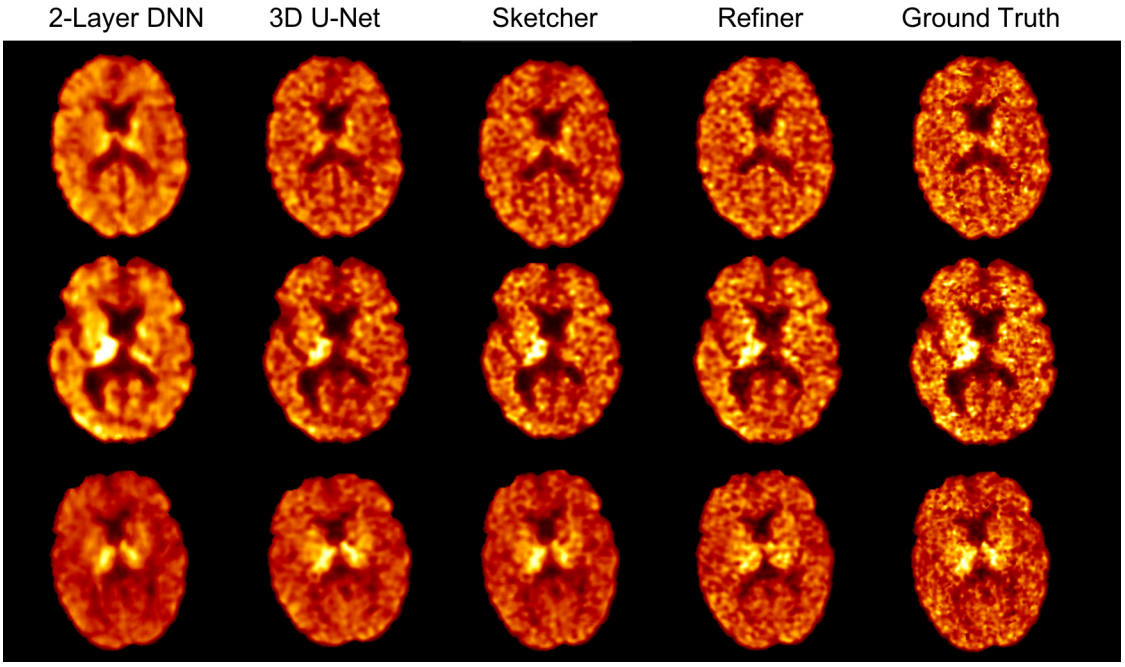


Fig. 4. Qualitative comparison of the results of our method ("Refined"), of a 2-layer DNN , of a 3D U-Net and of a single cGAN (corresponding to the Sketcher in our approach and denoted as "Sketch") to the ground truth.

Table 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

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 et al., 2016). This method involves the determination of a threshold to separate demyelinated 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 et al., 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 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 8 (B).

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 multi-modal 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.

Figure 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

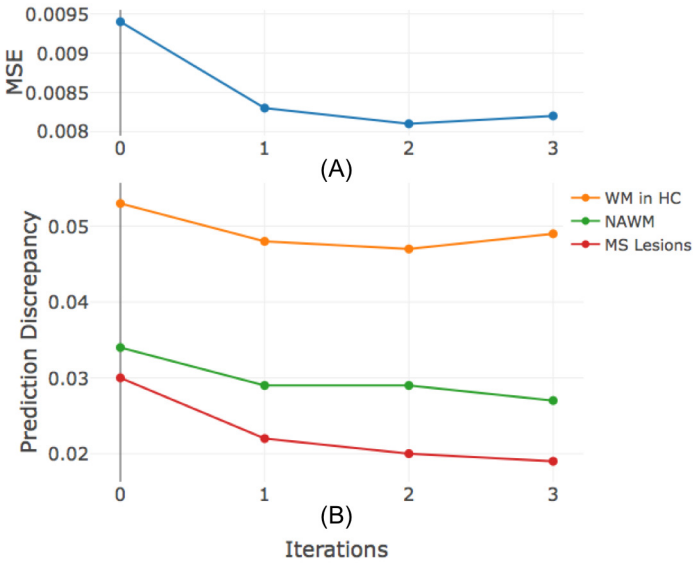


Fig. 5. Performance assessment with respect to different number of iterations. Note that the iteration 0 is our Sketcher and an additional Refiner is used for each new iteration.

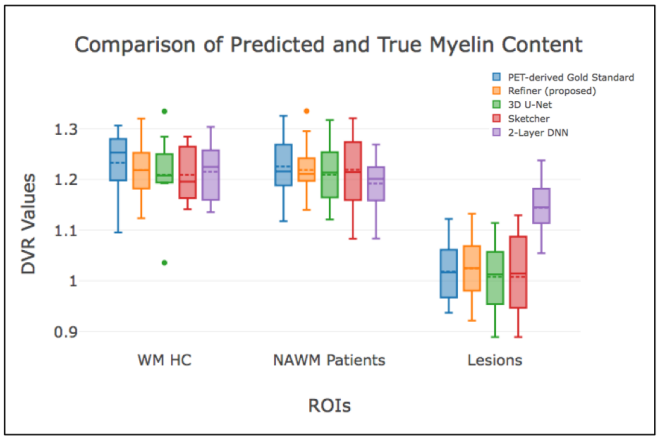


Fig. 6. Group level evaluation for all the methods. The box plots show the median (middle solid line), mean (middle dotted line) and min-max (below and above line) DVR for each ROI for PET-derived DVR parametric map used as gold standard (blue) and the prediction results from our method (yellow), 3D U-Net (green), Sketcher (red) and 2-Layer DNN (violet).

loss function is able to effectively shift the attention of the neural networks towards the MS lesions.

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 shows the corresponding image quality metrics (MSE and PSNR as defined in Eq. 8 and 9). 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. (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.

Table 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 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 ($p > 0.5$).

Table 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)

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 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

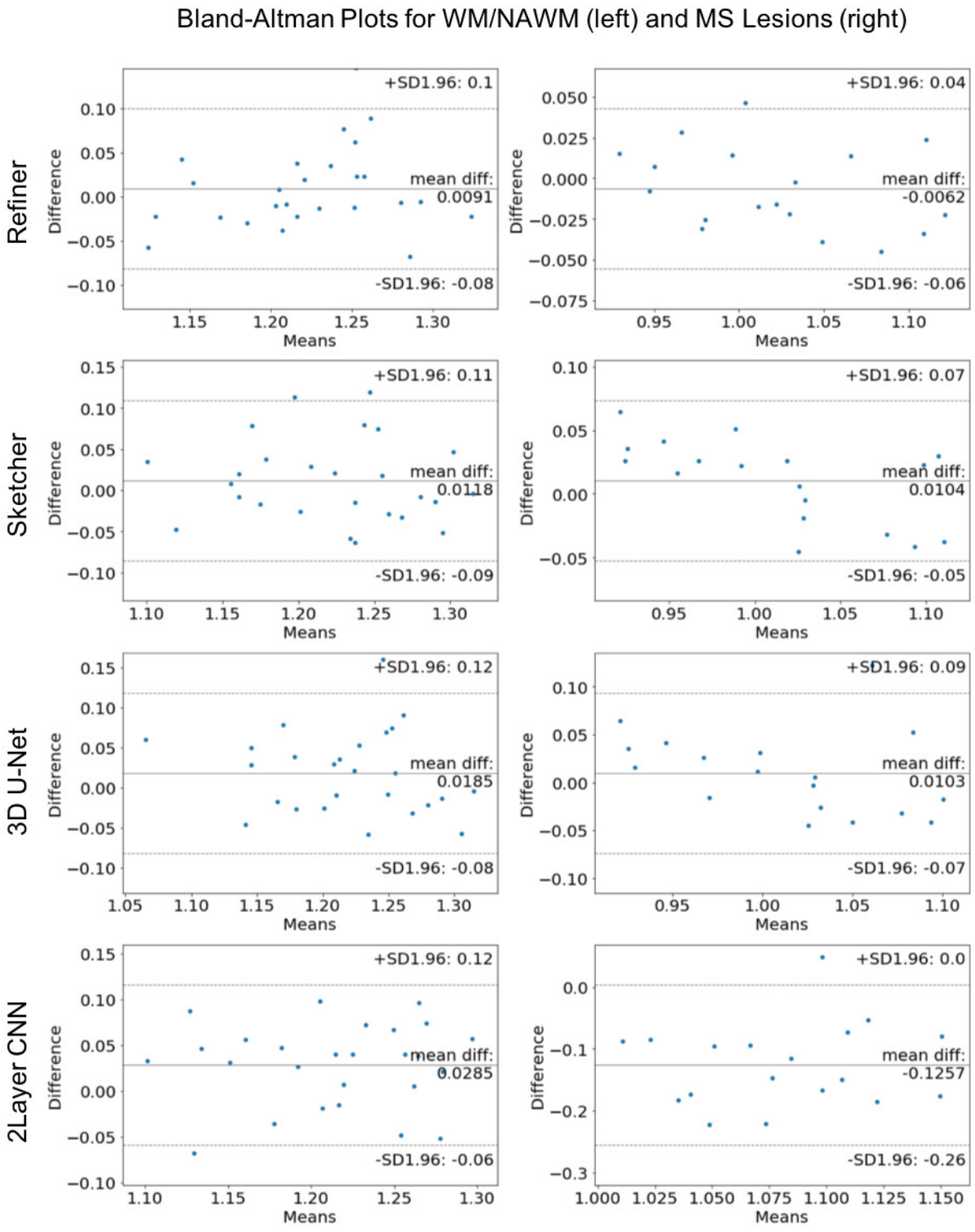
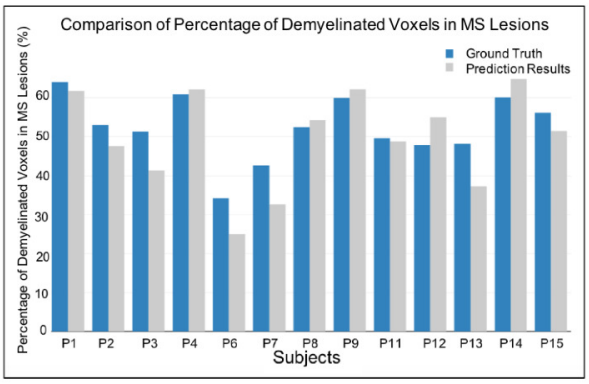


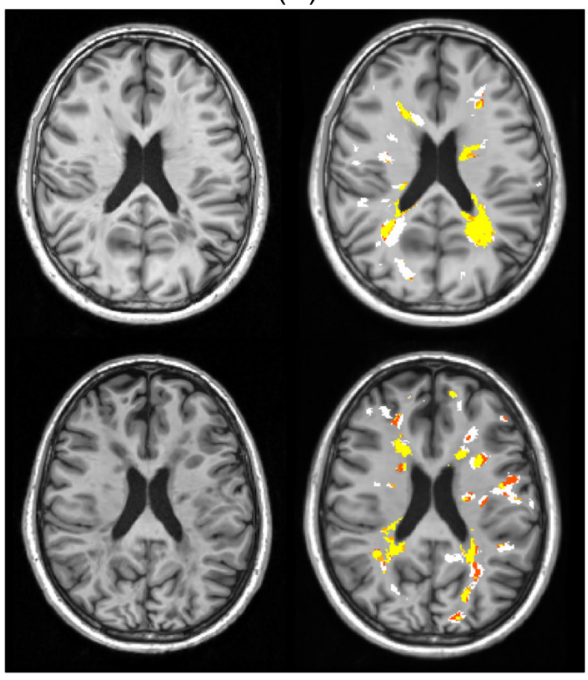
Fig. 7. Bland-Altman Plots for WM/NAWM (left) and MS lesions (right) at the individual level for all the methods.

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



(A)



(B)

Fig. 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.)

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. (2018); Li et al. (2014) and Pan et al. (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. (2016). Specifically, the cascade GANs designed in Wang et al. (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. 4 and 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. (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. (2018); Li et al. (2014) and Pan et al. (2018), only a single MRI pulse sequence is used for prediction, for example Sikka et al. (2018) and Li et al. (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. (2018) and Havaei et al. (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 et al., 2016) or the max fusion (Chartsias et al., 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. Fur-

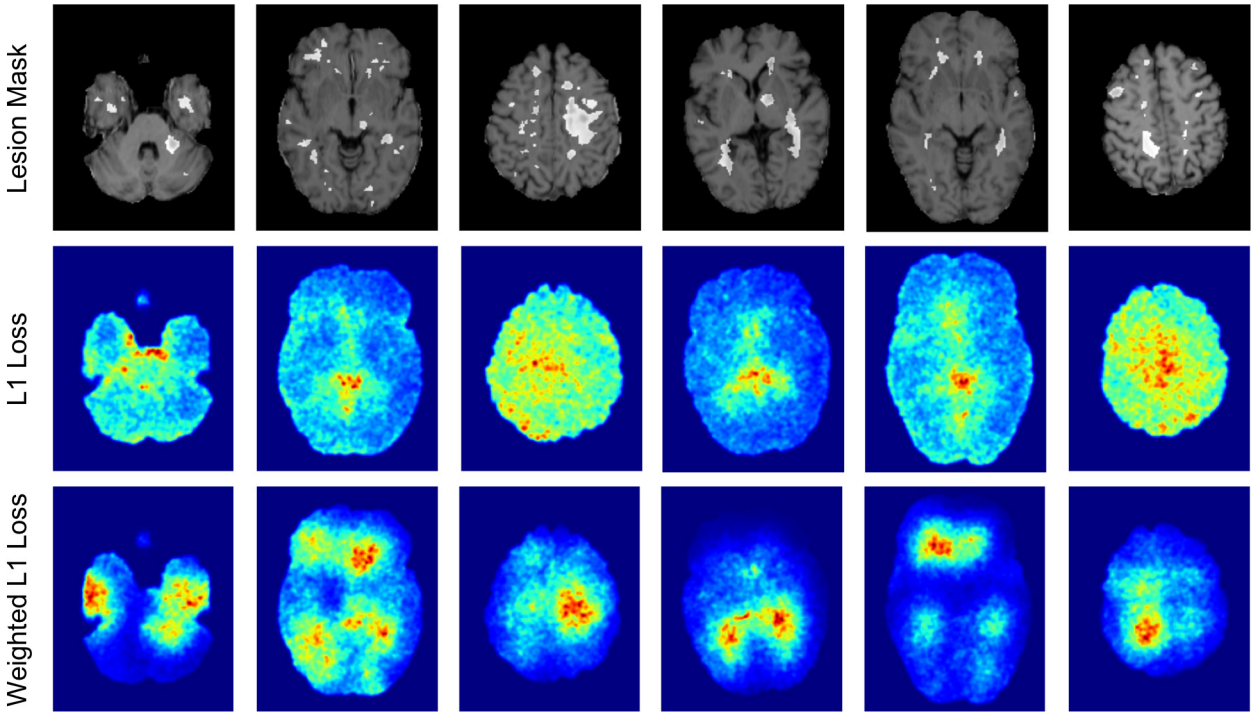


Fig. 9. The proposed visual attention saliency map. The white regions shown in first row are MS lesion masks. The second row shows some examples of the attention of neural networks when L1 loss is used as the traditional constraint in the loss function, without the specific weighting scheme that we proposed. 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.)

thermore, 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 et al., 2014) which only propagates positive gradients for positive activations; 2) *RELU Backpropagation* (Zeiler and Fergus, 2014) which only propagates positive gradients. Moreover, class activation maps (CAM) (Zhou et al., 2016) and Grad-CAM (Selvaraju et al., 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.

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.

Acknowledgments

The first author is funded by an Inria fellowship. The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) ANR-11-IDEX-004 (Agence Nationale de la Recherche-11- Initiative d'Excellence-004, project Learn-PETMR number SU-16-R-EMR-16), and from the "Contrat d'Interface Local" program (to Dr Colliot) from Assistance Publique-Hôpitaux de Paris (AP-HP). The clinical study was funded by specific grants from ELA (European Leukodystrophy Association, grant 2007-0481), and INSERM-DHOS (grant 2008-recherche clinique et translationnelle) and has been sponsored by APHP (Assistance Publique des Hôpitaux de Paris). Emilie Poirion has been funded by IUIS (Institut Universitaire dIngenerie pour la Santé, Sorbonne Université, and both Benedetta Bodini and Emilie Poirion received funding from fondation ARSEP).

References

Ben-Cohen, A., Klang, E., Raskin, S.P., Amitai, M.M., Greenspan, H., 2017. Virtual pet images from ct data using deep convolutional networks: Initial results, in: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (Eds.), Simulation and Synthesis in Medical Imaging, Springer International Publishing, Cham. pp. 49–57.

Bi, L., Kim, J., Kumar, A., Feng, D., Fulham, M., 2017. Synthesis of positron emission tomography (pet) images via multi-channel generative adversarial networks (gans), in: CMMI 2017, SWITCH 2017, RAMBO 2017, Springer. pp. 43–51.

Bodini, B., Veronese, M., Garca-Lorenzo, D., Battaglini, M., Poirion, E., Chardain, A., Freeman, L., Louapre, C., Tchikviladze, M., Papeix, C., Doll, F., Zalc, B., Lubetzki, C., Bottlaender, M., Turkheimer, F., Stankoff, B., 2016. Dynamic imaging of individual remyelination profiles in multiple sclerosis. Annals of Neurology 79, 726–738.

Burgos, N., Cardoso, M.J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C.J., Schott, J.M., Duncan, J.S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S., 2014. Attenuation Correction Synthesis for Hybrid PET-MR Scanners: Application to Brain Studies. IEEE Transactions on Medical Imaging 33, 2332–2341.

Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A., 2018. Multimodal mr synthesis via modality-invariant latent representation. IEEE Transactions on Medical Imaging 37, 803–814.

Choi, H., Lee, D.S., 2018. Generation of structural mr images from amyloid pet: Application to mr-less quantification. Journal of Nuclear Medicine 59, 1111–1117.

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

Compston, A., Coles, A., 2008. Multiple sclerosis. Lancet 372, 1502–1517.

Denton, E.L., Chintala, S., szlam, a., Fergus, R., 2015. Deep generative image models using a laplacian pyramid of adversarial networks, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 1486–1494.

Fischl, B., 2012. Freesurfer. Neuroimage 62, 774–781.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.

Han, X., 2017. Mr-based synthetic ct generation using a deep convolutional neural network method. Medical Physics 44, 1408–1419.

Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. Hemis: Hetero-modal image segmentation, in: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham. pp. 469–477.

Hofmann, M., Steinke, F., Scheel, V., Charpiat, G., Farquhar, J., Aschoff, P., Brady, M., Schlkopf, B., Pichler, B.J., 2008. Mri-based attenuation correction for pet/mri: A novel approach combining pattern recognition and atlas registration. Journal of Nuclear Medicine 49, 1875–1883.

Hong, W., Wang, Z., Yang, M., Yuan, J., 2018. Conditional generative adversarial network for structured domain adaptation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Huynh, T., Gao, Y., Kang, J., Wang, L., Zhang, P., Lian, J., Shen, D., 2016. Estimating CT Image From MRI Data Using Structured Random Forest and Auto-Context Model. IEEE Trans Med Imaging 35, 174–183.

Ioffe, S., Szegedy, C., 2015. 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, pp. 448–456.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2016. Image-to-image translation with conditional adversarial networks. arxiv .

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. NeuroImage 62, 782 – 790. 20 YEARS OF fMRI.

Jog, A., Carass, A., Pham, D.L., Prince, J.L., 2014. Random Forest FLAIR Reconstruction from T1, T2, and PD -Weighted MRI. Proc IEEE Int Symp Biomed Imaging 2014, 1079–1082.

Leynes, A.P., Yang, J., Wiesinger, F., Kaushik, S.S., Shanbhag, D.D., Seo, Y., Hope, T.A., Larson, P.E., 2018. 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. Journal of Nuclear Medicine 59, 852–858.

Li, R., Zhang, W., Suk, H.I., Wang, L., Li, J., Shen, D., Ji, S., 2014. Deep learning based imaging data completion for improved brain disease diagnosis, in: MICCAI 2014. Springer. volume 8675 of *LNCS*, pp. 305–312.

Liu, F., Jang, H., Kijowski, R., Bradshaw, T., McMillan, A.B., 2018. Deep learning mr imagingbased attenuation correction for pet/mr imaging. Radiology 286, 676–684.

Logan, J., Fowler, J.S., Volkow, N.D., Wang, G.J., Ding, Y.S., Alexoff, D.L., 1996. Distribution volume ratios without blood sampling from graphical analysis of pet data. Journal of Cerebral Blood Flow & Metabolism 16, 834–840.

Ma, K., Shu, Z., Bai, X., Wang, J., Samaras, D., 2018. Docunet: Document image unwarping via a stacked u-net, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Maspero, M., Savenije, M.H.F., Dinkla, A.M., Seevinck, P.R., Intven, M.P.W., Jurgenliemk-Schulz, I.M., Kerkmeijer, L.G.W., van den Berg, C.A.T., 2018. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. Physics in Medicine & Biology 63, 185001.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth

International Conference on 3D Vision (3DV), pp. 565–571.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. CoRR abs/1411.1784.

Pan, Y., Liu, M., Lian, C., Zhou, T., Xia, Y., Shen, D., 2018. Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for alzheimer's disease diagnosis, in: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham. pp. 455–463.

Rabinovici, G.D., Furst, A.J., O'Neil, J.P., Racine, C.A., Mormino, E.C., Baker, S.L., Chetty, S., Patel, P., Pagliaro, T.A., Klunk, W.E., Mathis, C.A., Rosen, H.J., Miller, B.L., Jagust, W.J., 2007. 11c-pib pet imaging in alzheimer disease and frontotemporal lobar degeneration. Neurology 68, 1205–1212.

Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X., 2017. SVF-Net: Learning Deformable Image Registration Using Shape Matching, in: MICCAI 2017, Springer International Publishing, Québec, Canada. pp. 266–274.

Ronneberger, O., P.Fischer, Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 234–241.

Roy, S., Carass, A., Shiee, N., Pham, D.L., Prince, J.L., 2010. MR contrast synthesis for lesion segmentation, in: Proc IEEE Int Symp Biomed Imaging, pp. 932–935.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. doi:10.1109/ICCV.2017.74.

Sevetlidis, V., Giuffrida, M.V., Tsaftaris, S.A., 2016. Whole image synthesis using a deep encoder-decoder network, in: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (Eds.), SASHIMI2016. Springer. volume 9968 of *LNCS*, pp. 127–137.

Sikka, A., Peri, S.V., Bathula, D.R., 2018. MRI to FDG-PET: Cross-modal synthesis using 3d u-net for multi-modal alzheimer's classification, in: Gooya, A., Goksel, O., Oguz, I., Burgos, N. (Eds.), Simulation and Synthesis in Medical Imaging, Springer International Publishing, Cham. pp. 80–89.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR abs/1312.6034.

Smith, S.M., 2002. Fast robust automated brain extraction. Human Brain Mapping 17, 143–155.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A., 2014. Striving for simplicity: The all convolutional net. CoRR abs/1412.6806.

Stankoff, B., Freeman, L., Aigrot, M.S., Chardain, A., Doll, F., Williams, A., Galanaud, D., Armand, L., Lehericy, S., Lubetzki, C., Zalc, B., Bottlaender, M., 2011. Imaging central nervous system myelin by positron emission tomography in multiple sclerosis using [methyl-11c]-2-(4-methylaminophenyl)- 6-hydroxybenzothiazole. Annals of Neurology 69, 673–680.

Theano Development Team, 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688. URL: http://arxiv.org/abs/1605.02688.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: Improved n3 bias correction. IEEE Transactions on Medical Imaging 29, 1310–1320.

Vemulapalli, R., Nguyen, H.V., Zhou, S.K., 2015. Unsupervised cross-modal synthesis of subject-specific scans, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 630–638.

Veronese, M., Bodini, B., Garca-Lorenzo, D., Battaglini, M., Bongarzone, S., Comtat, C., Bottlaender, M., Stankoff, B., Turkheimer, F.E., 2015. Quantification of [11c]pib pet for imaging myelin in the human brain: A testretest reproducibility study in high-resolution research tomography. Journal of Cerebral Blood Flow & Metabolism 35, 1771–1782.

Wang, C., Macnaught, G., Papanastasiou, G., MacGillivray, T., Newby, D., 2018. Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks, in: Gooya, A., Goksel, O., Oguz, I., Burgos, N. (Eds.), Simulation and Synthesis in Medical Imaging, Springer International Publishing, Cham. pp. 52–60.

Wang, Y., Zhang, L., van de Weijer, J., 2016. Ensembles of generative adversarial networks. CoRR abs/1612.00991. arXiv:1612.00991.

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Ayache, N., Stankoff, B., Colliot, O., 2018. Learning myelin content in multiple sclerosis from multimodal mri through adversarial training, in: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing. pp. 514–522.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I., 2017. Deep mr to ct synthesis using unpaired data, in: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (Eds.), Simulation and Synthesis in Medical Imaging, Springer International Publishing, Cham. pp. 14–23.

Xiang, L., Wang, Q., Nie, D., Zhang, L., Jin, X., Qiao, Y., Shen, D., 2018. Deep embedding convolutional neural network for synthesizing CT image from t1-weighted MR image. Medical Image Analysis 47, 31–44.

Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E., 2013. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 606–613.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham. pp. 818–833.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A., 2018. Context encoding for semantic segmentation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zheng, Q., Delingette, H., Duchateau, N., Ayache, N., 2018. 3D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Transactions on Medical Imaging 37, 2137–2148.

Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A., 2016. Learning Deep Features for Discriminative Localization. CVPR .