NeuroImage 196 (2019) 1–15

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Multi-branch convolutional neural network for multiple sclerosis lesion segmentation

Shahab Aslani [a,b,*], Michael Dayan [a,c], Loredana Storelli [d], Massimo Filippi [d], Vittorio Murino [a,e], Maria A. Rocca [d], Diego Sona [a,f,**]

[a] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, Italy
[b] Science and Technology for Electronic and Telecommunication Engineering, University of Genoa, Italy
[c] Human Neuroscience Platform, Fondation Campus Biotech Geneva, Switzerland
[d] Neuroimaging Research Unit, Institute of Experimental Neurology (INSPE), Division of Neuroscience, San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, Italy
[e] Dipartimento di Informatica, University of Verona, Italy
[f] NeuroInformatics Laboratory, Fondazione Bruno Kessler, Trento, Italy

ARTICLE INFO

Keywords:
Multiple sclerosis
Lesions
Brain
Multiple image modality
Segmentation
Convolutional neural network

ABSTRACT

In this paper, we present an automated approach for segmenting multiple sclerosis (MS) lesions from multi-modal brain magnetic resonance images. Our method is based on a deep end-to-end 2D convolutional neural network (CNN) for slice-based segmentation of 3D volumetric data. The proposed CNN includes a multi-branch down-sampling path, which enables the network to encode information from multiple modalities separately. Multi-scale feature fusion blocks are proposed to combine feature maps from different modalities at different stages of the network. Then, multi-scale feature upsampling blocks are introduced to upsize combined feature maps to leverage information from lesion shape and location. We trained and tested the proposed model using orthogonal plane orientations of each 3D modality to exploit the contextual information in all directions. The proposed pipeline is evaluated on two different datasets: a private dataset including 37 MS patients and a publicly available dataset known as the ISBI 2015 longitudinal MS lesion segmentation challenge dataset, consisting of 14 MS patients. Considering the ISBI challenge, at the time of submission, our method was amongst the top performing solutions. On the private dataset, using the same array of performance metrics as in the ISBI challenge, the proposed approach shows high improvements in MS lesion segmentation compared with other publicly available tools.

1. Introduction

Multiple sclerosis (MS) is a chronic, autoimmune and demyelinating disease of the central nervous system causing lesions in the brain tissues, notably in white matter (WM) (Steinman, 1996). Nowadays, magnetic resonance imaging (MRI) scans are the most common solution to visualize these kind of abnormalities owing to their sensitivity to detect WM damage (Compston and Coles, 2008).

Precise segmentation of MS lesions is an important task for understanding and characterizing the progression of the disease (Rolak, 2003). To this aim, both manual and automated methods are used to compute the total number of lesions and total lesion volume. Although manual segmentation is considered the gold standard (Simon et al., 2006), this method is a challenging task as delineation of 3-dimensional (3D) information from MRI modalities is time-consuming, tedious and prone to intra- and inter-observer variability (Sweeney et al., 2013). This motivates machine learning (ML) experts to develop automated lesion segmentation techniques, which can be orders of magnitude faster and immune to expert bias.

Among automated methods, supervised ML algorithms can learn from previously labeled training data and provide high performance in MS lesion segmentation. More specifically, traditional supervised ML methods rely on hand-crafted or low-level features. For instance, Cabezas et al. (2014) exploited a set of features, including intensity channels (fluid-attenuated inversion-recovery (FLAIR), proton density-weighted (PDw), T1-weighted (T1w), and T2-weighted (T2w)), probabilistic

* Corresponding author. Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, Italy..
** Corresponding author. Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, Italy..
E-mail addresses: shahab.aslani@iit.it (S. Aslani), diego.sona@iit.it (D. Sona).

https://doi.org/10.1016/j.neuroimage.2019.03.068
Received 16 November 2018; Received in revised form 23 March 2019; Accepted 28 March 2019
Available online 3 April 2019

tissue atlases (WM, grey matter (GM), and cerebrospinal fluid (CSF)), a map of outliers with respect to these atlases (Schmidt et al., 2012), and a set of low-level contextual features. A Gentleboost algorithm (Friedman et al., 2000) was then used with these features to segment multiple sclerosis lesions through a voxel by voxel classification.

During the last decade, deep learning methods, especially convolutional neural networks (CNNs) (LeCun et al., 1998), have demonstrated outstanding performance in biomedical image analysis. Unlike traditional supervised ML algorithms, these methods can learn by themselves how to design features directly from data during the training procedure (LeCun et al., 2015). They provided state-of-the-art results in different problems such as segmentation of neuronal structures (Ronneberger et al., 2015), retinal blood vessel extraction (Liskowski and Krawiec, 2016), cell classification (Han et al., 2016), brain extraction (Kleesiek et al., 2016), brain tumor (Havaei et al., 2017), tissue (Moeskops et al., 2016), and MS lesion segmentation (Valverde et al., 2017).

In particular, CNN-based biomedical image segmentation methods can be categorized into two different groups: patch-based and image-based methods. In patch-based methods, a moving window scans the image generating a local representation for each pixel/voxel. Then, a CNN is trained using all extracted patches, classifying the central pixel/voxel of each patch as a healthy or unhealthy region. These methods are frequently used in biomedical image analysis since they considerably increase the amount of training samples. However, they suffer of an increased training time due to repeated computations over the overlapping features of the sliding window. Moreover, they neglect the information over the global structure because of the small size of patches (Tseng et al., 2017).

On the contrary, image-based approaches process the entire image exploiting the global structure information (Tseng et al., 2017; Brosch et al., 2016). These methods can be further categorized into two groups according to the processing of the data: slice-based segmentation of 3D data (Tseng et al., 2017) and 3D-based segmentation (Brosch et al., 2016).

In slice-based segmentation methods, each 3D image is converted to its 2D slices, which are then processed individually. Subsequently, the segmented slices are concatenated together to reconstruct the 3D volume. However, in almost all proposed pipelines based on this approach, the segmentation is not accurate, most likely because the method ignores part of the contextual information (Tseng et al., 2017).

In 3D-based segmentation, a CNN with 3D kernels is used for extracting meaningful information directly from the original 3D image. The main significant disadvantage of these methods is related to the training procedure, which usually fits a large number of parameters with a high risk of overfitting in the presence of small datasets. Unfortunately, this is a quite common situation in biomedical applications (Brosch et al., 2016). To overcome this problem, recently, 3D cross-hair convolution has been proposed (Liu et al., 2017; Tetteh et al., 2018), where three 2D filters are defined for each of the three orientations around a voxel (each one is a plane orthogonal to X, Y, or Z axis). Then, the sum of the result of the three convolutions is assigned to the central voxel. The most important advantage of the proposed idea is the reduced number of parameters, which makes training faster than a standard 3D convolution. However, compared to standard 2D convolution (slice-based), still, there are three times more parameters for each layer, which increases the chance of overfitting in small datasets.

1.1. Related works

The literature offers some methods based on CNNs for MS lesion segmentation. For example, Vaidya et al. (2015) proposed a shallow 3D patch-based CNN using the idea of sparse convolution (Li et al., 2014) for effective training. Moreover, they added a post-processing stage, which increased the segmentation performance by applying a WM mask to the output predictions. Ghafoorian and Platel (2015) developed a deep CNN based on 2D patches in order to increase the number of the training samples and avoid the overfitting problems of 3D-based approaches. Similarly, in (Birenbaum and Greenspan, 2016), multiple 2D patch-based CNNs have been designed to take advantage of the common information within longitudinal data. Valverde et al. (2017) proposed a pipeline relying on a cascade of two 3D patch-based CNNs. They trained the first network using all extracted patches, and the second network was used to refine the training procedure utilizing misclassified samples from the first network. Roy et al. (2018) proposed a 2D patch-based CNN including two pathways. They used different MRI modalities as input for each pathway and the outputs were concatenated to create a membership function for lesions. Recently, Hashemi et al. (2018) proposed a method relying on a 3D patch-based CNN using the idea of a densely connected network. They also developed an asymmetric loss function for dealing with highly unbalanced data. Despite the fact that all the proposed patch-based techniques have good segmentation performance, they suffer from lacking global structural information. This means that global structure of the brain and the absolute location of lesions are not exploited during the segmentation.

In contrast, Brosch et al. (2016) developed a whole-brain segmentation method using a 3D CNN. They used single shortcut connection between the coarsest and the finest layers of the network, which enables the network to concatenate the features from the deepest layer to the shallowest layer in order to learn information about the structure and organization of MS lesions. However, they did not exploit middle-level features, which have been shown to have a considerable impact on the segmentation performance (Ronneberger et al., 2015).

1.2. Contributions

In this paper, we propose a novel deep learning architecture for automatic MS lesion segmentation consisting of a multi-branch 2D convolutional encoder-decoder network. In this study, we concentrated on whole-brain slice-based segmentation in order to prevent both the overfitting present in 3D-based segmentation (Brosch et al., 2016) and the lack of global structure information in patch-based methods (Ghafoorian et al., 2017; Valverde et al., 2017; Roy et al., 2018). We designed an end-to-end encoder-decoder network including a multi-branch downsampling path as the encoder, a multi-scale feature fusion and the multi-scale upsampling blocks as the decoder.

In the encoder, each branch is assigned to a specific MRI modality in order to take advantage of each modality individually. During the decoding stage of the network, different scales of the encoded attributes related to each modality, from the coarsest to the finest, including the middle-level attributes, were combined together and upconvolved gradually to get fine details (more contextual information) of the lesion shape. Moreover, we used three different (orthogonal) planes for each 3D modality as an input to the network to better exploit the contextual information in all directions. In summary, the main contributions in this work are:

- A whole-brain slice-based approach to exploit the overall structural information, combined with a multi-plane strategy to take advantage of full contextual information.
- A multi-level feature fusion and upsampling approach to exploit contextual information at multiple scales.
- The evaluation of different versions of the proposed model so as to find the most performant combination of MRI modalities for MS lesion segmentation.
- The demonstration of top performance on two different datasets.

2. Material

In order to evaluate the performance of the proposed method for MS lesion segmentation, two different datasets were used: the publicly available ISBI 2015 Longitudinal MS Lesion Segmentation Challenge dataset (Carass et al., 2017) (denoted as the ISBI dataset), and an

in-house dataset from the neuroimaging research unit (NRU) in Milan (denoted as the NRU dataset).

2.1. ISBI 2015 Longitudinal MS lesion segmentation challenge

The ISBI dataset included 19 subjects divided into two sets, 5 subjects in the training set and 14 subjects in the test set. Each subject had different time-points, ranging from 4 to 6. For each time-point, T1w, T2w, PDw, and FLAIR image modalities were provided. The volumes were composed of 182 slices with FOV = 182 × 256 and 1-mm cubic voxel resolution. All images available were already segmented manually by two different raters, therefore representing two ground truth lesion masks. For all 5 training images, lesion masks were made publicly available. For the remaining 14 subjects in the test set, there was no publicly available ground truth. The performance evaluation of the proposed method over the test dataset was done through an online service by submitting the binary masks to the challenge[1] website (Carass et al., 2017).

2.2. Neuroimaging research unit

The NRU dataset was collected by a research team from Ospedale San Raffaele, Milan, Italy.

It consisted of 37 MS patients (22 females and 15 males) with mean age 44.6 ± 12.2 years. The patient clinical phenotypes were 24 relapsing remitting MS, 3 primary progressive MS and 10 secondary progressive MS. The mean Expanded Disability status Scale (EDSS) was 3.3 ± 2, the mean disease duration was 13.1 ± 8.7 years and the mean lesion load was 6.2 ± 5.7 ml. The dataset was acquired on a 3.0 T Philips Ingenia CX scanner (Philips Medical Systems) with standardized procedures for subjects positioning.

The following sequences were collected: Sagittal 3D FLAIR sequence, FOV = 256 × 256, pixel size = 1 × 1 mm, 192 slices, 1-mm thick; Sagittal 3D T2w turbo spin echo (TSE) sequence, FOV = 256 × 256, pixel size = 1 × 1 mm, 192 slices, 1-mm thick; Sagittal 3D high resolution T1w, FOV = 256 × 256, pixel size = 1 × 1 mm, 204 slices, 1-mm thick.

For the validation of the NRU dataset, two different readers, with more than 5 years of experience in manual T2 hyperintense MS lesion segmentation performed the lesion delineation blinded to each other's results. We estimated the agreement between the two expert raters by using the Dice similarity coefficient (*DSC*) as a measure of the degree of overlap between the segmentations, and we found a mean *DSC* of 0.87. Differently from ISBI dataset, the two masks created by the two expert raters were used to generate a high quality "gold standard" mask by the intersection of the two binary masks from the two raters, which was used for all experiments with this dataset. This was to follow the common clinical practice of considering a single consensus mask between raters, which was particularly justified in our case due to the high *DSC* value between the two raters.

2.2.1. Ethical statement

Approval was received from the local ethical standards committee on human experimentation; written informed consent was obtained from all subjects prior to study participation.

3. Method

3.1. Data preprocessing

From the ISBI dataset, we selected the preprocessed version of the images available online at the challenge website. All images were already skull-stripped using Brain Extraction Tool (BET) (Smith, 2002), rigidly registered to the $1mm^3$ MNI-ICBM152 template (Oishi et al., 2008) using FMRIB's Linear Image Registration tool (FLIRT) (Jenkinson and Smith, 2001; Jenkinson et al., 2002) and N3 intensity normalized (Sled et al., 1998).

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

3.2. Network architecture

In this work, we propose a 2D end-to-end convolutional network based on the residual network (ResNet) (He et al., 2016). The core idea of ResNet is the use of identity shortcut connections, which allows for both preventing gradient vanishing and reducing computational complexity. Thanks to these benefits, ResNets have shown outstanding performance in computer vision problems, specifically in image recognition task (He et al., 2016).

We modified ResNet50 (version with 50 layers) to work as a pixel-level segmentation network. This has been obtained by changing the last prediction layer with other blocks and a dense pixel-level prediction layer inspired by the idea of the fully convolutional network (FCN) (Long et al., 2015). To exploit the MRI multi-modality analysis, we built a pipeline of parallel ResNets without weights sharing. Moreover, a multi-modal feature fusion block (MMFF) and a multi-scale feature upsampling block (MSFU) were proposed to combine and upsample the features from different modalities and different resolutions, respectively.

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

3.2.1. Input features preparation

For each MRI volume (and each modality), three different plane orientations (axial, coronal and sagittal) were considered in order to generate 2D slices along x, y, and z axes. Since the size of each slice depends on the orientation (axial = 182 × 218, coronal = 182 × 182, sagittal = 218 × 182), they were zero-padded (centering the brain) to obtain equal size (218 × 218) for each plane orientation. This procedure was applied to all three modalities. Fig. 1 illustrates the described procedure using FLAIR, T1w, and T2w modalities. This approach is similar to the one proposed in (Roth et al., 2014), where they used a 2.5D representation of 3D data.

3.2.2. Network architecture details

The proposed model essentially integrates multiple ResNets with other blocks to handle multi-modality and multi-resolution approaches, respectively. As can be seen in Fig. 2, the proposed network includes three main parts: downsampling networks, multi-modal feature fusion using MMFF blocks, and multi-scale upsampling using MSFU blocks.

In the downsampling stage, multiple parallel ResNets (without weights sharing) are used for extracting multi-resolution features, with each ResNet associated to one specific modality (in our experiments, we used FLAIR, T1w, and T2w). In the original ResNet50 architecture, the first layer is composed of a 7 × 7 convolutional layer with stride 2 to downsample the input by an order of 2. Then, a 3 × 3 max pooling layer with stride 2 is applied to further downsample the input followed by a bottleneck block without downsampling. Subsequently, three other bottleneck blocks are applied, each one followed by a downsampling convolutional layer with stride 2.

Therefore, ResNet50 can be organized into five blocks according to

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

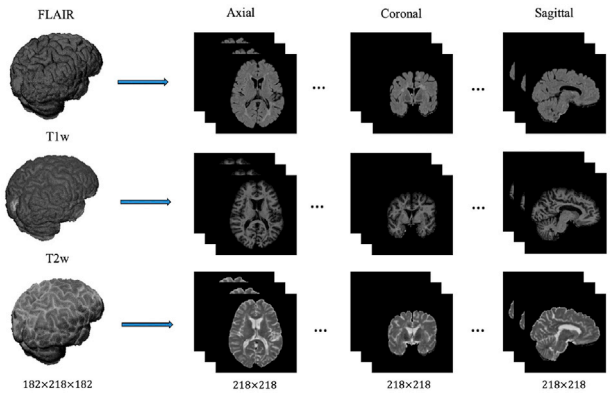

Fig. 1. Input features preparation. For each subject, three MRI modalities (FLAIR, T1w, and T2w) were considered. 2D slices related to the orthogonal views of the brain (axial, coronal and sagittal planes) were extracted from each modality. Since the size of extracted slices was different with respect to the plane orientations (axial = 182 × 218, coronal = 182 × 182, sagittal = 218 × 182), all slices were zero-padded while centering the brain so to obtain all slices with the same size (218 × 218), no matter their orientation.

the resolution of the generated feature maps (109 × 109, 54 × 54, 27 × 27, 14 × 14, and 7 × 7). Thanks to this organization, we can take advantage of the multi-resolution. Features with the same resolution from different modalities are combined using MMFF blocks as illustrated in Fig. 3(a). Each MMFF block includes 1 × 1 convolutions to reduce the number of feature maps (halving them), followed by 3 × 3 convolutions for adaptation. A simple concatenation layer is then used to combine the features from different modalities.

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

After the last MSFU block, a soft-max layer of size 2 is used to generate the output probability maps of the lesions. In our experiments the probabilistic maps were thresholded at 0.5 to generate binary classification for each pixel (lesion vs. non-lesion). It is important to mention that in all proposed blocks before each convolutional and upconvolutional layer, we use a batch normalization layer (Ioffe and Szegedy, 2015) followed by a rectifier linear unit activation function (Nair and Hinton, 2010). Size and number of feature maps in the input and output of all convolutional layers are kept the same.

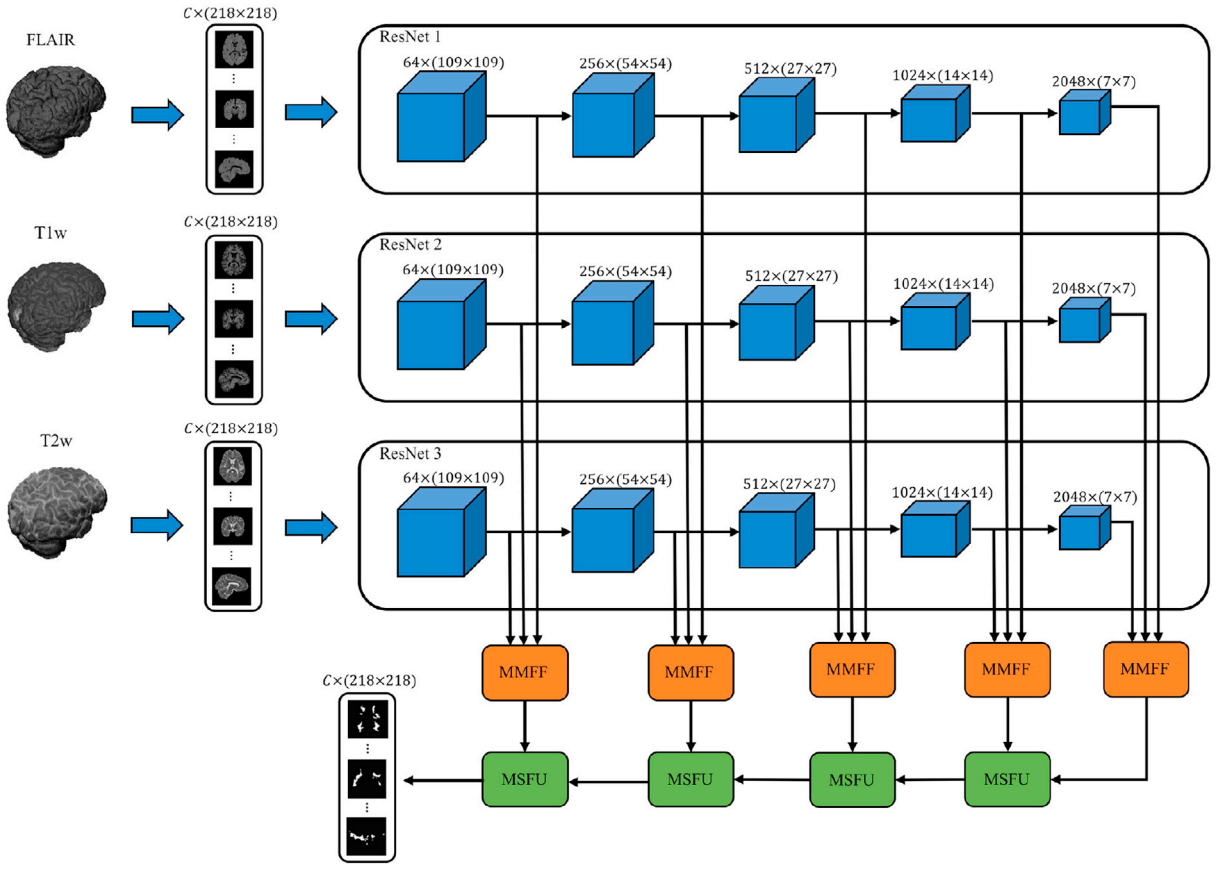

Fig. 2. General overview of the proposed method. Input data is prepared as described in Section 3.2.1, where volumes for each modality (FLAIR, T1w, and T2w) are described by slices (*C* is the total number of the slices along axial, coronal, and sagittal orientations, and 218 × 218 is their size after zero-padding). Data is presented in input by slices, and the model generates the corresponding segmented slices. The downsampling part of the network (blue blocks) includes three parallel ResNets without weight sharing, each branch for one modality (in this Figure, we used three modalities: FLAIR, T1w, and T2w). Each ResNet can be considered composed by 5 blocks according to the resolution of the representations. For example, the first block denotes 64 representations with resolution 109 × 109. Then, MMFF blocks are used to fuse the representations with the same resolution from different modalities. Finally, the output of MMFF blocks is presented as input to MSFU blocks, which are responsible for upsampling the low-resolution representations and for combining them with high-resolution representations.

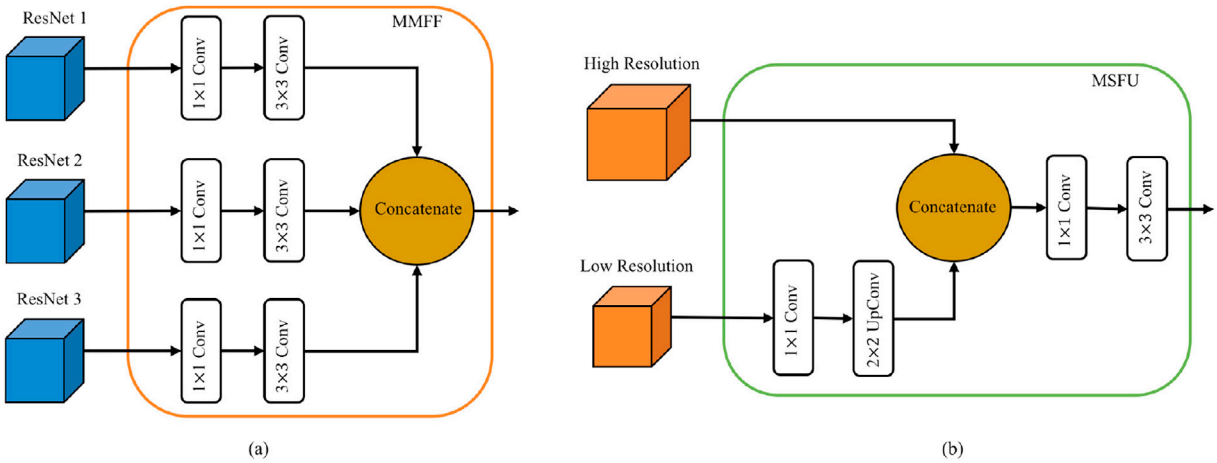

Fig. 3. Building blocks of the proposed network. a) MMFF block is used to combine representations from different modalities (FLAIR, T1w, and T2w) at the same resolution. b) MSFU block is used to upsample low-resolution features and combine them with higher-resolution features.

3.2.3. Implementation details

The proposed model was implemented in Python language[2] using Keras[3] (Chollet et al., 2015) with Tensorflow[4] (Abadi et al., 2015) backend. All experiments were done on a Nvidia GTX Titan X GPU. Our multi-branch slice-based network was trained end-to-end. In order to train the proposed CNN, we created a training set including the 2D slices from all three orthogonal views of the brain, as described in Section 3.2.1. Then, to limit extremely unbalanced data and omit uninformative samples, a training subset was determined by selecting only slices containing at least one pixel labeled as lesion. Considering that for each subject in the ISBI dataset, there were 4–6 recordings, the number of slices selected per subject ranged approximately from 1500 to 2000. In the NRU dataset, the number of slices ranged approximately from 150 to 300 per subject.

To optimize the network weights and early stopping criterion, the created training set was divided into training, and validation subsets, depending on the experiments described in the following Section (In all experiments, the split was performed on the subject base, to simulate a real clinical condition). We trained our network using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0001 multiplied by 0.95 every 400 steps. The size of mini-batches was fixed at 15 and each mini-batch included random slices from different orthogonal views. The maximum number of training epochs was fixed to 1000 for all experiments, well beyond the average converging rate. Fig. 5 illustrates an example of performance evolution during training of the network in terms of mean *DSC* (refer to 4.1 for details). Indeed, a performance plateau was systematically observed over all experiments before 1000 epochs. The best model was then selected according to the validation set. In the case shown on Fig. 5, the best performance was obtained at epoch 810. The training computation time for 1000 epochs was approximately 36 h.

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

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

where $p_i \in [0, \ldots, 1]$ is the predicted value of the soft-max layer and g_i is the ground truth binary value for each pixel i.

We slightly modified the original soft dice loss (Milletari et al., 2016) by replacing (-Dice) with (1-Dice) for visualization purposes. Indeed, the new equation returns positive values in the range $[0, \ldots, 1]$. This change does not impact the optimization.

3.2.4. 3D binary image reconstruction

Output binary slices of the network are concatenated to form a 3D volume matching the original data. In order to reconstruct the 3D image from the output binary 2D slices, we proposed a multi-planes reconstruction (MPR) block. Feeding each 2D slice to the network, we get as output the associated 2D binary lesion classification map. Since each original modality is duplicated three times in the input, once for each slice orientation (coronal, axial, sagittal), concatenating the binary lesion maps belonging to the same orientation results in three 3D lesion classification maps. To obtain a single lesion segmentation volume, these three lesion maps are combined via majority voting (the most frequent lesion classification are selected) as illustrated in Fig. 4. To justify the choice of majority voting instead of other label fusion methods, refer to Appendix B.

3.3. Data and code availability statement

The NRU dataset is a private clinical dataset and can not be made publicly available due to confidentiality. The code will be made available to anyone contacting the corresponding authors.

4. Experiments

4.1. Evaluation metrics

The following measures were used to evaluate and compare our model with other state-of-the-art methods.

- Dice Similarity Coefficient:

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

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

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

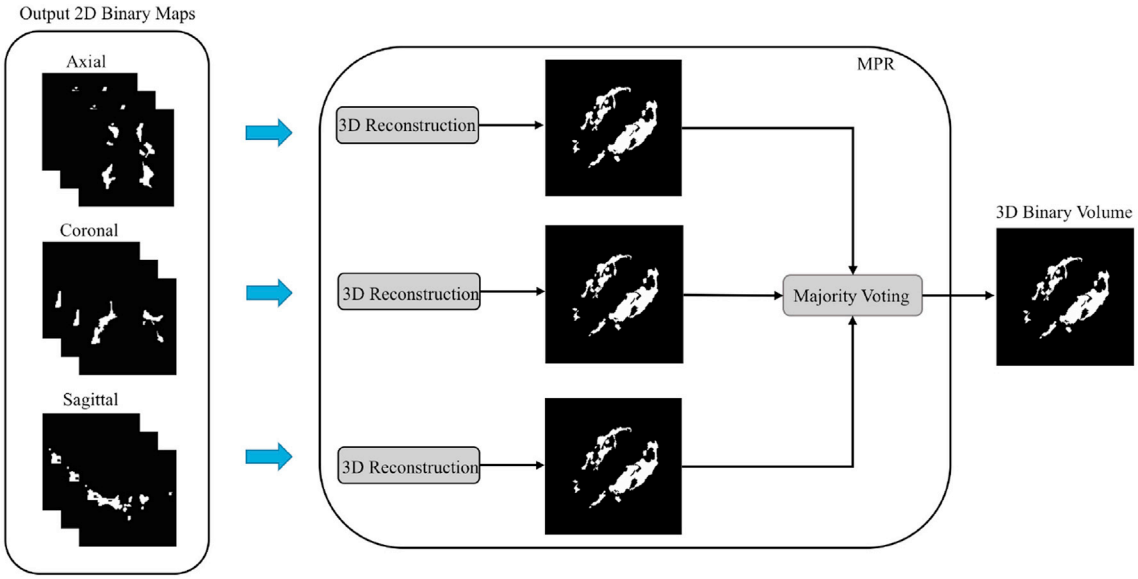

Fig. 4. The MPR block produces a 3D volumetric binary map by combining the 2D output binary maps of the network. First, the output 2D binary maps associated to each plane orientation (axial, coronal, and sagittal) are concatenated to create three 3D binary maps. Then, a majority vote is applied to obtain a single lesion segmentation volume.

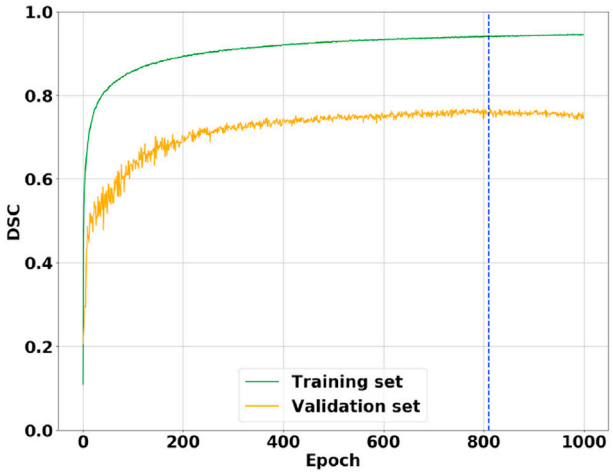

Fig. 5. Example of *DSC* metric dynamics (eq. (2)) during training on ISBI dataset. Experimentally, we found that a performance plateau was systematically reached before 1000 training epochs. To avoid overfitting, the best model was selected according to the validation set performance. In this specific experiment (training: subjects 1 to 4, validation: subject 5), the best model was selected based at epoch 810, which corresponded to the performance peak on validation set.

- Lesion-wise True Positive Rate:

$$LTPR = \frac{LTP}{RL} \tag{3}$$

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

- Lesion-wise False Positive Rate:

$$LFPR = \frac{LFP}{PL} \tag{4}$$

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

- Average Symmetric Surface Distance:

$$SD = \frac{1}{|N_{gt}| + |N_s|} \cdot \left(\sum_{x \in N_{gt}} \min_{y \in N_s} d(x,y) + \sum_{x \in N_s} \min_{y \in N_{gt}} d(x,y) \right) \tag{5}$$

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

- Hausdorff Distance:

$$HD = \max \left\{ \max_{x \in N_{gt}} \min_{y \in N_s} d(x,y), \max_{x \in N_s} \min_{y \in N_{gt}} d(x,y) \right\} \tag{6}$$

As described in (Carass et al., 2017), the ISBI challenge website provides a report on the submitted test set including some measures such as:

- Positive Prediction Value:

$$PPV = \frac{TP}{TP + FP} \tag{7}$$

- Absolute Volume Difference:

$$VD = \frac{|TP_s - TP_{gt}|}{TP_{gt}} \tag{8}$$

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

- Overall evaluation score:

$$SC = \frac{1}{|R| \cdot |S|} \cdot \sum_{R,S} \left(\frac{DSC}{8} + \frac{PPV}{8} + \frac{1 - LFPR}{4} + \frac{LTPR}{4} + \frac{Cor}{4} \right) \tag{9}$$

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

4.2. Experiments on the ISBI dataset

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

Since the ground truth was available only for the training set, in the first experiment, we ignored the official ISBI test set. We only considered data with available ground truth (training set with 5 subjects) as mentioned in (Brosch et al., 2016). To obtain a fair result, we tested our approach with a nested leave-one-subject-out cross-validation (3 subjects for training, 1 subject for validation and 1 subject for testing - refer to Appendix A for more details). To evaluate the stability of the model, this experiment was performed evaluating separately our method on the two sets of masks provided by the two raters.

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

4.3. Experiment on the NRU dataset

To test the robustness of the proposed model, we performed two experiments using the NRU dataset including 37 subjects. In the first experiment, we implemented a nested 4-fold cross-validation over the whole dataset (21 subjects for training, 7 subjects for validation and 9 subjects for testing - refer to A for more details). Since for each test fold we had an ensemble of four nested trained models, the prediction on each test fold was obtained as a majority vote of the corresponding ensemble. To justify the use of majority voting instead of other label fusion methods, we repeated the same experiment using different volume aggregation methods (refer to Appendix B for more details).

For comparison, we tested three different publicly available MS lesion segmentation software: OASIS (Automated Statistic Inference for Segmentation) (Sweeney et al., 2013), TOADS (Topology reserving Anatomy Driven Segmentation) (Shiee et al., 2010), and LST (Lesion Segmentation Toolbox)(Schmidt et al., 2012). OASIS generates the segmentation exploiting information from FLAIR, T1w, and T2w modalities, and it only requires a single thresholding parameter, which was optimized to obtain the best *DSC*. TOADS does not need parameter tuning and it only requires FLAIR and T1w modalities for segmentation. Similarly, LST works with FLAIR and T1w modalities only. However, it needs a single thresholding parameter that initializes the lesion segmentation. This parameter was optimized to get the best *DSC* in this experiment.

We also tested the standard 2D U-Net (Ronneberger et al., 2015), repeating the training protocol described in Appendix A. Indeed, we used the same training set as described in Section 3.2.1 and 3.2.3, with the difference that 2D slices from all modalities were aggregated in multiple channels. This network was trained using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0001 multiplied by 0.9 every 800 steps. For the sake of comparison, optimization was performed on the soft Dice Loss function (eq. (1)) (Milletari et al., 2016). To get the 3D volume from output binary slices of the network, we used the proposed MPR block as described in Section 3.2.4.

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

As outlined in Section 2.2, while for the ISBI dataset, we evaluated our method on two separate sets of masks, one for each rater, in the NRU dataset, we considered the manual consensus segmentation as a more robust gold standard against which to validate the proposed method. Nevertheless, to evaluate the stability of the model trained with the gold standard labeling, we also tested it separately on the two sets of masks (refer to Appendix C for more details).

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

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

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

5. Results

5.1. ISBI dataset

In the first experiment, we evaluated our model using three measures: *DSC*, *LTPR*, and *LFPR* to make our results comparable to those obtained in (Brosch et al., 2016; Maier and Handels, 2015; Aslani et al., 2019). Table 1 summarizes the results of the first experiment when comparing our model with previously proposed methods. The table shows the mean *DSC*, *LTPR*, and *LFPR*. As can be seen in that table, our method outperformed other methods in terms of *DSC* and *LFPR*, while the highest *LTPR* was achieved by our recently published method (Aslani et al., 2019). Fig. 6 shows the segmentation outputs of the proposed method for subject 2 (with high lesion load) and subject 3 (with low lesion load) compared to both ground truth annotations (rater 1 and rater 2).

In the second experiment, the official ISBI test set was used. Indeed, all 3D binary output masks computed on the test set were submitted to the ISBI website. Several measures were calculated online by the challenge website. Table 2 shows the results on all measures reported as a mean across raters. At the time of the submission, our method had an overall evaluation score of 92.12 on the official ISBI challenge web service,[5] making it amongst the top-ranked methods with a published paper or a technical report.

5.2. NRU dataset

Table 3 reports the results of the first experiment on NRU dataset

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

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

Method	Rater 1			Rater 2		
	DSC	LTPR	LFPR	DSC	LTPR	LFPR
Rater 1	–	–	–	0.7320	0.6450	0.1740
Rater 2	0.7320	0.8260	0.3550	–	–	–
Maier and Handels (2015) (GT1)	*0.7000*	0.5333	0.4888	*0.6555*	0.3777	*0.4444*
Maier and Handels (2015) (GT2)	*0.7000*	0.5555	*0.4888*	0.6555	0.3888	*0.4333*
Brosch et al. (2016) (GT1)	0.6844	*0.7455*	0.5455	0.6444	*0.6333*	0.5288
Brosch et al. (2016) (GT2)	0.6833	*0.7833*	0.6455	0.6588	*0.6933*	0.6199
Aslani et al. (2019) (GT1)	0.6980	**0.7460**	*0.4820*	0.6510	**0.6410**	0.4506
Aslani et al. (2019) (GT2)	0.6940	**0.7840**	0.4970	*0.6640*	**0.6950**	0.4420
Ours (GT1)	**0.7649**	0.6697	**0.1202**	**0.6989**	0.5356	**0.1227**
Ours (GT2)	**0.7646**	0.7002	**0.2022**	**0.7128**	0.5723	**0.1896**

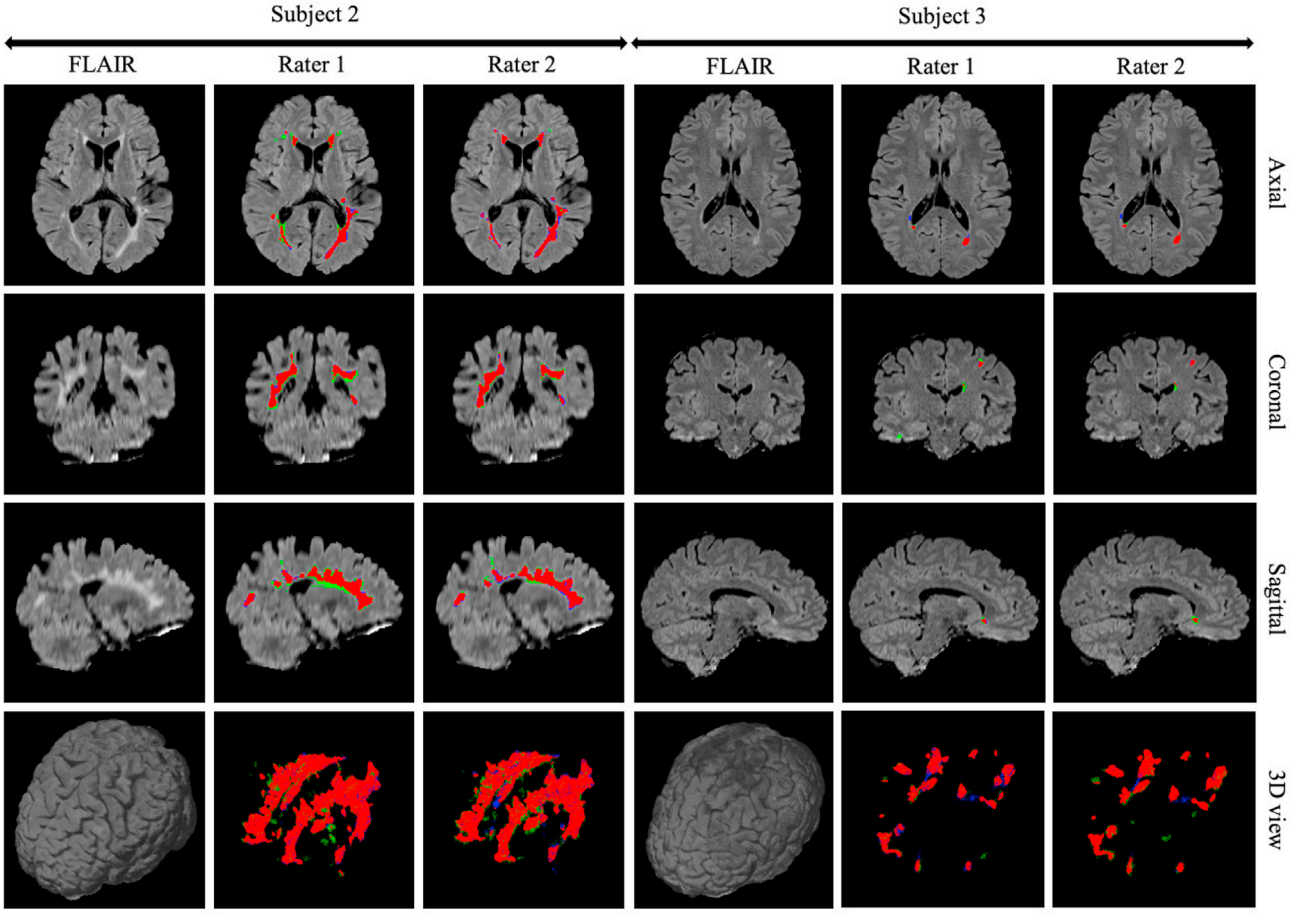

Fig. 6. Output segmentation results of the proposed method on two subjects of the ISBI dataset compared to ground truth annotations provided by rater 1 and rater 2. From left to right, the first three columns are related to subject 2 with high lesion load and reported *DSC* values of 0.8135 and 0.8555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

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

Fig. 7 features boxplots of the *DSC*, *LFPR*, *LTPR*, *PPV*, *VD*, *SD* and *HD* evaluation metrics obtained from the different methods and summarized in Table 3. This Figure shows statistically significant differences between model performances for most metrics and methods when compared to ours, after multiple comparison correction with the conservative Holm's method. The output segmentation of all methods applied to a random subject (with medium lesion load) can be seen with different plane orientations in Fig. 8.

Fig. 9 depicts the relationship between the volumes of all ground truth lesions and the corresponding estimated size for each evaluated method (one datapoint per lesion). With a qualitative evaluation, it can be seen that TOADS and OASIS methods tend to overestimate lesion volumes as many lesions are above the dashed black line, i.e., many

Table 2
Results related to the top-ranked methods (with published papers or technical reports) evaluated on the official ISBI test set and reported on the ISBI challenge website. *SC*, *DSC*, *PPV*, *LTPR*, *LFPR*, and *VD* are mean values across the raters. For detailed information about the metrics, refer to Section 4.1. Values in bold and italic refer to the metrics with the first-best and second-best performances, respectively.

Method	SC	DSC	PPV	LTPR	LFPR	VD
Hashemi et al. (2018)	**92.48**	0.5841	**0.9207**	0.4135	**0.0866**	0.4972
Ours	*92.12*	0.6114	*0.8992*	0.4103	*0.1393*	0.4537
Andermatt et al. (2017)	92.07	*0.6298*	0.8446	0.4870	0.2013	0.4045
Valverde et al. (2017)	91.33	**0.6304**	0.7866	0.3669	0.1529	**0.3384**
Maier and Handels (2015)	90.28	0.6050	0.7746	0.3672	0.2657	0.3653
Birenbaum and Greenspan (2016)	90.07	0.6271	0.7889	**0.5678**	0.4975	*0.3522*
Aslani et al. (2019)	89.85	0.4864	0.7402	0.3034	0.1708	0.4768
Deshpande et al. (2015)	89.81	0.5960	0.7348	0.4083	0.3075	0.3762
Jain et al. (2015)	88.74	0.5560	0.7300	0.3225	0.3742	0.3746
Sudre et al. (2015)	87.38	0.5226	0.6690	*0.4941*	0.6776	0.3837
Tomas-Fernandez and Warfield (2015)	87.01	0.4317	0.6973	0.2101	0.4115	0.5109
Ghafoorian et al. (2017)	86.92	0.5009	0.5491	0.4288	0.5765	0.5707

Table 3
Results related to the first NRU dataset experiment. Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for different methods. Values in bold and italic indicate the first-best and second-best results.

Method	DSC	PPV	LTPR	LFPR	VD	SD	HD
TOADS (Shiee et al., 2010)	0.5241	0.5965	**0.4608**	0.6277	*0.4659*	5.4392	13.60
LST (Schmidt et al., 2012)	0.3022	0.5193	0.1460	0.3844	0.6966	7.0919	14.35
OASIS (Sweeney et al., 2013)	0.4193	0.3483	0.3755	0.4143	2.0588	*3.5888*	18.33
U-NET (Ronneberger et al., 2015)	*0.6316*	*0.7748*	0.3091	*0.2267*	*0.3486*	3.9373	*9.235*
OURS	**0.6655**	**0.8032**	*0.4465*	**0.0842**	**0.3372**	**2.5751**	**6.728**

lesions are estimated larger than they really are. On the contrary, LST method tends to underestimate the lesion sizes. U-Net and our method, on the contrary, produced lesions with size more comparable to the ground truth. However, with a quantitative analysis, our model produced the slope closest to unity (0.9027) together with the highest Pearson correlation coefficient (0.75), meaning our model provided the stronger global agreement between estimated and ground truth lesion volumes (note that a better agreement between lesion volumes does not mean the segmented and ground truth lesions better overlap – the amount of overlap was measured with the *DSC*).

Table 4 shows the performance of the proposed model with respect to different combinations of modalities in the second experiment. The SB version of the proposed model used with one modality had noticeably better performance in almost all measures when using FLAIR modality. However, all modalities carry relevant information as better performance in most metrics was obtained when using a combination of modalities. In MB versions of the model, all possible two-branch and three-branch versions were considered. As shown in Table 4, two-branch versions including FLAIR modality showed a general better performance than the single-branch version using single modality. This emphasizes the importance of using FLAIR modality together with others (T1w and T2w). However, overall, a combination of all modalities in the three-branch version of the model showed the best general performance compared to the other versions of the network.

6. Discussion and conclusions

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

When having insufficient training data in deep learning based approaches, which is very common in the medical domain, transfer learning has demonstrated to be an adequate solution (Chen et al., 2015, 2016; Hoo-Chang et al., 2016). Not only it helps boosting the performance of the network but also it significantly reduces overfitting. Therefore, we used the parallel ResNet50s pre-trained on ImageNet as a multi-branch downsampling network while the other layers in MMFF and MSFU blocks were randomly initialized from a Gaussian distribution. We then fine-tuned the whole network on the given MS lesion segmentation task.

In brain image segmentation, a combination of MRI modalities overcomes the limitations of single modality approaches, allowing the models to provide more accurate segmentations (Kleesiek et al., 2016; Moeskops et al., 2016; Aslani et al., 2019). Unlike previously proposed deep networks (Brosch et al., 2016; Aslani et al., 2019), which stacked all modalities together as a single input, we designed a network with several downsampling branches, one branch for each individual modality. We believe that stacking all modalities together as a single input to a network is not an optimal solution since during the downsampling procedure, the details specific to the most informative modalities can vanish when mixed with less informative modalities. On the contrary, the multi-branch approach allows the network to abstract higher-level features at different granularities specific to each modality. Independently of the ground truth used for training and testing the model, results in Table 1 confirm our claim showing that a network with separate branches generated more accurate segmentations (e.g., $DSC = 0.7649$) than single-branch networks with all modalities stacked, as proposed by Brosch et al. (2016) (e.g., $DSC = 0.6844$) and Aslani et al. (2019) (e.g., $DSC = 0.6980$). Indeed, the mentioned methods (single-branch) generally obtained higher *LTPR* values (e.g., 0.7455 and 0.7460) than multi-branch (e.g., 0.6697). However, they also obtained very high *LFPR* values showing a significant overestimation of lesion volumes. The proposed method, instead, showed the best trade-off between *LTPR* and *LFPR*.

When examining the influence of different modalities, results in Table 4 demonstratesin Table 4 demonstrated that the most important modality for that the most important modality for MS lesion segmentation was FLAIR ($DSC > 0.65$). This is likely due to the fact that FLAIR sequences benefit from CSF signal suppression and hence provide a higher image contrast between MS lesions and the surrounding normal appearing WM. Using all modalities together in a SB network (by concatenating them as single multi-channel input) and in a MB network (each modality as single input to each branch) showed good segmentation performance. This could be due to the combination of modalities

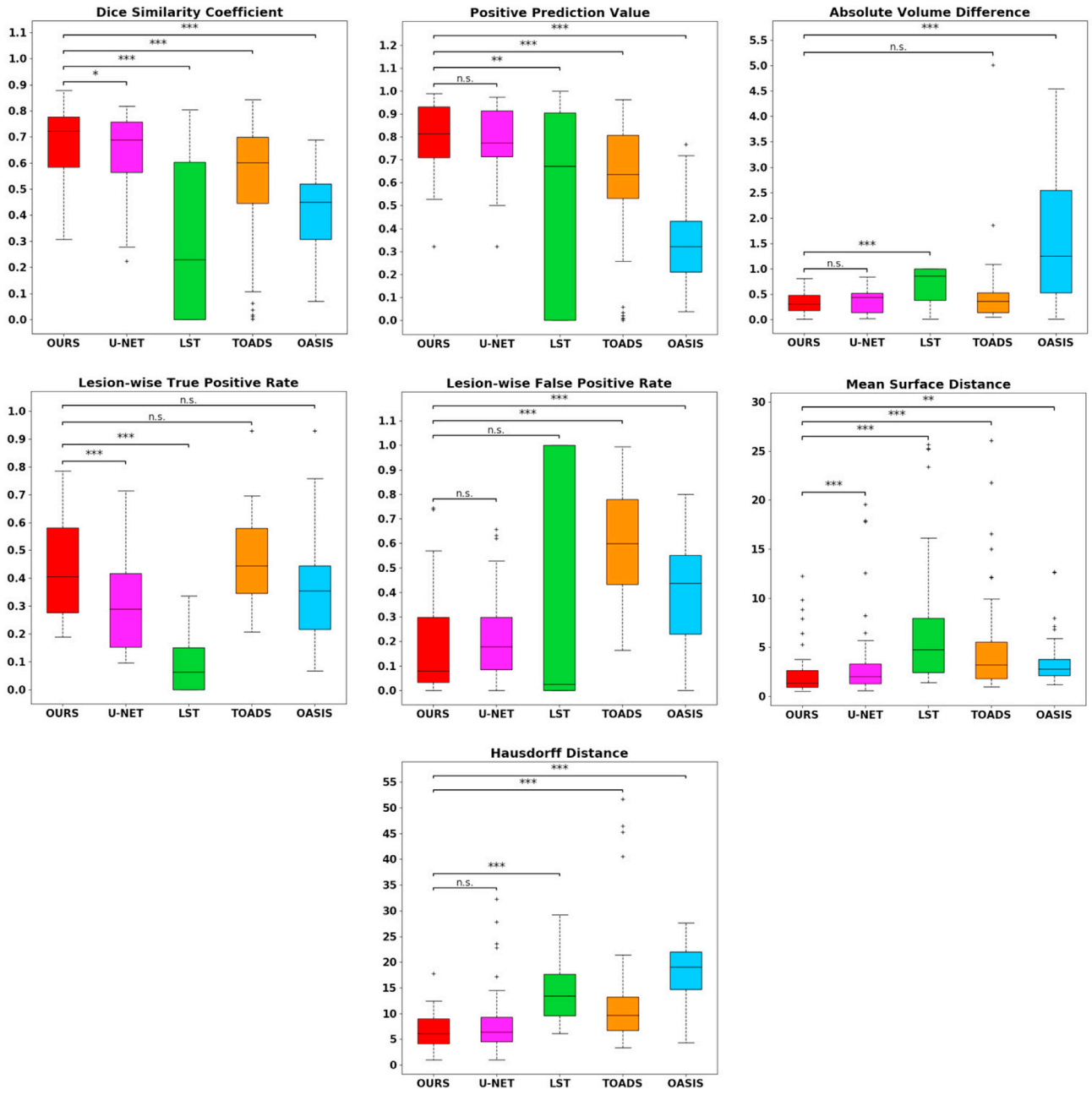

Fig. 7. Boxplots showing the performance of tested models with all measures on NRU dataset. Among all methods, the proposed one had the best trade-off between the lesion-wise true positive rate and lesion-wise false positive rate, while having the best mean value for dice similarity coefficient, positive prediction value, absolute volume differences, mean surface distance and hausdorff distance. Statistically significant differences between our method and the others were assessed using resampling statistics with multiple comparison correction. The significance threshold was set as $\alpha = 0.05$. p-values were annotated as follows: '*' for $p < 0.05$, '**' for $p < 0.005$, '***' for $p < 0.0005$, and 'n.s.' for non-significant values.

helping the algorithm identifying additional information regarding the location of lesions. However, supporting our claim that stacking all modalities together as a single input to the network is not an optimal solution, top performance, indeed, was obtained in most measures with the MB network when using all available modalities, as can be seen in Table 4.

In deep CNNs, attributes from different layers include different information. Coarse layers are related to high-level semantic information (category specific), and shallow layers are related to low-level spatial information (appearance specific) (Long et al., 2015), while middle layer attributes have shown a significant impact on segmentation performance (Ronneberger et al., 2015). Combining these multi-level attributes from the different stages of the network makes the representation richer than using single-level attributes, like in the CNN based method proposed by Brosch et al. (2016), where a single shortcut connection between the deepest and the shallowest layers was used. Our model, instead, includes several shortcut connections between all layers of the network, in order to combine multi-scale features from different stages of the network as inspired by U-Net architecture (Ronneberger et al., 2015). The results shown in Table 1 suggest that the combination of multi-level features during the upsampling procedure helps the network exploiting more contextual information associated to the lesions. This could explain why the performance of our proposed model ($DSC = 0.7649$) is higher than the method proposed by Brosch et al. (2016) ($DSC = 0.6844$).

Patch-based CNNs suffer from lacking spatial information about the lesions because of the patch size limitation. To deal with this problem, we

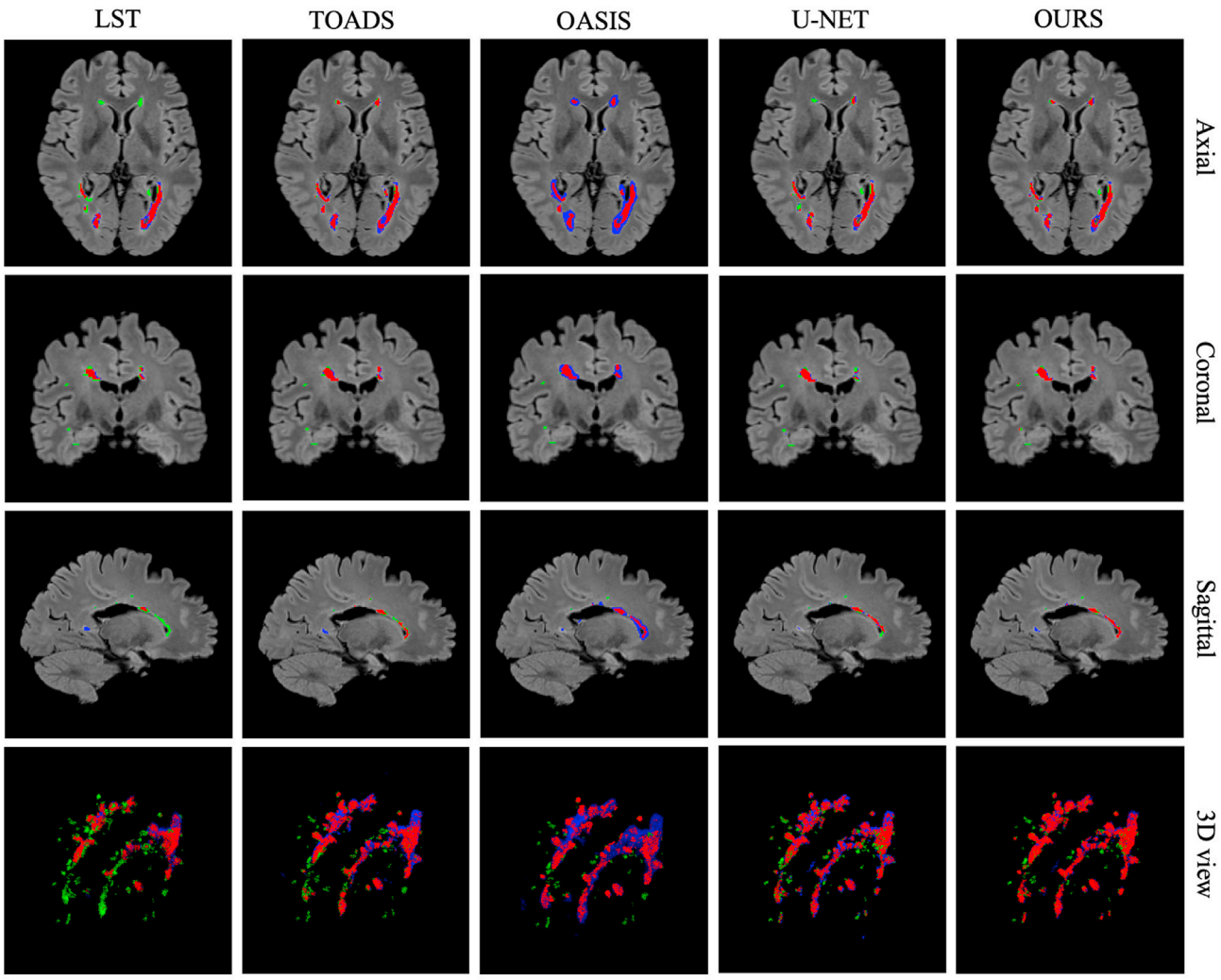

Fig. 8. Output segmentation results of the different methods for one subject with medium lesion load from the NRU dataset compared with ground truth annotation. Reported *DSC* values for TOADS, OASIS, LST, U-Net and our proposed method for this subject are 0.7110, 0.4266, 0.6505, 0.7290 and 0.7759, respectively. On all images, true positives, false negatives, and false positives are colored in red, green and blue, respectively.

proposed a whole-brain slice-based approach. Compared with patch-based methods (Valverde et al., 2017; Ghafoorian et al., 2017), we have shown that our model has better performance for most measures, as seen in Table 2. Although the CNN proposed by Valverde et al. (2017) had the highest *DSC* value among all, our method showed better performance regarding the *LTPR* and *LFPR*, which indicates that our model is robust in identifying the correct location of lesions. The method proposed by Birenbaum and Greenspan (2016) has been optimized to have the highest *LTPR*. However, their method showed significantly lower performance in *LFPR*. Compared with this method, our method has better trade-off between *LTPR* and *LFPR*.

As mentioned in (Carass et al., 2017), manual delineation of MS lesions from MRI modalities is prone to intra- and inter-observer variability, which explains the relatively low *DSC* between two experts delineating the same lesions (~0.73 for ISBI data as shown in Table 1). Automated methods are therefore expected to have a maximum performance in the same order of magnitude when comparing their generated segmentation with the rater's one. Accordingly, it is important to notice that, our model obtained a performance (*DSC*) close to the experts agreement, as can be seen in Table 1.

The proposed method also has some limitations. We observed that the proposed pipeline is slightly slow in segmenting a 3D image since segmenting whole-brain slices takes a longer time compared to other CNN-based approaches (Roy et al., 2018). The time required to segment a 3D image is proportional to the size of the image and is based on the computational cost of three sequential steps: input features preparation 3.2.1, slice-level segmentation 3.2.2, and 3D image reconstruction 3.2.4. In both the ISBI and NRU datasets, the average time for segmenting an input image with our model, including all 3 steps, was approximately 90 s.

A still open problem in MS lesion segmentation task is the identification of cortical and subcortical lesions. To this aim, we plan to use other MRI modalities such as double inversion recovery (DIR) sequences for the identification of cortical lesions, which benefits of the signal suppression from both CSF and WM. Moreover, we believe that introducing information from the tissue class could help improve the network identifying cortical, subcortical and white matter lesions. Therefore, we think that would be very promising to design a multi-task network for segmenting different parts of brain including different tissue types (WM, GM, CSF) and different types of MS lesions (including cortical lesions).

Since the assessment of the disease burden from MRI of MS patients requires the quantification of the volume of hyperintense lesions on T2-weighted images, the final goal of the method proposed was to obtain an automatic and robust MS lesion segmentation tool. This will be particularly useful to facilitate scaling advanced MS analysis based on myelin imaging (Dayan et al., 2017) or multi-modal characterization of white matter tracts (Dayan et al., 2016) to large datasets. The long term goal, more generally, is the translation of this automatic method into a clinical

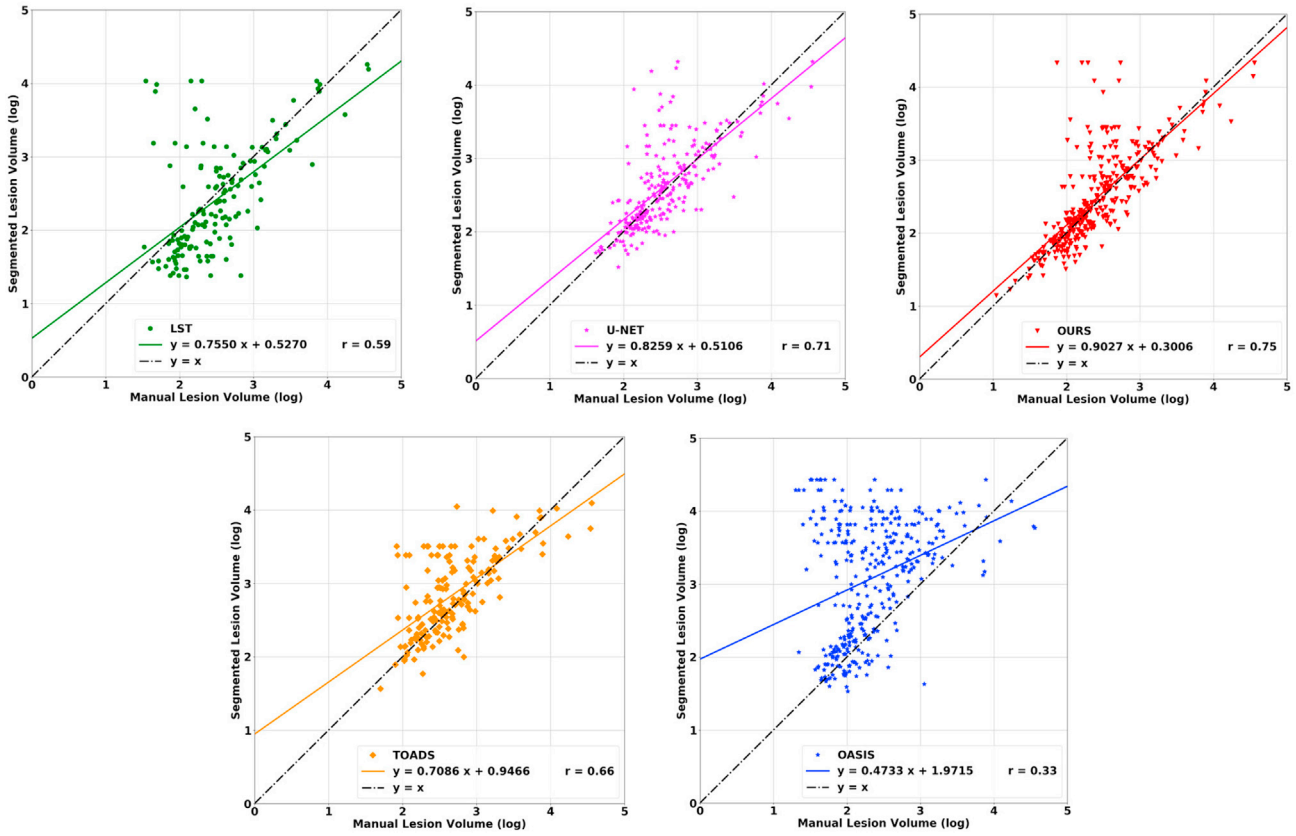

Fig. 9. Comparison of the lesion volumes produced by manual and automatic segmentation on the NRU dataset with different methods. Each point is associated with a single lesion. Colored (solid) lines indicate the correlation between manual and segmented lesion volumes. Black (dotted) lines indicate the ideal regression line. Slope, intercept, and Pearson's linear correlation (all with $p \ll 0.001$) between manual and estimated masks can also be seen for different methods.

Table 4
The proposed model was tested with different combinations of the three modalities in the second NRU dataset experiment. SB and MB denote the single-branch and multi-branch versions of the proposed model, respectively. Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for different methods. Values in bold and italic indicate the first-best and second-best values.

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

tool. However, to be fully ready for clinical applications, the method should be also validated on healthy subjects and in a longitudinal framework. The test on healthy subjects needs to be done to evaluate the amount of false positives generated by any approach on healthy brain scans. The experiments in a longitudinal framework are useful to assess the model reliability and capability to identify new, enlarged and stable lesions. Moreover, still exploiting ISBI dataset, which includes longitudinal data, we could focus on leveraging this information to boost the performance of segmentation.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgments

We acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Appendix A. Evaluation Protocols

This appendix includes 3 tables that describe the training procedures in details related to Sections 4.2 and 4.3.

Table A1 and Table A2 give detailed information about how we implemented training procedure on the ISBI dataset for the first and second experiments. Table A3 describes the first and second experiments. Table A3 describes the nested 4-fold cross-validation training procedure applied on the NRU dataset in the first experiment.

Table A.1
This table shows the implementation of first experiment in Section 4.2. In this experiment, we evaluated our model using the ISBI dataset with available ground truth (training set with 5 subjects). We implemented a nested leave-one-subject-out cross-validation (3 subjects for training, 1 subject for validation, and 1 subject for testing). The numbers indicate the subject identifier.

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

Table A.2
This table shows the implementation of the second experiment in Section 4.2. In this experiment, our model was evaluated using official ISBI test set including 14 subjects without publicly available ground truth. We trained our model doing a leave-one-subject-out cross-validation on whole training set (4 subject for training, 1 subject for validation, and 14 subject for testing). The numbers indicate the subject identifier.

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

Table A.3
This table gives detailed information regarding the training procedure for the first experiment in Section 4.3. In this experiment, we implemented a nested 4-fold cross-validation over the whole NRU dataset including 37 subjects. [A-B @ C-D] denotes subjects A to B and C to D.

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

Appendix B. Labels Aggregation

In order to aggregate the outcomes of ensembles of labeling, beyond majority voting, we tested alternative well known label fusion methods. Specifically, we repeated the first experiment on NRU dataset as described in Section 4.2 substituting the majority vote framework with averaging and STAPLE (Simultaneous Truth and Performance Level) (Warfield et al., 2004) methods, used to aggregate both the output volumes of the three plane orientations and the output volumes of the different models during cross-validation. Table B1 indicates the performance of each method. Overall, majority voting had better performance than other methods. Therefore, we selected this method for all experiments.

Table B.1
This table shows the results of the first experiment on the NRU dataset using our model as described in Section 4.2. We implemented the same experiment using different methods for fusing output volumes (when merging the outputs from each plane orientation, and also when merging the outputs of models from different cross-validation folds). Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for each method. Values in bold and italic indicate the first-best and second-best results.

Method	DSC	PPV	LTPR	LFPR	VD	SD	HD
Majority Voting	**0.6655**	*0.8032*	**0.4465**	*0.0842*	**0.3372**	2.575	**6.728**
Averaging	0.5883	**0.8391**	0.3220	**0.0788**	0.4625	3.216	*8.503*
STAPLE (Warfield et al., 2004)	*0.6632*	0.7184	*0.3989*	0.0802	*0.3883*	**2.330**	8.629

Appendix C. Rater Evaluation on NRU Dataset

In the first NRU dataset experiment, beyond verifying the quality of the proposed model on the ground truth generated from the consensus of two experts, we also compared the performance with the ground truth from each individual experts. The rationale behind the experiment was to assess the consistency of the system across raters. Table C1 shows the corresponding results. As expected from the high consensus between the masks provided by the two raters (as mentioned in Section 2.2), our trained model using the gold standard mask (derived from the two raters' masks) showed comparable results when evaluated with either raters' masks or the consensus mask as ground truth.

Table C.1
This table indicates the performance of our trained model in the NRU dataset first experiment when using different ground truth masks as testing. Mean values of *DSC*, *PPV*, *LTPR*, *LFPR*, *VD*, *SD* and *HD* were measured for each method. Values in bold and italic indicate the first-best and second-best results.

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

References

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

Andermatt, S., Pezold, S., Cattin, P.C., 2017. Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. In: International MICCAI Brainlesion Workshop. Springer, pp. 31–42.

Aslani, S., Dayan, M., Murino, V., Sona, D., 2019. Deep 2d encoder-decoder convolutional neural network for multiple sclerosis lesion segmentation in brain MRI. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, pp. 132–141. URL. https://doi.org/10.1007%2F978-3-030-11723-8_13.

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

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2016. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imaging 35 (5), 1229–1239.

Cabezas, M., Oliver, A., Valverde, S., Beltran, B., Freixenet, J., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Lladó, X., 2014. Boost: a supervised approach for multiple sclerosis lesion segmentation. J. Neurosci. Methods 237, 108–117.

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

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

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

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

Compston, A., Coles, A., 2008. Multiple sclerosis. Lancet 372 (9648), 1502–1517.

Dayan, M., Hurtado Rúa, S.M., Monohan, E., Fujimoto, K., Pandya, S., LoCastro, E.M., Vartanian, T., Nguyen, T.D., Raj, A., Gauthier, S.A., 2017. Mri analysis of white matter myelin water content in multiple sclerosis: a novel approach applied to finding correlates of cortical thinning. Front. Neurosci. 11, 284.

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

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

Friedman, J., Hastie, T., Tibshirani, R., et al., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. 28 (2), 337–407.

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

Ghafoorian, M., Platel, B., 2015. Convolutional neural networks for ms lesion segmentation, method description of diag team. In: Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, 1–2.

Han, X.-H., Lei, J., Chen, Y.-W., 2016. Hep-2 cell classification using k-support spatial pooling in deep cnns. In: International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 3–11.

Hashemi, S.R., Salehi, S.S.M., Erdogmus, D., Prabhu, S.P., Warfield, S.K., Gholipour, A., 2018. Tversky as a Loss Function for Highly Unbalanced Image Segmentation Using 3d Fully Convolutional Deep Networks. CoRR Abs/1803, p. 11078. URL. http://arxiv.org/abs/1803.11078.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.

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

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

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv Preprint arXiv:1502.03167.

Jain, S., Sima, D.M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., De Mey, J., Barkhof, F., Steenwijk, M.D., Daams, M., et al., 2015. Automatic segmentation and volumetry of multiple sclerosis brain lesions from mr images. Neuroimage: Clinic 8, 367–375.

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

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. CoRR Abs/1412, p. 6980. URL. http://arxiv.org/abs/1412.6980.

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

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Li, H., Zhao, R., Wang, X., 2014. Highly Efficient Forward and Backward Propagation of Convolutional Neural Networks for Pixelwise Classification. arXiv Preprint arXiv:1412.4526.

Liskowski, P., Krawiec, K., 2016. Segmenting retinal blood vessels with<? pub _newline?> deep neural networks. IEEE Trans. Med. Imaging 35 (11), 2369–2380.

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

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Maier, O., Handels, H., 2015. Ms lesion segmentation in mri with random forests. In: Proc. 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, 1–2.

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

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J., Išgum, I., 2016. Automatic segmentation of mr brain images with a convolutional neural network. IEEE Trans. Med. Imaging 35 (5), 1252–1261.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814.

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

Rolak, L.A., 2003. Multiple sclerosis: it is not the disease you thought it was. Clin. Med. Res. 1 (1), 57–60.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 520–527.

Roy, S., Butman, J.A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2018. Multiple Sclerosis Lesion Segmentation from Brain Mri via Fully Convolutional Neural Networks. arXiv Preprint arXiv:1803.09172.

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

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

Simon, J., Li, D., Traboulsee, A., Coyle, P., Arnold, D., Barkhof, F., Frank, J., Grossman, R., Paty, D., Radue, E., et al., 2006. Standardized mr imaging protocol for multiple sclerosis: consortium of ms centers consensus guidelines. Am. J. Neuroradiol. 27 (2), 455–461.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.

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

Sudre, C.H., Cardoso, M.J., Bouvy, W.H., Biessels, G.J., Barnes, J., Ourselin, S., 2015. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. IEEE Trans. Med. Imaging 34 (10), 2079–2102.

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

Tetteh, G., Efremov, V., Forkert, N.D., Schneider, M., Kirschke, J., Weber, B., Zimmer, C., Piraud, M., Menze, B.H., 2018. Deepvesselnet: Vessel Segmentation, Centerline Prediction, and Bifurcation Detection in 3-d Angiographic Volumes. arXiv Preprint arXiv:1803.09340.

Tomas-Fernandez, X., Warfield, S.K., 2015. A model of population and subject (mops) intensities with application to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imaging 34 (6), 1349–1361.

Tseng, K.-L., Lin, Y.-L., Hsu, W., Huang, C.-Y., 2017. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, pp. 3739–3746.

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

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

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.