Asymmetric Similarity Loss Function to Balance Precision and Recall in Highly Unbalanced Deep Medical Image Segmentation

Seyed Raein Hashemi[1,2], Seyed Sadegh Mohseni Salehi[1,3], *Student Member, IEEE,*
Deniz Erdogmus[3], *Senior Member, IEEE,* Sanjay P. Prabhu[1],
Simon K. Warfield[1], *Senior Member, IEEE*, and Ali Gholipour[1], *Senior Member, IEEE*

[1]Computational Radiology Laboratory, Boston Children's Hospital, and Harvard Medical School, Boston MA 02115
[2]Computer and Information Science Department, Northeastern University, Boston, MA, 02115
[3]Electrical and Computer Engineering Department, Northeastern University, Boston, MA, 02115

Fully convolutional deep neural networks have been asserted to be fast and precise frameworks with great potential in image segmentation. One of the major challenges in utilizing such networks raises when data is unbalanced, which is common in many medical imaging applications such as lesion segmentation where lesion class voxels are often much lower in numbers than non-lesion voxels. A trained network with unbalanced data may make predictions with high precision and low recall (sensitivity), being severely biased towards the non-lesion class which is particularly undesired in medical applications where false negatives are actually more important than false positives. Various methods have been proposed to address this problem including two step training, sample re-weighting, balanced sampling, and similarity loss functions. In this paper we propose a framework based on an asymmetric similarity loss function to mitigate the issue of data imbalance to achieve much better trade-off between precision and recall in training fully convolutional deep networks. To this end, we developed a patch-wise 3D densely connected network with an asymmetric loss function, where we used large overlapping image patches for intrinsic and extrinsic data augmentation, a patch selection algorithm, and a patch prediction fusion strategy based on B-spline weighted soft voting to take into account the uncertainty of prediction in patch borders. We applied this method to multiple sclerosis lesion segmentation based on the MSSEG 2016 and ISBI 2015 challenges, where we achieved average Dice similarity coefficient of 69.8% and 65.74%, respectively, using our proposed patch-wise 3D densely connected network. Our results show marked improvement over the results reported in the literature and those of an approach based on 3D U-Net in these challenges. Significant improvement in F_1 and F_2 scores and the area under the precision-recall curve was achieved in test using the asymmetric similarity loss layer and our 3D patch prediction fusion method. The asymmetric similarity loss function based on F_β scores generalizes the Dice similarity coefficient and can be effectively used with the patch-wise strategy developed here to train fully convolutional deep neural networks for highly unbalanced image segmentation.

***Index Terms*—Lesion segmentation, Asymmetric loss function, Convolutional neural network, DenseNet, Patch prediction fusion.**

This study was supported in part by the National Institutes of Health grants R01 NS079788 and R01 EB018988; and in part by a Technological Innovations in Neuroscience Award from the McKnight Foundation to A.Gholipour.
Corresponding author: S.R.Hashemi (email: hashemi.s@husky.neu.edu).

The trained model is available as a Docker image and can be pulled with this command: docker pull **raeinhashemi/msseg2016:v1**

I. Introduction

CONVOLUTIONAL neural networks have shown promising results in a wide range of applications including image segmentation. Recent medical image processing literature shows significant progress towards automatic segmentation of brain lesions [1], [2], tumors [3], [4], [5], and neuroanatomy [6], [7], [8] using 2D networks [3], [6], [9], and more recently using 3D network architectures [8], [2]. Fully convolutional networks (FCNs) with multi-scale skip connections, in particular, have shown great performance [9], [10], [11].

In this work, we considered automatic brain lesion segmentation in Multiple sclerosis (MS). MS is the most common disabling neurologic autoimmune disease resulting from recurrent attacks of inflammation in the central nervous system [12], [13]. Across the extensive literature for automated MS lesion segmentation, there are methods that try to alleviate the data imbalance issue by equal selection of training samples from each class [3], [14], whereas others propose using more persistent loss functions [1], [11], [15], both of which we combine together as a rigorous solution. As our first contribution in this work to deal with significantly unbalanced data, we investigate and compare the generality and performance of our proposed asymmetric loss function based on the F_β scores with the Dice similarity loss function recently proposed for medical image segmentation using FCNs [11].

In addition, we further diminish the problem of data imbalance by using patches that lead to relatively higher ratio of lesion versus non-lesion samples. Overlapping patches provide intrinsic data augmentation, make a better balance in data for training, and make the network adaptable for any size inputs with efficient memory usage in both test and training. We propose a patch prediction fusion strategy to take into account the prediction uncertainty in patch borders. In what follows, we review the state-of-the-art in MS lesion segmentation and the related work that motivated this study. Then we show two network architectures with our proposed loss function that generate accurate lesion segmentation compared to the literature according to several performance metrics.

II. Related Work

Many novel and genuine algorithms, methods, and models have been continuously developed and improved over the past

years on MS lesion segmentation. As the number of these methods grew, so did the desire for higher precision and more general solutions. In spite of the fact that there are lots of fully automated segmentation algorithms, the accuracy of these methods are not yet in an acceptable range, highlighting the difficulty of the problem. Therefore, lesion segmentation remains an active and important area of research.

The state-of-the-art MS lesion segmentation methods mostly use aggregations of skull stripping, bias correction, image registration, atlases, intensity feature information, data augmentation, and image priors or masks in training. The most recently proposed deep learning techniques for lesion segmentation include recurrent neural networks (RNN) with DropConnect [16], cascaded convolutional neural networks [17], [18], deep convolutional encoder networks [1], and independent image modality convolution pipelines [19]. There has also been other more classic supervised methods such as decision random forests [20], [21], non-local means [22], [23], and combined inference from patient and healthy populations [24]. One of the most recent techniques for the application of lesion segmentation, proposes the use of generalized dice overlap as a loss function [25] which assigns weights to different segmentation labels based on their quantity and volume in the training data. The other recent technique merges the two popular architectures of Unet and DenseNet while forming a hybrid structure [26] for liver and tumor segmentation.

In this study, we propose an asymmetric similarity loss function based on F_β scores to train deep fully convolutional neural networks using two network architectures: the U-net [15] due to its fast speed attribute [27] and DenseNet because of its deep and powerful infrastructure [28], both in a 3D manner. This work extends our preliminary report of using Tversky index [29] as a loss function for 3D U-net [30]. To the best of our knowledge this is the first study proposing a similarity loss function for precision and recall adjustments in training 3D deep fully convolutional networks for highly unbalanced data. Within our approach, we investigate the effects of asymmetry in the similarity loss function on whole-size as well as patch-size images with two different deep networks. In addition, we incorporate a soft weighted voting method, calculating weighted average of probabilities predicted by many augmented overlapping patches in an image. Our results show that this significantly improved lesion segmentation accuracy. Based on our experimental results, we strongly recommend the use of precision-recall balancing properties of asymmetric loss functions as a way to approach both balanced and unbalanced data in medical image segmentation where precision and recall may not have equal importance. We also propose a 3D patch-wise densely connected network with large overlapping patches and a patch prediction fusion method for best results.

III. Materials and Methods

A. Network Architecture

We designed and evaluated two fully convolutional neural networks with two different network architectures: 1) a 3D fully convolutional network [31], [32] based on the U-net architecture [15], and 2) a 3D densely connected network [28] based on the Dense-Net architecture [33]. To this end, we develop a 3D U-net and a 3D patch-wise Dense-Net while introducing an asymmetric loss layer based on F_β scores. The details of the network architectures are described next and we follow with the loss function formulation, and our proposed 3D patch prediction fusion method for the patch-wise network.

1) 3D U-net

We propose a 3D U-net with an asymmetric similarity loss layer [30]. This U-net style architecture is shown in Figure 1. It consists of a contracting and an expanding path (to the right and left, respectively). High-resolution features in the contracting path are concatenated with upsampled versions of global low-resolution features in the expanding path to help the network learn both local and global information. In the contracting path, padded $3 \times 3 \times 3$ convolutions are followed by ReLU non-linear layers. $2 \times 2 \times 2$ max pooling layers with stride 2 are applied after every two convolutional layers. The number of features is doubled after each downsampling by the max pooling layers. The expanding path contains $2 \times 2 \times 2$ transposed convolution layers after every two convolutional layers, and the resulting feature map is concatenated to the corresponding feature map from the contracting path. At the final layer a $1 \times 1 \times 1$ convolution with softmax activation is used to reach the feature map with depth of two, equal to the number of lesion and non-lesion classes.

2) 3D Patch-Wise Dense-Net

We propose a 3D patch-wise Dense-Net based on 3D DenseSeg [33] with overlapping patches, a new asymmetric similarity loss layer and a patch prediction fusion strategy. Figure 2 shows the schematic architecture of the 3D patch-wise Dense-Net. This Dense-Net style architecture consists of three initial $3 \times 3 \times 3$ convolutional layers followed by five dense blocks with a growth rate of 12. Growth rate refers to the increase amount in the number of feature maps after each layer in a dense block. In each dense block there are four $3 \times 3 \times 3$ convolutional layers preceding with $1 \times 1 \times 1$ convolutional layers referred to as bottlenecks [28], which have the purpose of reducing the number of input feature maps. Skip connections are made between all layers of each dense block. Aside from the last dense block, others are followed by a $1 \times 1 \times 1$ convolutional layer and a max pooling layer which are named transition blocks. Down sampling of stride two occurs in each transition block to reduce the feature map dimensionality for computational efficiency. Each of the convolutional layers is followed by batch normalization and ReLU activation layers. Dropout rate of 0.2 is only applied after $3 \times 3 \times 3$ convolutional layers within dense blocks. At the final layer a $1 \times 1 \times 1$ convolution with sigmoid output is used to reach the feature map with depth of one (lesion or non-lesion class).

Prior to proceeding to the main classifier, results of all dense blocks are upsampled using deconvolutional layers, using transpose matrices of convolutions. Afterwards, the results are concatenated and passed through the main classifier to calculate the probability map of the input patch. In the proposed architecture, fully convolutional layers are used instead of fully connected layers [34] to achieve much faster testing time. This architecture segments large 3D image patches. Therefore,

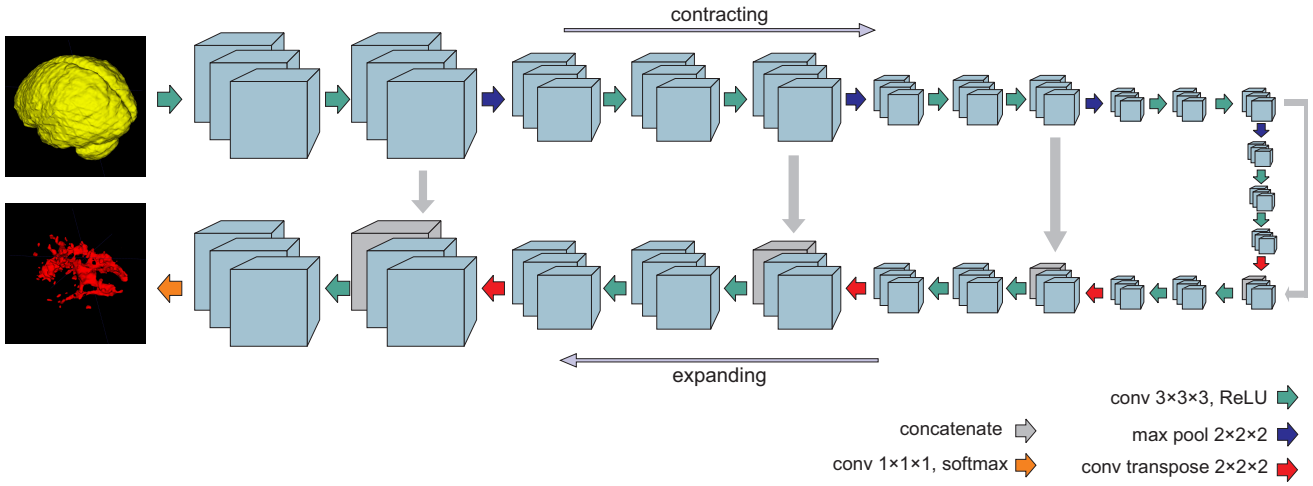

Figure 1. The 3D U-net style architecture with full-size images as inputs and skip connections between a contracting path and an expanding path.

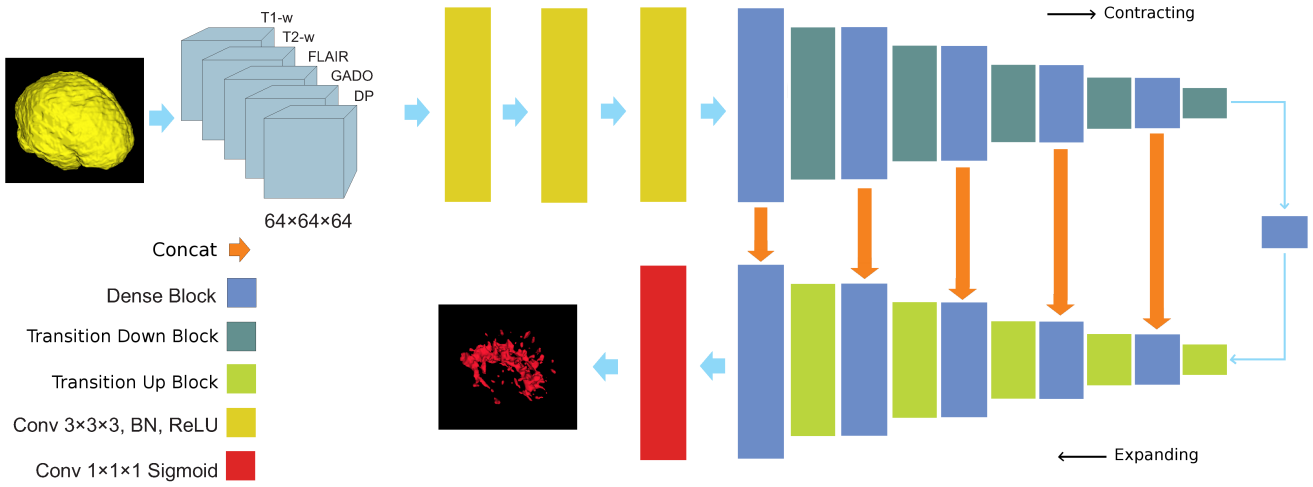

Figure 2. The 3D patch-wise Dense-Net style architecture with $64 \times 64 \times 64$ five channel input patches, consisting of five dense blocks and four convolutional layers with bottlenecks within each block. Overlapping patches of a full size image are used as inputs to this network for training and testing.

to segment any size input image, overlapping large patches (typically of size $64 \times 64 \times 64$ or $128 \times 128 \times 128$) extracted from the image are used as input to the network. These patches are augmented and their predictions are fused to provide final segmentation of a full-size input image. The loss layer, patch augmentation and patch prediction fusion, and the details of training are discussed in the sections that follow.

B. Asymmetric Similarity Loss Function

The output layers in our two networks consist of 1 plane. There is one plane for MS Lesion class. Lesion voxels are labeled as 1 and non-lesion voxels are labeld as zero. We applied sigmoid on each voxel in the last layer to form the last feature map. Let P and G be the set of predicted and ground truth binary labels, respectively. The Dice similarity coefficient D between P and G is defined as:

$$D(P, G) = \frac{2|PG|}{|P| + |G|} \tag{1}$$

Loss functions based on the Dice similarity coefficient have been proposed as alternatives to cross entropy to improve training 3D U-Net and other network architectures [11], [25]; however D, as the harmonic mean of precision and recall, weighs false positives (FPs) and false negatives (FNs) equally. It is a symmetric similarity loss function. To make a better adjustment of the weights of FPs and FNs (and achieve a better balance between precision and recall) in training fully convolutional deep networks for highly unbalanced data, where detecting small number of voxels in a class is crucial, we propose an asymmetric similarity loss function based on the F_β scores which is defined as:

$$F_\beta = (1 + \beta^2) \frac{precision \times recall}{\beta^2 \times precision + recall} \tag{2}$$

Equation (2) can be written as:

$$F(P, G; \beta) = \frac{(1 + \beta^2)|PG|}{(1 + \beta^2)|PG| + \beta^2|G \setminus P| + |P \setminus G|} \tag{3}$$

where $|P \setminus G|$ is the relative complement of G on P. To define the F_β loss function we use the following formulation:

$F_\beta =$

$$\frac{(1+\beta^2)\sum_{i=1}^{N} p_i g_i}{(1+\beta^2)\sum_{i=1}^{N} p_i g_i + \beta^2 \sum_{i=1}^{N}(1-p_i)g_i + \sum_{i=1}^{N} p_i(1-g_i)} \tag{4}$$

where in the output of the sigmoid layer, the p_i is the probability of voxel i be a lesion and $1-p_i$ is the probability of voxel i be a non-lesion. Additionally, the ground truth training label g_i is 1 for a lesion voxel and 0 for a non-lesion voxel. The gradient of the F_β in Equation (4) with respect to P is defined as $\nabla F_\beta = [\frac{\partial F_\beta}{\partial p_1}, \frac{\partial F_\beta}{\partial p_2}, ..., \frac{\partial F_\beta}{\partial p_N}]$ where each element of gradient vector can be calculated as:

$$\frac{\partial F_\beta}{\partial p_j} =$$

$$\frac{(1+\beta^2) g_j (\beta^2 \sum_{i=1}^{N}(1-p_i)g_i + \sum_{i=1}^{N} p_i(1-g_i))}{((1+\beta^2)\sum_{i=1}^{N} p_i g_i + \beta^2 \sum_{i=1}^{N}(1-p_i)g_i + \sum_{i=1}^{N} p_i(1-g_i))^2} \tag{5}$$

Considering this formulation we do not need to use weights to balance the training data. Also by adjusting the hyperparameter β we can control the trade-off between precision and recall (FPs and FNs). It is notable that the F_β index generalizes the Dice coefficient and the Tanimoto coefficient (as known as Jaccard index). In the case of $\beta = 1$ the F_β index simplifies to be the Dice similarity coefficient (F_1). Larger β weighs recall higher than precision (by placing more emphasis on false negatives). We hypothesize that using higher β in our asymmetric similarity loss function helps us shift the emphasis to decrease FNs and boost recall, therefore achieve better performance in terms of precision-recall trade-off.

C. 3D Patch Prediction Fusion

To use our 3D patch-wise Dense-Net architecture to segment a full-size input image (of any size), overlapping large patches (of size $64 \times 64 \times 64$ or $128 \times 128 \times 128$) are taken from the image and fed into the network. In both training and testing, patches are augmented, fed into the network, and their predictions are fused in a procedure that is described in this section. A network with smaller input patch size uses less memory. Therefore, to fit the $128 \times 128 \times 128$ size patches into the memory we used an extra $2 \times 2 \times 2$ convolution layer with stride 2 at the very beginning of our architecture to reduce the image size.

The amount of intersection area (overlap) between patches is adjustable. If we were to use 75% overlaps, the prediction time would be roughly an hour per 3D image. However, to keep the prediction time close to 5 minutes per image, we used 50% overlaps (stride of 1/2 of the patch size) on patch windows. Therefore, given the input image sizes of $128 \times 224 \times 256$, the algorithm produces $5 \times 8 \times 9$ patches per augmentation. There are four augmentations, the original image, and the three 180 degree rotations for each plane. Consequently, our model performs 1,440 patch predictions per 3D image (of the above-mentioned size) and 32 predictions per voxel.

The predictions from overlapping patches are fused to form the segmentation of the full-size image. In case of no overlap and no patch augmentation, each voxel on the original image has one predicted value, therefore predictions from tiled patches can just be tiled to produce the original image segmentation. However, this does not lead to the best results due to the lack of augmentation in test and training and also because patch predictions are less accurate in the patch borders due to incomplete image features in patch borders. This is shown in Figure 3 where lesions in the border of patches are not correctly segmented in the tiling method where no overlap between patches was used. In the second column, where patches with 50% overlap were used, each voxel received multiple predictions from overlapping patches.

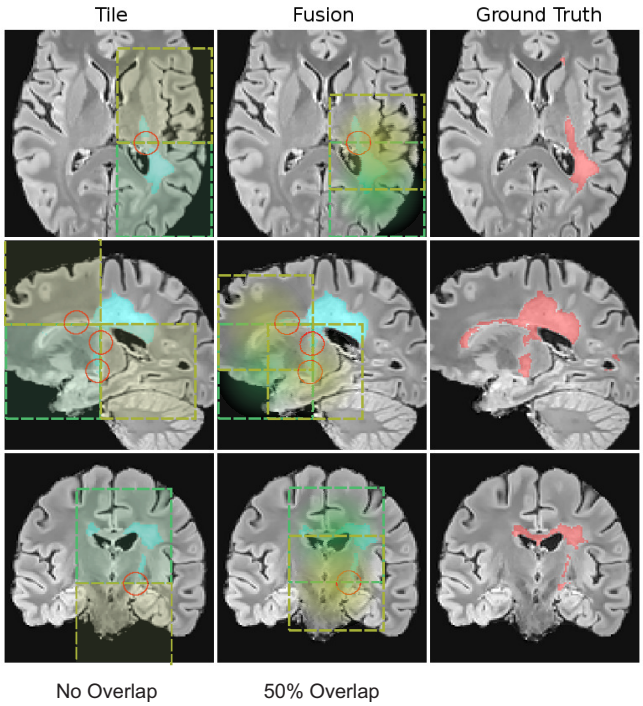

Figure 3. Patch selection of the fusion method compared to the patch tiling method. The predictions are based on the DenseNet model with $\beta = 1.5$. Voxels near patch borders get relatively lower accuracy predictions when a tiling approach is used, while for the fusion approach voxels near the border of one patch will be at the center of another patch resulting in a higher accuracy. The differences of predictions are shown with red circles.

To take into account the relative uncertainty of predictions near patch borders, we use a weighted soft voting approach to fuse patch predictions as opposed to the conventional voting (averaging) method, e.g. [35]. To this end, we calculate the relative weights of soft predictions using a second-order spline function at each patch center. This allows fusion of predictions from all overlapping and augmented patches while giving lower weights to predictions that are made by patches on their borders. With 50% overlap, voxels near the borders of one patch are near the center of another patch as is seen in Figure 3. In our experiments we compared different scenarios, in particular compared our proposed spline patch prediction fusion with uniform patch prediction fusion and patch tiling.

D. Datasets

We trained and evaluated our networks on data sets from the MS lesion segmentation (MSSEG) challenge of the 2016 Medical Image Computing and Computer Assisted Intervention conference [36] as well as the MS lesion segmenta-

tion challenge of the 2015 IEEE International Symposium on Biomedical Imaging (ISBI) conference [37]. T1-weighted magnetization prepared rapid gradient echo (MPRAGE), Fluid-Attenuated Inversion Recovery (FLAIR), Gadolinium-enhanced T1-weighted MRI, Proton Density (PD), and T2-weighted MRI scans of 15 subjects were used as five channel inputs for the MSSEG challenge, and T1-weighted MPRAGE, FLAIR, PD, and T2-weighted MRI scans of 5 subjects with a total of 21 stacks were used as four channel inputs for the ISBI challenge. In the MSSEG dataset, every group of five subjects were in different domains: 1) Philips Ingenia 3T, 2) Siemens Aera 1.5T and 3) Siemens Verio 3T. In the ISBI dataset, all scans were acquired on a 3.0 Tesla MRI scanner. Images of different sizes were all rigidly registered to a reference image of size $128 \times 224 \times 256$ for the MSSEG dataset. After registration, average lesion voxels per image was 15,500, with a maximum of 51,870 and a minimum of 647 voxels.

E. Training

We trained our two FCNs with asymmetric loss layers to segment MS lesions in MSSEG and ISBI datasets. Details of the training process of each network are described here.

1) 3D Unet

Our 3D U-Net was trained end-to-end. Cost minimization on 1000 epochs was performed on the MSSEG dataset using ADAM optimizer [38] with an initial learning rate of 0.0001 multiplied by 0.9 every 1000 steps. The training time for this network was approximately 4 hours on a workstation with Nvidia Geforce GTX1080 GPU.

2) 3D patch-wise Dense-Net

Our 3D patch-wise Dense-Net was trained end-to-end. Cost minimization on 5000 epochs (for the MSSEG dataset) and 1000 epochs (for the ISBI dataset) was performed using ADAM optimizer [38] with an initial learning rate of 0.0005 multiplied by 0.95 every 500 steps with a step growth rate of 2 every 16,000 steps. For instance, the first growth happens at the 16,000th step, where the interval of 500 would be multiplied by two. The training time for this network was approximately 18 hours (MSSEG) and 3 hours (ISBI) on a workstation with Nvidia Geforce GTX1080 GPU. The input patch size was chosen $64 \times 64 \times 64$ for the MSSEG images and $128 \times 128 \times 128$ for the ISBI images in a trade-off between accuracy of extracted features (field-of-view) in each patch and limitations on the GPU memory. The selected size appeared to be both effective and practical for comparisons.

Similarity loss functions (including the Dice similarity coefficient and our proposed asymmetric similarity loss) rely on true positive (TP) counts. The networks would not be able to learn if the TP value is zero leading to a zero loss value. Therefore, only patches with a minimum of 10 lesion voxels were selected for training the patch-wise DenseNet architecture. Nevertheless, equal number of patches was selected from each image. Therefore, the FCNs trained equally with the training data, although they may have had a more diverse pool on images with more number of lesion voxels.

F. Testing

In order to test the architectures properly, five-fold cross validation was used as the total number of subjects was very limited. For MSSEG dataset, each fold contained 3 subjects each from 3 different centers. For ISBI dataset, each fold contained 4 stacks from one subject (total of 5 subjects). In order to test each fold we trained the networks each time from the beginning using the other 4 folds containing images of 12 subjects (MSSEG) and 4 subjects with 4 stacks each (ISBI). After feeding forward the test subjects through the networks, voxels with computed probabilities of 0.5 or more were considered to belong to the lesion class and those with probabilities < 0.5 were considered non-lesion.

IV. Experiments and Results

We conducted experiments to evaluate the relative effectiveness of different networks, asymmetry in loss functions, and patch prediction fusion on lesion segmentation. In this section, first we describe the wide range of metrics used for evaluation, and then present the results of experiments on the two challenge datasets, where we compare our methods with the results reported in the literature.

A. Evaluation Metrics

To evaluate the performance of our networks and compare them against state-of-the-art methods in MS lesion segmentation, we calculate and report several metrics including those used in the literature and the challenges. This includes the Dice Similarity Coefficient (DSC) which is the ratio of twice the amount of intersection to the total number of voxels in prediction (P) and ground truth (G), defined as:

$$DSC = \frac{2\,|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN}$$

where TP, FP, and FN are the true positive, false positive, and false negative rates, respectively. We also calculate and report sensitivity (recall) defined as $\frac{TP}{TP+FN}$ and specificity defined as $\frac{TN}{TN+FP}$ and the F_2 score as a measure that is commonly used in applications where recall is more important than precision (as compared to F_1 or DSC):

$$F_2 = \frac{5TP}{5TP + 4FN + FP}$$

To critically evaluate the performance of lesion segmentation for the highly unbalanced (skewed) datasets, we use the Precision-Recall (PR) curve (as opposed to the receiver-operator characteristic, or ROC, curve) as well as the area under the PR curve (the APR score) [39], [40], [41]. For such skewed datasets, the PR curves and APR scores (on test data) are preferred figures of algorithm performance.

In addition to DSC and True Positive Rate (TPR, same as sensitivity or recall), seven other metrics were used in the ISBI 2015 challenge. These included the Jaccard index defined as:

$$Jaccard = \frac{TP}{TP + FP + FN};$$

the Positive Predictive Value (PPV) defined as the ratio of true positives to the sum of true and false positives:

$$PPV = \frac{TP}{TP + FP};$$

the lesion-wise true positive rate (LTPR), and lesion-wise false positive rate (LFPR), which are more sensitive in measuring the accuracy of segmentation for smaller lesions that are important to detect when performing early disease diagnosis [42]. LTPR is the ratio of true positives to the sum of true positives and false negatives, whereas LFPR is the ratio of false positives to the sum of false positives and true negatives, both only on lesion voxels:

$$LTPR = \frac{TP}{TP + FN} \quad , \quad LFPR = \frac{FP}{FP + TN};$$

the Volume Difference (VD) defined as the absolute difference in volumes divided by the volume of ground truth:

$$VD = \frac{Vol(Seg) - Vol(GT)}{Vol(GT)},$$

where GT and Seg denote ground truth and predicted segmentation, respectively; the average segmentation volume which is the average of all segmented lesion volumes; and the average symmetric Surface Difference (SD) which is the average of the distance (in millimetres) from the predicted lesions to the nearest GT lesions plus the distance from the GT lesions to the nearest predicted lesions [37]. A value of $SD = 0$ would correspond to identical predicted and ground truth lesions. An overall score is also calculated in ISBI2015 challenge based on a combination of these metrics; however, it has been mentioned [37] that this single score does not necessarily represent the best criteria.

B. Results

1) Evaluation on the MSSEG dataset

To evaluate the effect of the asymmetric loss function in making the trade-off between precision and recall, and compare it with the Dice loss function (which is the harmonic mean of precision and recall) in MS lesion segmentation, we trained our FCNs with different β values on the MSSEG dataset. Note that $\beta = 1$ in Equation (3) corresponds to the Dice loss function. For better interpretability to choose β values, we rewrite Equation (3) as

$$F(P, G; \beta) = \frac{|PG|}{|PG| + \frac{\beta^2}{(1+\beta^2)}|G \setminus P| + \frac{1}{(1+\beta^2)}|P \setminus G|} \quad (6)$$

Based on this equation, we chose βs so that the coefficient of $|G \setminus P|$ (false negatives) spanned over 0.5 to 0.9 with an interval of 0.1 in our tests. The performance metrics are reported in Table I. These results show that 1) the balance between sensitivity and specificity was controlled by the parameters of the loss function; 2) according to all combined test measures (i.e. DSC, F_2, and APR score), the best results were obtained from the FCNs trained with $\beta = \sqrt{\frac{7}{3}} \sim 1.5$, which performed better than the FCNs trained with the Dice loss function corresponding to $\beta = 1$; 3) the results obtained from 3D patch-wise DenseNet was much better than the results obtained from 3D U-net; and 4) our proposed spline fusion of patch predictions led to improved performance of the patch-wise DenseNet with tiling and uniform patch prediction fusion. Overall, the best results were obtained with the 3D patch-wise DenseNet with asymmetric loss at $\beta = 1.5$, and spline-weighted soft voting for patch prediction fusion.

Table I
PERFORMANCE METRICS (ON THE MSSEG TEST SET) FOR DIFFERENT VALUES OF THE HYPERPARAMETER β USED IN TRAINING THE 3D U-NET ON FULL-SIZE IMAGES, AND 3D PATCH-WISE DENSENET WITH DIFFERENT PATCH PREDICTION FUSION METHODS. THE BEST VALUES FOR EACH METRIC HAVE BEEN HIGHLIGHTED IN BOLD. AS EXPECTED, IT IS OBSERVED THAT HIGHER β LED TO HIGHER SENSITIVITY (RECALL) AND LOWER SPECIFICITY. THE COMBINED PERFORMANCE METRICS, IN PARTICULAR APR, F_2 AND DSC INDICATE THAT THE BEST PERFORMANCE WAS ACHIEVED AT $\beta = 1.5$. NOTE THAT FOR HIGHLY UNBALANCED (SKEWED) DATA, THE APR AND F_2 SCORE ARE PREFERRED FIGURES OF ALGORITHM PERFORMANCE.

	3D U-Net				
β value	DSC	Sensitivity	Specificity	F_2 score	APR
1.0	53.42	49.85	**99.93**	51.77	52.57
1.2	54.57	55.85	99.91	55.47	54.34
1.5	**56.42**	56.85	99.93	**57.32**	**56.04**
2.0	48.57	61.00	99.89	54.53	53.31
3.0	46.42	**65.57**	99.87	56.11	51.65
	3D patch-wise DenseNet + Tiling				
β value	DSC	Sensitivity	Specificity	F_2 score	APR
1.0	67.53	68.55	**99.95**	66.02	70.5
1.5	**68.18**	74.1	99.93	**68.5**	**71.86**
3.0	62.55	**75.98**	99.91	67.03	67.75
	3D patch-wise DenseNet + Uniform Fusion				
β value	DSC	Sensitivity	Specificity	F_2 score	APR
1.0	68.81	75.28	**99.94**	69.91	72.15
1.5	**68.99**	79.97	99.90	**71.96**	**73.08**
3.0	63.05	**83.55**	99.89	70.65	69.85
	3D patch-wise DenseNet + Spline Fusion				
β value	DSC	Sensitivity	Specificity	F_2 score	APR
1.0	**70.3**	74.49	**99.95**	70.45	73.3
1.5	69.8	78.58	99.92	**71.6**	**73.59**
3.0	64.34	**81.02**	99.91	70.58	70.13

Figures 4 and 5 show the effect of different hyper-parameter (β) values on segmenting a subject with high density of lesions and a subject with very few lesions, respectively. The improvement by using the asymmetric loss function was specifically significant in cases with very small number of lesion voxels as we can see in Figure 5. Independent of the network architecture, training with the Dice loss function ($\beta = 1$), resulted in a high number of false negatives as many lesions were missed. Note that a high value of $\beta = 3$ also resulted in a drop in performance. Figure 6 shows the PR curves for three β levels for the 3D U-Net and the 3D patch-wise DenseNet with tiling, uniform fusion, and spline weighted fusion of patch predictions. As it can be seen in the PR curves (Figure 6) and APR results in Table I for different architectures, the best results corresponding to a good trade-off between sensitivity (recall) and specificity was achieved using the asymmetric loss function with $\beta = 1.5$. Figure 7 shows the boxplots of Dice, sensitivity, and specificity for the four networks trained with the loss function with different β levels. Although, $\beta = 1.5$ slightly decreased specificity, it led to a significant improvement in sensitivity (Figure 7)

and the APR, F_1 and F_2 scores (Table I). We further discuss the significance of these results in the MSSEG data in the Discussion section.

2) Results on the ISBI challenge

The results of our 3D patch-wise DenseNet trained with the asymmetric loss function with $\beta = 1.5$ and the patch selection and spline-weighted patch prediction fusion on the ISBI 2015 challenge is shown in Table II. As demonstrated in the table, we ranked higher than the top five teams in 6 out of 9 evaluation metrics, with DSC and Jaccard index, TPR, LTPR, SD, and average segmentation volume among them; and ranked second according to the ISBI 2015 overall score. We note that while our main goal was to achieve high recall (sensitivity - TPR), which was accomplished as we argued that recall was more important than PPV in this application, we also achieved higher DSC than other methods, which was unexpected but showed that the data imbalance was effectively addressed and the trained network performed well on the test set. Figure 8 shows the true positive, false negative, and false positive voxels overlaid on axial views of the baseline scans of two patients with high and low lesion loads (top and bottom rows, respectively) from our cross-validation folds in the ISBI challenge experiments. These results show low rate of false negatives in challenging cases.

V. Discussion

With our proposed 3D patch-wise DenseNet method we achieved improved precision-recall trade-off and high average DSC scores of 69.8% and 65.74% which are better than the highest ranked techniques examined on the MSSEG2016 and ISBI2015 challenges, respectively. In the MSSEG2016 challenge the 1st ranked team [43] reported an average DSC of 67%, and the 4th ranked team [44] reported an average DSC of 66.6%. In the ISBI2015 challenge we ranked higher than the top five teams in 6 out of 9 evaluation metrics (Table II). We achieved an improved performance by using a 3D patch-wise DenseNet architecture together with the asymmetric similarity loss function and our patch prediction fusion method.

Experimental results in MS lesion segmentation show that all performance evaluation metrics (on the test data) improved by using an asymmetric similarity loss function rather than using the Dice similarity coefficient in the loss layer. While the loss function was deliberately designed to weigh recall higher than precision (at $\beta = 1.5$), consistent improvements in all test performance metrics including DSC and F_2 scores on the test set indicate improved generalization through this type of training. Compared to DSC which weighs recall and precision equally, and the ROC analysis, we consider the area under the PR curves (APR, shown in Figure 6) the most reliable performance metric for such highly skewed data [41], [39].

For consistency in comparing to the literature on these challenges we reported all performance metrics, in particular DSC, sensitivity, and specificity for MSSEG, and nine metrics as well as the overall score for ISBI. We note that for such highly unbalanced (skewed) data the area under the PR curve (APR) is considered a better performance figure than the area under ROC curve; and recall (TPR), the F_2 scores, and in particular the LTPR are more important figures than PPV, the F_1 score (DSC), and the LFPR. Expert manual segmentation of the full extent of lesions (used as ground truth) is very challenging. The detection of small lesions, on the other hand, is very important; therefore lesion detection measures, such as LTPR and LFPR are often considered more important metrics compared to DSC. In particular, LTPR, which counts the ratio of true positives to the sum of true positives and false negatives, is considered a key performance metric. We achieved the highest LTPR among other methods in the ISBI2015 challenge test data.

VI. Conclusion

We introduced a new asymmetric similarity loss function based on F_β scores, that generalizes the Dice similarity coefficient, to achieve improved trade-off between precision and recall in segmenting highly unbalanced data via deep learning. To this end, we added our proposed loss layer to two state-of-the-art 3D fully convolutional deep neural networks based on the DenseNet [28] and U-net architectures [15]. To work with any-size 3D input images and achieve intrinsic data augmentation and balanced sampling to train our DenseNet architecture with similarity loss functions, we proposed a patch selection and augmentation strategy, and a patch prediction fusion method based on spline-weighted soft voting. We achieved marked improvements in several important evaluation metrics by our proposed method in two competitive challenges. To put the work in context, we reported average DSC, F_2, and APR scores of 69.8, 71.6, and 73.59 for the MSSEG challenge, and average DSC, Jaccard and Sensitivity (TPR) scores of 65.74, 50.04 and 66.77 for the ISBI challenge respectively, which indicate that our approach performed better than the latest methods applied in MS lesion segmentation [36], [37], [16], [17], [43], [44]. Based on these results, we recommend the use of asymmetric similarity loss functions within our proposed method based on large overlapping patches and patch prediction fusion to achieve better precision-recall balancing in highly unbalanced medical image segmentation applications.

References

[1] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1–1, 2016.

[2] K. Kamnitsas, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[3] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[4] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[5] C. Wachinger, M. Reuter, and T. Klein, "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy," *NeuroImage*, 2017.

[6] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

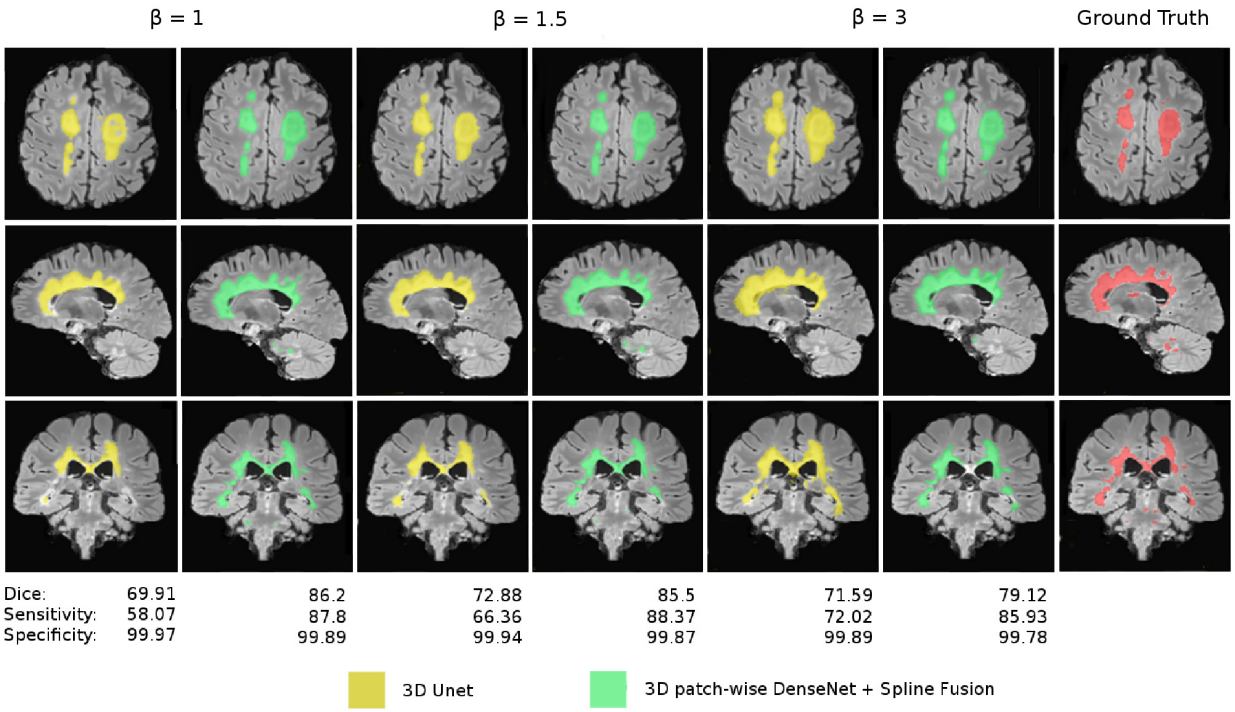

Figure 4. The effect of different weights on FP and FN imposed by the asymmetric loss function on a case with extremely high density of lesions. Axial, sagittal, and coronal sections of images have been shown and the Dice, sensitivity, and specificity values of each case are shown underneath the corresponding column. The best results were obtained at $\beta = 1.5$ with our proposed 3D patch-wise DenseNet with spline patch prediction fusion.

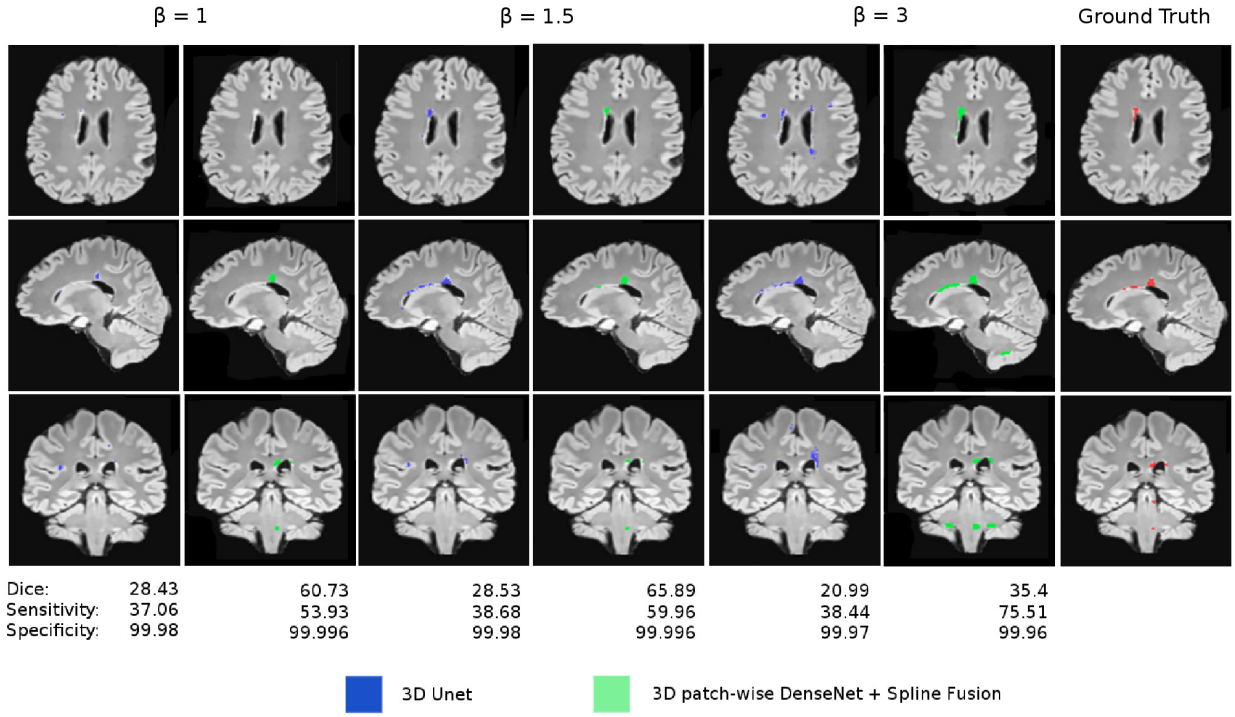

Figure 5. The effect of different weights on FP and FN imposed by the asymmetric loss function on a case with extremely low density of lesions. Axial, sagittal, and coronal sections of images have been shown and the Dice, sensitivity, and specificity values of each case are shown underneath the corresponding column. The best results were obtained at $\beta = 1.5$ with our proposed 3D patch-wise DenseNet with spline patch prediction fusion.

Table II
THE FIVE TOP RANKING TEAMS OF THE ISBI 2015 LONGITUDINAL MS LESION SEGMENTATION CHALLENGE WITH AVERAGE METRICS OF CHALLENGE SCORE, DICE COEFFICIENT, JACCARD COEFFICIENT, POSITIVE PREDICTIVE VALUE (PPV), SENSITIVITY (TPR), LESION TPR BASED ON LESION COUNT (LTPR), LESION FPR BASED ON LESION COUNT (LFPR), VOLUME DIFFERENCE (VD), AVERAGE SYMMETRIC SURFACE DIFFERENCE (SD) AND AVERAGE SEGMENTATION VOLUME. AVERAGE MANUAL VOLUME OF THE TWO RATERS IN THE CHALLENGE WAS **15,648**. OUR PROPOSED METHOD (IMAGINE) ACHIEVED BETTER RESULTS IN 6 OUT OF 9 EVALUATION METRICS COMPARED TO THE OTHER METHODS.

	Score	DSC	Jaccard	PPV	TPR	LTPR	LFPR	VD	SD	Avg Segmentation Volume
asmsl [16]	**92.076**	62.98	47.38	**84.46**	53.69	48.7	20.13	40.45	3.65	10532
IMAGINE (proposed)	91.523	**65.74**	**50.04**	71.39	**66.77**	**50.88**	21.93	37.27	**2.88**	**14429**
nic vicorob test	91.44	64.28	48.52	79.24	57.02	38.72	15.46	**32.58**	3.44	10269
VIC TF FULL	91.331	63.04	47.21	78.66	55.46	36.69	**15.29**	33.84	3.56	10740
MIPLAB v3	91.267	62.73	47.13	79.96	54.98	45.39	23.17	35.85	2.91	10181

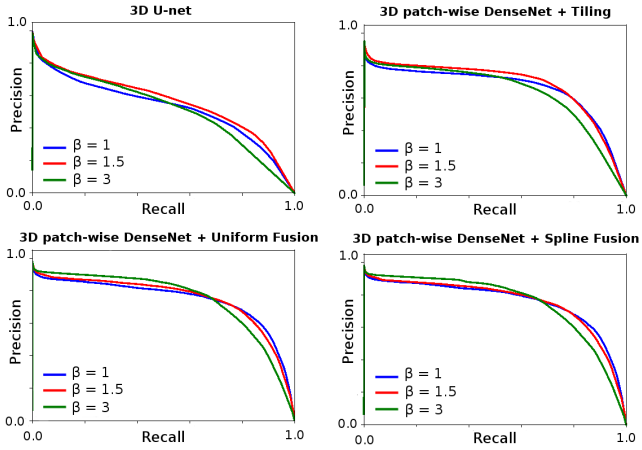

Figure 6. PR curves for all test set obtained by the four examined approaches with different loss function β values. The best results based on the precision-recall trade-off were always obtained at $\beta = 1.5$ and not with the Dice loss function ($\beta = 1.0$); although the difference was less significant when we used large overlapping patches with our patch selection and patch prediction fusion methods that contributed to achieve better balanced sampling of data and improved fusion of augmented data at training and test. The combination of the asymmetric loss function and our 3D patch-wise DenseNet with spline patch prediction fusion generated the best results (Table I).

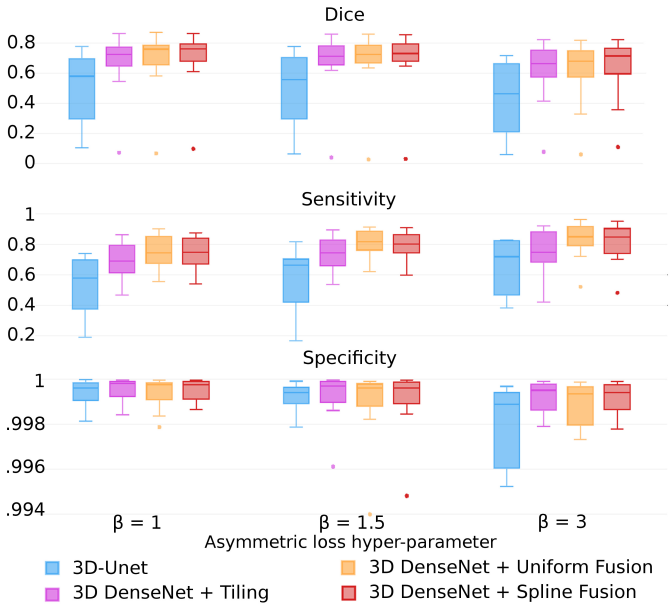

Figure 7. Boxplots of the evaluation scores: Dice, sensitivity, and specificity for the four examined approaches. Overall, these results show that our DenseNet model with the asymmetric loss function and spline patch prediction fusion made the best trade-off between sensitivity and specificity and generated the highest Dice coefficients among all methods.

[7] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[8] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, 2017.

[9] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (Auto-Net) for brain extraction in magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, 2017.

[10] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

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

[12] L. Steinman, "Multiple sclerosis: A coordinated immunological attack against myelin in the central nervous system," *Cell*, vol. 85, no. 3, pp. 299–302, 1996.

[13] L. A. Rolak, "Multiple sclerosis: It's not the disease you thought it was," *Clinical Medicine and Research*, vol. 1, no. 1, pp. 57–60, 2003.

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

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[16] S. Andermatt, S. Pezold, and P. C. Cattin, "Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units," *Brain Lesion (BrainLes) workshop of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 31–42, 2017.

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

[18] P. F. Christ, F. Ettlinger, F. Grun, M. E. A. Elshaer, J. Lipkov, S. Schlecht, F. Ahmaddy, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, F. Hofmann, M. D'Anastasi, S.-A. Ahmadi, G. Kaissis, J. Holch, W. Sommer, R. Braren, V. Heinemann, and B. Menze, "Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2016.

[19] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Hetero-modal image segmentation," *Springer*, pp. 469—-47, 2016.

[20] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images," *NeuroImage*, vol. 57, no. 2, pp. 378–390, 2011.

[21] A. Jesson and T. Arbel, "Hierarchical mrf and random forest segmentation of ms lesions and healthy tissues in brain mri," in *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*,

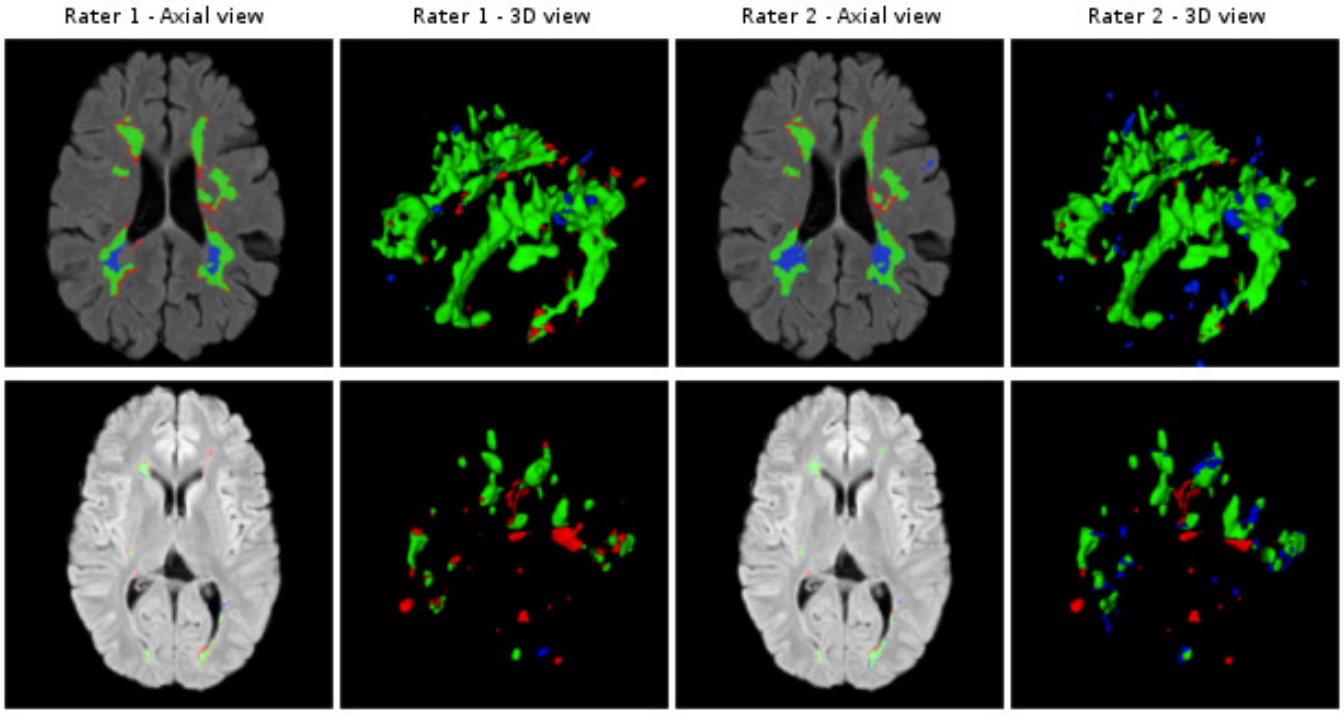

Figure 8. ISBI 2015 results. High volume of lesion (top row) and Low volume of lesion (bottom row) segmentation results compared to both manual segmentations (GTs) for baseline scans of patient 2 and patient 3, respectively. Computed DSC scores of 82.35 and 79.62 (top row), as well as 71.9 and 74.47 (bottom row) was calculated for raters 1 and 2 respectively. True positives, false negatives and false positives are colored in the order of green, blue and red.

2015, pp. 1–2.

[22] N. Guizard, P. Coupé, V. S. Fonov, J. V. Manjón, D. L. Arnold, and D. L. Collins, "Rotation-invariant multi-contrast non-local means for ms lesion segmentation," *NeuroImage*, vol. 8, pp. 376–389, 2015.

[23] M. J. Fartaria, A. Roche, A. Şorega, K. O'Brien, G. Krueger, B. Maréchal, P. Sati, D. S. Reich, T. Kober, M. B. Cuadra, and C. Granziera, "Automated detection of white matter and cortical lesions in early stages of multiple sclerosis," *Journal of Magnetic Resonance Imaging*, vol. 43, pp. 1445–1454, 2016.

[24] X. Tomas-Fernandez and S. K. Warfield, "A model of population and subject (mops) intensities with application to multiple sclerosis lesion segmentation," *IEEE transactions on medical imaging*, vol. 34, no. 6, pp. 1349–1361, 2015.

[25] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *arXiv:1707.03237*, 2017.

[26] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "A model of population and subject (mops) intensities with application to multiple sclerosis lesion segmentation," *arXiv:1709.07330v2*, 2017.

[27] S. S. M. Salehi, S. R. Hashemi, C. Velasco-Annis, A. Ouaalam, J. A. Estroff, D. Erdogmus, S. K. Warfield, and A. Gholipour, "Real-time automatic fetal brain extraction in fetal mri by deep learning," *IEEE International Symposium on Biomedical Imaging*, 2018.

[28] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[29] A. Tversky, "Features of similarity." *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[30] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image seg- mentation using 3d fully convolutional deep networks," *International Workshop on Machine Learning in Medical Imaging*, 2017.

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

[32] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[33] T. D. Bui, J. Shin, and T. Moon, "3d densely convolution networks for volumetric segmentation," *arXiv:1709.03199v5*, 2017.

[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[35] J. Bernal, K. Kushibar, M. Cabezas, S. Valverde, A. Oliver, and X. Lladó, "Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging," *arXiv preprint arXiv:1801.06457*, 2018.

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

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

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

[39] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 451–466.

[40] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[41] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[42] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. Louis Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical Image Analysis*, no. 1, pp. 1–18, 2013.

[43] R. McKinley, T. Gundersen, F. Wagner, A. Chan, R. Wiest, and M. Reyes, "Nabla-net: a deep dag-like convolutional architecture for biomedical image segmentation: application to white-matter lesion segmentation in multiple sclerosis," *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, pp. 46–52, 2016.

[44] F. Vera-Olmos, H. Melero, and N. Malpica, "Random forest for multiple sclerosis lesion segmentation," *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, pp. 90–95, 2016.