Semi-Supervised Multiple Evidence Fusion for Brain Tumor Segmentation

Ling Huang [a,b], Su Ruan [b], Thierry Denœux [a,c]

[a] *Heudiasyc, CNRS, Université de technologie de Compiègne, France*
[b] *Quantif, LITIS, University of Rouen Normandy, France*
[c] *Institut universitaire de France, France*

Abstract

The performance of deep learning-based methods depends mainly on the availability of large-scale labeled learning data. However, obtaining precisely annotated examples is challenging in the medical domain. Although some semi-supervised deep learning methods have been proposed to train models with fewer labels, only a few studies have focused on the uncertainty caused by the low quality of the images and the lack of annotations. This paper addresses the above issues using Dempster-Shafer theory and deep learning: 1) a semi-supervised learning algorithm is proposed based on an image transformation strategy; 2) a probabilistic deep neural network and an evidential neural network are used in parallel to provide two sources of segmentation evidence; 3) Dempster's rule is used to combine the two pieces of evidence and reach a final segmentation result. Results from a series of experiments on the BraTS2019 brain tumor dataset show that our framework achieves promising results when only some training data are labeled.

Keywords: Machine learning, medical image segmentation, information fusion, deep learning, Dempster-Shafer theory, Brain tumor segmentation

1. Introduction

The precise diagnosis of the disease and the accurate delineation of the target lesion are critical for helping radiotherapy and improving the clinical treatment effect. Important progress in computer vision tasks, such as image classification and segmentation, have been made thanks to the development of feature representation through deep neural network models such as, e.g., residual or dense connection [47, 26], attention-gated connection [27], multiple scales features integration by model ensemble [27], transformer mechanism [17], etc. Though the study of better feature representation in deep learning shows promising performance, there still remains a big gap between experimental performance and clinical application, due to the requirement of labeled training data, the inability to model imperfect information, and the limited information provided by a single source of evidence.

First, acquiring a big labeled training dataset is particularly challenging in the medical domain; this issue has become one of the bottlenecks of learning-based approaches. Region labeling in medical image segmentation tasks requires domain knowledge, skilled expertise,

and careful delineation of boundaries. The contradiction between the increasing demand for segmentation accuracy on the one hand, and the shortage of perfect (precise and reliable) annotations on the other hand has so far limited the performance of learning-based medical image segmentation methods. In recent years, many methods have been developed to address the scarce annotation problem, such as self-training [30], adversarial training [23], co-training [36], and clustering [31], in which only partially labeled or unlabeled data are used. It should be noted that training a deep learning model with unsupervised learning is more challenging and cannot always meet high precision requirements. Therefore, researchers now focus increasingly on semi-supervised learning.

Second, probabilistic deep networks, such as UNet [38], SegResNet [35] or DesNet [47] have limitations when it comes to quantifying prediction uncertainty. Most models are trained to approximate conditional class probabilities using the softmax transformation. However, the predictions of deep networks are not always reliable in regions containing uncertain information, such as the pixels close to the lesion boundary; as a result, probabilistic deep neural models are often over-confident for these pixels, resulting in a high risk of misclassification.

Third, a single source of information provides limited information, resulting in high uncertainty. Multiple sources of information can provide complementary information, making it possible to improve segmentation performance. Fusion methods, such as majority voting and averaging, are not robust in case of strongly conflicting information sources, making the fusion result unreliable. Thus, an effective fusion strategy to combine different sources of evidence is important for improving segmentation accuracy.

In this work, we address the above challenges using Dempster-Shafer theory (DST) [8] [39] and semi-supervised learning [1]. The main idea is to address the annotation requirements through the design of semi-supervised learning and to decrease the uncertainty caused by the lack of annotations with evidential segmentation and evidential fusion. In particular, compared with our previous conference paper [20], we now use a slighter encoding-decoding feature extraction module with a multi-fiber unit connection instead of a residual unit connection and we use a class-independent Dice loss function to optimize the model. Moreover, we evaluate the method through complete experiments on the BraTS2019 dataset[2]. We first perform a sensitivity analysis by evaluating the impact of hyperparameters, i.e., the number of prototypes used to model segmentation uncertainty and the percentage of labeled training data. We then carry out a comparative analysis of segmentation performance with the state-of-the-art in the cases of fully supervised learning and semi-supervised learning. The contributions of this work can be summarized as follows:

- We propose a semi-supervised learning algorithm based on similarity constraints between two outputs, allowing the model to maintain a good performance when trained with fewer labels;

[1]This paper is an extended version of the short paper presented at the 2021 International Symposium on Biomedical Imaging (ISBI 2021) [20].

[2]`https://www.med.upenn.edu/cbica/brats2019/data.html`.

- We use a softmax transformation to generate segmentation probabilities and an evidential segmentation network to generate degrees of belief quantifying segmentation uncertainty;
- We propose a multiple evidence fusion strategy to combine the evidence from probability and mass distribution with Dempster's rule;
- The whole model is optimized with a class-independent Dice loss in the case of training with labels and optimized with mean square loss in the case of training without labels.

The rest of the paper is organized as follows: Section 2 starts with a brief reminder of Dempster-Shafer theory, the evidential neural network, and medical image segmentation with semi-supervised learning. The proposed framework is then introduced in Section 3. Section 4 shows experimental results on the BraTS2019 dataset. Section 5 gives conclusions and perspectives.

2. Related work

The theory of belief functions will first be summarized in Section 2.1. The evidential neural network model, a classifier based on DST used in our approach, will then be recalled in Section 2.2. Finally, semi-supervised medical image segmentation will be reviewed in Section 2.3.

2.1. Dempster-Shafer theory

Dempster-Shafer theory, first introduced by Dempster [8] and Shafer [39] is a generalization of Bayesian probability theory. The DST is more flexible and suitable under weaker conditions [40], i.e., imperfect (uncertain, imprecise, partial) information. Compared to probabilistic approaches, the DST model has more degrees of freedom, making it possible to represent severe uncertainty [13].

Representation of evidence. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ be a finite set of hypotheses about some question, called the frame of discernment. Evidence about this question can be represented by a mass function m from the power set 2^Ω to $[0, 1]$, such that

$$\sum_{A \subseteq \Omega} m(A) = 1,$$

and $m(\emptyset) = 0$. The method used to generate mass functions in this paper will be introduced in Section 2.2. Any subset $A \subseteq \Omega$ such that $m(A) > 0$ is called a focal set of m. The mass $m(\Omega)$ represents the degree of ignorance about the problem. When all focal sets are singletons, the mass function is said to be *Bayesian*. The information provided by a mass function m can be represented by a belief function Bel or a plausibility function Pl from 2^Ω to $[0, 1]$ defined, respectively, as

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{1a}$$

and

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}), \tag{1b}$$

for all $A \subseteq \Omega$; in (1b), $\bar{A}$ denotes the complement of A. The quantity $Bel(A)$ can be interpreted as a degree of support to A, while $Pl(A)$ can be seen as a measure of lack of support given to the complement of A.

Dempster's rule. Two mass functions m_1 and m_2 derived from two independent items of evidence can be combined by Dempster's rule [39] as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \tag{2}$$

for all $A \subseteq \Omega, A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) = 0$. In (2), κ represents the degree conflict between m_1 and m_2 defined as

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \tag{3}$$

Decision-making. After combining all the available evidence in the form of a mass function, it is necessary to make a decision. Decision methods based on DST are reviewed in [12]. In classification, the most usual decision rule is to select the class with maximum plausibility.

DST has been applied in the medical image domain and achieved great success in uncertainty modeling, and evidence fusion, such as dose painting [28], disease diagnoses [21], tumor segmentation [31][20][19], etc. A review of DST-based medical image segmentation can be found in [22]. In this paper, we construct the semi-supervised evidence fusion model under the framework of DST.

2.2. Evidential neural network (ENN)

In [11], Denœux proposed an evidential neural network classifier (ENN) [10] using a three-layer fully connected network. The first layer is composed of I radial basis function (RBF) units, each of which computes an activation based on the Euclidean distance between input vector x and a prototype p_i,

$$a_i = \exp(-\gamma_i \|x - p_i\|^2), \tag{4}$$

where γ_i is a positive scale parameter. The second layer then represents the evidence of each prototype p_i by a mass function m_i defined as

$$m_i(\{\omega_k\}) = \alpha_i u_{ik} a_i, \quad k = 1, \ldots, K \tag{5a}$$

$$m_i(\Omega) = 1 - \alpha_i a_i, \tag{5b}$$

where $\alpha \in (0, 1)$ is a parameter associated with prototype p_i and $\Omega = \{\omega_1, \ldots, \omega_K\}$ is the set of classes. Parameter u_{ik} is the membership degree of prototype p_i to class ω_k; the following equality holds: $\sum_{k=1}^{K} u_{ik} = 1$. The degree of belief that x belongs to class ω_k based on the

evidence of prototype p_i is, thus, proportional to the degree of membership of p_i to ω_k, and to a decreasing function of the distance between x and p_i. Finally, the third layer combines the evidence from the I prototypes using Dempster's rule (2). The network output is the mass function m defined as $m = \bigoplus_{i=1}^{I} m_i$. The focals sets of m are the singletons $\{\omega_k\}$ for $k = 1, \ldots, K$ and Ω. After fusion, input vector x is assigned to the most plausible class ω_{k^*}, with $k^* = \arg\max_k m(\{\omega_k\})$.

In the ENN model, the parameters to be learnt are the prototypes p_i, their membership value u_{ik}, as well as parameters α_i and γ_i. Let ψ denote the vector of all parameters. The network is trained by minimizing the regularized loss function

$$L(\psi) = \sum_{k=1}^{K} (pl_k - y_k)^2 + \lambda \sum_{i=1}^{I} \alpha_i, \tag{6}$$

where pl_k is the output plausibility of class ω_k, $y_k = 1$ if the true class is ω_k and $y_k = 0$ otherwise, and λ is a regularization coefficient. The idea of applying the ENN model to features extracted by a convolutional neural network (CNN) was first proposed by Tong et al. in [42]. In this paper, we use ENN as one of the segmentation models to obtain segmentation results with a measure of uncertainty.

2.3. Semi-supervised medical image segmentation

Techniques for semi-supervised medical image segmentation can be divided into three classes: graph-constrained [45], self-learning [30, 33], and generative adversarial learning methods [34, 41]. As an example in the first category, the authors of [45] used a rectangle as a soft constraint by transforming it into a Euclidean distance map and predicting object masks with a convolutional neural network. Self-learning methods use techniques such as an auxiliary model to generate similar pseudos labels so as to guide to training process. In [4], Baur et al. lifted the concept of auxiliary manifold embedding for semi-supervised learning by using the labeled data to optimize the network with a primary loss function; they further augmented the training batches with unlabelled samples and prior knowledge. In [30], Li et al. employed a transformation consisting of a self-ensembling model to enhance the regularization effect for pixel-level predictions. Generative adversarial training has become widely used for semi-supervised learning because of its strong feature simulation ability. In [34], Mondal et al. designed a few-shot 3D multi-modality medical image segmentation method with a generative adversarial network (GAN). The authors used a $K + 1$ class prediction method to guide GAN output plausible predictions for actual unlabelled data by restricting its output for fake examples.

Compared with graph-constrained methods, self-learning and adversarial learning methods have gained popularity in recent brain tumor segmentation research as they do not require additional annotation information for training. In [33], Min et al. proposed a self-training framework for semi-supervised brain tumor segmentation. A hierarchical distillation was used to generate reliable pseudo labels for unlabeled data, which were mixed with manual labels to retrain a two-stream mutual attention network. In [41], Sun et al. proposed an adversarial training-based semi-supervised brain tumor segmentation model composed of

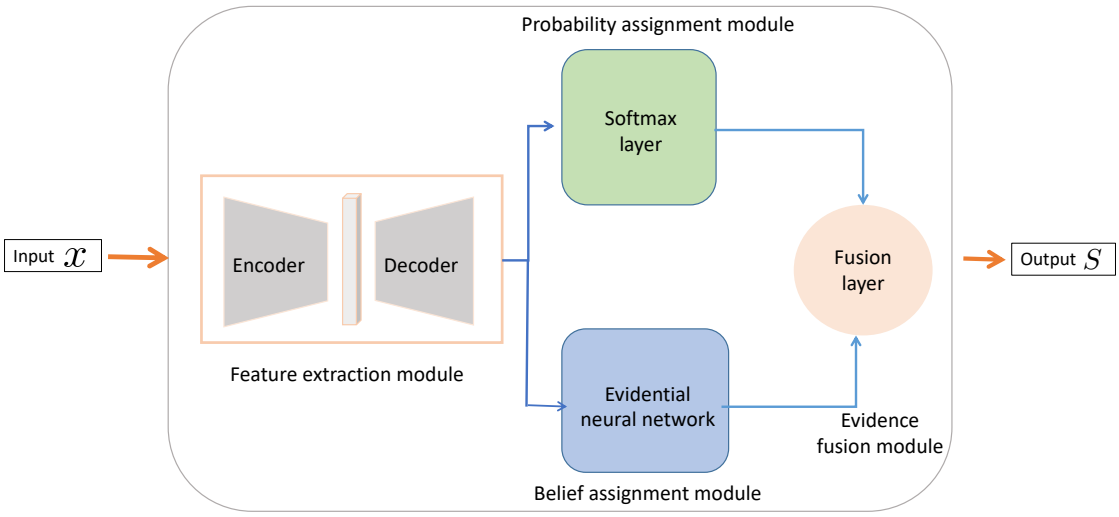


Figure 1: Overall flowchart of our proposal, composed of four modules for feature extraction, probability assignment, basic belief assignment, and evidence fusion.

a segmenter, a generator, and a discriminator. The discriminator learns the tumor boundary with the label maps produced by the segmenter and the supplementary label maps synthesized by the generator.

The above approaches focus on training a segmentation model with partially labeled training data. Though experimental results are promising, only a few authors have studied the uncertainty caused by the low quality of the images and the lack of annotations. In this paper, we adopt the main idea of self-training and use the image information to construct a semi-supervised brain tumor segmentation framework. We propose to use a probabilistic segmentation module and an evidential segmentation module in parallel. The available evidence is then combined using Dempster's rule to decrease segmentation uncertainty.

3. Proposed framework

The proposed SEFNet architecture is composed of four modules for feature extraction, probability/belief calculation, and evidence fusion. Figure 1 shows the overall flowchart of our proposal. The first step is to extract features from input images x. Second, two modules are used in parallel to map features into probabilities and mass functions. The probability assignment module uses the softmax transformation to map the extracted features to probabilities. The belief assignment module uses the ENN model recalled in Section 2.2 to map the extracted features to mass functions. Third, the two sources of evidence, probabilities and mass functions, are combined by Dempster's rule in the multiple evidence fusion module. This architecture will be described in greater detail in Section 3.1. The semi-supervised learning algorithm will be introduced in Section 3.2.

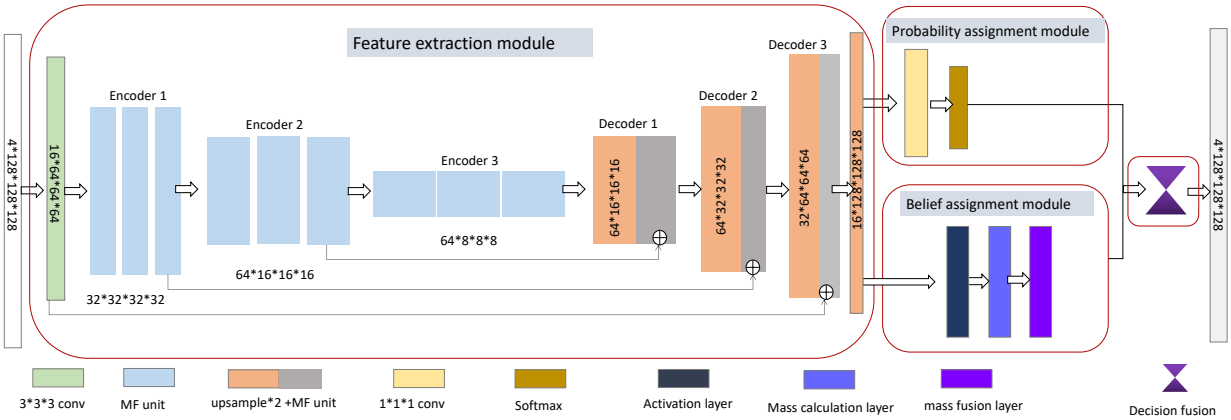


Figure 2: Proposed evidential medical image segmentation framework.

3.1. Evidential segmentation framework with multiple evidence fusion

Figure 2 shows the detailed evidential segmentation framework. The input is composed of four MRI modalities. (Here, we show an example of input data of size $4 \times 128 \times 128 \times 128$.)

Feature extraction. As the focus of this paper is to improve the segmentation performance by modeling prediction uncertainty and combing evidence from different classifiers, we selected UNet, a basic medical image segmentation model, as our feature extraction module with some small modifications. Our proposed framework could alternatively be applied to any state-of-the-art feature extraction models such as, e.g., DenseCNN [26], attentionCNN [27], transformerUNet [17], uResNet [14], to obtain better performance. As shown in Figure 2, we apply a multi-fiber unit in the encoding stage to achieve multi-scale representation. In the decoding stage, the high-resolution features from the encoding stage are concatenated with the upsampled features, which makes the whole feature extraction module similar to UNet[38]. Figure 3(a) shows an example of a multi-fiber unit with a multiplexer layer. Such a unit slices the residual unit into M parallel and separated fibers with the multiplexer layer, which enables information sharing among parallel fibers. Let V_{in}, V_{mid}, and V_{out} denote the number of input channels, middle channels, and output channels, respectively. The total number of connections for the residual unit and multi-fiber units is, respectively, $V_{in} \times V_{mid} + V_{mid} \times V_{out}$ and $(V_{in} \times V_{mid} + V_{mid} \times V_{out})/M$.

Probability and belief assignment modules. The *probability assignment* module comprises a $1 \times 1 \times 1$ projection layer followed by a softmax layer, which maps the feature vectors into probabilities directly. The output of the probability assignment module is denoted as p_{CNN}, and we assume that each voxel belongs to one of four classes denoted as $\{0, 1, 2, 4\}$. The *belief assignment* module is based on the ENN model recalled in Section 2.2; it comprises three layers: an RBF layer that computes distance-based activations using (4), a mass calculation layer that computes mass functions using (5), and a combination layer that combined mass functions derived from prototypes using Dempster's rule (2).

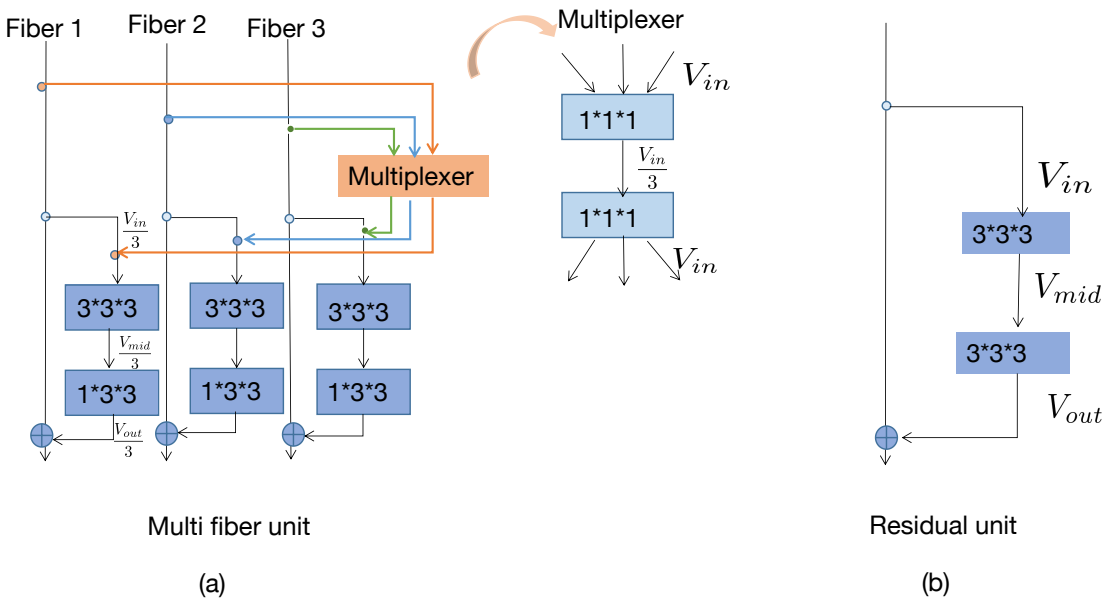


Figure 3: Illustration of the multi-fiber and residual units. (a) A multi-fiber unit with a multiplexer for transferring information across separated fibers (example of three fibers here) [7], (b) A single fiber with residual unit [18].

Evidence fusion. The objective of this module is to make a final segmentation decision. The decision based on several information sources can be expected to be more accurate and reliable than using a single source of information. In our case, if only part of the training data is labeled, the uncertainty is higher than it is in the fully supervised case. To increase the segmentation performance, we propose an additional *evidence fusion* layer to combine evidence from the probability and belief assignment modules. Here, the voxel-wise output probability distributions p_{CNN} from the probability assignment module can be seen as a Bayesian mass functions, which can be combined with the voxel-wise output mass functions m_{ENN} from the belief assignment module using Dempster's rule (2). The combined mass functions are Bayesian, and are given by

$$(p_{CNN} \oplus m_{ENN})(\{\omega_k\}) = \frac{p_{CNN}(\omega_k) pl_{ENN}(\omega_k)}{\sum_{l=1}^{K} p_{CNN}(\omega_l) pl_{ENN}(\omega_l)}, \quad k = 1, \ldots, K, \tag{7}$$

where $pl_{ENN}(\omega_k) = m_{ENN}(\{\omega_k\}) + m_{ENN}(\Omega)$ is the plausibility of class ω_k derived from mass function m_{ENN}.

3.2. Semi-supervised learning

We propose a semi-supervised learning algorithm to optimize the framework when only part of the training data is labeled, with the aim of obtaining an accuracy as close as possible to that of a fully supervised learning method. The general idea is that similar images are expected to produce similar classification or segmentation results even if some transformations have been performed because the relevant characteristics are preserved despite the

transformation. During each learning epoch, a transformed copy x_t of each input image x is computed using one of several transformations, namely, random intensity change, Gaussian blur and exponential noise. This transformation operation will be described in Section 4.1. Two loss functions are proposed for training data with and without labels.

Training with labels. We train the network with the labeled data using the class-independent Dice loss $\mathsf{loss1}$, which measures the overlap region between the output S and the ground truth G:

$$\mathsf{loss1} = \sum_{k=1}^{K} \left(1 - 2 \frac{\sum_{n=1}^{N_1} S_{kn} G_{kn}}{\sum_{n=1}^{N_1} S_{kn} + G_{kn}} \right), \tag{8}$$

where $G_{kn} = 1$ if voxel n belongs to class k, and $G_{kn} = 0$ otherwise, S_{kn} represents the estimated probability that voxel n belongs to class k, and N_1 is the number of labeled voxels.

Training without labels. For the data without labels, we use the mean squared error loss, denoted here as $\mathsf{loss2}$, to optimize the feature representation by minimizing the difference between the original output S and the transformed output S_t:

$$\mathsf{loss2} = \frac{1}{2N_2K} \sum_{k=1}^{K} \sum_{n=1}^{N_2} \left\| S_{kn} - S_{kn}^{t} \right\|^2, \tag{9}$$

where S_{kn}^{t} is the estimated probability that voxel n of the transformed image x_t belongs to class k, and N_2 is the number of unlabeled voxels.

4. Experimental results

In this section, we present numerical experiments to verify the effectiveness of the proposed model. In Section 4.1, we introduce the dataset, the preprocessing of the data, the parameter setting, and the evaluation criteria. An analysis of the sensitivity to hyper-parameters and a comparative analysis of segmentation performance are then reported, respectively, in Sections 4.2 and 4.3.

4.1. Experimental settings

Dataset. The experiment data are those of the Brain Tumor Segmentation (BraTS) 2019 challenge [32, 2, 3]. The dataset consists of 335 cases of patients for training, 125 cases for validation and 166 cases for test. Since the test set has not been made available after the BraTS challenge, we used the official validation set to test our model, i.e., we used it as a test set. For each patient, we have four kinds of MR sequences: T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR). Each of them has a volume of $155 \times 240 \times 240$. For training data, each case was annotated into three heterogeneous histological sub-regions: peritumoral edema (ED, label 2), necrotic core and non-enhancing tumor (NRC/NET, label 1), and enhancing tumor (ET, label 4). The background is marked as label 0. Figure 4 shows an example of the four modalities and

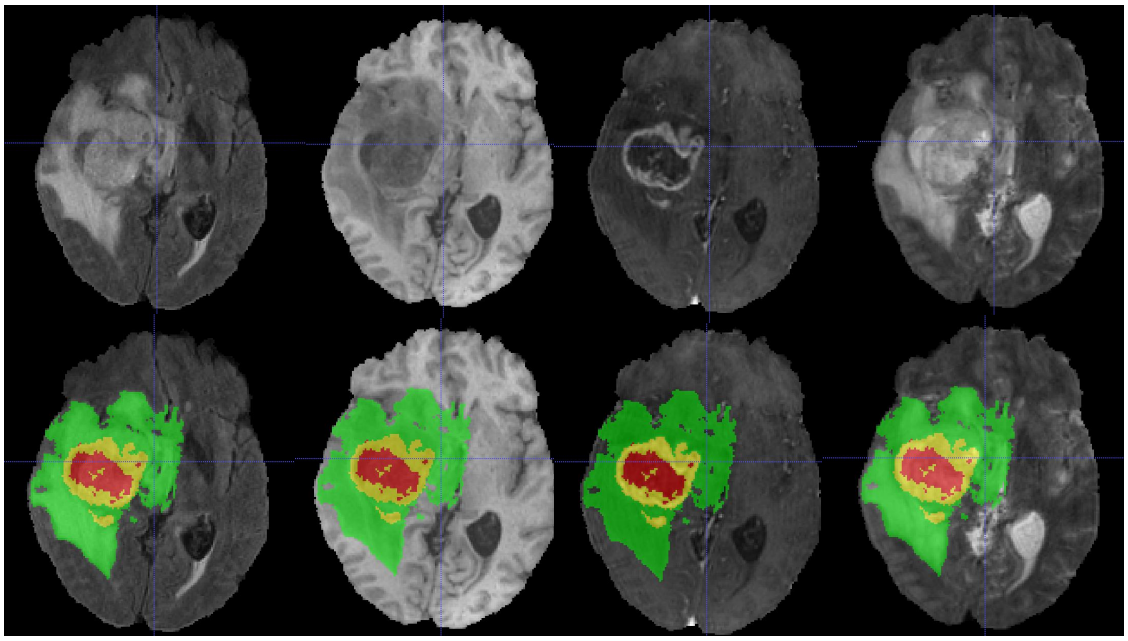


Figure 4: Example of one patient from BraTS2019 dataset. The first and second rows show MR images and MR images with labeled tumor masks, respectively. From left to right: FLAIR, T1, T1Gd, T2. Labels 1, 2, and 4 are marked in red, green, and yellow, respectively.

the corresponding tumor region. The evaluation was based on the segmentation accuracy of three overlap regions: enhancing tumor (ET, label 4), tumor core (TC, the composition of label 1 and 4), and whole tumor (WT, the composition of label 1, 2, and 4). For validation data, only four modalities of MR sequence information are available.

We used five-fold cross-validation to train our SEFNet model. During training, we randomly divided the BraTS2019 training set into five equal-size datasets. The training process was then repeated five times, using each of the five datasets exactly once as the validation set. The cross-validation segmentation performance is the average performance of the five models. For a fair comparison with the state-of-the-art, we fine-tuned the best-performing model with the full training set and tested the performance on the validation set. The segmentation performances were assessed by the online evaluation server CBICA's Image Processing Portal [3].

Pre-processing. Before feeding the data into the framework, several pre-processing methods were used to process the input data. We first applied intensity normalization to each MRI modality from each patient independently by subtracting the mean and dividing by the standard deviation of the brain region only. Moreover, to prevent overfitting, we used four types of data augmentation. First, we applied a random intensity shift between $[-0.1, 0.1]$ and random intensity scaling between $[0.9, 1.1]$ to the MRI data. Second, we randomly cropped the MRI data from $155 \times 240 \times 240$ to $128 \times 128 \times 128$. Third, we used random

[3] `https://ipp.cbica.upenn.edu/`.

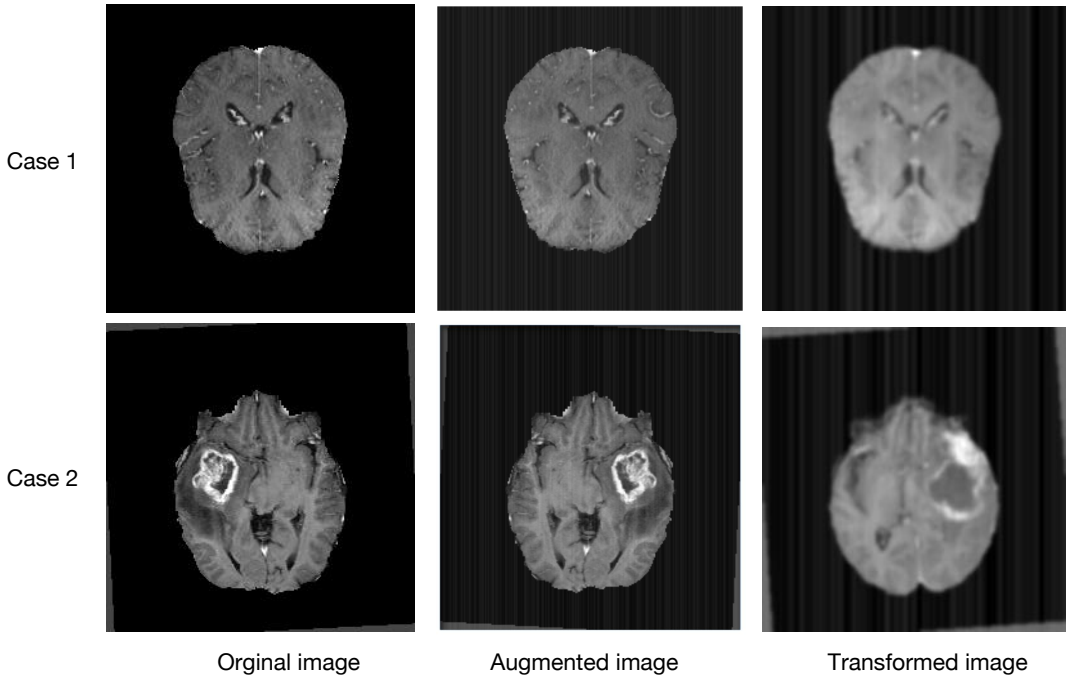


Figure 5: Example of pre-processed images. From left to right: the original image, the augmented image and the transformed image, respectively.

rotation with a rotation angle of 10. Finally, we used random mirror flipping for MRI data along each 3D axis with a flip probability of 50%. The data augmentation operation was applied during each training epoch. Since the data augmentation operation is randomly chosen, the input image x in each training epoch varies.

For semi-supervised training, we used additional transformations of each preprocessed input x. We first applied random intensity change on the input with the shift between $[-0.2, 0.2]$ and scaling between $[0.9, 1.1]$. We then added Gaussian Blur with a standard deviation of 3 to the image. Finally, we added exponential noise with an exponent of 3 to the input. After transformation, for each input image x, we obtained a transformed image x_t. Figure 5 shows two examples of input images before and after prepossessing. (To better show the difference between images, here we only show the aligned images without random cropping). Compared with the original image, the augmented image is randomly flipped with a small intensity change and image rotation. Compared with augmented input x, the transformed input x_t is more blurry and noisy.

Parameter settings. For the feature extraction module, the spatial dimension and input channel were set, respectively, to 3 and 4. The number of channels (filters) of the input layer was set to 16. The number of channels of MF units were set to $32, 64, 64$ for the three encoders and each MF unit had 16 parallel fibers.

For the belief assignment module, parameters α and γ were initialized at 0.5 and 0.01. The prototypes p_i as well as the membership degrees u_{ik} were initialized randomly from

Xavier uniform [16] distributions. Regularization coefficient λ in (6) was set to 0 because regularization is done through data augmentation, as explained above. The impact of the number of prototypes will be discussed in Section 4.2.

The maximum training epoch was set to 300. The training process was stopped when the performance did not increase in 20 epochs. The Adam optimization algorithm with batch size 8 was used to train the model. The initial learning rate was set to 0.001 at the beginning and decayed with an adjusted learning rate

$$lr_q = lr_0 \left(1 - \frac{q}{N_q}\right)^{0.9}, \tag{10}$$

where q is an epoch counter and N_q is the maximum number of epochs. The model with the best cross-validation performance was saved as the final model for testing. All experiments were performed using Python and the PyTorch framework on a desktop computer with a 2.20GHz Intel(R) Xeon(R) CPU E5-2698 v4 and a Tesla V100-SXM2 graphics card with 32 GB GPU memory.

Evaluation criteria. We used two performance criteria: the Dice score and the Hausdorff distance, to measure the segmentation performance. The Dice score is defined as

$$\mathsf{Dice(S,G)} = \frac{2\,|S \cap G|}{|S| + |G|}, \tag{11}$$

where G denotes the manually labeled region, S the segmented region, and $S \cap G$ the overlap area between S and G; the notation $|\cdot|$ denotes the area of a region. The Hausdorff distance is defined as

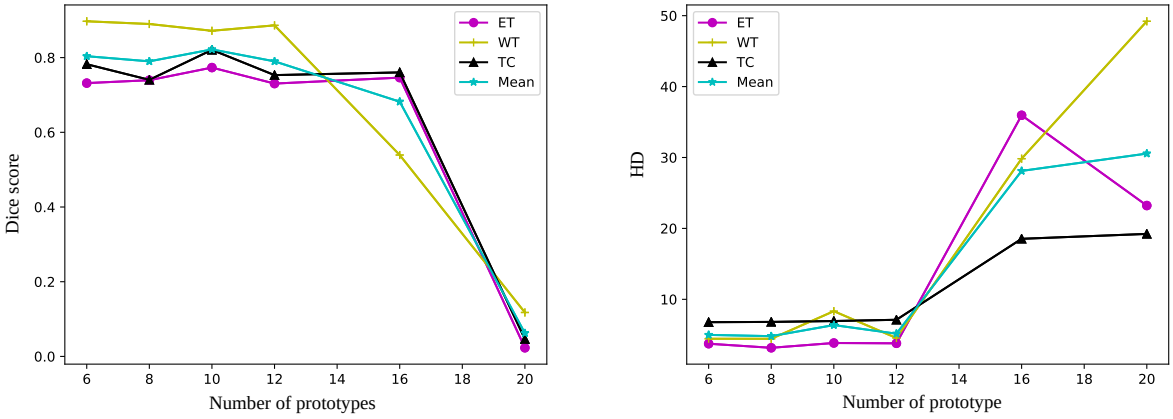
$$\mathsf{Hausdorff} = \max\left(\max_{i \in S} \min_{j \in G} d(i,j), \max_{j \in G} \min_{i \in S} d(i,j)\right), \tag{12}$$

where d represents the Euclidean distance between voxels of S and G. For each patient, we separately computed these two indices for the three classes and then averaged indices over the patients.

4.2. Sensitivity analysis

Number of prototypes. The number of prototypes in the belief assignment module is an important hyper-parameter that may impact segmentation performance. We trained the model with 6, 8, 10, 12, 16, and 20 prototypes. Figure 6 shows the corresponding results. We can see that, when the number of prototypes is comprised between 6 and 12, the performance is stable. SEFNet achieves the best performance in terms of Dice score and Hausdorff Distance with 10 prototypes. With more than 12 prototypes, the performance decreases. In the following experiments, the number of prototypes in SEFNet was fixed to 10.

Proportion of labeled data. To evaluate the impact of the proportions of the labeled training data, we trained our model with 100%, 70%, 50%, and 30% labeled data. The cross-validation performance is reported in Table 1 in terms of Dice score and Hausdorff distance



(a) (b)

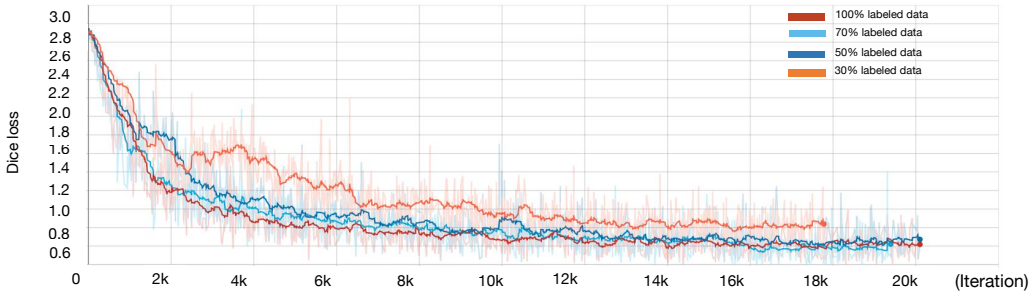
Figure 6: Values of the (a) Dice score (the higher the better), (b) Hausdorff distance (HD) (the lower the better) of ET, TC, WT with the different number of prototypes.

Table 1: Cross-validation performances of SEFNet on the BraTS2019 dataset with different proportions of labeled data.

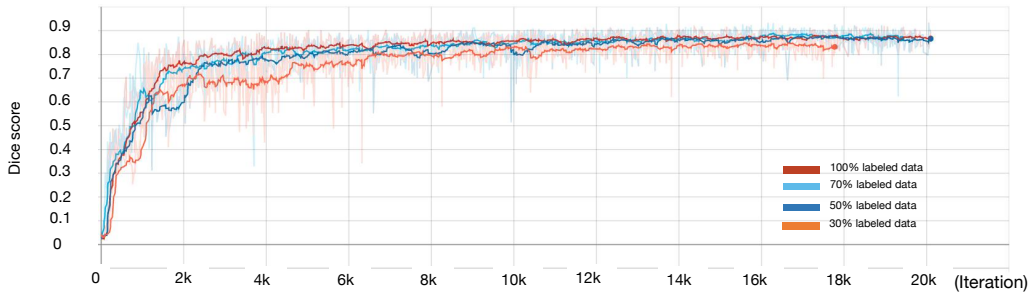
Label	Dice Score (± SD)			Hausdorff distance (± SD)		
proportion	ET	WT	TC	ET	WT	TC
100 %	0.783±0.227	0.906±0.056	0.805±0.220	3.481±4.516	4.618±3.829	6.978±7.595
70 %	0.762±0.215	0.905±0.056	0.814±0.201	5.151±7.254	6.219±11.037	6.507±7.732
50 %	0.727±0.261	0.904±0.066	0.806±0.215	5.115±6.149	4.858±5.399	6.672±7.529
30 %	0.713±0.247	0.897±0.070	0.799±0.215	5.904±7.214	5.554±6.867	6.459±6.489

with the corresponding standard deviations (SD) among the evaluated cases. Compared with 100% training data labeled, the Dice score decreased by 2.1%, 5.6%, and 7%, respectively, when only 70%, 50%, and 30% training data are labeled. With different proportions of the labeled training data, the model shows comparable performance according to the Hausdorff Distance. Compared with WT segmentation, the Dice score standard deviations are quite high for ET and TC segmentation. Two main reasons can explain this phenomenon: the segmentation between ET and NRC/NET is challenging; some cases do not contain ET or NRC/NET, while the model segment DE into ET or NRC/NET class. Figure 7 shows plots of the Dice loss and Dice score for different proportions of training labels during training. With the decrease in training labels, the segmentation performance only decreases slightly. Table 2 shows the segmentation performance on the online validation dataset with the five corresponding models obtained by cross-validation. With 70%, 50%, and 30% labeled training data, the Dice score decreased by 2.9%, 4.2%, and 7.5%, respectively, compared with the fully supervised case. These results demonstrate the effectiveness of our semi-supervised learning scheme.

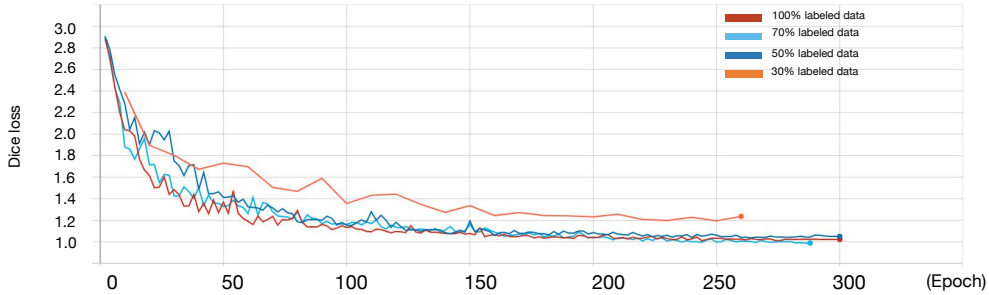
Furthermore, we compared the distribution of the Dice score among 125 validation cases in our model SEFNet and the baseline model MFNet, under the different proportions of



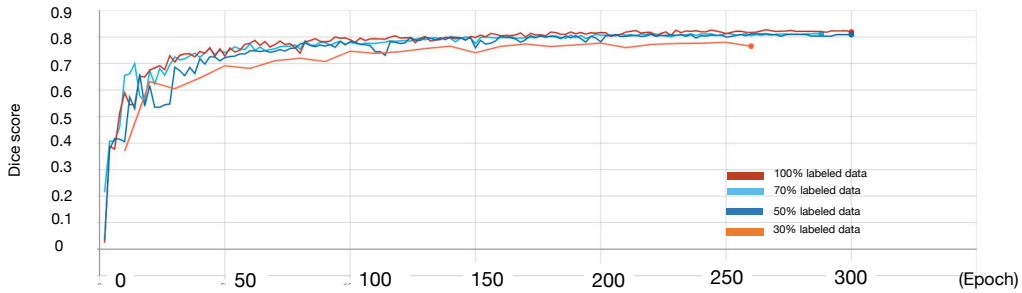
(a)



(b)



(c)



(d)

Figure 7: Plots of Dice loss and Dice score during training. (a) Mean training Dice loss, (b) Mean Dice training score during, (c) Mean cross-validation Dice loss, (d) Mean cross-validation Dice score.

Table 2: Performances of SEFNet on the BraTS2019 validation dataset with different proportions of labeled data.

Label	Dice Score (± SD)			Hausdorff distance (± SD)		
proportion	ET	WT	TC	ET	WT	TC
100%	0.763±0.254	0.883±0.095	0.808±0.201	4.592±8.401	5.983±6.202	7.671±10.608
70%	0.734±0.286	0.884±0.110	0.767±0.242	5.021±8.929	5.718±7.929	8.427+11.799
50%	0.721±0.294	0.877±0.129	0.777±0.227	7.204±17.002	7.942±15.235	11.830±19.809
30%	0.688±0.289	0.887±0.113	0.744±0.267	8.609±17.533	7.448±12.585	12.236±19.690

labeled training data. The results are reported in Figure 8. To simplify the comparison, we only show the mean value of the Dice score here. The boxplot displays the data based on a five-number summary: the minimum (the lowest data point excluding any outliers), the maximum (the largest data point excluding any outliers), the median, and the 25% and 75% percentile. As shown in Figure 8, SEFNet yields better performance than MFNet with different proportions of labeled training data. When 100% training data are labeled, both SEFNet and MFNet achieve high segmentation accuracy. With the decreasing proportion of labeled training data, the SEFNet becomes increasingly superior to MFNet. Compared with MFNet, when only 30% training data are labeled, SEFNet yields around 4%, 5%, and 8% increase of Dice score for, respectively, ET, WT, and TC.

4.3. Comparative analysis of segmentation performance

Comparison with fully-supervised methods. We first compared our results with the state-of-the-art under fully-supervised learning on the BraTS2019 validation set. The comparison is presented in Table 3. We highlight the best results in bold characters and underline the second-best results. SEFNet achieved Dice scores of 0.793, 0.868, 0.861, and 0.841, respectively, for ET, WT, TC, and the mean over the three regions. Compared with MFNet, this corresponds to an increase of 4%, 6.2%, and 4.2% of Dice score in ET, TC and the mean, respectively. Also, SEFNet surpasses most of the reported methods, i.e., UNet, attention UNet, and MCNet. The performance of SEFNet is not as good as the top one solution of the BraTS2019 challenge segmentation task, which uses a two-stage cascaded U-Net [25] (double model), with marginal performance gaps of 0.9%, 4.1%, and 0.3% for ET, WT, and TC in Dice score, respectively. The two-stage cascaded U-Net framework uses one UNet for coarse segmentation and another one for accurate segmentation, which makes it possible to increase the segmentation accuracy to a certain degree. However, the computation cost is high and the reported memory requirement is over 12 GB during the experiment with a batch size of 1.

Figure 9 presents a visual comparison of the brain tumor segmentation results obtained from different slices. From left to right, we can see the ground truth (GT), the segmentation result of the baseline method (MFNet), our proposal (SEFNet), and the difference maps between MFNet and GT on the one hand and between SEFNet and GT on the other hand. The white points in the difference map indicate wrong-segmented voxels. We highlight the regions with fewer misclassified voxels by red circles in Figure 9, where there are fewer white points. Compared with MFNet, SEFNet can generate more precise segmentation results.

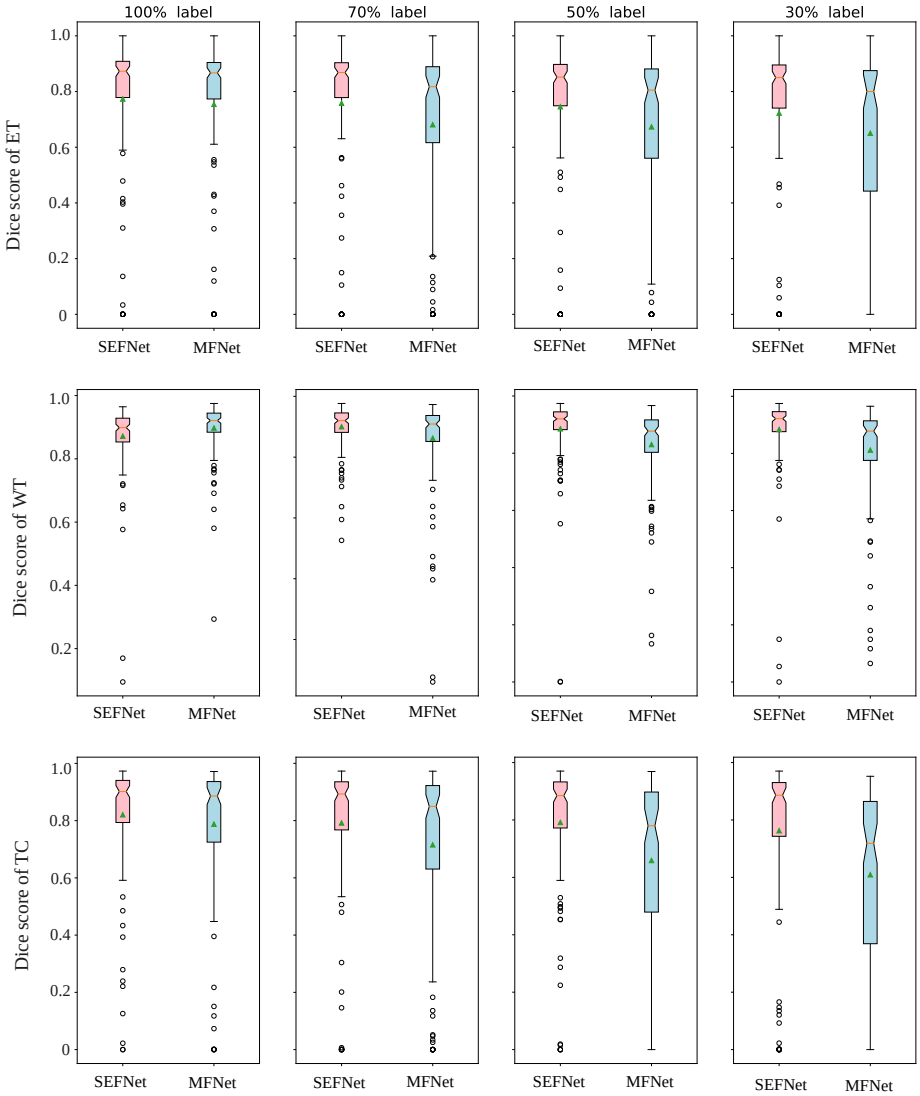


Figure 8: Dice score of ET, TC, and WT with different percentages of labeled training data. From left to right: the results of training with 100%, 70%, 50%, and 30% labels. The first, second, and third rows show the Dice score of ET, WT, and TC, respectively. The pink and light-blue boxplots represent the results of our proposal (SEFNet) and the baseline model (MFNet), respectively. The orange line and green triangle represent the median and mean values of the Dice score, respectively.

Table 3: Performance comparison on the BraTS2019 validation set (fully-supervised learning). The best results are in bold, and the second best results are underlined.

Methods	Dice score				Hausdorff distance			
	ET	WT	TC	Mean	ET	WT	TC	Mean
SEFNet (ours)	<u>0.793</u>	0.868	<u>0.861</u>	<u>0.841</u>	5.616	8.329	6.618	6.854
MFNet [5]	0.753	0.880	0.765	0.799	<u>4.872</u>	8.022	9.706	7.562
DMFNet [5]	0.756	0.890	0.799	0.815	5.069	6.531	7.454	6.351
3D-UNet [43]	0.737	0.894	0.807	0.812	5.994	<u>5.567</u>	7.357	<u>6.342</u>
Dense-UNeT[1]	0.600	0.700	0.630	0.643	11.690	14.330	17.100	14.373
AttentionUNet [24]	0.704	<u>0.898</u>	0.792	0.798	7.050	6.290	8.760	7.370
MCNet [29]	0.771	0.886	0.813	0.823	6.232	7.409	<u>6.033</u>	6.558
Two-stage cascaded U-Net [25]	**0.802**	**0.909**	**0.864**	**0.858**	**3.145**	**4.263**	**5.439**	**4.282**

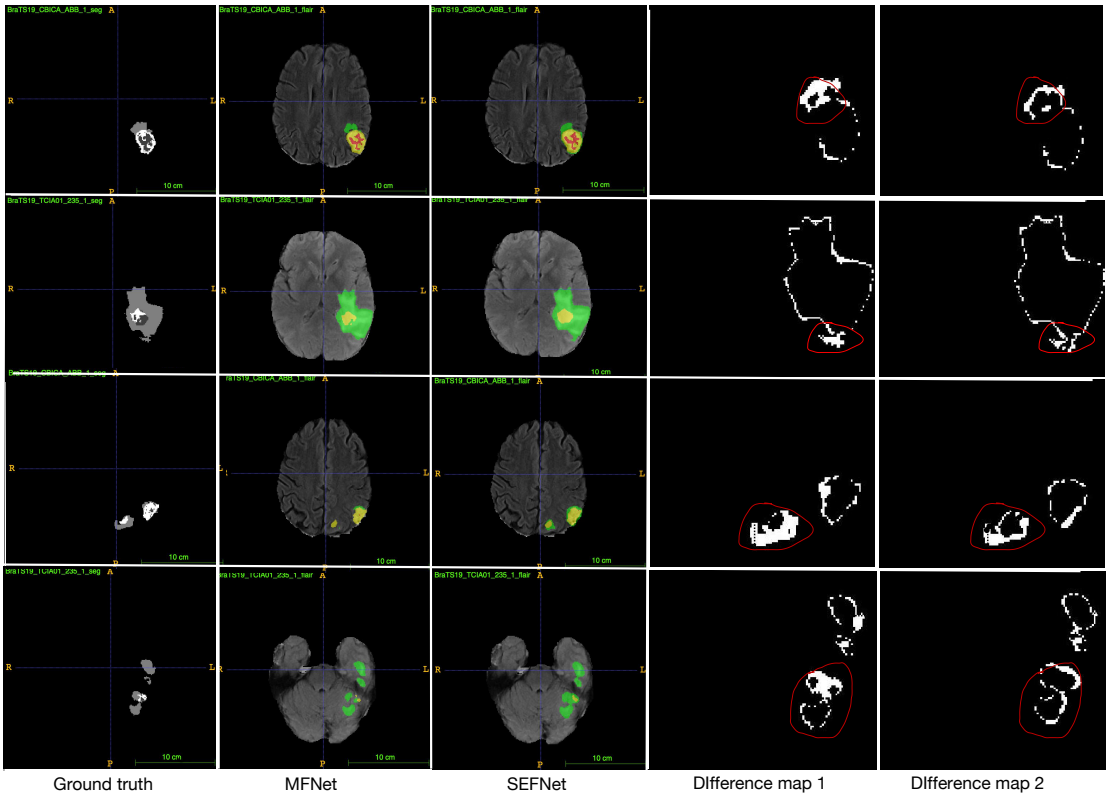


Figure 9: Example segmentation results on the BraTS2019 training dataset. The figures in rows show the results of different axial slices from left to right: the ground truth (GT), the segmentation results of MFNet and SEFNet, and the difference maps between GT and the corresponding segmentation map. The points in white indicate where the segmented result is wrong. Labels 1, 2, and 4 are marked in red, green, and yellow, respectively. Moreover, we highlight the regions with fewer misclassified voxels by red circles in difference maps.

Table 4: Performance comparison with state-of-the-art semi-supervised methods on the BraTS datasets, with 50% labeled data. Symbol $*$ indicates results that are not available from the published papers.

Dataset	Method	# Training instances	# Test instances	Dice score			
				ET	WT	TC	Mean
BraTS2019	SEFNet (ours)	285	125	0.721	**0.877**	**0.777**	**0.792**
BraTS2018	MASSL [6]	200	50	*	0.770	*	*
	SAM-GAN[44]	285	66	*	*	*	0.751
BraTS2017	Transfer-UNet[46]	285	46	0.734	0.690	0.631	0.685
	PGAN [41]	285	46	**0.751**	0.711	0.649	0.703
BraTS2015	Transfer-UNet [46]	140	80	0.633	0.616	0.642	0.630
	PGAN [41]	140	80	0.668	0.652	0.667	0.662
	TSMAN [33]	244	30	*	*	*	0.707

Comparison with semi-supervised methods. We also compared the segmentation performance with the state-of-the-art under semi-supervised learning. Comparing our results with those of other semi-supervised methods is difficult because these methods have been tested on BraTS datasets from different years (2015, 2017, 2018, or 2019), resulting in the various training and validation set components. Also, the different definitions of training labels make the comparison difficult. In this work, we simplified the analysis by only comparing the results when 50% of the training data are labeled. As we can see from Table 4, SEFNet achieves the best performance on validation data with the reported 0.792 mean Dice score on the BraTS2019 dataset. Compared with the performance of MASSL and SAM-GAN on the BraTS2018 dataset, SEFNet yields an increase of 10% and 4.1% in the Dice score of WT and Mean, respectively. Compared with the best performance of PGAN on the BraTS2017 dataset, SEFNet shows an increase of 16.6%, 12.8%, and 8.9% in the Dice score of WT, TC, and Mean, respectively. Compared with the performance of TSMAN on the BraTS2015 dataset, SEFNet yields an increase of 9% in the mean Dice score. This comparison provides empirical evidence for the effectiveness of SEFNet in semi-supervised learning tasks.

5. Conclusion

In this paper, we have presented a semi-supervised evidence fusion framework (SEFNet) for medical image segmentation. In the SEFNet framework, we compute two pieces of segmentation evidence: probability functions generated by a softmax layer, and mass functions generated by an evidential neural network module. Dempster's rule is then used to fuse the two pieces of evidence and to decrease segmentation uncertainty. For labeled images, we used the supervised class-independent Dice loss to guide the training process. For unlabeled images, we used information constraints through image transformation operations to provide training guidance. Quantitative and qualitative results on the BraTS2019 dataset show that using evidence fusion with semi-supervised learning makes it possible to efficiently deal with segmentation uncertainty and to achieve only a small performance degradation with 50% of labeled data, as compared to the fully-supervised case.

This work can be continued in several directions. First, it would be interesting to study the evidence from the belief assignment module in greater detail; in particular, the masses

assigned to the frame of discernment could be meaningful and exploited in the decision. Second, one of the potential issues in this work may be the dependence of the sources of information. The two pieces of evidence obtained from the probability and belief assignment modules are not independent because they are derived from the same features. The use of alternative combination rules such as the cautious rule [9] or even parameterized families of rules [37] could be investigated. In the future, we also plan to study the reliability of segmentation results with the expected calibration error [15] or reliability diagrams. Finally, the multimodal evidence fusion framework could be extended to other medical data analysis tasks, such as biological indicators analysis, survival prediction, and gene-disease correction prediction.

Acknowledgements

This work was carried out in the framework of the Labex MS2T, which was funded by the French Government through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). The first author was supported by the China Scholarship Council (No. 201808331005).

References

[1] Agravat, R.R., Raval, M.S.: Brain tumor segmentation and survival prediction. In: International MICCAI Brain lesion Workshop. pp. 338–348. Springer, Shenzhen, China (Oct, 2019)

[2] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)

[3] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge. arXiv preprint arXiv:1811.02629 (2018)

[4] Baur, C., Albarqouni, S., Navab, N.: Semi-supervised deep learning for fully convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 311–319. Springer, Quebec, Canada (Sep, 2017)

[5] Chen, C., Liu, X., Ding, M., Zheng, J., Li, J.: 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 184–192. Springer, Shenzhen, China (Oct, 2019)

[6] Chen, S., Bortsova, G., Juárez, A.G.U., van Tulder, G., de Bruijne, M.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–465. Springer, Shenzhen, China (Oct, 2019)

[7] Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Multi-fiber networks for video recognition. In: Proceedings of the european conference on computer vision (ECCV). pp. 352–367. Munich, Germany (Sep, 2018)

[8] Dempster, A.P.: Upper and lower probability inferences based on a sample from a finite univariate population. Biometrika **54**(3-4), 515–528 (1967)

[9] Denœux, T.: Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. Artificial Intelligence **172**, 234–264 (2008)

[10] Denœux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE transactions on systems, man, and cybernetics **25**(5), 804–813 (1995)

[11] Denœux, T.: A neural network classifier based on Dempster-Shafer theory. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **30**(2), 131–150 (2000)
[12] Denoeux, T.: Decision-making with belief functions: a review. International Journal of Approximate Reasoning **109**, 87–110 (2019)
[13] Denœux, T., Dubois, D., Prade, H.: Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In: Marquis, P., Papini, O., Prade, H. (eds.) A Guided Tour of Artificial Intelligence Research, vol. 1, chap. 4, pp. 119–150. Springer Verlag (2020)
[14] Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M.d.C., Dickie, D.A., Wardlaw, J., et al.: White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. NeuroImage: Clinical **17**, 918–934 (2018)
[15] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
[16] Hanin, B., Rolnick, D.: How to start training: The effect of initialization and architecture. arXiv preprint arXiv:1803.01719 (2018)
[17] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
[18] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. Jun, 770–778. Las Vegas, NV, USA (2016)
[19] Huang, L., Ruan, S., Decazes, P., Denœux, T.: Lymphoma segmentation from 3D PET-CT images using a deep evidential network. International Journal of Approximate Reasoning **149**, 39–60 (2022)
[20] Huang, L., Ruan, S., Denœux, T.: Belief function-based semi-supervised learning for brain tumor segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 160–164. Nice, France (2021). https://doi.org/10.1109/ISBI48211.2021.9433885
[21] Huang, L., Ruan, S., Denoeux, T.: Covid-19 classification with deep neural network and belief functions. In: The Fifth International Conference on Biological Information and Biomedical Engineering (BIBE 2021). pp. 1–4. Hangzhou, China (2021)
[22] Huang, L., Ruan, S., Thierry, D.: Application of belief functions to medical image segmentation: A review. Information fusion **91**, 737–756 (2023)
[23] Ian, G., Jean, P., Mehdi, M., Xu, B., David, W., Sherjil, O., Aaron, C., Yoshua, B.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680. Montréal, Canada (Dec, 2014)
[24] Islam, M., Vibashan, V., Jose, V.J.M., Wijethilake, N., Utkarsh, U., Ren, H.: Brain tumor segmentation and survival prediction using 3D attention UNet. In: International MICCAI Brainlesion Workshop. pp. 262–272. Springer, Shenzhen, China (Oct, 2019)
[25] Jiang, Z., Ding, C., Liu, M., Tao, D.: Two-stage cascaded U-Net: 1st place solution to BraTS challenge 2019 segmentation task. In: International MICCAI Brainlesion Workshop. pp. 231–241. Springer, Shenzhen, China (Oct,2019)
[26] Karthik, R., Menaka, R., Hariharan, M., Won, D.: Ischemic lesion segmentation using ensemble of multi-scale region aligned CNN. Computer Methods and Programs in Biomedicine **200**, 105831 (2021)
[27] Karthik, R., Radhakrishnan, M., Rajalakshmi, R., Raymann, J.: Delineation of ischemic lesion from brain MRI using attention gated fully convolutional network. Biomedical Engineering Letters **11**(1), 3–13 (2021)
[28] Lelandais, B., Ruan, S., Denœux, T., Vera, P., Gardin, I.: Fusion of multi-tracer PET images for dose painting. Medical image analysis **18**(7), 1247–1259 (2014)
[29] Li, X., Luo, G., Wang, K.: Multi-step cascaded networks for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 163–173. Springer, Shenzhen, China (Oct, 2019)
[30] Li, X., Yu, L., Chen, H., Fu, C., Heng, P.: Transformation consistent self-ensembling model for semi-supervised medical image segmentation. arXiv preprint arXiv:1903.00348 (2019)
[31] Lian, C., Ruan, S., T, D., Li, H., Vera, P.: Spatial evidential clustering with adaptive distance metric

for tumor segmentation in FDG-PET images. IEEE Transactions on Biomedical Engineering **65**(1), 21–30 (2017)

[32] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)

[33] Min, S., Chen, X., Zha, Z.J., Wu, F., Zhang, Y.: A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 4578–4585 (2019)

[34] Mondal, A., Dolz, J., Desrosiers, C.: Few-shot 3D multi-modal medical image segmentation using generative adversarial learning. arXiv preprint arXiv:1810.12241 (2018)

[35] Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. pp. 311–320. Springer, Granada, Spain (Sep, 2018)

[36] Peng, J., Guillermo, E., Marco, P., Christian, D.: Deep co-training for semi-supervised image segmentation. Pattern Recognition p. 107269 (2020)

[37] Quost, B., Masson, M.H., Denœux, T.: Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. International Journal of Approximate Reasoning **52**(3), 353–374 (2011)

[38] Ronneberger, O., Fischer, P., net. Brox, T.: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany (Oct, 2015)

[39] Shafer, G.: A mathematical theory of evidence, vol. 42. Princeton University Press (1976)

[40] Sun, L., Wang, Y.: A multi-attribute fusion approach extending Dempster-Shafer theory for combinatorial-type evidences. Expert Systems with Applications **96**, 218–229 (2018)

[41] Sun, Y., Zhou, C., Fu, Y., Xue, X.: Parasitic GAN for semi-supervised brain tumor segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1535–1539. IEEE, Taipei, Taiwan (Sep, 2019)

[42] Tong, Z., Xu, P.and Denœux, T.: An evidential classifier based on Dempster-Shafer theory and deep learning. Neurocomputing **450**, 275–293 (2021)

[43] Wang, F., Jiang, R., Zheng, L., Meng, C., Biswal, B.: 3D U-net based brain tumor segmentation and survival days prediction. In: International MICCAI Brainlesion Workshop. pp. 131–141. Springer, Shenzhen, China (Oct, 2019)

[44] Xi, N.: Semi-supervised attentive mutual-info generative adversarial network for brain tumor segmentation. In: 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–7. IEEE, Wellington, New Zealand (Nov, 2019)

[45] Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep grabcut for object selection. In: 28th British Machine Vision Conference, BMVC 2017. BMVA Press, London, UK (Sep, 2017)

[46] Zeng, G., Yang, X., Li, J., Yu, L., Heng, P.A., Zheng, G.: 3D U-net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3D MR images. In: International workshop on machine learning in medical imaging. pp. 274–282. Springer, Quebec City, Quebec, Canada (Jun, 2017)

[47] Zhang, R., Zhao, L., Lou, W., Abrigo, J.M., Mok, V.C., Chu, W.C., Wang, D., Shi, L.: Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional DenseNets. IEEE transactions on medical imaging **37**(9), 2149–2160 (2018)