MTANS

Chen, Gaoxiang; Ru, Jintao; Zhou, Yilin; Rekik, Islem; Pan, Zhifang; Liu, Xiaoming

Published in:
NeuroImage

DOI:
10.1016/j.neuroimage.2021.118568

Publication date:
2021

Licence:
CC BY-NC-ND

Document Version
Publisher's PDF, also known as Version of record

NeuroImage 244 (2021) 118568

Contents lists available at ScienceDirect

NeuroImage

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

MTANS: Multi-Scale Mean Teacher Combined Adversarial Network with Shape-Aware Embedding for Semi-Supervised Brain Lesion Segmentation

Gaoxiang Chen [a], Jintao Ru [a], Yilin Zhou [b], Islem Rekik [c,d], Zhifang Pan [a,*], Xiaoming Liu [e], Yezhi Lin [a], Beichen Lu [a], Jialin Shi [a]

[a] The First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China
[b] The State University of New York at Stony Brook, NY 11794, USA
[c] BASIRA Lab, Faculty of Computer and Informatics, Istanbul Technical University, 34469 Istanbul, Turkey
[d] School of Science and Engineering, Computing, University of Dundee, Dundee DD1HN, UK
[e] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China

ARTICLE INFO

Keywords:
Brain Lesion Segmentation
Semi-Supervised Learning
MRI
Deep Learning

ABSTRACT

The annotation of brain lesion images is a key step in clinical diagnosis and treatment of a wide spectrum of brain diseases. In recent years, segmentation methods based on deep learning have gained unprecedented popularity, leveraging a large amount of data with high-quality voxel-level annotations. However, due to the limited time clinicians can provide for the cumbersome task of manual image segmentation, semi-supervised medical image segmentation methods present an alternative solution as they require only a few labeled samples for training. In this paper, we propose a novel semi-supervised segmentation framework that combines improved mean teacher and adversarial network. Specifically, our framework consists of (i) a student model and a teacher model for segmenting the target and generating the signed distance maps of object surfaces, and (ii) a discriminator network for extracting hierarchical features and distinguishing the signed distance maps of labeled and unlabeled data. Besides, based on two different adversarial learning processes, a multi-scale feature consistency loss derived from the student and teacher models is proposed, and a shape-aware embedding scheme is integrated into our framework. We evaluated the proposed method on the public brain lesion datasets from ISBI 2015, ISLES 2015, and BRATS 2018 for the multiple sclerosis lesion, ischemic stroke lesion, and brain tumor segmentation respectively. Experiments demonstrate that our method can effectively leverage unlabeled data while outperforming the supervised baseline and other state-of-the-art semi-supervised methods trained with the same labeled data. The proposed framework is suitable for joint training of limited labeled data and additional unlabeled data, which is expected to reduce the effort of obtaining annotated images.

1. Introduction

Automatic segmentation of magnetic resonance images (MRI) is a fundamental problem and challenge in the field of medical image analysis. Image segmentation can provide important quantitative measures for lesion grading, classification, and disease diagnosis. Accurate medical image segmentation can further assist clinicians in evaluating the treatment response to related diseases and providing a reliable basis for surgical planning and rehabilitation strategies (Kaus et al., 2001).

In recent years, computer-aided automatic segmentation frameworks for brain lesion images such as multiple sclerosis, ischemic stroke and brain tumor have achieved significant advances (Zhang et al., 2019, Akkus et al., 2017, Chen et al., 2020, Kamnitsas et al., 2017). However, most existing brain lesion segmentation methods, especially those based on deep learning, relied on a large number of high-quality labeled data. It was always time-consuming and expensive to produce accurate voxel-level annotations of medical images for training deep learning models on a particular clinical task. Besides, such segmentations might suffer from inter- and intra-annotator (e.g., clinician) variability. Hence, ideally one would design an automated deep learning architecture to accurately segment medical images using a few labeled samples.

To circumvent the need for labeled data, unsupervised learning has been proposed for medical image labeling (Dalca et al., 2018). However, due to the very low segmentation accuracy, such fully unsupervised approaches might not only fail to provide reliable automated clinical diagnoses of patients but also be agnostic to complex anatomical structures or lesions with large variability in shape and size. As another solution, weakly-supervised learning (Ahn and Kwak, 2018, Huang et al., 2018, Lu et al., 2017, Song et al., 2019) did not require voxel-level labeled

* Corresponding author.
E-mail address: panzhifang@wmu.edu.cn (Z. Pan).

https://doi.org/10.1016/j.neuroimage.2021.118568.
Received 18 July 2021; Accepted 7 September 2021
Available online 8 September 2021.

data but used image-level labeled data instead as the weak supervised signal in the network training. Nevertheless, the image-level annotations or boundary boxes for 3D medical images also need domain knowledge and were expensive to acquire. The application of weakly-supervised learning models in medical imaging was still limited. Besides, semi-supervised learning methods (Cheplygina et al., 2019) struck a balance between cumbersome supervision and no-supervision, which presents a new lead for designing hybrid medical data analysis methods without the need of time-consuming labels.

The application of semi-supervised learning in image segmentation has attracted significant attention. Papandreou et al. (Papandreou et al., 2015) proposed a semi-supervised method using deep convolutional neural networks that required image-level annotations and bounding boxes for semantic segmentation. Hong et al. (Hong et al., 2015) used a few labeled samples and a large number of weakly class annotations to train separate classification and segmentation networks and transfer class information between the networks. Similarly, the segmentation network was also trained by combining image-level weak annotations (Lee et al., 2019, Wei et al., 2018). In addition to the unlabeled data, these methods also required image-level signals to assist semi-supervised learning.

With the development of generative adversarial networks (GAN) (Goodfellow et al., 2014), some methods based on GAN have been proposed for image semantic segmentation by only using unlabeled data. Souly et al. (Souly et al., 2017) expanded the training data using a generator network that produced images to remove the dependence on the weakly annotations for auxiliary training. Similarly, Sun et al. (Sun et al., 2019) introduced GAN into the brain tumor segmentation task, and its network was composed of a segmentor, a generator and a discriminator. The discriminator could better learn the boundary information of the brain tumor through the label maps from the segmentor and the fake label maps from the generator. However, with such methods, the generated image examples may not be realistic enough to help the training process. Zhang et al. (Zhang et al., 2017) proposed a deep adversarial network (DAN) without producing additional data, in which the discriminator was used for evaluating the segmentation results of labeled images and unlabeled ones to distinguish them. To better use the discriminator to improve performance, Hung et al. (Hung et al., 2018) proposed an adversarial learning strategy that the supervised model was regarded as a generator while training a discriminator to determine the quality of the segmentation results, and the reliable results were used as pseudo-labels to achieve the self-training scheme. Nie et al. (Nie et al., 2018) further combined the adversarial network based on (Hung et al., 2018) with a sample attention mechanism that could automatically select unlabeled data. As the current state-of-the-art method for semi-supervised medical image segmentation, Li et al. (Li et al., 2020) used the adversarial network to capture shape-aware features with signed distance maps (SDM) (Dangi et al., 2019, Xue et al., 2019) and imposed constraints on the segmentation output of unlabeled data. However, these methods have not yet explored the multi-task training of the discriminator.

The methods based on consistency training (Laine and Aila, 2017, Miyato et al., 2019, Ouali et al., 2020, Tarvainen and Valpola, 2017) have gained success in semi-supervised learning, and are further explored for semi-supervised medical segmentation. The idea is that the prediction results remain consistent after adding noise to the input data. Specifically, the mean teacher model (Tarvainen and Valpola, 2017) was a consistency-based method, which encouraged the segmentation results of two models (student model and teacher model) with the same network architecture to be consistent for the same unlabeled input with different noises, and improved the performance of semi-supervised learning by averaging the model weights. Then, this consistency regularization was extended for MR segmentation (Perone and Cohen-Adad, 2018). Peng et al. (Peng et al., 2020) further proposed deep co-training that encouraged different classifiers to output consistency predictions while increasing the diversity of models based on adversarial samples. The disadvantage of this method was that it needed to train multiple segmentation networks simultaneously and combine multiple segmentation results in the test stage, which required greater computational resources. In addition, Cui et al. (Cui et al., 2019) adapted the mean teacher model to the segmentation task of ischemic stroke lesions. Yu et al. (Yu et al., 2019) further proposed improved consistency loss under the guidance of uncertainty maps for semi-supervised segmentation. Such applications showed the effectiveness of the mean teacher model for the segmentation of binary medical images and have the potential to be further improved to make better use of the unlabeled data. Recently, Mittal et al. (Mittal et al., 2019) proposed a dual-branch framework with a branch of GAN-based supervised segmentation network and another branch of mean teacher-based semi-supervised classification network, that was the state-of-the-art semi-supervised segmentation method for natural images. This work demonstrated the complementarity of the mean teacher model and the adversarial learning model. However, in such a framework, both models were trained separately, and the network fusion was required to combine the output of the two models.

To solve these shortcomings, inspired by the related works, we propose a novel semi-supervised learning framework that deeply integrates the adversarial network into the improved multi-scale mean teacher for brain lesion segmentation. Our framework consists of a student model, a teacher model, and a discriminator, all of which adopt convolutional neural networks (CNNs). The student and teacher models based on the same segmentation network are trained to produce the segmentation probability maps and SDM. According to the principle of consistency training, these two models encourage their segmentation maps to be consistent. However, unlike the previous work (Cui et al., 2019, Yu et al., 2019) that directly calculated the consistency loss between the segmentation probability maps of the student model and teacher model, a new consistency loss derived from the segmentation regions is proposed in our framework. First, we multiply the segmentation results from two models with the same input images, obtaining two sets of segmentation regions, which represent the lesion regions of the original MRI corresponding to the segmentation results. Then, the two sets of region images are passed to the discriminator for similarity comparison. After extracting hierarchical image features from multi-layer convolution modules, the multi-scale feature consistency loss is finally calculated to represent the similarity between the outputs of the student and teacher models. Also, in our framework, the shape-aware embedding scheme is introduced by an adversarial loss based on the discriminator. Through learning the shape information from SDM of labeled and unlabeled data, geometric constraints are imposed on the segmentation results, which can effectively guide the learning of the student model. In the training process, the parameters of the teacher model are updated according to the student model by using the exponential moving average (Tarvainen and Valpola, 2017) (EMA) strategy.

The major contributions of our work can be articulated as follows:

1) We propose a multi-scale consistency strategy for semi-supervised segmentation. Compared with previous consistency loss, which is only computed between the segmentation results of the student and teacher models, the new loss function pushes both models to map their segmentation results to the lesion regions of the original image, thereby incorporating voxel-level regularization information and further improving the performance of teacher-student co-learning.
2) A joint training framework based on two different adversarial learning processes is explored. On one hand, the discriminator in the proposed framework is used for supervised adversarial learning, forcing the segmentation probability maps from the student model to be closer to the ground truth by maximizing a multi-scale loss function. On the other hand, the same discriminator is used to distinguish the SDM from labeled and unlabeled data for implementing the shape-aware embedding scheme through another adversarial learning.

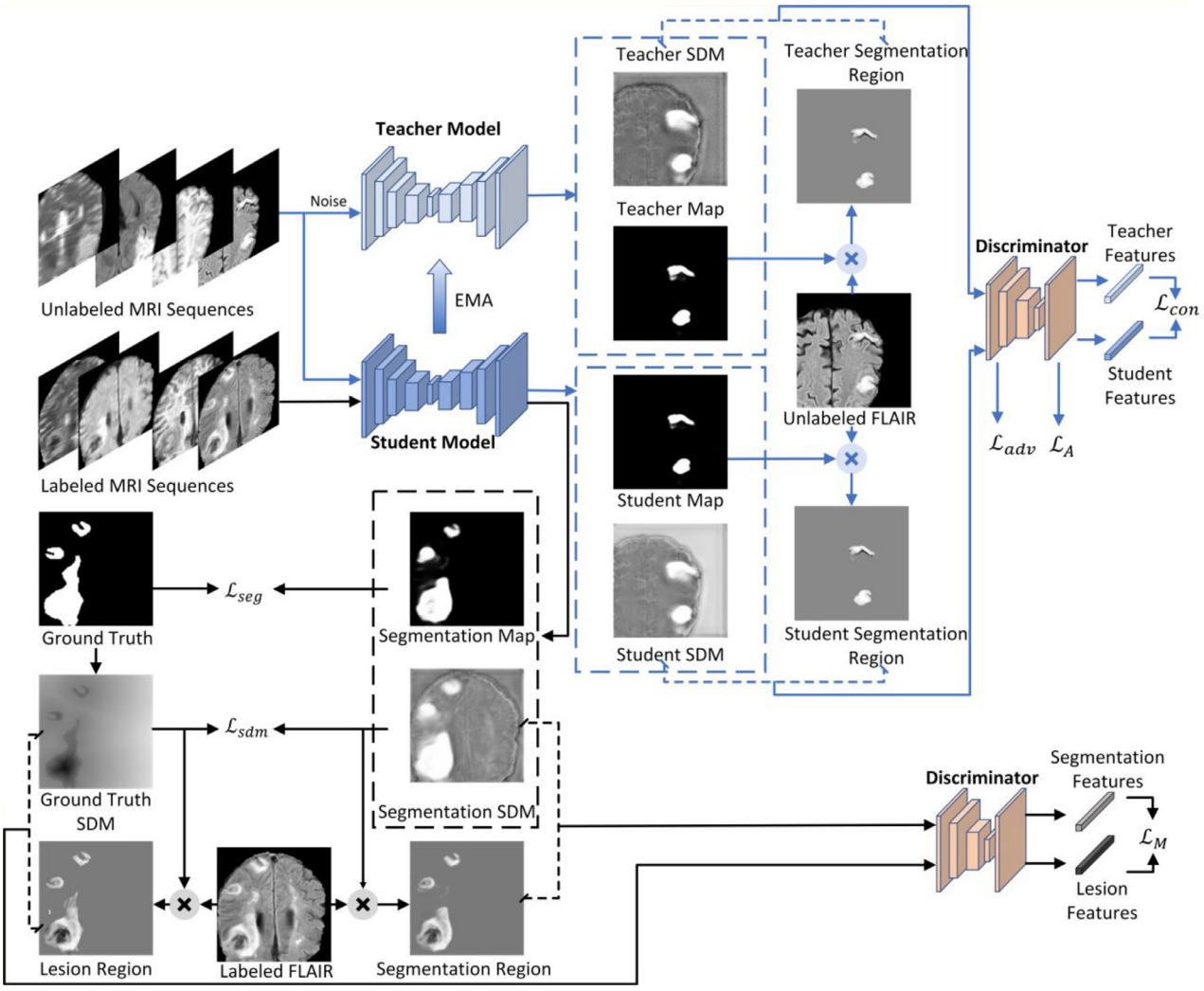

Fig. 1. The overview of our proposed semi-supervised framework for brain lesion segmentation using multimodal MRI. The student model and the teacher model both produce the segmentation probability maps and signed distance maps (SDM), while the segmentation regions and SDM serve as inputs to the discriminator. It is worth noting that the same discriminator in two training processes is represented as two discriminators in this figure. The blue solid lines and the black solid lines represent the processing flow of unlabeled and labeled images, respectively (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

3) We have conducted extensive experiments on three different multi-modal brain lesion segmentation datasets, including ISBI 2015, ISLES 2015, and BRATS 2018. Compared with related state-of-the-art semi-supervised segmentation methods, our framework can efficiently leverage the unlabeled data in each task to improve the segmentation quality, demonstrating its stability and generalizability.

2. Methods

The overview of the proposed framework for semi-supervised 3D brain lesion segmentation applied to multimodal MRI of brain tumor is shown in Fig. 1. Our framework is mainly composed of two networks: a segmentation network for building student model and teacher model, and a discriminator as an adversarial network. Aiming at the training of the student and teacher models on both labeled and unlabeled images, a multi-scale consistency achieved by adversarial learning is proposed. Besides, the shape-aware feature learning is further embedded to constrain the segmentation results.

Given two sets of images, the labeled images X_l and the unlabeled images X_u, the size of the entire training set is N, where the number of labeled and unlabeled images is L and U, respectively. The entire training set can be expressed as the set $S = \{X_n, Y_l\}$, comprising the total images X_n and the ground truth Y_l corresponding to X_l. $X_n = \{X_l, X_u\} = \{x_1, \ldots, x_L, x_{L+1}, \ldots, x_{L+U}\} \in R^{H\times W\times D\times N}$, $Y_l = \{y_1, \ldots, y_L\} \in R^{H\times W\times D\times C\times L}$. $H \times W \times D$ denotes the size of each image, where H, W, D represent the height, width and depth, respectively. The number of label classes in each segmentation task is C.

2.1. Multi-scale consistency

The multi-scale mean teacher is one of the fundamental parts of the proposed framework, with an improved consistency training strategy. Similar to the original architecture of the student and teacher models (Tarvainen and Valpola, 2017), our framework also contains a student model S and a teacher model T, which have the same CNNs structure for segmentation. During the training stage, the original mean teacher optimized two kinds of losses, one is the segmentation loss based on labeled images, the other is the consistency loss, which was generally calculated directly based on the output probability maps of the student model and teacher model. To enforce the consistency training, the clean unlabeled images were fed into the student model while the same one with additional Gaussian noise was input to the teacher model simultaneously. Based on the assumption of the consistency strategy, these two models

were expected to produce similar segmentation results when trained on clean and noisy samples.

In our framework, different from the previous methods, a discriminator A for adversarial learning is introduced as an important component where we further propose a new consistency loss based on multi-scale features extracted from this discriminator. Specifically, given the segmentation maps of unlabeled images generated from the student model and teacher model, we overlay them with the original input images to produce segmentation regions.

As shown in Fig. 1, these two sets of segmentation regions are generated from voxel-by-voxel multiplication of the input MRI and the segmentation probability maps, which can be regarded as the student segmentation regions and the teacher segmentation regions, respectively. In our consistency training, these two segmentation regions are encouraged to be similar instead of only considering the consistency of the probability maps like the original mean teacher model.

Since CNNs can effectively learn image features with multi-layer scales. To better measure the consistency of segmentation regions, the hierarchical features of the segmentation regions from the CNNs-based discriminator are extracted and concatenated at multiple layers. Then, the multi-scale features of two inputs from the corresponding network layers are compared by computing the difference between the student segmentation regions and the teacher ones.

More formally, the multi-scale loss (Xue et al., 2018) calculated based on the hierarchical features from the discriminator is regarded as our proposed new consistency loss $\mathcal{L}_{con}$:

$$\mathcal{L}_{con} = \sum_{h,w,d} \delta_{mae}\left(A\left(X_u \otimes S_{seg}\left(X_u\right)\right)^{(h,w,d)}, A\left(X_u \otimes T_{seg}\left(X_u\right)\right)^{(h,w,d)}\right) \quad (1)$$

where $S_{seg}(\cdot)$ and $T_{seg}(\cdot)$ represent the segmentation probability maps from the student model and the teacher model, respectively. $\otimes$ indicates the voxel-by-voxel multiplication operation of two images, thus $X_u \otimes S_{seg}(X_u)$ and $X_u \otimes T_{seg}(X_u)$ denote the student segmentation regions and teacher segmentation regions that are obtained by multiplying the same unlabeled input image and two corresponding segmentation probability maps. And δ_{mae} is defined as:

$$\delta_{mae}\left(A_f(X), A_f\left(X'\right)\right) = \frac{1}{K}\sum_{i=1}^{K}\left|A_f(X)^i - A_f\left(X'\right)^i\right| \quad (2)$$

where K is the number of network layers in the discriminator and $A_f(X)^i$ is the feature vector output at the i-th layer. The purpose of optimizing $\mathcal{L}_{con}$ is that the probability map produced by S is closer to that of T, and to better learn the distribution of unlabeled images.

2.2. Shape-aware feature learning

To further improve the model performance using unlabeled images, we also implement a shape-aware embedding scheme based on our discriminator. Therefore, the function of the proposed segmentation network is expanded, which can generate not only the segmentation probability maps but also the 3D signed distance maps (SDM). Specifically, the tanh activation function is used in the final output layer of the student model (Xue et al., 2019) to obtain SDM. In our framework, each voxel point in the SDM image is assigned a value, which indicates the distance from the point to the closest point on the surface of the target lesion.

First, from the labeled images, we can effectively learn the representation of shape-aware features and the loss $\mathcal{L}_{sdm}$ based on SDM can be formulated as follows:

$$\mathcal{L}_{sdm} = \sum_{h,w,d} \delta_{mse}\left(S_{sdm}\left(X_l\right)^{(h,w,d)}, Z_l^{(h,w,d)}\right) \quad (3)$$

where δ_{mse} denotes the commonly used mean square error loss. $S_{sdm}(X_l)$ represents the SDM of labeled images generated from the student model. Z_l is the SDM derived from the corresponding ground truth Y_l.

For utilizing unlabeled images to constrain the segmentation results of the student model, we employ SDM-based adversarial training between unlabeled images and labeled images to better learn and encode the shape features of the target object.

Thus, the SDM and the corresponding segmentation regions will simultaneously serve as the inputs of the discriminator. Specifically, for all input images X_n, in addition to the hierarchical features generated for the student and teacher models, the discriminator will also produce the SDM related output $A_{sdm}(X_n)$ only for the student model.

In general, the discriminator generates multi-scale features corresponding to the unlabeled images, so that the student and teacher models are consistent, and it is also used to force the SDM output of the unlabeled and labeled images from the student model to be consistent.

On one hand, the discriminator generates multi-scale features corresponding to the unlabeled images to improve the consistency training of the student and teacher models. On the other hand, discriminator-based adversarial learning is used to force the SDM output of the unlabeled and labeled images from the student model to be consistent.

2.3. Network training

In the adversarial training process of our framework, the student model is forced to generate SDM to fool the discriminator, while the discriminator is trained to distinguish between the input SDM from labeled images or unlabeled images so that the information we learn can be closer to the geometric shape of the ground truth. To train the discriminator network, we minimize the following spatial cross-entropy loss $\mathcal{L}_A$ for the discriminator defined as:

$$\mathcal{L}_A = \sum_{h,w,d} \delta_{bce}\left(A_{sdm}\left(X_l\right)^{(h,w,d)}, 1\right) + \delta_{bce}\left(A_{sdm}\left(X_u\right)^{(h,w,d)}, 0\right) \quad (4)$$

where δ_{bce} is the binary cross-entropy loss. $A_{sdm}(X_l)$ and $A_{sdm}(X_u)$ represent the outputs of the discriminator corresponding to the SDM generated by the labeled images and unlabeled images, respectively. During the training phase, the discriminator is encouraged to give the SDM inputs that are produced from the labeled images higher scores, while the SDM inputs of the unlabeled images correspond to lower scores. This loss is used to train the discriminator to separate the unlabeled SDM from the labeled SDM distribution more precisely.

For the student model, the multi-class cross-entropy is adopted as the supervised segmentation loss, bring the segmentation results closer to the distribution of the ground truth. Also, the dice loss (Isensee et al., 2018) is integrated into this segmentation loss. More specifically, this voxel-wise loss between the probability maps from the student model and the corresponding ground truth is given as:

$$\mathcal{L}_{seg} = \sum_{h,w,d}\left(\delta_{mce}\left(S_{seg}\left(X_l\right)^{(h,w,d)}, Y_l\right) + \delta_{dc}\left(S_{seg}\left(X_l\right)^{(h,w,d)}, Y_l\right)\right) \quad (5)$$

where δ_{mce} and δ_{dc} is the multi-class cross-entropy loss and the dice loss, Y_l is the one-hot encoded ground truth vector.

Also, an adversarial loss $\mathcal{L}_{adv}$ that is given by discriminator is calculated:

$$\mathcal{L}_{adv} = -\sum_{h,w,d} \delta_{bce}\left(A_{sdm}\left(X_u\right)^{(h,w,d)}, 0\right) \quad (6)$$

As in (Goodfellow et al., 2014), when training the segmentation network and updating the parameters, we replace the term of $\mathcal{L}_{adv}$ with $+\sum_{h,w,d} \delta_{bce}(A_{sdm}(X_u)^{(h,w,d)}, 1)$, which is used to maximize the probability that the SDM corresponding to the unlabeled images is considered as the distribution that is generated by the ground truth.

Finally, we optimize the segmentation network with the total loss $\mathcal{L}_S$ that can be defined as the sum of the four losses described above:

$$\mathcal{L}_S = \mathcal{L}_{seg} + \lambda_{sdm}\mathcal{L}_{sdm} + \lambda_{con}\mathcal{L}_{con} + \lambda_{adv}\mathcal{L}_{adv} \quad (7)$$

$\mathcal{L}_{seg}$ represents the sum of multi-class cross-entropy and dice loss, and $\mathcal{L}_{sdm}$ represents the shape-aware mean square loss, both of which

are based on labeled images. $\mathcal{L}_{con}$ and $\mathcal{L}_{adv}$ represent the multi-scale consistency loss and the adversarial loss that are computed with unlabeled images respectively. λ_{sdm}, λ_{con} and λ_{adv} are the corresponding weighting coefficients to balance the relative importance of the proposed losses.

During the training stage, to force the output of the student model to be more reliable, we also introduce a supervised adversarial training process based on another multi-scale feature loss. This loss was calculated from the ground truth and segmentation results when training with labeled images. Similar to $\mathcal{L}_{con}$, we multiply the input labeled image with the segmentation results of the student model and the ground truth ones, to produce the segmentation regions and the real lesion regions of the original MRI, respectively. Next, we input these region images to the discriminator separately, and another multi-scale feature loss $\mathcal{L}_M$ is obtained:

$$\mathcal{L}_M = \sum_{h,w,d} \delta_{mae}\Big(A\big(X_l \otimes S_{seg}(X_l)\big)^{(h,w,d)}, A\big(X_l \otimes Y_l\big)^{(h,w,d)}\Big) \tag{8}$$

where $A(X_l \otimes Y_l)$ represents the hierarchical features of the real lesion regions extracted from the discriminator. Thus, our training objective of the student model and discriminator can be jointly described as a min-max process, which can be written as:

$$\min_{\theta_S}\max_{\theta_A} \mathcal{L}\big(\theta_S, \theta_A\big) = \mathcal{L}_S + \mathcal{L}_A + \mathcal{L}_M \tag{9}$$

Overall, the student model S and discriminator A in our framework are trained by backpropagation using the loss $\mathcal{L}$. In the alternating training process, given a fixed A, S aims to minimize the loss $\mathcal{L}_S$ and $\mathcal{L}_M$ for the parameters θ_S. Next, we fix S, while A aims to minimize the loss $\mathcal{L}_A$ and maximize the loss $\mathcal{L}_M$ for the parameters θ_A.

Besides, in every training step j, the parameters of the teacher model θ_T are updated based on the parameters θ_S using the exponential moving average (EMA). This update strategy can be defined as:

$$\theta_T(j) = \alpha\theta_T(j-1) + (1-\alpha)\theta_S \tag{10}$$

where α is the hyperparameter that controls the EMA decay.

3. Experiments

The proposed architecture was evaluated on three public datasets of 3D MRI for brain lesion segmentation tasks, including multiple sclerosis lesion segmentation, ischemic stroke lesion segmentation, and brain tumor segmentation.

3.1. Datasets

3.1.1. Multiple sclerosis lesion

Firstly, the dataset of the ISBI longitudinal multiple sclerosis lesion segmentation challenge (ISBI 2015) (Carass et al., 2017) was selected to evaluate the performance of our proposed framework on brain lesion segmentation. In this dataset, a total of 21 images from 5 patients with different time points are available as training data. Since the longitudinal image information was not considered in our experiment, each time-point was treated as a separate training image. Each time-point image in the training data corresponds to two manual segmentation labels that are annotated by two different raters. Thus, the training data were finally considered to be 42 images to make full use of each label. The unseen test data contains 14 patients with 4 to 6 time-points, resulting in 61 images. The images of training and test data both contain four different MRI modalities: FLAIR, MPRAGE, T2 and Proton Density (PD). In our semi-supervised settings, we first randomly split the training data into 35 scans as a training set and 7 scans as a test set, then considered 20% (7 scans) of the training set as labeled images and the remaining 80% (28 scans) as unlabeled images.

3.1.2. Ischemic stroke lesion

The ischemic stroke lesion dataset from MICCAI 2015 (ISLES 2015) (Maier et al., 2017) contains 28 labeled MRI scans of ischemic stroke lesion cases. Each scan contains four MRI modalities: T1, Diffusion-Weighted Imaging (DWI), T2, and FLAIR. The images also have been preprocessed by experts. We split 28 scans into 20 scans and 8 scans for training and testing. To evaluate models trained with different ratios of the training set, we used 10% (2 scans) and 20% (4 scans) of the training set as labeled input images and the corresponding remaining as unlabeled images. Besides, due to the size limit of this dataset, an additional cross-validation experiment under 10% of the semi-supervised settings was performed to make the proposed method more convincing. Specifically, 18 scans of the dataset were randomly taken as unlabeled images, and 5-fold cross-validation was applied on the remaining 10 scans.

3.1.3. Brain tumor

Then, we extended our experiment on multi-class imbalanced data, with the brain tumor segmentation dataset at MICCAI 2018 (BRATS 2018) (Menze et al., 2015). It consists of 285 training MRI scans, which are randomly grouped into a training set with 228 scans and a testing set with 57 scans. Each scan of the patient contains four MRI modalities: T1, T2, FLAIR, and post-contrast T1-weighted (T1c). Further details about preprocess steps that have been performed on this dataset can be found in (Menze et al., 2015). We also randomly drew nearly 10% (22 scans) and 20% (45 scans) from the whole training set as labeled images and the remaining data as unlabeled images. To verify the generalization of our model, we also evaluated the trained models on 66 unseen test data. The goal of this task is to evaluate three tumor regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET).

In each experiment, the baseline model was trained without unlabeled data, and other semi-supervised methods used the same data settings as our framework.

3.2. Evaluation metrics

For different brain lesion datasets, we used evaluation methods and metrics consistent with each segmentation challenge. In the evaluation of ISBI 2015, the Dice, positive predictive value (PPV), true positive rate (TPR), lesion false positive rate (LFPR), lesion true positive rate (LTPR), and the Pearson's correlation coefficient of the volumes (VC) were calculated to describe the difference between the segmentation results from the methods and the ground truth from two human rates. To better evaluate important metrics, the methods were ranked based on the website score (WS), which was computed independently by the challenge website[1], and can be described as the total weighted score of the above metrics:

$$WS = \frac{1}{|R|}\frac{1}{|S|}\sum_{R,S}\left(\frac{Dice}{8} + \frac{PPV}{8} + \frac{1-LFPR}{4} + \frac{LTPR}{4} + \frac{VC}{4}\right) \tag{14}$$

where S is the set of all subjects, R is the set of all raters. A method with a WS score of 90 is considered to be comparable to the performance of human raters (Carass et al., 2017).

In addition to the commonly used Dice, Precision, Sensitivity, and Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD) was also evaluated in ISLES 2015 (Maier et al., 2017). As for the metrics of the BRATS dataset, we used Dice, Specificity, Sensitivity, and HD95 (Menze et al., 2015), which can be calculated from the online evaluation system[2].

3.3. Network Architecture and implementation

In our experiments, for the segmentation network of the proposed framework, the patch-based 3D U-Net modified from (Kao et al., 2019) was employed, which can process 3D input patches of $128 \times 128 \times 128$ voxels. The network has an encoder path and a decoder path composed

[1] https://smart-stats-tools.org/lesion-challenge-2015/

[2] https://ipp.cbica.upenn.edu/

Table 1
Quantitative comparison for the performance of the supervised baseline trained with 20% and 100% labeled set and semi-supervised methods trained with 20% labeled and 80% unlabeled set on the ISBI 2015 training data.

Lab/All	Method	Dice(%)	PPV(%)	TPR(%)	LFPR(%)	LTPR(%)
	baseline	77.46±7.06	79.81±13.04	77.23±11.02	51.65±18.19	79.67±14.69
	MT	77.37±6.36	**80.62±13.52**	76.80±11.86	42.76±19.56	81.63±11.68
20%	UA-MT	77.12±6.59	79.43±13.81	77.49±11.95	39.19±20.86	81.72±11.68
	SASSNet	77.39±7.18	75.95±14.93	**82.19±11.09**	49.07±15.23	**82.33±10.38**
	MTAN	77.11±6.43	80.27±13.29	76.60±11.97	37.34±17.77	80.85±11.37
	MTANS	**78.51±6.24**	79.32±13.16	79.58±8.05	**12.67±11.50**	76.87±11.01
100%	baseline	81.60±6.35	78.28±11.27	86.32±5.77	43.06±12.11	88.53±5.77

of four context modules with the convolutional layer. The number of filters in the layers of the encoder-decoder path are 32, 64, 128, and 256, respectively. Moreover, the tanh activation is added to the final 3D convolution block to form an SDM module. The discriminator is the 3D version extended from (Xue et al., 2018), which consists of 6 convolutional layers for downsampling and a multilayer perceptron for binary classification.

For better comparison, the 3D U-Net (Kao et al., 2019) backbone was employed as a supervised baseline, which trained with the same labeled data as other semi-supervised methods on all three tasks. As the semi-supervised methods that were related to our framework, the mean teacher (MT) was trained with the same segmentation network architecture and parameter settings. Besides, the segmentation networks in UA-MT (Yu et al., 2019) and SASSNet (Li et al., 2020) were also replaced by 3D U-Net.

In the experiments, for MT, UA-MT and our framework, both the student model and teacher model were evaluated for better comparison. Thus, when testing the unseen data of ISBI 2015 and BRATS 2018, the student model or teacher model which performed better in the training data, was selected for testing.

In particular, two versions of our framework, MTAN and MTANS, were trained to test the strategies adopted in our framework. MTAN denotes the combination of the MT model and adversarial learning with introducing the multi-scale feature consistency loss for training, which is the main component of our framework. MTANS represents the further implementation of the MTAN with shape-aware embedding.

The implementation of our proposed framework was developed using PyTorch. All models were trained on NVIDIA GeForce RTX 2080 Ti GPU with 11GB of RAM. The maximum number of training epochs was fixed to 600, and the training time of MTANS was nearly 20.5 hours on ISBI 2015, 10 hours on ISLES 2015 and 163 hours on BRATS 2018.

For the segmentation network, AMSGrad optimizer (Reddi et al., 2018) was adopted for all models, both with an initial learning rate of 3×10^{-4} and a weight decay of 3×10^{-5}. For the discriminator, we used the Stochastic Gradient Descent (SGD) (Bottou, 2010), with the initial learning rate of 1×10^{-4}, the momentum of 0.5 and the weight decay of 1×10^{-4} for all three tasks. For the three hyperparameters involved in our proposed framework, λ_{con}, λ_{sdm} and λ_{adv}, we first fixed λ_{con} to be 0.1, which is the same as the original Mean-Teacher model and UA-MT. Then we built a series of experiments for hyperparameter tuning on the ISLES 2015 dataset to determine λ_{sdm} and λ_{adv} as 0.3 and 0.1 respectively. In the other two brain lesion datasets, ISBI 2015 and BRATS 2018, our framework also used the same hyperparameters and initial weights. The detailed update strategy for each step was the same as (Li et al., 2020). Besides, the EMA decay α was 0.99 (Tarvainen and Valpola, 2017).

For the processing of the datasets, we have not used any image augmentation during the training of the evaluated methods in our experiments. Since the preprocessed versions of training and test images in this challenge have been provided for experiments, we only applied N4 bias field correction (Tustison et al., 2010) for all images of ISLES 2015 and BRATS 2018. Then, z-score normalization was performed on each image of three datasets as another preprocessing step. The random crop strategy was used to produce the 3D patch-wise images as input for training. In the testing stage, we fed the uncropped original images into the trained model and obtained the segmentation labels. The source code of our proposed framework is available at https://github.com/wzcgx/MTANS.

4. Results

4.1. Multiple sclerosis lesion segmentation

We first evaluated the performance of the proposed framework and other comparison methods for multiple sclerosis lesion segmentation on the ISBI 2015. The supervised 3D U-Net trained with labeled training data is used as a baseline model, several recent semi-supervised segmentation methods, including MT, UA-MT and SASSNet were selected to compare with our proposed methods. In this experiment, in addition to the baseline trained with the same labeled data (20%, 7 labeled) as the semi-supervised methods, a fully supervised baseline with all labeled data (100%, 35 labeled) was also trained.

First, as shown in Table 1, model training and testing experiments were conducted on ISBI 2015 training data. We can observe that from the perspective of various metrics, there was no comparison semi-supervised method has particularly outstanding performance on this dataset, which meant achieving higher scores than other methods on all metrics. Among them, the proposed MTANS has the best performance in Dice compared with other semi-supervised methods and even achieved a lower LFPR score than the full supervised baseline. More specifically, according to the quantitative segmentation results shown in Fig. 2, our method generated relatively fewer false positives than other comparison methods.

Table 2 shows the quantitative results of unseen test data obtained by the baseline and the methods trained with 20% labeled training set. Through further comparison among the results between semi-supervised methods, the role of the proposed multi-scale feature consistency and shape-aware embedding in our framework can be investigated. As one of the state-of-the-art semi-supervised segmentation methods, SASSNet achieved the best scores in Dice, its WS score was worse than our proposed methods due to the poor performance in LFPR. In comparison, the performance of MTAN in PPV and LFPR was better than SASSNet. As seen in Fig. 3, when comparing the results of MTAN and MTANS, it is noticeable that MTANS further significantly improved the LFPR score of MTAN thanks to the shape-aware embedding.

As a combination of the metrics, the score of WS shows that the comprehensive performance of the methods. It is worth noting that our proposed MTAN and MTANS achieved high scores of 89.39 and 90.86, both were superior to other comparative semi-supervised methods and baseline. In particular, the score of MTANS was higher than the baseline model trained by the 100% training set, with a score of 89.77.

4.2. Ischemic stroke lesion segmentation

In the experiments of the proposed semi-supervised method for ischemic stroke segmentation, we analyzed the segmentation performance and effect of our framework under different settings. In Table 3,

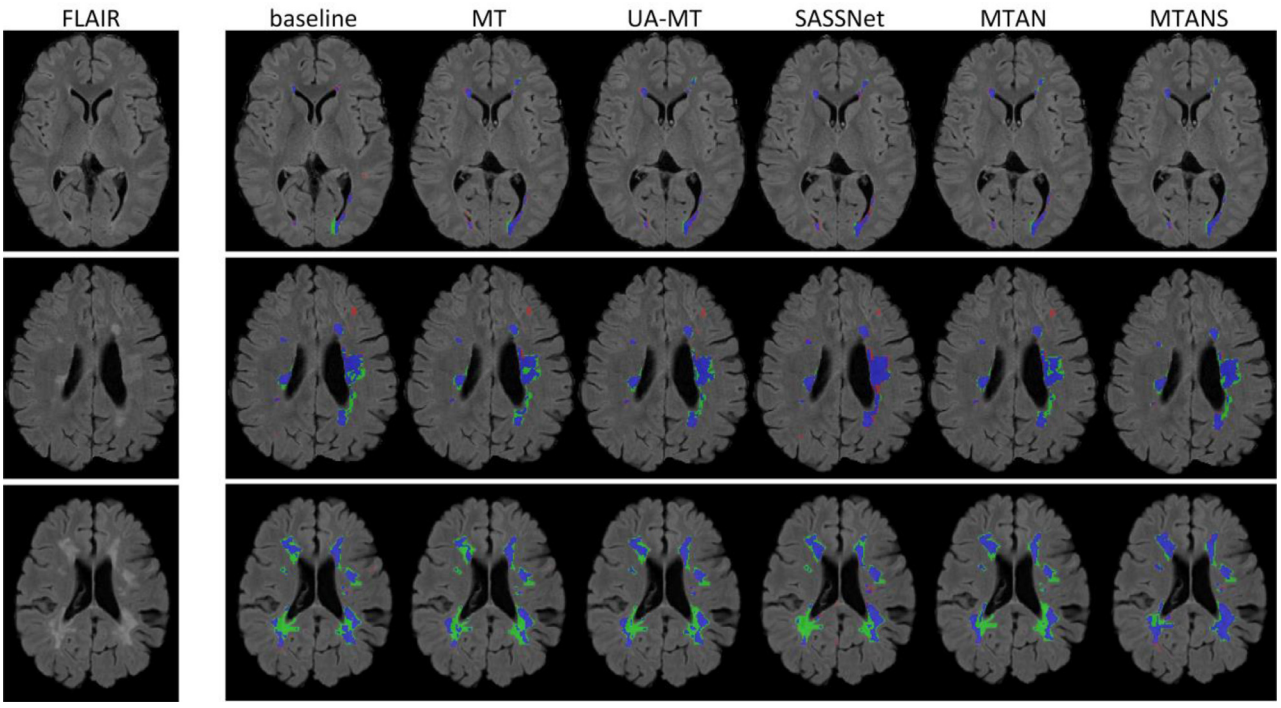

Fig. 2. Examples of three cases from the ISBI 2015 dataset. The segmentation results of each method trained with 20% labeled set are overlapped with the ground truth. The true positives, false negatives and false positives of the result images are colored in blue, green and red, respectively (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 2
Quantitative comparison for the performance of proposed semi-supervised methods, supervised baseline and other semi-supervised methods on the ISBI 2015 unseen test data.

Lab/All	Method	WS	Dice(%)	PPV(%)	TPR(%)	LFPR(%)	LTPR(%)	VC
	baseline	88.48	53.81±12.22	78.71±14.44	43.14±14.78	54.99±20.27	45.03±19.49	0.8325
	MT	88.92	54.15±13.12	79.21±15.53	43.68±15.89	48.94±20.50	45.40±20.53	0.8127
20%	UA-MT	89.33	56.00±13.92	80.99±14.71	45.58±17.15	46.44±21.19	**46.17±21.28**	0.8197
	SASSNet	89.16	**56.89±12.57**	76.76±17.08	**47.95±14.90**	46.71±19.22	44.64±20.94	**0.8431**
	MTAN	89.39	53.80±14.29	81.88±15.63	42.48±16.02	43.79±21.54	44.42±20.67	0.8291
	MTANS	**90.86**	53.12±15.03	**84.26±16.61**	41.21±15.68	**15.09±16.49**	34.29±18.89	0.8301
100%	baseline	89.77	61.75±13.70	78.13±16.16	53.80±17.08	51.63±20.32	53.11±22.33	0.8719

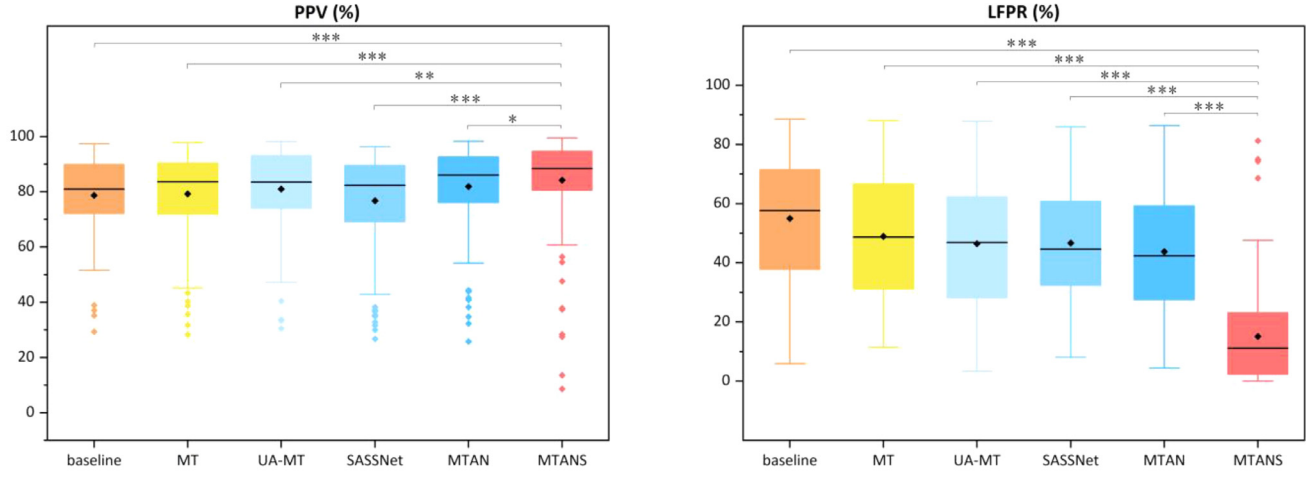

Fig. 3. Box plot of positive predictive value (PPV) and lesion false positive rate (LFPR) for the segmentation results of ISBI test data when trained with 7 scans as labeled images. We have performed a paired student's t-test between the proposed MTANS and other models to calculate p-values. * denotes $p < 0.05$, ** denotes $p < 0.005$, and *** denotes $p < 0.0005$.

Table 3
Quantitative evaluation of our methods and other comparison methods on the ISLES 2015 dataset under the two ratios of labeled training set.

Lab/All	Method	Dice(%)	Precision(%)	Sensitivity(%)	ASSD	HD
10%	baseline	55.47±25.48	57.24±25.20	67.04±3.23	7.70±8.35	61.29±18.45
	MT	55.39±32.58	59.70±34.64	67.19±38.17	9.65±15.63	48.84±26.02
	UA-MT	57.45±26.65	61.25±23.73	68.42±34.55	7.58±8.17	61.95±18.88
	SASSNet	59.63±24.12	63.33±30.03	**74.18±23.55**	5.37±2.75	39.08±22.17
	MTAN	56.16±26.96	65.62±26.05	66.89±33.03	7.10±5.73	70.00±13.24
	MTANS	**61.66±20.71**	**72.52±20.74**	68.10±31.51	**4.65±2.58**	**37.39±22.81**
20%	baseline	59.21±25.37	69.42±23.72	65.63±29.59	4.31±2.07	37.56±15.58
	MT	64.73±18.60	**75.75**±19.24	64.78±26.86	**3.16±1.40**	38.92±25.26
	UA-MT	64.41±21.84	72.59±18.67	68.82±27.85	3.56±1.38	47.65±21.25
	SASSNet	63.99±24.64	60.47±26.19	79.38±16.61	5.30±4.26	49.48±23.30
	MTAN	60.34±25.41	68.62±26.01	69.55±28.41	4.16±2.27	45.87±16.25
	MTANS	**69.08±12.56**	67.97±21.65	**79.51±17.42**	3.51±2.43	**29.75±14.95**

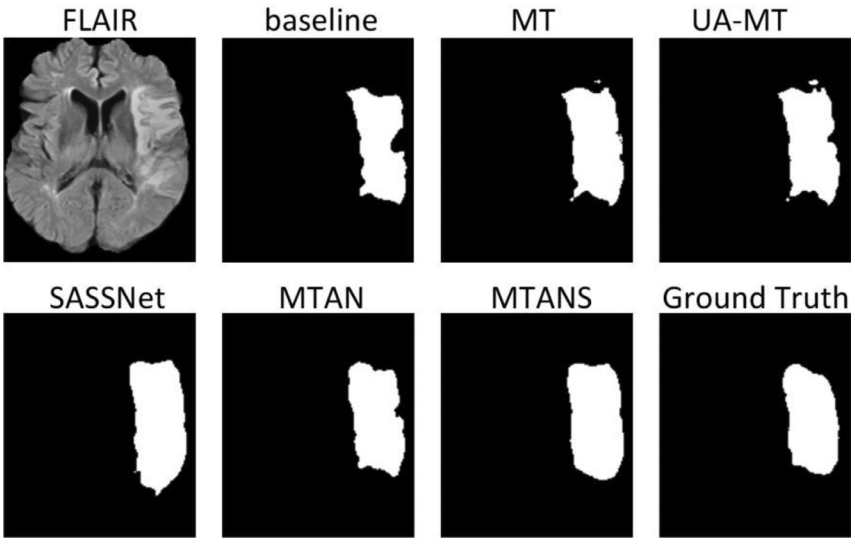

Fig. 4. Qualitative results of segmentation examples obtained by proposed MTAN and MTANS models, and other comparison methods that all trained with only 10% labeled training set on ISLES 2015 dataset.

we present the performance of the models trained on the ISLES 2015 dataset.

Specifically, we first experimented with 10% images of the training set that were regarded as the labeled samples, and the remaining images of the training set that were considered as the unlabeled samples. The proposed MTAN achieved a Dice score of 56.16%, only higher than baseline and MT, but as far as Precision is concerned, the performance of MTAN obtained a better score of 65.62% than comparable semi-supervised methods. Additionally, we can observe that our MTANS has the best semi-supervised performance, with higher Dice and Precision, and lower ASSD and HD measurements. Qualitative results of the several models trained with 10% labeled set on ISLES 2015 are shown in Fig. 4.

Then, we increased the ratio of the labeled training set up to 20% to find out the effect of different labeled and unlabeled data on the segmentation performance. As seen in Fig. 5, with the increase of labeled data, the performance of UA-MT and SASSNet in this experiment was worse than that of the MT model. In contrast, MTANS achieved a Dice score of 69.08%, which was nearly 10% higher than the baseline.

Finally, we performed 5-fold cross-validation on all methods under the 10% setting. The proposed MTAN obtained a better Dice score than UA-MT, which is one of the state-of-the-art methods based on consistency training. In addition, MTAN also achieved better performance in Sensitivity and HD. Although the best score of Precision was obtained, the overall performance of UA-MT on this dataset was not ideal. Considering the Dice and HD scores of MTAN, we notice that its performance was better than the consistency-based methods. Overall, the results in Table 4 show that MTANS was still the best semi-supervised method, and compared to other methods, it has outstanding performance on almost all metrics in this experiment.

4.3. Brain tumor segmentation

We further evaluated the performance of our semi-supervised framework for the segmentation of multi-class lesions in brain tumor images. Tables 5 and 6 present the evaluation performance of our framework and other methods on the BRATS 2018 training data under 10% and 20% experiment settings, respectively. The visual segmentation examples of models trained with 10% setting can be found in Fig. 6, and Fig. 7 shows the detailed boxplot for this experiment.

First of all, when our MTAN only used 10% of the labeled training set, it outperformed other methods only in Specificity. However, the shape-aware embedding still shows an all-around improvement to our MTAN. It can be observed from Fig. 7 that our MTANS was better than MTAN and other methods in Dice, Sensitivity and HD scores of the whole tumor and tumor core regions. We can also observe from Table 6 that not all semi-supervised methods could boost the performance while the labeled training data increases and the unlabeled data decreases, especially for the segmentation of the tumor core region, the Dice of MT, UA-MT, and our MTAN were all lower than the baseline. And for the Dice measurement of the enhancing tumor region, UA-MT performed poorly, while MTAN achieved the highest segmentation score. Besides, MTANS obtained higher scores on most evaluation items than other methods, proving the effectiveness of our semi-supervised framework in this task. Specifically, under both experimental settings, it is especially noticeable when comparing MTANS with baseline, the Dice increased from 79.74% to 83.03%, 81.04% to 84.86% for whole tumor regions, 66.90% to 71.79%, 72.53% to 74.15% for the tumor core. With the comparison of both Sensitivity and HD95, MTANS still achieved the best performance than other semi-supervised methods in these two regions.

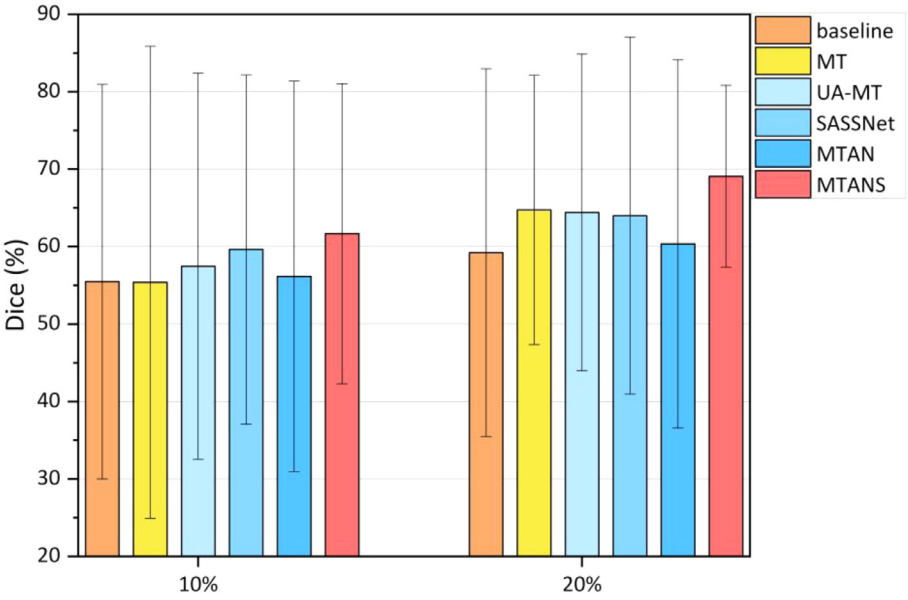

Fig. 5. The bar plot of the performance of mean and standard deviation measured by Dice for baseline, MT, UA-MT, SASSNet, our proposed MTAN and MTANS trained with 10% and 20% labeled training set on ISLES 2015 dataset.

Table 4
Quantitative comparison of our proposed framework and other methods all trained with 10% labeled set, using 5-fold cross-validation on ISLES2015 dataset.

Method	Dice(%)	Precision(%)	Sensitivity(%)	ASSD	HD
baseline	58.41±24.03	70.14±28.37	57.16±27.32	8.32±11.65	55.13±21.71
MT	59.26±23.95	74.28±25.72	57.49±28.53	7.14±10.34	51.87±24.92
UA-MT	59.54±22.27	**77.40±22.68**	57.51±27.55	6.02±5.82	51.40±20.66
SASSNet	61.91±21.25	67.63±27.02	65.54±20.45	7.31±8.41	48.21±31.51
MTAN	61.85±21.13	74.74±22.40	60.33±25.64	7.23±15.33	48.32±22.05
MTANS	**64.41±18.40**	68.43±21.49	**68.11±22.31**	**4.93±4.45**	**40.39±23.70**

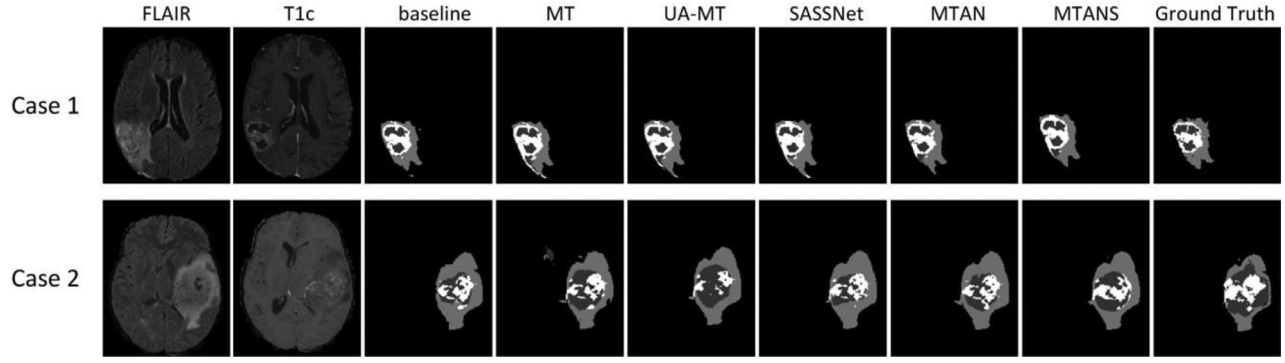

Fig. 6. Qualitative brain tumor segmentation results of two cases from BRATS 2018 dataset achieved by the supervised baseline, comparison semi-supervised methods and the proposed methods that all trained with 10% labeled data.

We further verified the generalizability of all models by directly applying the trained model to the unseen test data. Tables 7 and 8 show the detailed quantitative experiment results of each method on the BRATS 2018 unseen test data. Overall, we can conclude that in terms of the number of best results obtained in these metrics, our proposed MTANS performed better than comparable methods. As seen in Table 7, the Dice of our method on the whole tumor is better than other comparison methods when only trained with 10% labeled data. It is worth noting that the Dice score of MTANS in the whole tumor region is 85.68%, which is even better than the results of other methods trained with 20% labeled data in Table 8, except for UA-MT. In addition, our method has the best score of Specificity and Sensitivity in this region. And in the experiments of 20% setting, MTANS achieved the best results on more metrics, especially the improvement of the Dice and Sensitivity in the tumor core region.

5. Discussion

We have evaluated the performance of our proposed semi-supervised framework and other comparison methods, with their applicability in three different brain lesion segmentation tasks. As one of the relevant semi-supervised methods, the consistency-based model has been evaluated and compared in detail. In the first segmentation task, the proposed multi-scale consistency loss shown more comprehensive performance. The website score of our MTAN was better than methods based on original consistency loss, uncertainty-based consistency loss, and shape-aware semi-supervised strategy, as shown in Table 2. Also, we can observe that these consistency-based models may not be stable enough from the further ischemic stroke lesion segmentation experiments. As seen in Table 3, although MT performed better than UA-MT and MTAN in Dice under 20% experimental setting, it could not outperform base-

Table 5
Evaluation results using four metrics obtained by our methods trained with 10% labeled training set on BRATS 2018 training dataset and the comparison with supervised baseline and other semi-supervised methods.

Method	Dice (%)			Specificity (%)			Sensitivity (%)			HD95 (mm)		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
baseline	79.74±17.75	66.90±26.22	60.23±32.07	99.75±0.38	99.90±1.41	99.96±0.58	82.10±17.00	68.13±28.25	62.87±33.67	18.71±21.97	20.51±25.10	20.21±31.81
MT	80.19±15.95	71.09±24.59	**64.71±28.74**	99.72±0.43	99.90±0.15	99.95±0.08	84.16±14.72	72.14±26.14	67.36±30.51	16.98±17.99	17.28±22.84	18.32±30.00
UA-MT	81.18±17.29	70.72±25.16	62.79±30.74	99.73±0.43	99.91±0.14	**99.96±0.06**	85.68±15.68	71.69±26.85	63.78±33.39	14.32±18.12	16.52±22.38	18.86±31.32
SASSNet	82.95±13.81	71.25±24.20	64.17±26.97	99.75±0.32	99.91±0.12	99.93±0.08	87.32±12.20	71.42±27.24	**70.55±30.29**	14.37±18.83	14.70±20.46	18.87±31.42
MTAN	81.14±16.32	70.04±25.10	62.42±30.24	**99.77±0.32**	**99.92±0.10**	**99.96±0.06**	84.03±16.59	69.59±27.43	63.19±32.56	14.62±16.99	15.27±21.23	19.25±31.23
MTANS	**83.03±17.44**	**71.79±24.62**	60.99±28.82	**99.77±0.32**	99.91±0.11	99.94±0.10	**87.66±16.02**	**74.38±28.09**	64.16±32.60	**12.38±18.95**	**12.75±17.75**	**16.67±26.59**

Table 6
Evaluation results using four metrics obtained by our methods trained with 20% labeled set on BRATS 2018 training dataset and the comparison with supervised baseline and other semi-supervised methods.

Method	Dice (%)			Specificity (%)			Sensitivity (%)			HD95 (mm)		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
baseline	81.04±18.33	72.53±25.18	64.83±28.98	99.68±0.75	99.90±0.15	99.94±0.08	84.51±16.58	74.70±27.27	**70.30±31.97**	13.99±17.55	15.23±20.23	16.50±28.52
MT	81.72±18.83	71.86±25.65	64.95±29.53	99.72±0.60	99.91±0.18	99.95±0.06	84.30±15.77	71.38±27.35	67.97±31.80	14.25±18.61	15.32±21.36	14.75±27.62
UA-MT	83.52±13.10	71.29±25.37	64.30±30.27	**99.87±0.24**	99.91±0.17	**99.97±0.05**	81.08±16.07	71.27±27.26	63.17±31.32	10.01±11.83	13.45±19.60	15.71±28.41
SASSNet	83.93±14.98	72.68±24.71	65.00±29.07	99.77±0.49	99.85±0.25	99.95±0.07	85.14±13.18	75.40±25.49	67.41±31.88	12.03±16.91	12.75±17.54	16.98±27.86
MTAN	82.56±15.72	70.82±25.16	**65.93±29.61**	99.85±0.23	**99.94±0.09**	99.96±0.05	81.68±16.76	69.39±28.05	67.18±31.36	13.65±18.16	12.94±16.67	16.18±28.42
MTANS	**84.86±14.43**	**74.15±22.99**	65.05±27.85	99.82±0.22	99.89±0.15	99.95±0.05	**86.97±14.33**	**77.05±24.83**	68.82±30.43	**9.16±11.86**	**11.23±13.99**	**14.20±24.66**

Table 7
Segmentation performance of proposed semi-supervised methods on the BRATS 2018 unseen test data and the comparison with supervised baseline and other semi-supervised methods that all trained with 10% labeled set.

Method	Dice (%)			Specificity (%)			Sensitivity (%)			HD95 (mm)		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
baseline	82.77±14.73	66.94±26.37	70.22±27.31	97.13±12.35	97.83±12.27	98.23±12.28	86.27±17.42	70.98±28.89	76.17±25.08	16.74±21.33	18.08±21.44	26.85±78.45
MT	82.65±14.19	69.64±27.12	**72.40±25.20**	97.01±12.45	98.04±12.28	98.21±12.28	87.00±16.54	70.69±29.65	76.79±24.24	15.98±18.06	16.41±20.37	21.52±65.91
UA-MT	83.93±12.79	69.75±28.21	72.18±26.11	98.72±1.78	**99.66±0.44**	**99.77±0.26**	88.56±16.02	70.60±29.98	76.75±25.84	14.46±19.34	**14.24±17.74**	8.45±15.44
SASSNet	84.20±10.81	**70.22±28.42**	71.88±26.49	98.61±1.85	99.65±0.47	99.68±0.34	89.65±14.00	70.86±30.89	**80.97±23.34**	**13.29±17.37**	15.89±20.65	8.81±16.01
MTAN	83.74±13.56	68.10±27.72	72.00±26.69	97.18±12.25	98.02±12.28	98.26±12.28	87.54±16.15	68.17±29.97	74.89±26.52	13.62±17.83	21.58±48.80	25.29±78.87
MTANS	**85.68±8.70**	68.80±27.30	69.63±27.55	**98.79±1.15**	99.42±0.89	99.75±0.30	**90.39±11.38**	**73.18±29.75**	74.90±26.70	13.87±20.38	15.18±21.19	**7.69±12.96**

Table 8
Segmentation performance of proposed semi-supervised methods on the BRATS 2018 unseen test data and the comparison with supervised baseline and other semi-supervised methods that all trained with 20% labeled set.

Method	Dice (%)			Specificity (%)			Sensitivity (%)			HD95 (mm)		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
baseline	85.27±9.35	71.97±26.70	72.64±25.97	98.48±3.88	99.35±1.14	99.68±0.38	88.87±10.75	76.68±27.53	**80.37±23.07**	10.36±12.07	11.59±12.03	6.28±9.90
MT	84.34±15.92	73.55±25.45	72.53±25.37	**98.66±4.49**	**99.45±1.75**	99.64±0.83	86.29±17.56	74.61±27.84	78.82±22.34	9.19±8.92	**10.97±14.23**	**5.59±9.73**
UA-MT	85.80±13.33	73.61±25.77	71.83±26.09	97.69±12.32	98.13±12.28	98.28±12.29	85.47±15.67	73.70±27.38	73.58±25.22	8.95±13.90	22.41±64.90	22.96±78.50
SASSNet	85.15±16.29	73.32±27.10	71.82±26.14	97.14±13.13	97.80±12.41	98.17±12.28	86.75±17.85	76.34±27.79	78.65±25.61	**8.17±13.80**	22.81±65.13	18.54±65.15
MTAN	84.45±16.70	73.26±26.13	**73.40±24.58**	97.33±12.82	98.11±12.29	98.25±12.28	85.01±19.01	72.63±28.12	77.11±23.77	9.38±14.50	18.24±48.02	18.80±65.44
MTANS	**86.35±8.14**	**73.66±24.82**	71.68±26.10	**98.66±2.72**	99.43±0.84	**99.73±0.34**	**91.15±10.07**	**78.00±25.84**	77.79±24.29	**8.17±9.88**	12.10±15.53	5.69±8.80

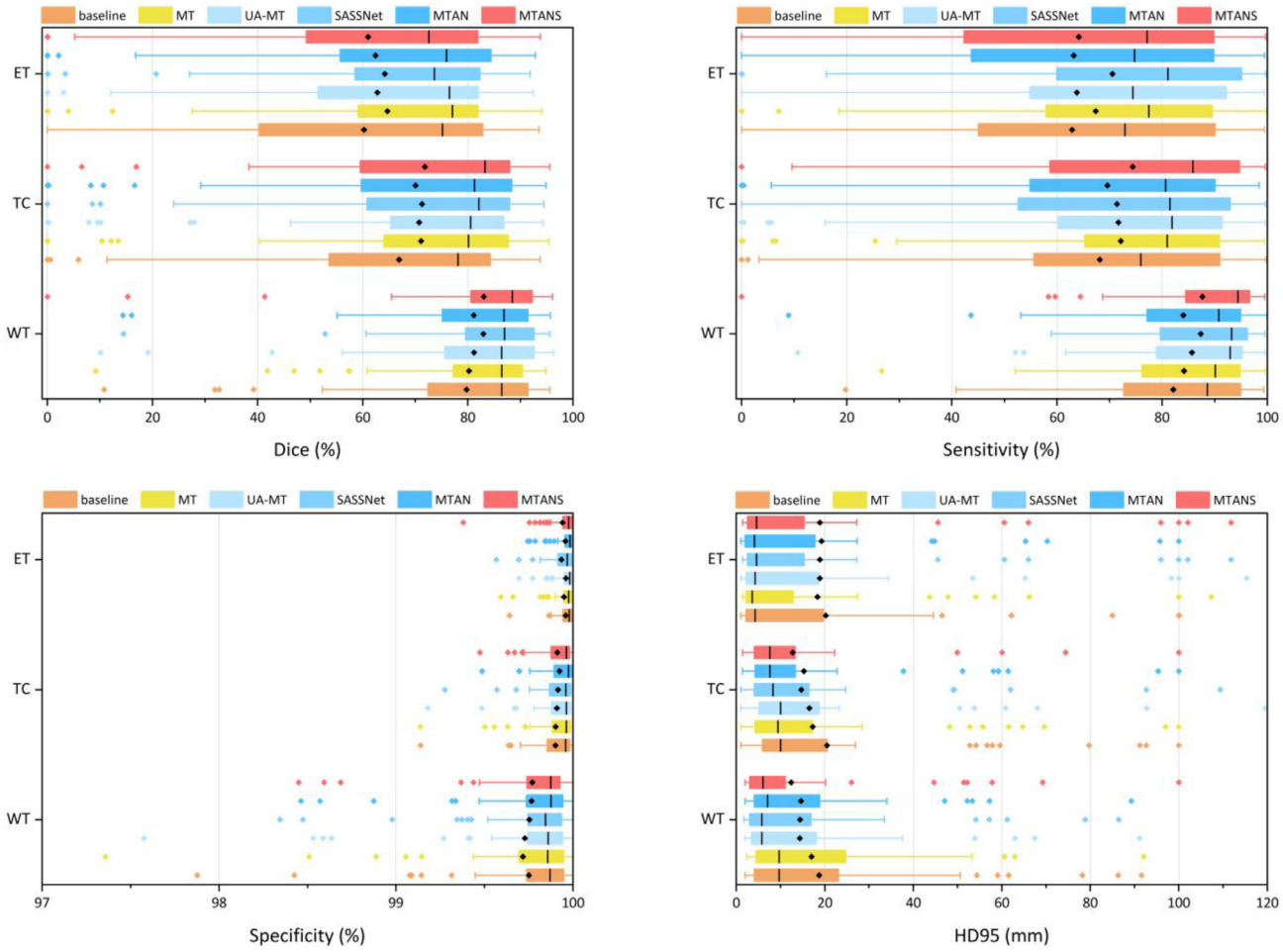

Fig. 7. Box plot of the experimental results of our proposed MTAN and MTANS models, and other comparison methods on BRATS 2018 training dataset, when training with 10% labeled set.

line under 10% experimental setting with fewer labeled data and more unlabeled data. In contrast, when the ratio of the labeled training set is smaller, the multi-scale consistency loss could effectively use more unlabeled data to improve the original consistency loss, and was comparable to the uncertainty-based model. Similarly, for the segmentation of brain tumor with more classes, as shown in Tables 7 and 8, the Dice performance of MT was better than that of UA-MT in enhancing tumor segmentation but was worse than baseline in the segmentation of the whole tumor region.

The semi-supervised application of the adversarial network mainly benefits from the ability of the discriminator to provide extra supervision for the unlabeled data. In our framework, the proposed multi-scale consistency loss requires a modality showing obvious brain lesions, and this loss is designed for capturing the semantic information of labeled data through adversarial learning and achieving anatomical consistency of unlabeled data based on the student and teacher models. As shown in Table 2, Fig. 3 and Table 4, The overall performance of MTAN on ISBI 2015 and cross-validation results on ISLES 2015 were both slightly better than UA-MT.

Our experiments also show that the combination of the proposed consistency loss and shape-aware embedding based on another adversarial learning is obvious for performance improvement. As shown in Tables 1 and 2, the proposed MTANS has achieved the best LFPR scores on both training data and test data from ISBI 2015. According to the low LFPR of MTAN and the high TPR of SASSNet, we interpret the best performance of MTANS in LFPR as the effective combination of the reduction of false positives due to multi-scale consistency learning and the increase of true positives due to shape-aware learning. The quantitative results in Fig. 2 also show that MTAN generated relatively few false positives, while SASSNet generated relatively more true positives. In general, the human-level performance of ISBI 2015 can be achieved in our framework with only 7 labeled images for training. Similarly, for ischemic stroke and brain tumor, the lesion area is uncertain and its shape is irregular, but the performance can also be improved by applying shape constraint combined with our consistency strategy. In the experiments of ISLES 2015 and the segmentation of whole tumor regions on BRATS 2018, MTANS achieved the best results on more metrics compared with other methods.

In clinical applications, the labeling of 3D medical images often requires efforts and time from experts, and the existing automatic labeling tools often need large-scale labeled data for training. Semi-supervised learning allows experts to label only a small amount of data and the tools can be trained with the remaining unlabeled data. As our experimental results show, the segmentation performance of semi-supervised methods using extra unlabeled data is better than baseline using only labeled data, so how to train models with both labeled and unlabeled data more effectively is one of the important research directions of semi-supervised learning. Compared with the related work, our method shows stable and better performance in three experimental datasets without changing the architecture and hyperparameters. Therefore, this framework has the potential to be used as a universal tool for the annotation of brain lesion images.

One of the limitations of our semi-supervised segmentation framework is that although its overall performance was better than other com-

parison semi-supervised methods, it could not achieve the best scores of all metrics in the three segmentation tasks. In our future work, since the proposed framework is extensible, we would consider replacing the segmentation network with the current state-of-the-art models on the three challenge datasets to improve the performance. In addition, we will collect more clinical or cross-modal brain lesion images as unlabeled data to further validate the effectiveness of the proposed framework. Besides, the influence of the different ratios between unlabeled data and labeled data on the performance of semi-supervised learning also needs to be further studied.

6. Conclusions

In this paper, a novel semi-supervised framework for joint training of multi-scale mean teacher and improved adversarial network for multi-modal brain lesion segmentation is presented. Based on two kinds of adversarial learning, we embed a shape-aware strategy into the student and teacher models which also integrate the proposed multi-scale consistent regularization. Three public datasets related to 3D brain lesion segmentation were used to evaluate the performance of our semi-supervised framework on multiple sclerosis lesion, ischemic stroke lesion, and brain tumor segmentation tasks. Compared with the supervised methods trained with the same labeled data, the proposed framework improved the segmentation results, and the overall performance was also better than current state-of-the-art consistency training and shape-aware learning methods for semi-supervised medical image segmentation. Our work is expected to reduce the need for large-scale labeling in medical imaging and be served as an auxiliary tool to produce annotations for unlabeled data by only using a small amount of labeled data.

Declaration of Competing Interest

None.

Credit Author Statement

Gaoxiang Chen: Conceptualization, Methodology, Software, Writing-Original draft preparation. **Jintao Ru:** Software, Validation. **Yilin Zhou:** Visualization, Investigation. **Islem Rekik:** Writing-Reviewing and Editing. **Zhifang Pan:** Supervision, Project administration, Funding acquisition. **Xiaoming Liu:** Validation. **Yezhi Lin:** Writing-Reviewing and Editing, Funding acquisition. **Beichen Lu:** Formal analysis. **Jialin Shi:** Validation.

Data and Code Availability Statement

The datasets used for the experiments are available on the Longitudinal Multiple Sclerosis Lesion Segmentation Challenge website[3], ISLES Challenge 2015 website[4]and BraTS 2018 website[5].

The source code is available at https://github.com/wzcgx/MTANS.

Ethics Statement

The data used in this work is downloaded from publicly available websites of the ISBI Longitudinal MS Lesion Segmentation Challenge, ISLES 2015 and BRATS 2018.

[3] https://smart-stats-tools.org/lesion-challenge-2015
[4] https://www.smir.ch/ISLES/Start2015
[5] https://ipp.cbica.upenn.edu/categories/brats2018

Acknowledgments

This work was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LY21F020030 and No. LY16F030010, Wenzhou Science & Technology Bureau under Grant No.2018ZG016, and the National Natural Science Foundation of China under Grant No. 11901437.

References

Ahn, J., Kwak, S., 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4981–4990. doi:10.1109/CVPR.2018.00523.

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J., 2017. Deep learning for brain MRI Segmentation: state of the art and future directions. J. Digit. Imaging 30, 449–459. doi:10.1007/s10278-017-9983-4.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. doi:10.1007/978-3-7908-2604-3_16.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., Jorge Cardoso, M., Cawley, N., Ciccarelli, O., Wheeler-Kingshott, C.A.M., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick, O., Barillot, C., Tomas-Fernandez, X., Warfield, S.K., Vaidya, S., Chunduru, A., Muthuganapathy, R., Krishnamurthi, G., Jesson, A., Arbel, T., Maier, O., Handels, H., Iheme, L.O., Unay, D., Jain, S., Sima, D.M., Smeets, D., Ghafoorian, M., Platel, B., Birenbaum, A., Greenspan, H., Bazin, P.L., Calabresi, P.A., Crainiceanu, C.M., Ellingsen, L.M., Reich, D.S., Prince, J.L., Pham, D.L., 2017. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. Neuroimage 148, 77–102. doi:10.1016/j.neuroimage.2016.12.064.

Chen, G., Li, Q., Shi, F., Rekik, I., Pan, Z., 2020. RFDCR: automated brain lesion segmentation using cascaded random forests with dense conditional random fields. Neuroimage 211, 116620. doi:10.1016/j.neuroimage.2020.116620.

Cheplygina, V., de Bruijne, M., Pluim, J.P.W., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296. doi:10.1016/j.media.2019.03.009.

W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, C. Ye, 2019. Semi-supervised brain lesion segmentation with an adapted mean teacher model, in: Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture noTes in bioinforMatics). pp. 554–565. 10.1007/978-3-030-20351-1_43

Dalca, A.V., Guttag, J., Sabuncu, M.R., 2018. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 9290–9299. doi:10.1109/CVPR.2018.00968.

Dangi, S., Linte, C.A., Yaniv, Z., 2019. A distance map regularized CNN for cardiac cine MR image segmentation. Med. Phys. 46, 5637–5651. doi:10.1002/mp.13853.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial Networks. In: Proceedings of the First 12 Conferences Advances in Neural Information Processing Systems, pp. 2672–2680 arXiv:1406.2661 .

Hong, S., Noh, H., Han, B., 2015. Decoupled deep neural network for semi-supervised semantic segmentation. In: Proceedings of the First 12 Conferences Advances in Neural Information Processing Systems, pp. 1495–1503 arXiv:1506.04924 .

Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J., 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7014–7023. doi:10.1109/CVPR.2018.00733.

W.C. Hung, Y.H. Tsai, Y.T. Liou, Y.Y. Lin, M.H. Yang, 2018. Adversarial learning for semi-supervised semantic segmentation. arXiv:1802.07934

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H., 2018. Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. Lect. Notes Comput. Sci. 287–297. doi:10.1007/978-3-319-75238-9_25.

Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78. doi:10.1016/j.media.2016.10.004.

Kao, P.Y., Ngo, T., Zhang, A., Chen, J.W., Manjunath, B.S., 2019. Brain tumor segmentation and tractographic feature extraction from structural MR images for overall survival prediction. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science, Vol 11384. Springer, Cham doi:10.1007/978-3-030-11726-9_12.

Kaus, M.R., Warfield, S.K., Nabavi, A., Black, P.M., Jolesz, F.A., Kikinis, R., 2001. Automated segmentation of MR images of brain tumors. Radiology 218, 586–591. doi:10.1148/radiology.218.2.r01fe44586.

Laine, S., Aila, T., 2017. Temporal ensembling for semi-supervised learning. In: Proceedings of the 5th International Conference on Learning Representations, ICLR 2017-Conference Track arXiv:1610.02242 .

Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S., 2019. Ficklenet: weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5262–5271. doi:10.1109/CVPR.2019.00541.

Li, S., Zhang, C., He, X., 2020. Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Lecture Notes in Computer Science (Including

Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Science and Business Media Deutschland GmbH, pp. 552–561. doi:10.1007/978-3-030-59710-8_54.

Lu, Z., Fu, Z., Xiang, T., Han, P., Wang, L., Gao, X., 2017. Learning from weak and noisy labels for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 486–500. doi:10.1109/TPAMI.2016.2552172.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H.L., Havaei, M., Iftekharuddin, K.M., Jodoin, P.M., Kamnitsas, K., Kellner, E., Korvenoja, A., Larochelle, H., Ledig, C., Lee, J.H., Maes, F., Mahmood, Q., Maier-Hein, K.H., McKinley, R., Muschelli, J., Pal, C., Pei, L., Rangarajan, J.R., Reza, S.M.S., Robben, D., Rueckert, D., Salli, E., Suetens, P., Wang, C.W., Wilms, M., Kirschke, J.S., Krämer, U.M., Münte, T.F., Schramm, P., Wiest, R., Handels, H., Reyes, M., 2017. ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. 35, 250–269. doi:10.1016/j.media.2016.07.009.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34, 1993–2024. doi:10.1109/TMI.2014.2377694.

Mittal, S., Tatarchenko, M., Brox, T., 2019. Semi-supervised semantic segmentation with high- and low-level consistency. IEEE Trans. Pattern Anal. Mach. Intell 1–1. doi:10.1109/tpami.2019.2960224.

Miyato, T., Maeda, S.I., Koyama, M., Ishii, S., 2019. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 41, 1979–1993. doi:10.1109/TPAMI.2018.2858821.

Nie, D., Gao, Y., Wang, L., Shen, D., 2018. ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 370–378. doi:10.1007/978-3-030-00937-3_43 11073 LNCS.

Ouali, Y., Hudelot, C., Tami, M., 2020. Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 12671–12681. doi:10.1109/CVPR42600.2020.01269.

Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1742–1750. doi:10.1109/ICCV.2015.203.

Peng, J., Estrada, G., Pedersoli, M., Desrosiers, C., 2020. Deep co-training for semi-supervised image segmentation. Pattern Recognit. doi:10.1016/j.patcog.2020.107269.

Perone, C.S., Cohen-Adad, J., 2018. Deep semi-supervised segmentation with weight-averaged consistency targets. In: Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 12–19. doi:10.1007/978-3-030-00889-5_2.

Reddi, S.J., Kale, S., Kumar, S., 2018. On the convergence of Adam and beyond. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)-Conference Track, pp. 1–23.

Song, C., Huang, Y., Ouyang, W., Wang, L., 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3131–3140. doi:10.1109/CVPR.2019.00325.

Souly, N., Spampinato, C., Shah, M., 2017. Semi supervised semantic segmentation using generative adversarial network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5689–5697. doi:10.1109/ICCV.2017.606.

Sun, Y., Zhou, C., Fu, Y., Xue, X., 2019. Parasitic GAN for semi-supervised brain tumor segmentation. In: Proceedings-International Conference on Image Processing, ICIP, pp. 1535–1539. doi:10.1109/ICIP.2019.8803073.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1196–1205 arXiv:1703.01780 .

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320. doi:10.1109/TMI.2010.2046908.

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S., 2018. Revisiting dilated convolution: a simple approach for weakly- and semi-supervised semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7268–7277. doi:10.1109/CVPR.2018.00759.

Y. Xue, H. Tang, Z. Qiao, G. Gong, Y. Yin, Z. Qian, C. Huang, W. Fan, X. Huang, 2019. Shape-aware organ segmentation by predicting signed distance maps. 10.1609/aaai.v34i07.6946

Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X., 2018. SegAN: adversarial network with multi-scale L 1 loss for medical image segmentation. Neuroinformatics 16, 383–392. doi:10.1007/s12021-018-9377-x.

Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) doi:10.1007/978-3-030-32245-8_67.

Zhang, H., Valcarcel, A.M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R.T., Hett, K., Oguz, I., 2019. Multiple Sclerosis Lesion Segmentation with Tiramisu and 2.5D Stacked Slices. Springer International Publishing Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) doi:10.1007/978-3-030-32248-9_38.

Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z., 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 408–416. doi:10.1007/978-3-319-66179-7_47.