IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 15, 2022, accepted February 24, 2022, date of publication March 4, 2022, date of current version March 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156898

Multi-Modality Reconstruction Attention and Difference Enhancement Network for Brain MRI Image Segmentation

XIANGFEN ZHANG, YAN LIU, QINGYI ZHANG, AND FEINIU YUAN, (Senior Member, IEEE)
The College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China

Corresponding authors: Xiangfen Zhang (xiangfen@shnu.edu.cn) and Feiniu Yuan (yfn@shnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 61862029 and Project 62171285, and in part by the General Research Fund of Shanghai Normal University under Project KF2021100 and Project Sk201220.

ABSTRACT An important prerequisite for brain disease diagnosis is to segment brain tissues of Magnetic Resonance Imaging (MRI) into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF). To improve performance, we propose a Multi-modality Reconstruction Attention and Difference Enhancement Network (MRADE-Net). We first stack three inputs of multiple MRI modalities along axial, sagittal, and coronal axes to form three enlarged volumes. Then we adopt global average pooling along each axis, fully connected layers, and activation functions to produce three orthogonal 2D coefficient maps, which are used to reconstruct a 3D attention map for weighting the three inputs. These weighted inputs are added together to generate a feature map of Multi-modality Reconstruction Attention (MRA). Similarly, we present Single-modality Reconstruction Attention (SRA) for improving feature representation abilities in middle stages. In addition, the difference of encoding features between adjacent layers is used to compensate the loss of spatial information caused by down-sampling. Experimental results show that the proposed approach is more effective than the existing state-of-the-art segmentation methods and its performance is verified on several datasets.

INDEX TERMS Attention mechanism, feature difference, image segmentation, MRI, multi-modality.

I. INTRODUCTION

MRI is a pervasive and non-invasive imaging technology. Compared with other imaging technologies, it has many advantages, including higher contrast for different soft tissues, higher spatial resolution and more complete analysis of the organs in different dimensions [1]. Nowadays, MRI image segmentation is widely used in the field of medical brain image processing. It not only helps to detect brain lesions such as tumors, Alzheimer's disease and Parkinson's disease, but also contributes to analyze physiological changes of the aging brain [2].

In the last decades, many automatic image segmentation methods have been proposed, including traditional approaches and approaches based on deep learning. Traditional methods use various image filters to achieve feature extraction of target images. For example, in the task of MRI image segmentation, Kitrungrotsakul *et al.* [3] used organ boundary to segment lungs in medical images. Although these methods can achieve medical segmentation to a certain extent, they are easily affected by noise, resulting in poor segmentation accuracy.

Recently, with the rapid development of convolutional neural networks (CNNs) [4], the development of medical image segmentation methods based on deep learning have been extremely rapid. For instance, inspired by the local receptive field in CNNs model, Huang *et al.* [5] proposed a breast ultrasound image segmentation method based on semantic classification and stitching. This method combines neural network with traditional algorithms to improve the segmentation accuracy. Zhou *et al.* [6] proposed a deep learning based semi-automatic segmentation method to segment the media-adventitia and lumen-intima boundaries. Especially, they proposed a dynamic convolutional neural network to classify the small blocks generated by the norm line sliding window of the initial contour. Following the development

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Magno.

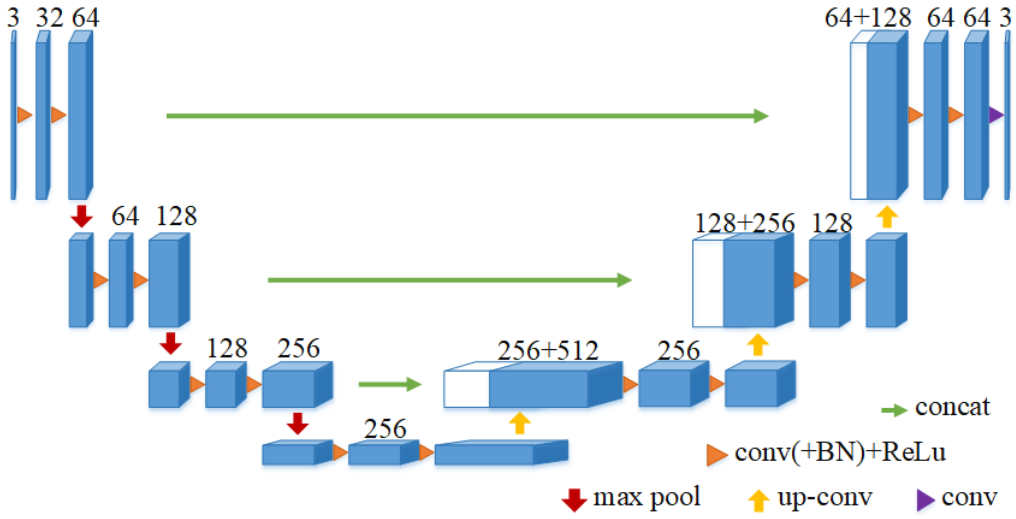

FIGURE 1. 3D-UNet architecture.

trend of CNNs, researchers utilize various CNNs to learn image feature representation from medical images. Fully convolutional networks (FCN) [7] can achieve pixel level classification, and has been widely used in complex segmentation tasks. Inspired by FCN, many FCN-based algorithms are presented. For example, Tian *et al.* [8] proposed an end-to-end deep fully convolutional neural network (PSNet) to segment the prostate automatically. Zhou *et al.* [9] introduced FCN in the 2.5D approach to segment organs based on 3D CT images. Nie *et al.* [10] proposed CC-3-D-FCN to segment the multi-modality neural images of 11 healthy babies. Christ *et al.* [11] proposed a cascaded FCN method to superimpose multiple FCNs together, and the prediction map obtained by each FCN will be sent to the next FCN for context feature extraction, thereby improving the segmentation accuracy. Ronneberger *et al.* [12] used the idea of FCN and proposed the U-Net architecture. It can effectively extract the representative features of medical images for segmentation under reasonable network depth. However, these methods may fail to capture some features that are more related to the segmentation object and ignore the information redundancy in concatenation. They are prone to degrading the performance of predicting the boundary between different classes.

In this paper, we propose a MRADE-Net for brain tissue segmentation on MRI images. By designing a Feature Difference Module (FDM), we can effectively reduce information redundancy and enhance the robustness of features at the same time. We propose an SRA module for middle stages to focus on more important information, thereby we can further enhance the feature representation capabilities of the network. Single modality data cannot provide sufficient tissue information, so we integrate multi-modality images (T1, T1-IR, and T2-FLAIR) into feature extraction. In addition, we design a Multi-level Deep Supervision Module (MDSM) in the segmentation network, leading to a joint objective function that supervises the feature extraction in the middle stages of the network. In conclusion, the main contributions of this article are as follows:

- To compensate for the loss of spatial information, we adopt a FDM to effectively minimize redundant information, which is produced by feature concatenation between encoders and decoders at different levels.
- We propose MRA to efficiently fuse information from different MRI modalities and enhance feature robustness. Three inputs of different modalities are stacked along different directions to produce three volumes, which are globally pooled to obtain three 2D maps containing contextual information about multiple modalities. Then, a 3D coefficient map is reconstructed by spanning the three 2D maps for achieving attention features. Thus, we can effectively fuse different modality MRI images and simultaneously implement attention mechanism. In addition, we design SRA to improve feature representation capabilities for middle features.
- The multi-level deep supervision mechanism is introduced in the training stage to accelerate the convergence speed of our network and optimize the mapping ability of the network to extract powerful features.

The remainder of this paper is as follows. Section II reviews related work. The details of the proposed method are shown in Section III. In Section IV, we introduce the materials and the experimental results. A discussion of the proposed network is proposed in section V.

II. RELATED WORK

A. 3D-UNet SEGMENTATION NETWORK

As shown in Fig. 1, 3D-UNet is a neural network with symmetric encoder and decoder [13]. In 3D-UNet, the feature of the encoder will be input to the decoder with equal resolution to provide the essential high-resolution features throughout

the network. The 3D-UNet take 3D volumes as input and processes them with 3D operations, such as 3D convolution, 3D max pooling, etc. This way can retrieve 3D spatial features to improve segmentation accuracy.

Based on 3D-UNet, Sun *et al.* [14] proposed an improved 3D-UNet with volumetric feature recalibration layer, called SW-3D-Unet to make full use of the spatial contextual features on inter-plane level. Yi-Jie *et al.* [15] presented 3D RU-Net for ROI localization and intra-region segmentation by using the ROI obtained from the encoder, and multiple levels of ROI are cut out from the regional features of the encoder, thus expanding the applicable volume size and effective sensing domain.

Although these deep learning based automated segmentation methods have achieved good performance on segmenting medical images, they are incapable of obtaining refined spatial information or reducing the information redundancy caused by concatenation. Therefore, we design the MRADE-Net, which regards the 3D-UNet as the baseline and introduces multi-modality reconstruction attention mechanisms and multi-modality fusion strategy in the input layer to perform the segmentation task.

B. ATTENTION MECHANISM

In human cognition, the information obtained from different senses is weighted by attention mechanism [16], [17]. This attention mechanism makes human beings to selectively focus on important information [18]. Inspired by this, the Google Deep Mind team proposed attention mechanism when performing the image classification task, which sparked an upsurge in attention mechanism research. For example, SENet [19] was proposed to recalibrate channeled feature responses adaptively by explicitly modeling the channel connection. Residual Attention Network [20] was constructed by stacking attention modules, which generate attention perception features. Specially, CBAM [21] is a lightweight architecture that uses both spatial and channel-wise attention to improve DNN performance. In addition to channel attention and spatial attention, some researchers also use other attention mechanisms. For example, Sun *et al.* [22] proposed a new stack attention U-Net (SAUN) for segmenting left ventricle.

Recently, numerous methods apply attention mechanism to medical image segmentation. Kaul *et al.* [23] proposed FocusNet, which integrates attention into FCN and realizes medical image segmentation through feature maps generated by convolutional encoder. Inspired by these attention mechanisms, we propose SRA to improve feature representation capabilities.

C. MULTI-MODALITY FUSION

In medical image analysis, due to the fact that multi-modality (such as T1, T1-IR, T2-FLAIR, etc.) data can provide complementary information for medical research, the fused information of multi-modality is widely used to achieve multi-tissue segmentation [24] and lesion segmentation [25], [26]. Deep learning based multi-modality image segmentation networks can be classified into layer-level fusion network, decision-level fusion network and input-level fusion network [27]. In the layer-level fusion network, images in each modality are implemented as input for training the enhancement network, and then these learned individual features will be fused in the network layer. The layer-level fusion network can effectively integrate and make use of multi-modality images [28], [29]. In the decision-level fusion segmentation network, the segmentation result of each modality image is obtained through its own separate segmentation network, and the respective segmentation results are then combined to achieve the final segmentation [30], [31]. The input-level fusion network [2], [32] usually stacks the multi-modality image in the channel dimension to get the fused feature that may be used to train the segmentation network.

Considering that the input-level fusion strategy can retain the original image information to the maximum extent and learn the inherent features of the image, we rely on the input-level fusion strategy to fully exploit the feature representation from multi-modal images. Specifically, in order to lay more emphasis on the vital information, we propose MRA to efficiently fuse information from different MRI modalities in the input-level fusion strategy.

III. THE PROPOSED METHOD

A. OVERVIEW OF THE PROPOSED METHOD

The Multi-modality Reconstruction Attention and Difference Enhancement Network (MRADE-Net) is an end-to-end trainable network, which contains three parts: multi-modality fusion part, an encoder part and a decoder part. The framework of our method is shown in Fig. 2. We have three inputs of MRI modalities (T1, T1-IR, T2-FLAIR) for the multi-modality fusion part. Each modality data is first processed by three convolution layers to extract features, and then the extracted features from the three inputs with different modalities are sent to the MRA for obtaining the fused 3D features as the input of the encoder part. Each down-sampling stage of the encoder part consists of a Max Pooling layer, two groups of convolutions, rectified linear unit (ReLU) and Batch Normalization (BN) layers. In the decoder part, we design a symmetrical structure similar to the encoder part. The difference is that we use a de-convolution layer to up-sample feature maps instead of max-pooling.

Furthermore, we also apply the SRA in both encoding and decoding stages of the overall network structure, in order to focus on more useful information of brain tissue. In particular, the FDM generates the differential information between two adjacent levels of the decoder, which is concatenated to the same level layer of the decoder. Thus, we can simultaneously reduce information redundancy and remedy spatial information loss due to down-sampling. As for the network training, we use MDSM to supervise four middle feature maps for further optimizing network parameters and improving segmentation accuracy.

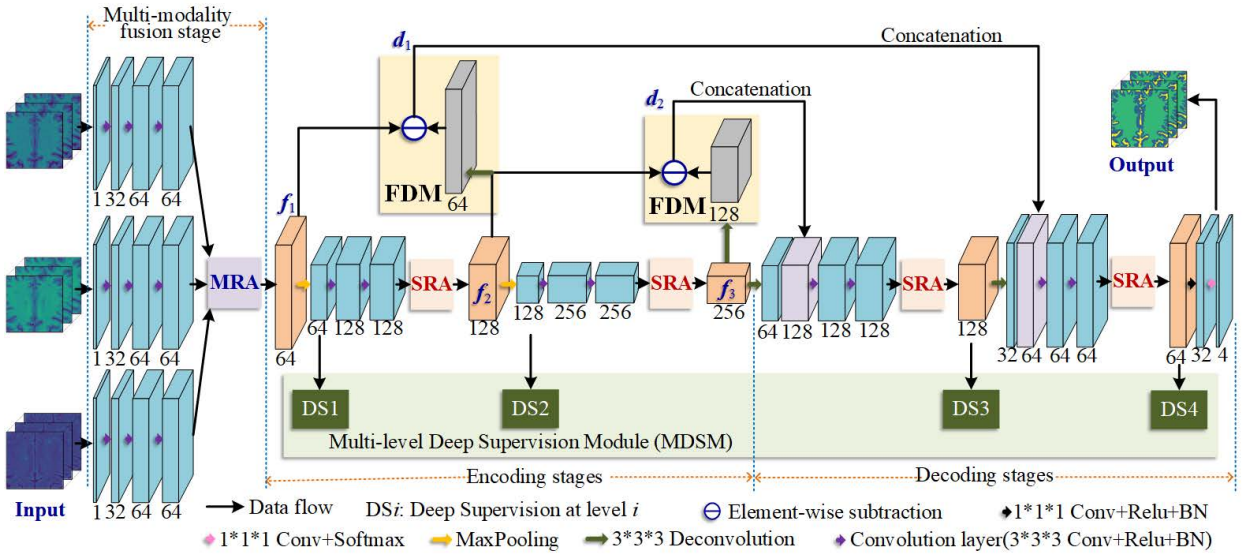

FIGURE 2. The network structure of Multi-modality Reconstruction Attention and Difference Enhancement Network (MRADE-Net).

B. MULTI-MODALITY RECONSTRUCTION ATTENTION (MRA)

To fuse information from multiple modalities, we propose a module of Multi-modality Reconstruction Attention (MRA) to fully exploit detail information and facilitate subsequent detail segmentation.

The structure of the MRA is shown in Fig. 3. The inputs to the MRA contain three MRI volumes with T1, T1-IR, and T2-FLAIR modalities. First, we respectively stack the inputs of three modalities along the Sagittal, Coronal, and Axial directions to generate three enlarged 3D feature volumes. Then, a global max-pooling is used to flatten each enlarged 3D volume along the corresponding direction to produce a 2D feature map, which contains abundant spatial and modality information. In fact, this process is just equivalent to the projection of data. Each flattened feature maps are processed by fully connected layers and activation functions to generate three 2D coefficient maps that are orthogonal to each other.

Then, we use a simple spanning method to reconstruct a 3D coefficient map, which is used to weight the three original inputs with different modalities for obtaining three maps of attention features. Finally, the three attention maps are added element-wisely to generate a fused attention feature map as the output of the MRA. In this way, we can not only reduce the number of parameters but also improve feature robustness.

C. SINGLE-MODALITY RECONSTRUCTION ATTENTION (SRA)

In the field of medical image segmentation, spatial detail information is of great significance to assist medical diagnosis. Inspired by the attention strategy of volumetric feature recalibration layer in SW-3D-Unet [14], we propose the Single-modality Reconstruction Attention (SRA) to maximize the ability to capture 3D spatial contextual information and refine the detailed features of the brain MRI images. The flowchart of SRA is shown in Fig. 4.

Supposed that the 3D feature tensor $\boldsymbol{f} \in R^{I\times J\times K}$ is the input of the SRA, where I, J, K denote its length, width, and height, respectively. To capture 3D spatial statistical contextual information on a chosen axis d, we use the global average pooling along the specified axis d. This process is just the projection of data along a given direction. The subscript of global average pooling indicates the projection axis d. The procedures of global average pooling along the three axes are formally defined as:

$$A'(i,k) = GAP_a(\boldsymbol{f}) = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{f}(i,j,k), \tag{1}$$

$$C'(j,k) = GAP_c(\boldsymbol{f}) = \frac{1}{I}\sum_{i=1}^{I}\boldsymbol{f}(i,j,k), \tag{2}$$

$$S'(i,j) = GAP_s(\boldsymbol{f}) = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{f}(i,j,k), \tag{3}$$

where GAP_d denotes the global average pooling along axis $d \in \{a, c, s\}$(a, c, s represent the axial, coronal and sagittal directions, respectively), and $A'(i,k) \in R^{I\times 1\times K}$, $C'(j,k) \in R^{1\times J\times K}$, and $S'(i,j) \in R^{I\times J\times 1}$ are the concentrated feature maps by squeezing the entire 3D tensor $\boldsymbol{f}$ along the axial, coronal and sagittal axes.

Then, we use two fully connected layers to filter and activate **A**', **C**', and **S**' respectively for obtaining three coefficient vectors. After that, we reshape these vectors back to the same size as $A'(i,k)$, $C'(j,k)$, $S'(i,j)$ respectively to produce three coefficient maps of **A**, **C**, and **S**, as shown in Fig. 4. The maps of **A**, **C**, and **S** are factually 2D tensors, which are denoted as the weight tensors in three directions. We reconstruct a three-dimensional coefficient attention volume $\boldsymbol{w} \in R^{I\times J\times K}$

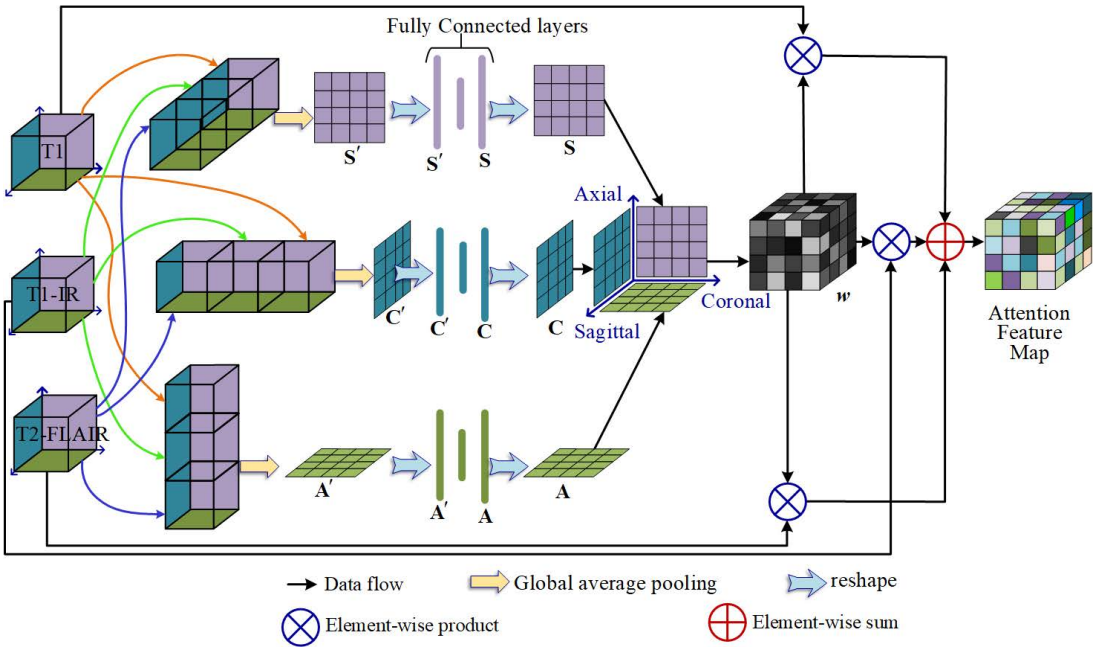

FIGURE 3. Multi-modality Reconstruction Attention (MRA).

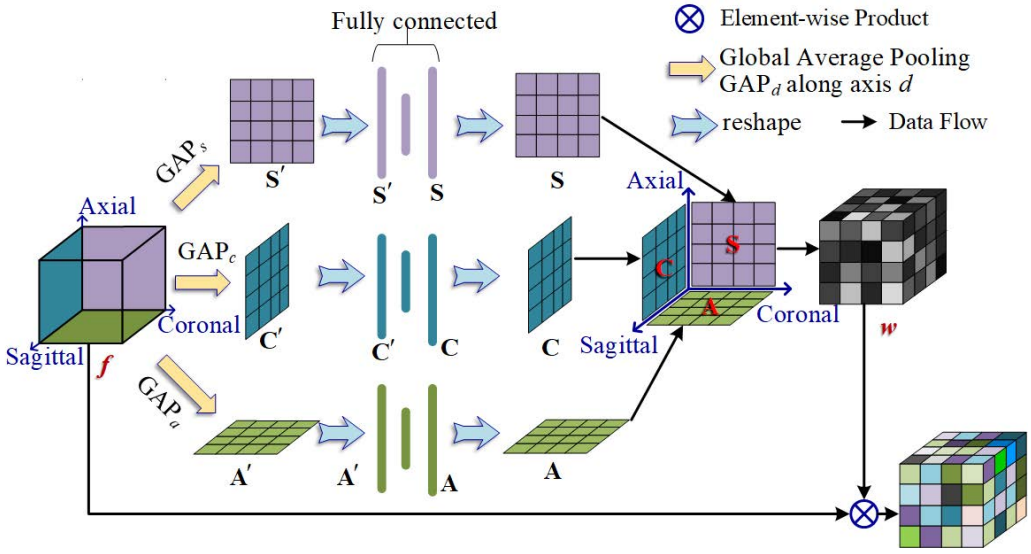

FIGURE 4. Single-modality Reconstruction Attention (SRA).

by simply multiplying each element value, formulated as follows:

$$\boldsymbol{w}(i,j,k) = A(i,k)C(j,k)S(i,j), \tag{4}$$

where $\boldsymbol{w}(i,j,k)$ represents the weight value of a given voxel (i,j,k).

Finally, we multiply the input tensor $\boldsymbol{f}$ by the attention weight tensor $\boldsymbol{w}$ to generate the final attention feature map. By reconstructing a 3D attention volume to weight the input tensor, we can enhance representation ability of our network to further optimize the segmentation result, especially for the spatial details in some boundary regions where multiple brain tissues are adjacent to each other.

D. FEATURE DIFFERENCE MODULE (FDM)

In the 3D U-net [10] structure, the encoder and decoder at the same resolution level have short connections, mainly to compensate for the detail information loss during the down-sampling process. However, directly connecting encoding features to corresponding decoding ones leads to too much redundant information, which greatly reduces the efficiency of networks. To solve that problem, we design a Feature

Difference Module (FDM) to reduce redundant information and improve network efficiency. The structure of the FDM is as shown in Fig. 2, and the transfer function of the FDM is defined as:

$$\boldsymbol{d}_{l-1} = \boldsymbol{f}_{l-1} - deconv(\boldsymbol{f}_l), \tag{5}$$

where $\boldsymbol{f}_{l-1}$ and $\boldsymbol{f}_l$ are two inputs to the FDM that are feature maps from two adjacent levels, $deconv(.)$ denotes a deconvolution operation for up-sampling feature maps in a learnable manner, and $\boldsymbol{d}_{l-1}$ is the output of the FDM. The output $\boldsymbol{d}_{l-1}$ of the FDM is connected to the feature map of the same resolution in the decoding stage.

Compared with straight connections in the 3D-UNet, our method concatenates only the difference information of two adjacent levels in encoding stages to the corresponding level of decoding stages. This allows us to utilize lost information during down-sampling, so we can greatly reduce redundant information, accelerate learning speed and improve robustness of our network.

E. MULTI-LEVEL DEEP SUPERVISION MODULE (MDSM)

It has been demonstrated in many networks, such as GoogLeNet [33], that deep supervision structures have significantly improved network performance. The central idea of deep supervision is to implement direct supervision in some early hidden layers of networks.

Inspired by the in-depth supervision mechanism, we design a Multi-level Deep Supervision Module (MDSM), which contains four deep supervision blocks (DS), named as DS1, DS2, DS3, and DS4, as shown in Fig. 2. DS1 and DS2 are placed in the encoder part, DS3 and DS4 are placed in the decoder part. In every DS block, we use deconvolution and classification layers to produce a predicted result, then a loss can be computed from the predicted result and its corresponding ground truth. The loss function during training can be described as:

$$\ell = \lambda l_0 + \mu_1(l_1 + l_2) + \mu_2(l_3 + l_4), \tag{6}$$

refer to (6), ℓ represents the integrated loss function of the training network, λ, μ_1, and μ_2 are coefficients for regulating the relative importance of each loss, l_0 denotes the loss of the output layer, l_1, l_2, l_3, and l_4 represent the losses of DS1, DS2, DS3, and DS4, respectively. The expression of l_0 is shown in (7), and the expression of l_1, l_2, l_3, and l_4 are shown in (8).

$$l_0 = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M}\omega_c\left[y_{ic}\log(p_{ic}) + (1-y_{ic})\log(1-p_{ic})\right], \tag{7}$$

$$l_n = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M}\omega_c\left[y_{ic}\log(s_{ic}^n) + (1-y_{ic})\log(1-s_{ic}^n)\right]. \tag{8}$$

Refer to (7), M represents the number of tissue categories, N is the number of pixels, the weight of category c is a coefficient ω_c, p_{ic} represents the predictive probability of pixel i belonging to category, and y_{ic} is the ground truth of pixel i corresponding to p_{ic}. If a pixel i belongs to the category of c, then $y_{ic} = 1$, otherwise $y_{ic} = 0$.

Refer to (8), S_{ic}^n represents the predicted probability of pixel i belonging to class c in the nth DS block ($n = 1, 2, 3, 4$).

IV. EXPERIMENTS

A. MATERIALS

This study evaluates the proposed method on the public datasets, the 2013 MICCAI MRBrainS Challenge dataset (MRBrainS13) [34], and the Internet Brain Segmentation Repository dataset (IBSR18) [35]. More details can be found as follows.

1) The MRBrainS13 dataset contains 5 labeled samples and 15 testing samples collected by scanning diabetic patients with different degrees of white matter lesions, which includes three modalities (T_1, T_1-IR, and T_2-FLAIR), and the voxel size of them is 0.958mm×0.958mm×3.00mm. All images are corrected by deviation, and the ground truth is manually segmented by medical experts.

2) The IBSR18 dataset contains 18 real T1 MRI data which are obtained from healthy subjects. Each real MRI data is 256 ×256 ×128. The MRI scans and the manual expert segmentations are provided by the Center for Morphometric Analysis at Massachusetts General Hospital.

B. IMPLEMENTATION

Our experiments were executed on Windows platform with a Nvidia GPU of NVIDIA GTX 1080Ti. The open-source framework TensorFlow [36] serves as the experimental environment for the proposed method. We use the dichotomy method to carry out parameter selection, and finally determine a set of optimal experimental parameters: $\lambda = 0.86$, $\mu_1 = 0.01$, $\mu_2 = 0.06$. The initial learning rate of the network model is 0.001, and the number of training iterations is 17000. In our experiments, we randomly cut the MRI brain image to the size of $32 \times 32 \times 32 \times n$ and then take them as the input of the network, where n is the number of MRI modalities ($n = 3$ for the T_1, T_1-IR, and T_2-FLAIR in this work). In the test section, we use the trained model to segment the test image and select a sliding window with an overlap of 8. The probability map is generated by the softmax layer in the model, which is used for label prediction.

In the experiment, we mainly use three evaluations indicators to measure the segmentation performance of different methods.

The first one is Dice coefficient (Dice), which is defined as:

$$Dice = 1 - \frac{2\,|P \cap G|}{|P| + |G|}. \tag{9}$$

where, P and G denote the prediction result and the ground truth respectively.

Absolute volume difference (AVD) is the second metric to measure the segmentation performance of algorithms, which

TABLE 1. Objective ablation experimental results on the validation set. 'Ours w/o SRA' refers to our proposed MRADE-Net without SRA.

Evaluation indicators	Dice (%)			AVD			HD			ACC (%)	AUC
tissue	WM	GM	CSF	WM	GM	CSF	WM	GM	CSF		
MRADE-Net (Ours)	91.95	89.14	84.54	1.82	6.17	10.51	0.66	1.12	1.12	89.46	0.853
Ours w/o SRA	91.60	88.92	84.42	3.16	5.27	10.29	0.68	1.13	1.13	88.17	0.836
Ours w/o MRA	90.94	88.45	81.77	5.60	3.76	14.93	1.00	1.00	1.05	87.53	0.822
Ours w/o FDM	91.28	88.41	83.76	3.26	5.95	12.32	1.01	1.04	1.05	87.65	0.795
Ours w/o MDSM	90.84	88.63	84.23	3.53	3.57	8.20	0.78	1.15	1.11	87.22	0.847

is defined as follows:

$$AVD = \frac{|V_P - V_G|}{V_G} \times 100\%, \tag{10}$$

where, V_P is the volume of the predicted segmentation, and V_G is the volume of the ground truth. The smaller the AVD value is, the better the model performs.

The last metric is Hausdorff distance (HD), which describes the degree of similarity between two sets of points, and is defined as:

$$H(P, G) = \max\{h(P, G), h(G, P)\}, \tag{11}$$

where, $h(P, G)$ and $h(G, P)$ represent the single-term hausdorff distance, and the expression are as follows:

$$h(P, G) = \max(p \in P) \min(g \in G) \|p - g\|, \tag{12}$$

$$h(G, P) = \max(g \in G) \min(p \in P) \|g - p\|, \tag{13}$$

where, $\|\cdot\|$ is the distance paradigm between two point-sets.

In addition, in the ablation experiment, we will also use the accuracy (ACC), receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) as evaluation indicators.

In our experiments, we first verify the effectiveness of the proposed SRA, FDM, MRA, and MDSM modules through ablation study. Then, we compare the segmentation results of MRADE-Net with other four segmentation networks including 3D-UNet, Hyperdense-Net, MMAN and SW-3D-Unet based on MRBrainS13 dataset. Meanwhile, in order to verify the segmentation performance of the proposed network on segmenting single modality MRI image, we also make a contrast of the evaluating indicator of MRADE-Net and other six networks (U-Net, Residual U-Net, Inception U-Net, Residual Inception U-Net, SegNet, and MhURI) on IBSR18 dataset.

C. ABLATION STUDY

As analyzed in section III.C, the SRA can weight and refine the detail information of brain tissues, MRA has the ability to fuse the unique features of each modality, the FDM module can reduce information loss during down sampling, and the MDSM makes the network convergent faster and optimizes the ability of the network to extract features.

In this section, we use experiments to verify the contribution of the SRA, MRA, FDM, and MDSM modules to the segmentation performance of the MRADE-Net. On the MRBrainS13 dataset, we carry out comparative experiments by removing each one of the four parts mentioned above. Table. 1 is the result we obtained on the validation set.

As is shown in Table. 1, after respectively removing each one of SRA, MRA, FDM, and MDSM from the proposed architecture, the performance of the network declines accordingly. Specifically, when only SRA layer is removed from the MRADE-Net, the Dice of WM and GM are decreased from 91.95% and 89.14% to 91.60% and 88.92% respectively. It indicates that the SRA layer is beneficial to the network for the reason that it accumulates more useful information during the segmentation task. Under the circumstances that we neglect only the MRA layer, the AVD-score of WM and CSF are increased from 1.82mm and 10.51mm to 5.60mm and 14.93mm respectively, which proves that the proposed MRA can extract and utilize rich image features in different modalities.

After removing the FDM, except for HD of GM and CSF, all other evaluation indicators decline. The FDM is able to reduce the redundancy of information while ensuring the supplement of lost information, and thus the segmentation result is worse than the proposed method with FDM. What is more, we remove only MDSM module from the network to perform experiments to test the effectiveness of the MDSM module. As is shown in Table. 1, MRADE-Net has the highest ACC value, which also shows that the network has the highest accuracy. The reason why MDSM can further boost the performance is because it can not only learn the consistency between the features learned from different layers, but also has the ability to adjust parameters in real time.

In addition, the ROC curve corresponding to the ablation experiment is shown in Fig. 5. We can see that whenever we remove a module, the fluctuation range of the ROC curve becomes larger, indicating that the removal of the module has a negative impact on the segmentation results. The area AUC is shown in Table. 1. It can also be seen that the AUC decreases slightly whenever a module is removed, which also proves the effectiveness of the proposed module.

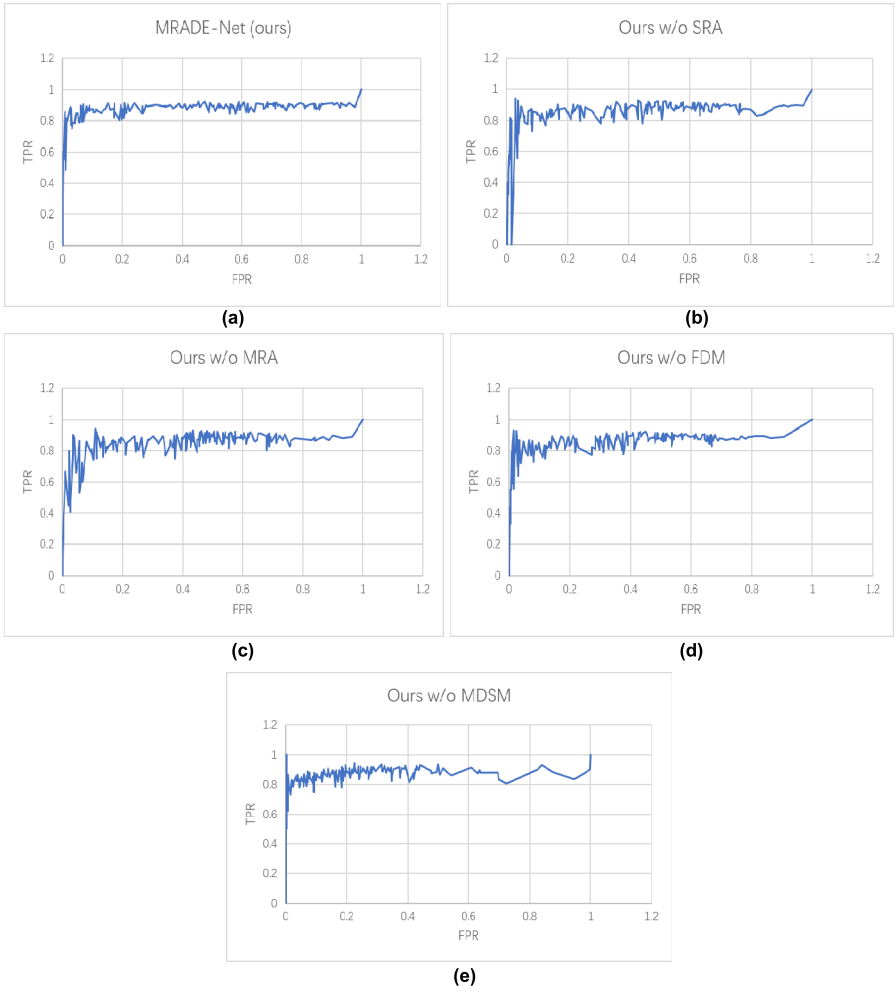

FIGURE 5. ROC curves of ablation experiments.

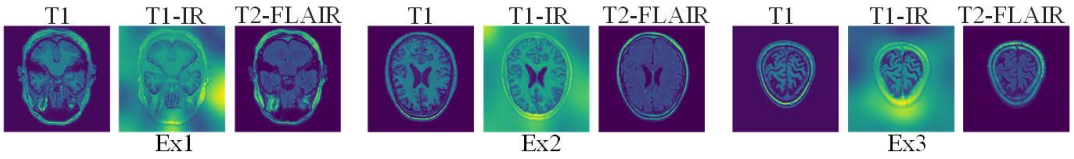

FIGURE 6. The original images of three samples.

To sum up, the experimental results have shown that SRA, FDM, MRA, and MDSM can play their effectiveness perfectly. With their mutual cooperation, the MRADE-Net can effectively guide the training of the fused feature of multi-modality, thereby acquiring more accurate results.

In Fig. 6, we present the original images of three samples, and their segmentation results in the ablation experiment are shown in Fig. 7. In Fig. 7, we present three segmentation examples, where the red boxes highlight specific differences. It is noteworthy that, when the three kinds of brain tissues are very close and small, it may lead to misclassification. Take Ex2 as an example, the brain tissues in the red box are very small and difficult to segment. However, we can see from Fig. 7 that the MRADE-Net model can achieve a more accurate distinction.

D. COMPARATIVE STUDY ON MRBRAINS13 DATASET

In this part, we compared our method with four other state-of-the-art brain tissue segmentation algorithms, including 3D-UNet [13], Hyperdense-Net [28], MMAN [37], and SW-3D-Unet [14]. The experimental statistics are shown in Table. 2, which lists the segmentation results of the proposed MRADE-Net and other different methods on the MRBrainS13 dataset.

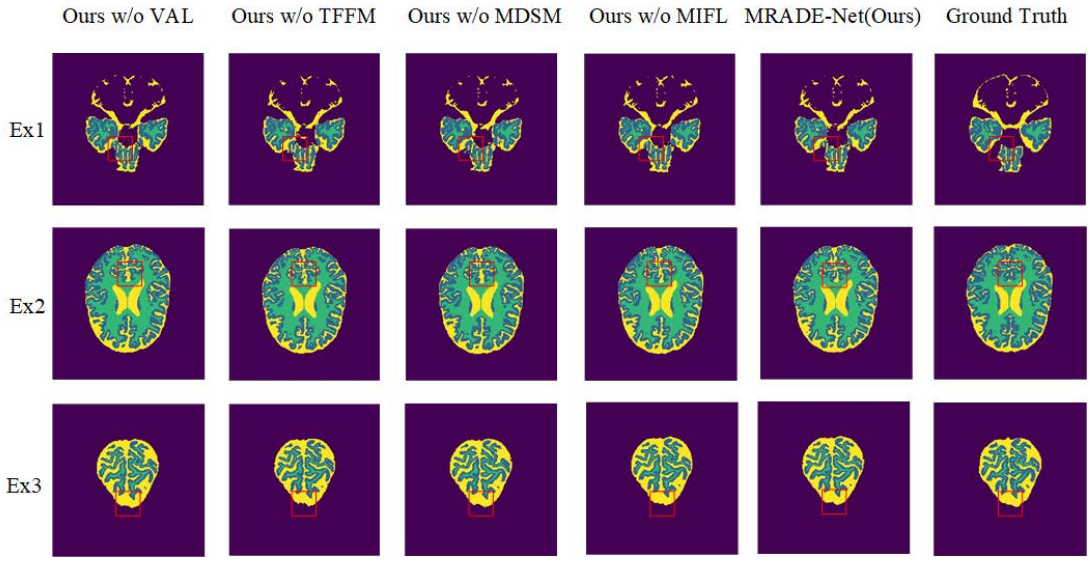

FIGURE 7. Qualitative results of ablation experiments on three MRI examples. 'w/o' means without.

TABLE 2. Segmentation result of the MRADE-Net and other excellent methods on MRBrainS13 dataset.

Evaluation Metric	Dice (%)			AVD			HD		
tissue	WM	GM	CSF	WM	GM	CSF	WM	GM	CSF
3D-UNet [13]	88.86	85.44	83.47	6.47	6.60	8.63	1.95	1.58	2.22
Hyperdense-Net [28]	89.46	86.33	83.42	6.03	6.19	7.31	1.78	1.34	2.26
MMAN [37]	89.70	86.40	84.86	6.28	5.72	6.75	1.88	1.38	2.03
SW-3D-Unet [14]	88.86	86.56	85.53	7.10	6.45	6.16	1.83	1.50	1.90
MRADE-Net (Ours)	91.95	89.14	84.54	1.82	6.17	10.51	0.66	1.12	1.12

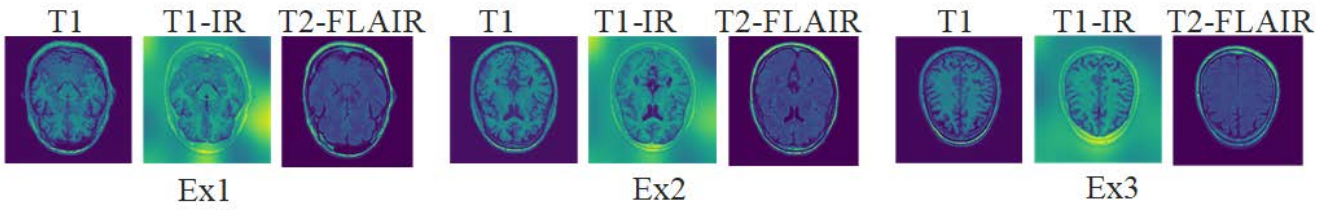

FIGURE 8. The original images of three samples.

As illustrated in Table. 2, the proposed MRADE-Net achieves the best performance in brain tissue segmentation. For example, our MRADE-Net method achieves the highest Dice coefficient of WM (i.e., 91.95%), which is obviously much better than the 3D U-Net method (i.e., 88.86%).

Compared with other methods (such as Hyperdense-Net, MMAN, and SW-3D-Unet), the proposed method has achieved significant improvements. In general, the Dice coefficient of MRADE-Net on WM is 3.09% higher than the counterparts', 3D-UNet and SW-3D-Unet, and gets 0.91mm improvement on Hausdorff Distance of CSF compared with MMAN. Besides, compared with 3D-UNet, Hyperdense-Net, MMAN and SW-3D-Unet, the proposed method achieves better results in terms of AVD value on WM.

In this part, we also give three examples, whose original images and segmented images are shown in Fig. 8 and Fig. 9 respectively. As shown in Fig. 9, the red box marks the partial differences of segmentation by different methods. By comparing the distinction of segmentations of different methods, we can see that the segmentation performance of MRADE-Net is generally better than the existing segmentation network.

E. EXPERIMENTAL RESULTS ON IBSR18 DATASET

In this part, in order to verify that our network structure is suitable for segmenting single-modality data, we compared the obtained results versus U-Net [12], Residual U-Net [38], Inception U-Net [39], Residual Inception U-Net [40],

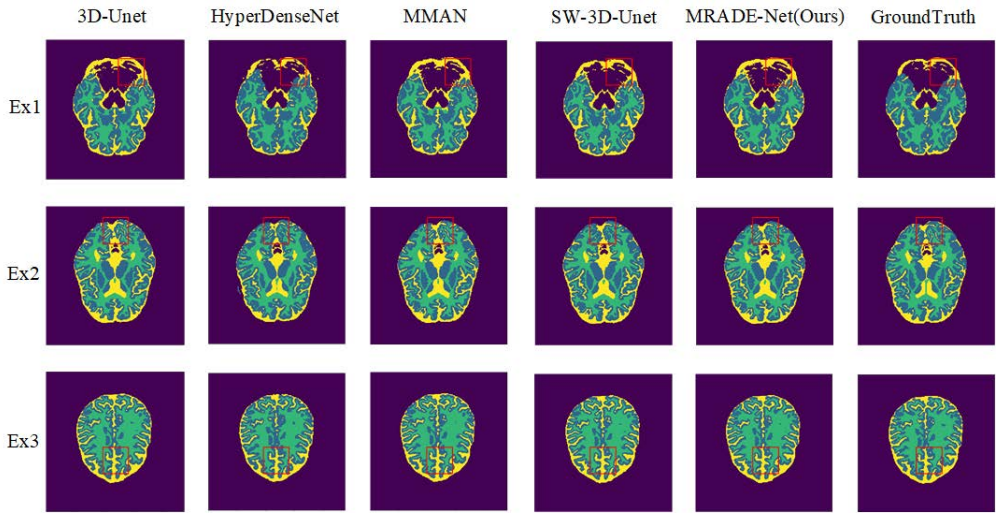

FIGURE 9. Comparison between the 3D-UNet, the HyperDenseNet, the MMAN, the SW-3D-UNet, and proposed method on MRBrainS13 dataset.

TABLE 3. Segmentation performance analysis of our network and some state-of-the-art deep learning networks on IBSR18 dataset.

Evaluation Metric	Dice (%)			AVD			HD		
tissue	WM	GM	CSF	WM	GM	CSF	WM	GM	CSF
U-Net [12]	84.62	87.66	73.20	4.16	4.31	27.16	7.32	7.24	8.36
Residual U-Net [38]	84.68	87.98	74.12	5.42	3.98	29.04	6.95	7.30	7.83
Inception U-Net [39]	85.04	88.12	74.46	4.03	3.57	27.86	5.33	6.25	7.85
Residual Inception U-Net [40]	85.08	88.22	74.90	4.46	4.87	26.38	4.58	6.09	5.89
SegNet [41]	82.76	86.26	72.80	4.23	4.52	28.89	8.42	8.49	14.83
MhURI [42]	87.14	90.42	76.34	3.89	4.35	26.54	5.26	4.37	6.23
MRADE-Net (Ours)	94.30	96.36	76.02	6.16	3.25	35.79	1.40	1.49	1.02

SegNet [41], and MhURI [42] on IBSR18 dataset. Table. 3 shows the segmentation results.

From Table. 3, we can see that the average Dice coefficients on three brain tissues yielded by MRADE-Net are 0.9636, 0.9430, and 0.7602, respectively, which are larger than those acquired by U-Net (0.8766, 0.8462, and 0.7320) and SegNet (0.8626, 0.8276, and 0.7280). Although MRADE-Net gets slightly lower value than MhURI on the Dice Score of CSF, the Dice Score value of WM and CSF is 5.94% and 7.16% higher, respectively. To sum up, the experimental results clearly show that the segmentation performance of the proposed network is better than that of the state-of-the-art networks.

V. CONCLUSION

In this article, we design a novel MRADE-Net to segment brain tissue based on multi modalities of MRI images (T_1, T_1-IR, and T_2-FLAIR). In the MRADE-Net, we apply a Multi-modality Reconstruction Attention to fuse the features of multi-modality MRI and focus on more important features effectively. We use an attention mechanism named SRA to weight the voxels in 3D features. More importantly, in order to reduce information loss and redundancy, we propose a FDM module between encoder and decoder. In addition, we apply Multi-level Deep Supervision Module when training the proposed network. Experiments show that the MRADE-Net can segment brain tissue more accurately.

REFERENCES

[1] M. Veluchamy and B. Subramani, "Brain tissue segmentation for medical decision support systems," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 1851–1868, Jun. 2020, doi: 10.1007/s12652-020-02257-8.

[2] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, no. 1, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[3] T. Kitrungrotsakul, X.-H. Han, and Y.-W. Chen, "Liver segmentation using superpixel-based graph cuts and restricted regions of shape constrains," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 3368–3371.

[4] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 17, 2021, doi: 10.1109/tpami.2021.3059968.

[5] Q. Huang, Y. Huang, Y. Luo, F. Yuan, and X. Li, "Segmentation of breast ultrasound image with semantic classification of superpixels," *Med. Image Anal.*, vol. 61, no. 1, Apr. 2020, Art. no. 101657, doi: 10.1016/j.media.2020.101657.

[6] R. Zhou, A. Fenster, Y. Xia, J. D. Spence, and M. Ding, "Deep learning-based carotid media-adventitia and lumen-intima boundary segmentation from three-dimensional ultrasound images," *Med. Phys.*, vol. 46, no. 7, pp. 3180–3193, May 2019, doi: 10.1002/mp.13581.

[7] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[8] Z. Tian, L. Liu, Z. Zhang, and B. Fei, "PSNet: Prostate segmentation on MRI based on a convolutional neural network," *J. Med. Imag.*, vol. 5, no. 2, pp. 1–7, Jan. 2018, doi: 10.1117/1.JMI.5.2.021208.

[9] X. Zhou, R. Takayama, S. Wang, T. Hara, and H. Fujita, "Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method," *Med. Phys.*, vol. 44, no. 10, pp. 5221–5233, Oct. 2017, doi: 10.1002/mp.12480.

[10] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123–1136, Mar. 2019, doi: 10.1109/TCYB.2018.2797905.

[11] P. F. Christ, M. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, and M. D'Anastasi, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Proc. MICCAI*, Istanbul, Turkey, 2016, pp. 415–423.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234–241.

[13] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, Istanbul, Turkey, 2016, pp. 424–432.

[14] L. Sun, W. Ma, X. Ding, Y. Huang, D. Liang, and J. Paisley, "A 3D spatially weighted network for segmentation of brain tissue from MRI," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 898–909, Apr. 2020, doi: 10.1109/TMI.2019.2937271.

[15] Y.-J. Huang, Q. Dou, Z.-X. Wang, L.-Z. Liu, Y. Jin, C.-F. Li, L. Wang, H. Chen, and R.-H. Xu, "3-D RoI-aware U-Net for accurate and efficient colorectal tumor segmentation," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5397–5408, Nov. 2021, doi: 10.1109/TCYB.2020.2980145.

[16] H. Feldman and K. J. Friston, "Attention, uncertainty, and free-energy," *Frontiers Hum. Neurosci.*, vol. 4, pp. 215–238, Dec. 2010, doi: 10.3389/fnhum.2010.00215.

[17] Z. Deng, L. Zhu, X. Hu, C.-W. Fu, X. Xu, Q. Zhang, J. Qin, and P.-A. Heng, "Deep multi-model fusion for single-image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 2453–2462.

[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998, doi: 10.1109/34.730558.

[19] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[21] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Munich, Germany, 2018, pp. 3–19.

[22] X. Sun, P. Garg, S. Plein, and R. J. Geest, "SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging," *Med. Phys.*, vol. 48, no. 4, pp. 1750–1763, Apr. 2021, doi: 10.1002/mp.14752.

[23] C. Kaul, S. Manandhar, and N. Pears, "Focusnet: An attention-based fully convolutional network for medical image segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Venice, Italy, Apr. 2019, pp. 455–458.

[24] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 547–556.

[25] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3D biomedical segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3739–3746.

[26] C. Ma, G. Luo, and K. Wang, "Concatenated and connected random forests with multiscale patch driven active contour model for automated brain tumor segmentation of MR images," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1943–1954, Aug. 2018, doi: 10.1109/TMI.2018.2805821.

[27] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vols. 3–4, Sep./Dec. 2019, Art. no. 100004, doi: 10.1016/j.array.2019.100004.

[28] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019, doi: 10.1109/TMI.2018.2878669.

[29] J. Dolz, C. Desrosiers, and I. B. Ayed, "IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet," in *Proc. CSI*, Granada, Spain, 2018, pp. 130–143.

[30] D. Nie, L. Wang, Y. Gao, and D. Shen, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Prague, Czech Republic, Apr. 2016, pp. 1342–1345.

[31] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, and B. Glocker, "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *Proc. MICCAI*, Quebec City, QC, Canada, 2017, pp. 15–64.

[32] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, no. 1, pp. 61–78, Feb. 2017, doi: 10.1016/j.media.2016.10.004.

[33] S. Christian, L. Wei, J. Yangqing, P. Sermanet, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[34] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, B. Marcel, W. H. Bouvy, D. B. Jeroen, A. Amir, D. B. Marleen, C. Aaron, and E. B. Ayman, "MRBrainS challenge: Online evaluation framework for brain image segmentation in 3T MRI scans," *Comput. Intell. Neurosci.*, vol. 2015, Jan. 2015, Art. no. 813696, doi: 10.1155/2015/813696.

[35] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 153–163, Feb. 2012, doi: 10.1109/TMI.2011.2163944.

[36] B. Pang, E. Nijkamp, and Y. N. Wu, "Deep learning with tensorflow: A review," *J. Educ. Behav. Stat.*, vol. 45, no. 2, pp. 227–248, 2020, doi: 10.3102/1076998619872761.

[37] J. Li, Z. L. Yu, Z. Gu, H. Liu, and Y. Li, "MMAN: Multi-modality aggregation network for brain segmentation from MR images," *Neurocomputing*, vol. 358, pp. 10–19, Sep. 2019, doi: 10.1016/j.neucom.2019.05.025.

[38] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018, doi: 10.1109/LGRS.2018.2802944.

[39] D. E. Cahall, G. Rasool, N. C. Bouaynaya, and H. M. Fathallah-Shaykh, "Inception modules enhance brain tumor segmentation," *Frontiers Comput. Neurosci.*, vol. 13, pp. 44–52, Jul. 2019, doi: 10.3389/fncom.2019.00044.

[40] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: Separable 3D U-Net for brain tumor segmentation," in *MICCAI*, Granada, Spain, 2019, pp. 358–368.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017, doi: 10.1109/TPAMI.2016.2644615.

[42] G. P. Hosal, T. Chowdhury, A. Kumar, A. K. Bhadra, J. Chakraborty, and D. Nandi, "MhURI: A supervised segmentation approach to leverage salient brain tissues in magnetic resonance images," *Comput. Methods Programs Biomed.*, vol. 200, no. 1, Mar. 2021, Art. no. 105841, doi: 10.1016/j.cmpb.2020.105841.

XIANGFEN ZHANG received the M.S. degree from the Nanjing University of Aeronautics and Astronautics, in 2001, and the Ph.D. degree from Shanghai Jiao Tong University, in 2008. She is currently an Associate Professor with The School of Information, Mechanical and Electrical Engineering, Shanghai Normal University. Her research interests include image processing and information fusion.

QINGYI ZHANG is currently pursuing the master's degree with Shanghai Normal University. Her research interests include computer vision and image processing.

YAN LIU is currently pursuing the master's degree with Shanghai Normal University. Her research interests include deep learning and medical image processing.

FEINIU YUAN (Senior Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, in 2004. He is currently a Professor with The School of Information, Mechanical and Electrical Engineering, Shanghai Normal University. His research interests include artificial intelligence, deep learning, pattern recognition, medical image processing, 3D reconstruction, and virtual reality.