



Multimodal Glioma Image Segmentation Using Dual Encoder Structure and Channel Spatial Attention Block

Run Su^{1,2}, Jinhui Liu^{1,2*}, Deyun Zhang³, Chuandong Cheng^{4,5,6} and Mingquan Ye⁷

¹ Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China, ² Science Island Branch of Graduate School, University of Science and Technology of China, Hefei, China, ³ School of Engineering, Anhui Agricultural University, Hefei, China, ⁴ Department of Neurosurgery, The First Affiliated Hospital of University of Science and Technology of China, Hefei, China, ⁵ Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China, ⁶ Anhui Province Key Laboratory of Brain Function and Brain Disease, Hefei, China, ⁷ School of Medical Information, Wannan Medical College, Wuhu, China

OPEN ACCESS

Edited by:

Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

Reviewed by:

Mengran Zhou,
Anhui University of Science and
Technology, China
Huaqing Zhu,
Anhui Medical University, China

*Correspondence:

Jinhui Liu
jhlui@iim.ac.cn

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Neuroscience

Received: 22 July 2020

Accepted: 15 September 2020

Published: 28 October 2020

Citation:

Su R, Liu J, Zhang D, Cheng C and
Ye M (2020) Multimodal Glioma Image
Segmentation Using Dual Encoder
Structure and Channel Spatial
Attention Block.
Front. Neurosci. 14:586197.
doi: 10.3389/fnins.2020.586197

Multimodal medical images provide significant amounts of complementary semantic information. Therefore, multimodal medical imaging has been widely used in the segmentation of gliomas through computational neural networks. However, inputting images from different sources directly to the network does not achieve the best segmentation effect. This paper describes a convolutional neural network called F-S-Net that fuses the information from multimodal medical images and uses the semantic information contained within these images for glioma segmentation. The architecture of F-S-Net is formed by cascading two sub-networks. The first sub-network projects the multimodal medical images into the same semantic space, which ensures they have the same semantic metric. The second sub-network uses a dual encoder structure (DES) and a channel spatial attention block (CSAB) to extract more detailed information and focus on the lesion area. DES and CSAB are integrated into U-Net architectures. A multimodal glioma dataset collected by Yijishan Hospital of Wannan Medical College is used to train and evaluate the network. F-S-Net is found to achieve a dice coefficient of 0.9052 and Jaccard similarity of 0.8280, outperforming several previous segmentation methods.

Keywords: medical image fusion, glioma segmentation, fully convolutional neural networks, DES, CSAB, F-S-Net

1. INTRODUCTION

Gliomas, which arise from the canceration of gliocyte in the brain and myelon, are the most common form of cancer in the skull, accounting for 80% of malignant brain tumors (Ostrom et al., 2014). The incidence ranges from 3 to 8 per 100,000 people and the fatality rate is high. Hence, the early diagnosis and treatment of gliomas are very important. The presence of gliomas can also cause complications such as increased intracranial pressure, brain edema, brain hernia, and psychosis. The size, location, and type of a glioma are determined by segmenting the affected region from other normal brain tissue. Accurate segmentation plays an important role in the diagnosis and treatment of gliomas. However, manual delineation practices not only require significant anatomical knowledge, but are also expensive, time consuming, and inaccurate. The automatic segmentation of gliomas would allow doctors to detect the growth of brain tumors earlier and provide additional information for the generation of treatment plans. Bi et al. (2019) believed that

artificial intelligence could improve the role of current standard diagnostic imaging technology by refining the preoperative classification of brain tumors above the level achievable by experts. Automatic segmentation based on computer-assisted intervention provides a steady solution for the treatment of gliomas, and is an effective tool in reducing the time required for the accurate detection, location, and delineation of tumor regions. Hence, it is necessary to automatically segment gliomas from medical images.

In recent years, methods based on deep learning (LeCun et al., 2015) have made significant breakthroughs in image classification (Krizhevsky et al., 2012; Rawat and Wang, 2017), image segmentation (Badrinarayanan et al., 2017; Garcia-Garcia et al., 2017), object detection (Ren et al., 2015; Zhao et al., 2019), object tracking (Li et al., 2018; Ristani and Tomasi, 2018), image captioning (Anderson et al., 2018; Hossain et al., 2019), and other fields (Hu et al., 2020). These breakthroughs have promoted the development of deep learning methods in the field of medical image analysis (Litjens et al., 2017; Altaf et al., 2019; Esteva et al., 2019). One of the best-known architectures for medical image segmentation is U-Net, initially proposed by Ronneberger et al. (2015), in which the backbone is a fully convolutional network (FCN) (Long et al., 2015). U-Net has received widespread attention from researchers in the field of medical image segmentation, and many improvements to U-Net have since been proposed (Alom et al., 2018; Oktay et al., 2018; Zhou et al., 2018). For example, Milletari et al. (2016) proposed V-Net for processing 3D medical images, whereby residual learning is employed to improve the convergence speed of the network and random nonlinear transformation and histogram matching are used for data augmentation. Milletari et al. also proposed the dice loss technique based on dice coefficients. Cheng et al. (2019) obtained a multilevel glioma segmentation network by combining an attention mechanism and atrous convolution with 3D U-Net. Chen et al. (2018b) used 3D U-Net and separable 3D convolution to build a separable 3D U-Net architecture. A multiscale masked 3D U-Net was proposed by Xu et al. (2018). The input to their network is a superimposed multiscale map, and multiscale information is obtained from the 3D ASPP layer.

Although methods based on deep learning have been widely used in this field, the current approaches have some disadvantages. Usually, researchers combine multimodal or multisequence medical images to obtain better segmentation accuracy (Kamnitsas et al., 2017b; Chen et al., 2018b; Xu et al., 2018; Zhao et al., 2018; Cheng et al., 2019). The multimodal medical images are input directly into the network for learning. However, the semantic conflicts between multimodal medical images cannot be completely avoided, and these may have a certain impact on the segmentation results. The method of image fusion can integrate valuable information from multimodal medical images, and the fusion results are typically more comprehensive than the original images (Liu et al., 2017). To date, there have been few reports on the segmentation of gliomas based on multimodal medical image fusion.

Another disadvantage of existing methods is that U-Net variants do not improve the basic architecture of U-Net. In

particular, the features of the medical images are extracted by a single encoder. This means there may be a loss of feature information. Therefore, it is necessary for networks to obtain and retain more useful features.

In this paper, we propose F-S-Net, which combines image fusion technology to obtain images with richer semantic information. F-S-Net consists of two sub-networks: a fusion sub-network and a segmentation sub-network. The fusion sub-network projects images obtained from computed tomography (CT) and magnetic resonance imaging (MRI) into the same semantic space for fusion. Compared with the original images, the fused image contains more semantic information for segmentation. To improve the segmentation performance, the segmentation sub-network uses a dual encoder structure (DES) and a channel spatial attention block (CSAB) to perform image segmentation. Based on the U-Net architecture, DES and CSAB use different sizes of convolution kernel to extract more effective features and focus on the lesion area. In the process of skip-connection, a 1×1 convolution and a concatenation operation are used to achieve better feature fusion. This method is conducive to feature extraction and utilization, and can achieve good performance. DES and CSAB are integrated into the networks based on the U-Net framework, and are found to improve the segmentation result. Experiments show that the cascaded networks proposed in this paper achieve better performance than existing approaches.

The contributions of this study are as follows:

1. A DES is constructed by increasing the width of the encoder. The proposed structure uses convolution kernels of different sizes to extract more effective features from images.
2. Our CSAB is constructed by combining channel attention and spatial attention mechanisms in the U-Net architecture. The proposed attention mechanism can be easily integrated into other networks that use the U-Net framework.
3. The proposed F-S-Net is formed by combining two sub-networks. One sub-network fuses CT and MRI images to enhance the semantic information of the images, while the other is used to segment gliomas accurately from the fused image.
4. Clinical glioma imaging data were collected from Yijishan Hospital of Wannan Medical College. The labels of each image were annotated by professional medical staff. The collected dataset provides a valuable tool for further research.
5. Extensive comparison experiments were conducted based on the collected dataset to demonstrate that the proposed method obtains the best segmentation performance among several deep segmentation methods.

2. RELATED WORK

Convolutional neural networks (CNNs) are a common architecture for glioma segmentation, especially the encoder-decoder model.

Wang et al. (2017) trained each tumor sub-region by using networks with similar architectures and cascading these

networks. The input to each network was the output from the previous network. However, some loss of global information might be caused by the way the gliomas are progressively segmented. Kamnitsas et al. (2017a) reported better results using ensembles of multiple models and architectures (EMMA). In particular, EMMA combined the DeepMedic (Kamnitsas et al., 2017b), FCN, and U-Net models and synthesized their segmentation results. The strong performance of EMMA helped Kamnitsa et al. to win the BraTS Challenge in 2017. However, EMMA does not offer end-to-end training, and the final result is affected by the accumulation of errors. Unlike most researchers, Isensee et al. (2018) demonstrated that competitive performance could be achieved with a few minor modifications to a generic U-Net. They reduced the number of feature maps before sampling from the decoder, and used additional training data to produce some improvements in terms of tumor enhancement. Myronenko (2018) won the BraTS 2018 challenge with a segmentation network based on the encoder-decoder architecture. An asymmetric encoder is used to extract features, and then two decoders segment the brain tumor and reconstruct the input image, respectively. The first decoder outputs the segmentation results from three tumor sub-regions, while the second uses a variational auto-encoder (VAE) to reconstruct the input image. The VAE branch only reconstructs the input images during the training stage. Jiang et al. (2019) achieved the best results in the 2019 BraTS challenge. They proposed a U-Net-based cascade network that is divided into two stages. In the first stage, a variant of U-Net produces an unshaped result. In the second stage, improved performance is obtained by increasing the width of the decoder. In fact, their network uses two decoders that are structurally similar, but have some differences in their up-sampling procedures: one decoder uses deconvolution while the other uses trilinear interpolation. Although multimodal medical imaging has been widely used in glioma segmentation, few researchers have considered the processing of multimodal medical images. This is a clear gap in the research, as the results might be affected by the different semantic information contained in multimodal medical images.

3. METHODS

This section describes the proposed F-S-Net architecture in detail. F-S-Net consists of two sub-networks, a fusion sub-network and a segmentation sub-network. The fusion sub-network uses multimodal images to obtain more detailed medical images with a wealth of semantic information. After processing the corresponding CT and MRI images, the fusion results are input to the segmentation sub-network. The segmentation sub-network uses a dual encoder architecture to extract detailed features from the lesion area. Different sizes of convolutional kernel are used to process images on parallel paths. At the same time, an attention mechanism is integrated into the CSAB module among the skip-connection processing. The final result is obtained by segmenting the fused results.

3.1. F-S-Net

Multimodal medical images have been widely used in medical image analysis tasks. As multimodal images contain different semantic information, image fusion technology is used to map the semantic information from the multimodal images to the same semantic space, including image structure information and edge information. Therefore, F-S-Net incorporates medical image fusion technology. The proposed network architecture is shown in **Figure 1**.

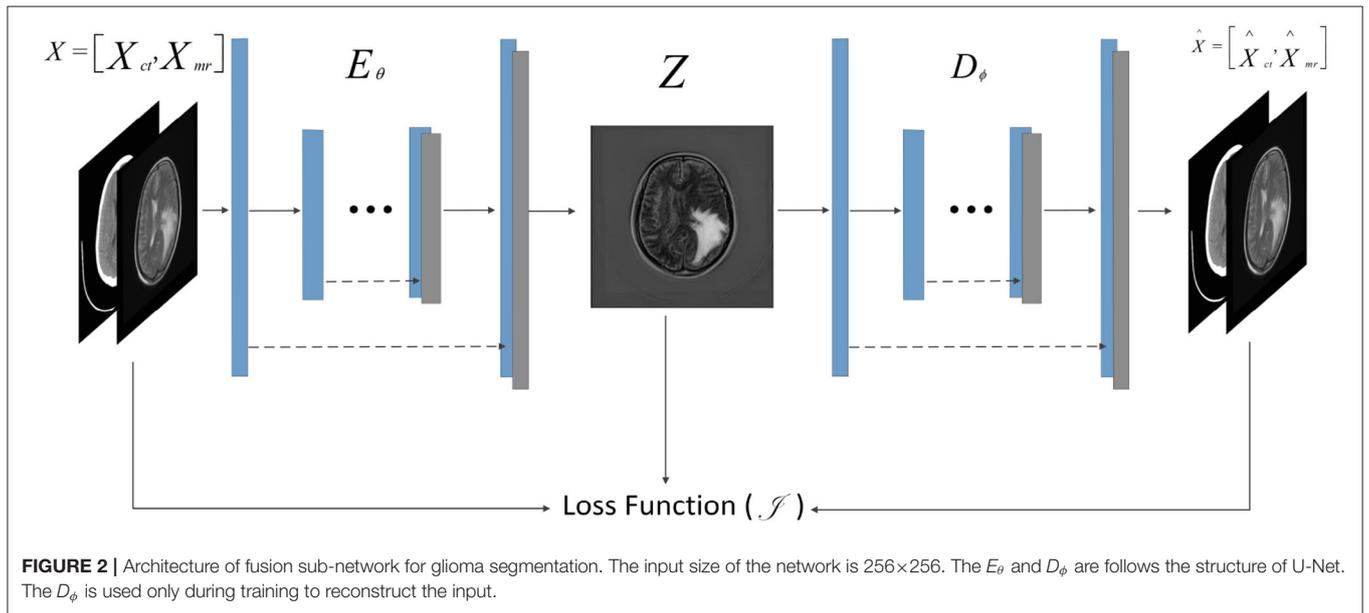
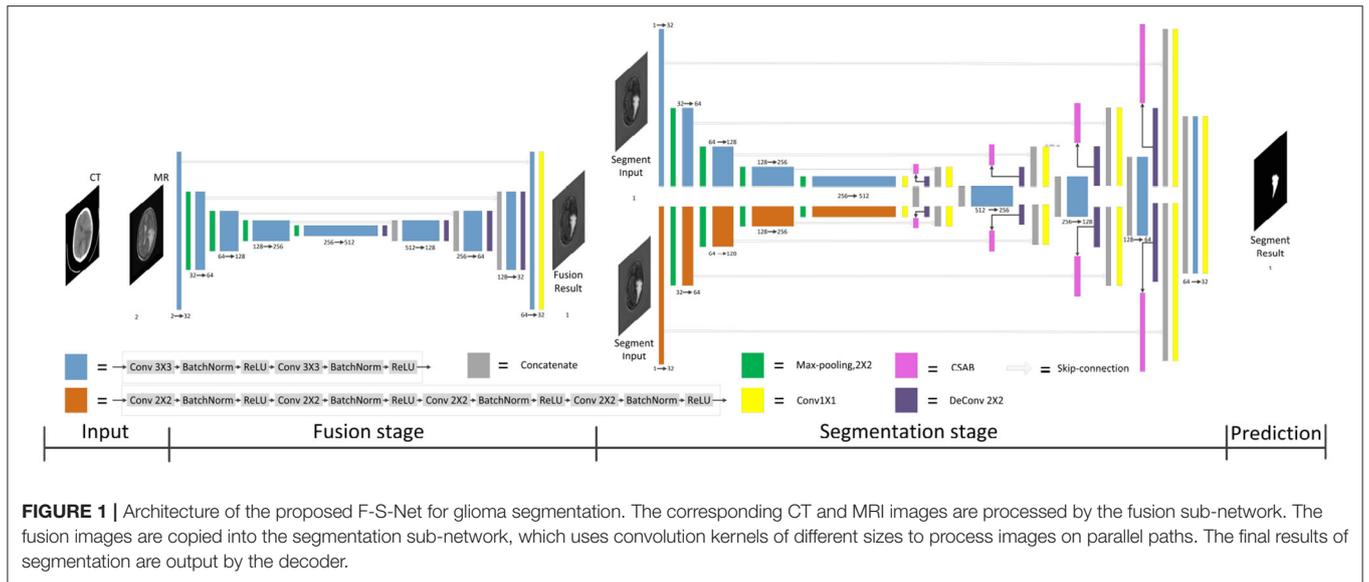
F-S-Net is divided into two stages. In the first stage, the fusion sub-network is used to fuse CT and MRI images. As the semantic information from various multimodal images is combined, this process provides more detailed medical images for segmentation networks. In the second stage, the fused image is input into the segmentation sub-network. The CSAB and DES modules are used in the segmentation sub-network based on the U-Net architecture. **Figure 2** shows the structure of the fusion sub-network (Fan et al., 2019). The E_θ and D_ϕ of fusion sub-network are follows the structure of U-Net. E_θ is used to generate the fusion results. D_ϕ is used to reconstruct the input. The loss value is determined by the input, fusion results, and reconstruction results. The loss function of the fusion sub-network has been modified by us. The details of the loss function are described in section 3.4. D_ϕ is used during the training stage. The segmentation sub-network architecture is a typical encoder-decoder structure, as shown in **Figure 3**. The segmentation sub-network consists of two encoders (left side) and a decoder (right side). The two encoders use convolution kernels of different sizes. In the skip-connection process, the attention mechanism is used to enable the network to extract the features of a specific area and perform feature fusion. The decoder is the same as in U-Net. The network takes input images of 256×256 pixels, and outputs images of the same size. The network can obtain more comprehensive and consistent medical images, and perform better segmentation tasks, after multimodal image fusion. The results are generated by minimizing the loss value.

3.2. Channel Spatial Attention Block

The attention mechanism is derived from the study of human vision. In computer vision, the attention mechanism allows the system to ignore irrelevant information and focus on important information. Combining channel attention, spatial attention, and the structural features of U-Net gives the CSAB module. This module enhances the salient features of the up-sampling process by applying an attention weight to the high- and low-dimensional features. The proposed structure is shown in **Figure 4**. The input feature maps x and g are scaled using the attention coefficient (α_3) computed in CSAB. Areas of concern are selected by analyzing the different types of attention weights provided by x and g .

Given an intermediate feature map $x, g \in R^{C \times H \times W}$ as input, CSAB obtains two intermediate 1D channel attention weights $\alpha_1, \alpha_2 \in R^{C \times 1 \times 1}$ and an intermediate 2D spatial attention weight $\alpha_3 \in R^{1 \times H \times W}$. **Figure 4** describes the calculation for each attention module. The overall attention process can be summarized as:

$$g^l = \alpha_1(g) \otimes g \quad (1)$$



$$x^l = \alpha_2(x) \otimes x \tag{2}$$

$$f = \alpha_3(g^l, x^l) \otimes x^l \otimes g^l \tag{3}$$

$$F = w(\text{Cat}[f, x]) + b \tag{4}$$

where \otimes denotes element-wise multiplication. F is the final output obtained by 1×1 convolution after fusing f and feature x .

Four different spatial context descriptions, g_{max} , g_{avg} , x_{max} , and x_{avg} , are obtained using average pooling and maximum pooling operations on the feature map. These four characteristics are entered into a small network for further processing. The output feature vectors of the small network are merged using a concatenation operation. Finally, the channel attention weights $\alpha_1(g)$ and $\alpha_2(x)$ are obtained after the dimension has been reduced by 1×1 convolution. The channel attention is calculated as follows:

$$g_{max} = \text{MaxPool}(g) \tag{5}$$

$$g_{avg} = \text{AvgPool}(g) \tag{6}$$

3.2.1. Channel Attention Block

The channel attention weight is produced from high- and low-dimensional features using the relationship among the features.

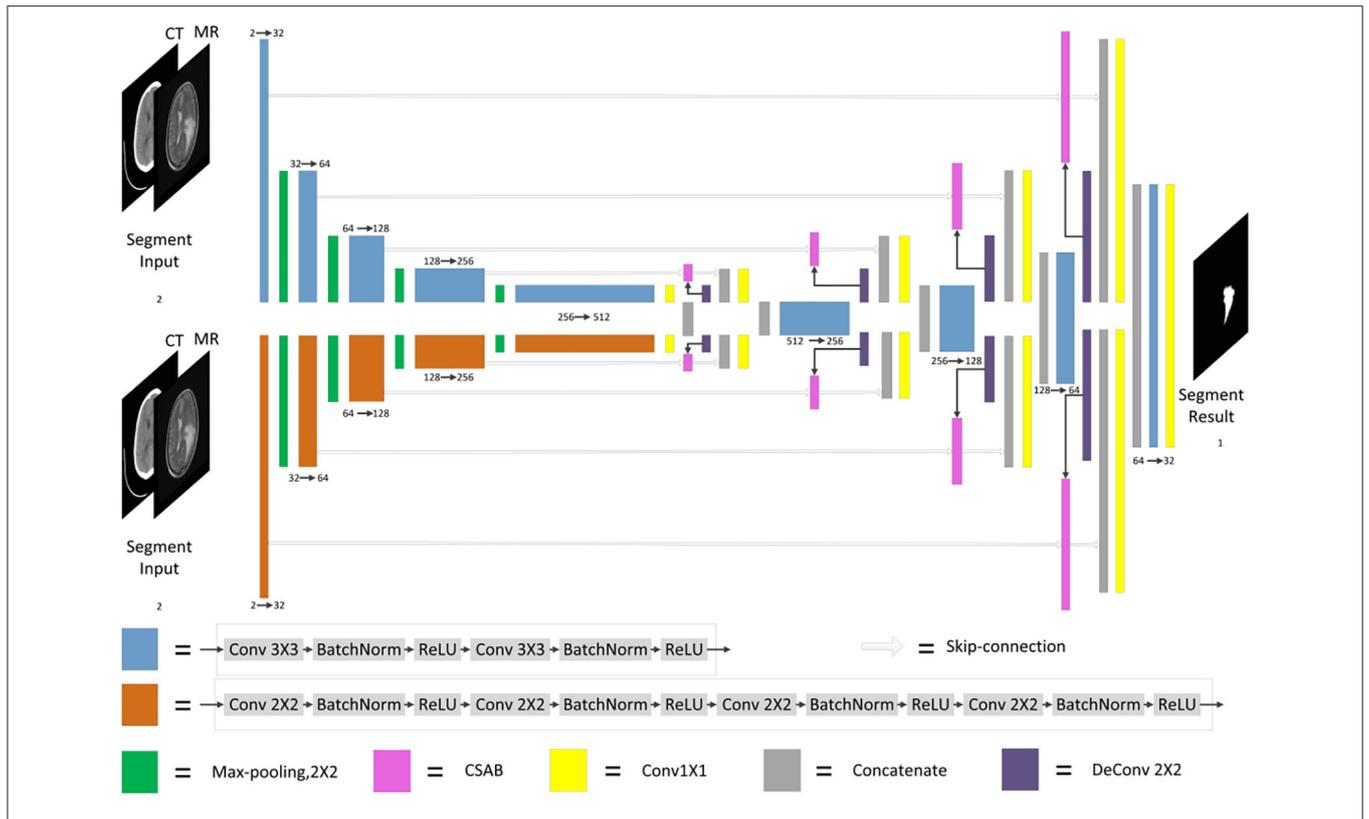


FIGURE 3 | Architecture of segmentation sub-network for glioma segmentation. The input size of the network is 256×256. Each different box denotes the different operations. The number of channels is denoted on the boxes. The parallel pathways process each feature using different sizes of convolution kernel, which are combined at the end of the encoder.

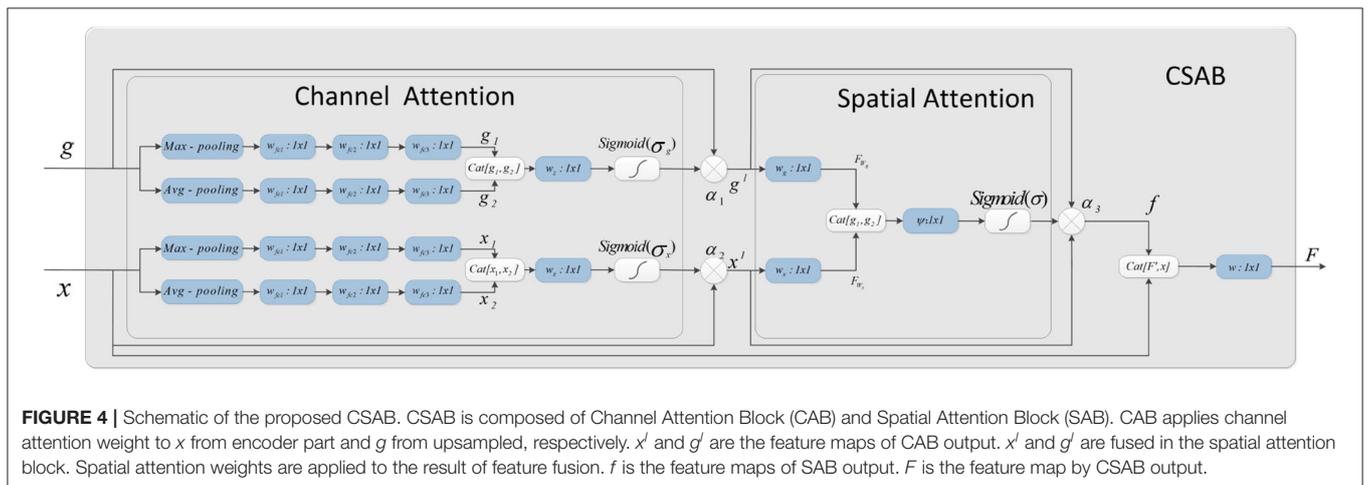


FIGURE 4 | Schematic of the proposed CSAB. CSAB is composed of Channel Attention Block (CAB) and Spatial Attention Block (SAB). CAB applies channel attention weight to x from encoder part and g from upsampled, respectively. x' and g' are the feature maps of CAB output. x' and g' are fused in the spatial attention block. Spatial attention weights are applied to the result of feature fusion. f is the feature maps of SAB output. F is the feature map by CSAB output.

$$x_{max} = \text{MaxPool}(x) \tag{7}$$

$$x_{avg} = \text{AvgPool}(x) \tag{8}$$

$$g_1 = w_{fc3}(w_{fc2}(w_{fc1}(g_{max}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{9}$$

$$g_2 = w_{fc3}(w_{fc2}(w_{fc1}(g_{avg}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{10}$$

$$x_1 = w_{fc3}(w_{fc2}(w_{fc1}(x_{max}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{11}$$

$$x_2 = w_{fc3}(w_{fc2}(w_{fc1}(x_{avg}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{12}$$

$$\alpha_1(g) = \sigma_g(w_z(\text{Cat}[g_1, g_2]) + b_z) \tag{13}$$

$$\alpha_2(x) = \sigma_x(w_z(\text{Cat}[x_1, x_2]) + b_z) \tag{14}$$

where σ_g and σ_x denote the sigmoid function, $W_{fc1} \in R^{C/8 \times 1 \times 1}$, $W_{fc2} \in R^{C/8 \times 1 \times 1}$, $W_{fc3} \in R^{C \times 1 \times 1}$, and $W_z \in R^{C \times 1 \times 1}$. W_{fc1} , W_{fc2} , W_{fc3} , and W_z denote the weight of each convolution. The rectified linear units (ReLU) activation function is followed by W_{fc1} , W_{fc2} , and W_{fc3} .

3.2.2. Spatial Attention Block

The spatial attention map is generated from $\alpha_1(g)$ and $\alpha_2(x)$ using the relationship among the features. The attention coefficient, $\alpha_3 \in [0, 1]$, suppresses the expression of irrelevant regions in the input. In addition, the attention coefficient can highlight features that are useful for the task.

In the spatial attention block, the high- and low-dimensional features are subjected to 1×1 convolution to obtain two features: $F_{W_g} \in R^{C \times H \times W}$ and $F_{W_x} \in R^{C \times H \times W}$. The concatenation operation then performs feature fusion. Finally, the spatial attention map of $\alpha_3 \in R^{1 \times H \times W}$ is generated by 1×1 convolution. The output of the spatial attention block (SAB) is the element-wise multiplication of the input feature graph and the attention coefficient. The spatial attention is calculated as follows:

$$F_{W_g} = w_g(g^l) + b_g \quad (15)$$

$$F_{W_x} = w_x(x^l) + b_x \quad (16)$$

$$f = \alpha_3(g^l, x^l) \otimes x^l \otimes g^l = \sigma(\psi(\text{Cat}[F_{W_g}, F_{W_x}]) + b) \otimes x^l \otimes g^l \quad (17)$$

where σ denotes the sigmoid function. W_g , W_x , and ψ represent the convolution kernel weights, and b_g , b_x , and b are the bias terms.

3.3. Dual Encoder Structure

The DES is developed by extending the encoder of U-Net. Two different encoders are used to extract features from images, and the convolution kernel size of the two encoders is different. One encoder has a convolution kernel size of 3×3 , while the other has a convolution kernel size of 2×2 . The encoder with a convolution kernel size of 3×3 is consistent with U-Net. Each layer consists of two 3×3 convolutions, followed by batch normalization (BN) and ReLU activation. The encoder with a convolution kernel of 2×2 is different from that of U-Net. Each layer consists of four 2×2 convolutions, each followed by BN and ReLU activation. The padding of the four 2×2 convolutions is 0101. The number of initial filters is 32. More feature information is obtained from images that use convolution kernels of different sizes. In addition, more significant information will be input to the decoder through the parallel paths design.

As the encoder has been expanded, it is necessary to fuse the features of each path when the features are input into the decoder. The output of CSAB is fused with the features obtained by up-sampling. Then, 1×1 convolution is used to reduce the dimension of the fused features. Finally, the processed features are input into the decoder. The two features from the encoder are processed separately. This approach is conducive to the integration of low- and high-dimensional information. The experimental results of the optimization procedure demonstrate

the effectiveness of our structure. The structure designed in this study is shown in **Figure 5**.

Let X_1 and X_2 be features extracted by the encoder. F_1 and F_2 are the features output by CSAB, respectively, and g is the feature obtained after up-sampling. F_1 and F_2 are fused with g , and the features connected by skip-connection are subjected to 1×1 convolution for dimension reduction, resulting in x_{13} and x_{23} . These two features are fused after dimensionality reduction to obtain X , which is input to the decoder. X is computed as follows:

$$X = \text{Cat}[x_{13}, x_{23}] \quad (18)$$

where F_1 , F_2 , x_{13} , and x_{23} are given by:

$$F_1 = \text{Att}(g, x_1) \quad (19)$$

$$F_2 = \text{Att}(g, x_2) \quad (20)$$

$$x_{13} = W_{x_1}(\text{Cat}[F_1, g]) + b_1 \quad (21)$$

$$x_{23} = W_{x_2}(\text{Cat}[F_2, g]) + b_2 \quad (22)$$

DES has two advantages. First, the convolution kernels of the two encoders are 3×3 and 2×2 , respectively. This strategy can extract more different features, which is beneficial to the segmentation task. Secondly, the features processed during the skip-connection ensure more complete information fusion. We have not made any major changes to the U-Net architecture. Therefore, our DES can be extended to most networks that are based on the U-Net architecture.

3.4. Loss Function

The loss function consists of three terms:

$$L_{total} = 0.02 * (L_{MSE} + L_{SSIM}) + L_{BCE} \quad (23)$$

L_{MSE} is the mean squared error (MSE) loss between the reconstructed output I_i and the input image O_i :

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^N (I_i - O_i)^2 \quad (24)$$

where N is the number of epochs.

L_{SSIM} is calculated as:

$$L_{SSIM} = \frac{1}{N} \sum_{i=1}^N (1 - \text{SSIM}(O_i, F_i)) \quad (25)$$

where $\text{SSIM}(\cdot)$ represents the structural similarity between two images (Wang et al., 2004). F_i represents the fused image.

L_{BCE} is the binary cross-entropy (BCE) loss applied to the segmentation output P_i and the segmentation mask T_i :

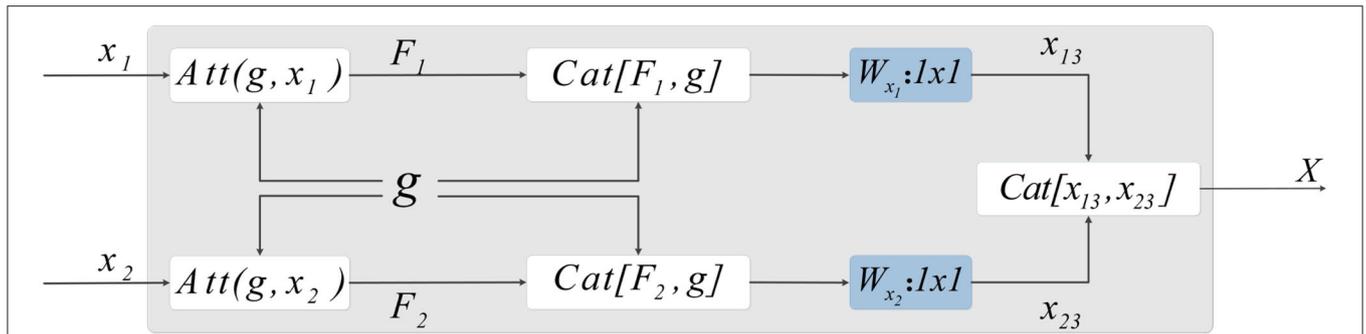


FIGURE 5 | Method of skip-connection. First, the attention weight is applied by CSAB to the output characteristics of the encoder. The features obtained from the up-sampling are then fused with the features after applying the attention weight. Subsequently, 1 × 1 convolution is used to reduce the dimension of the fused features. The final output is obtained by concatenating the two features.

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^N (T_i \log(P_i) + (1 - T_i) \log(1 - P_i)) \quad (26)$$

L_{MSE} and L_{SSIM} are the loss functions of the fusion sub-network, and L_{BCE} is the loss function for the segmentation sub-network. Since calculations of loss function is different, the loss functions must be balanced. The proposed model is trained with $\eta = 1$ and $\gamma = 1$. η represents the loss weight of the fusion sub-network. γ represents the loss weight of the segmentation sub-network. The loss curves are shown in **Figure 6**, from which we can learn that fusion loss is bigger than segmentation. To balance the loss weights between fusion and segmentation sub-networks, the loss weight in Equation (23) are set to $\eta = 0.02$ and $\gamma = 1$.

4. RESULTS

4.1. Experimental Environment

A 12 GB NVIDIA Titan X (Pascal) was used for training and evaluation. The system was running Windows 10 with an Intel Xeon CPU with 64 GB RAM. The program was written on Pycharm and is based on the Pytorch (Paszke et al., 2019) framework.

4.2. Dataset

The dataset contains clinical imaging data from 26 patients with brain gliomas examined at Yijishan Hospital of Wannan Medical College. The clinical image data consist of CT and T2-weighted MRI scans from glioma patients, of which nine images were acquired from low-grade glioma patients and 17 images were obtained from high-grade glioma patients. These are brain scans before treatment. After slicing the data, 860 pieces of CT and MRI images were obtained. Registration was completed after slicing. In addition, an expert was invited from the First Affiliated Hospital of the University of Science and Technology of China to manual delineate the whole tumor area. The data are shown in **Figure 7**.

Data augmentation was used to improve the generalization ability and robustness of the models. As the image size may

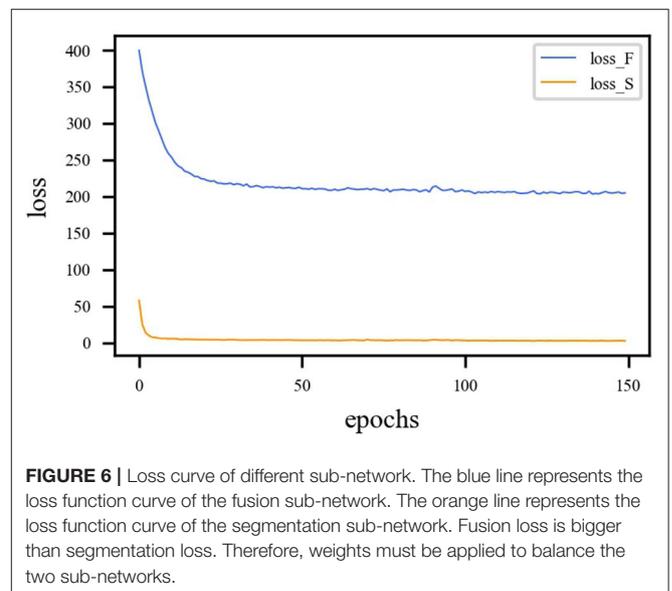


FIGURE 6 | Loss curve of different sub-network. The blue line represents the loss function curve of the fusion sub-network. The orange line represents the loss function curve of the segmentation sub-network. Fusion loss is bigger than segmentation loss. Therefore, weights must be applied to balance the two sub-networks.

change after data augmentation, the images were resampled to 256 × 256 pixels. Finally, the dataset was randomly divided into a training dataset (60%), validation dataset (20%), and test dataset (20%).

4.3. Evaluation Measures

The accuracy rate (ACC), positive predictive value (PPV), Jaccard similarity (JS), and dice coefficient (DC) were used as evaluation indexes. These metrics were calculated as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (27)$$

$$PPV = \frac{TP}{TP + FP} \quad (28)$$

$$JS = \frac{TP}{FP + TP + FN} \quad (29)$$

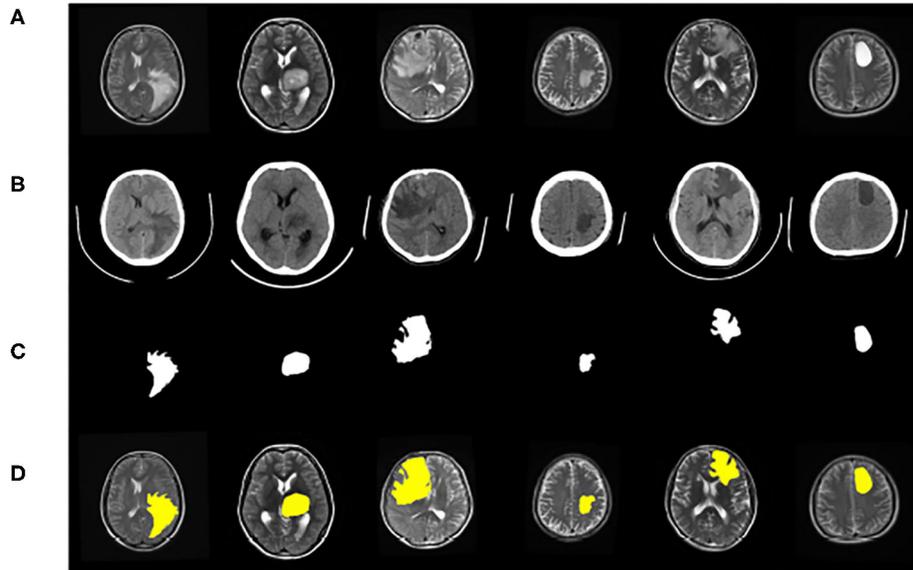


FIGURE 7 | Example of image modalities and ground truth in the multimodal glioma dataset. **(A)** Shows a head scan CT. **(B)** Shows a T2-weighted MRI. **(C)** Shows the ground truth. **(D)** Shows the merge result of **(B,C)**.

TABLE 1 | JS and DC for F-S-Net with different numbers of kernels and optimizers.

Number of convolution kernels	JS	DC
Adam + (32)	0.8070	0.8922
Adabound + (32)	0.8172	0.8975
SGD + (32)	0.7936	0.8839
Adabound + (16)	0.8040	0.8902

$$DC = \frac{2 * TP}{2 * TP + FP + FN} \quad (30)$$

where TP (true positive) represents the number of foreground pixels that are correctly classified as foreground (tumor region), TN (true negative) represents the number of background pixels that are correctly classified as background (non-tumor region), FP (false positive) represents the number of background pixels that are correctly identified as foreground, and FN (false negative) represents the number of foreground pixels that are incorrectly classified as background.

ACC is used to represent the classification accuracy of the classifier. PPV represents the proportion of true positives in all positive cases. JS reflects the ratio of the common area of the matched element to the split result. Any imprecise segmentation, whether under- or over-segmentation, will cause the JS to decrease. DC calculates the similarity between the prediction results and the ground truth to evaluate the performance of the model.

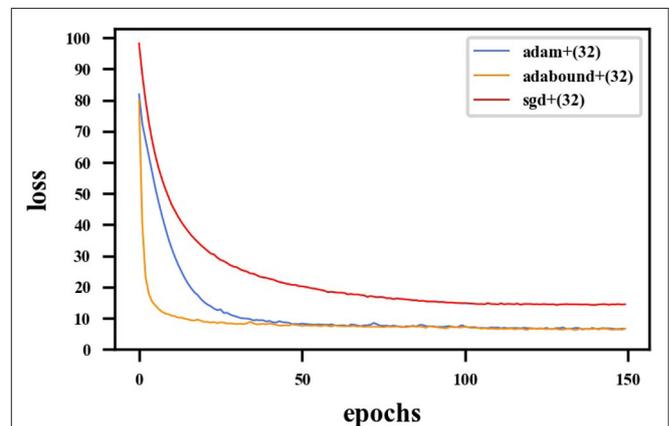


FIGURE 8 | Loss curve for F-S-Net with different optimizers. The red line represents loss function curve of the SGD optimizer. The blue line represents the loss function curve of the Adam optimizer. The orange line represents the loss function curve of the Adabound optimizer. The optimizer of Adabound has the fastest rate of convergence.

4.4. Training Optimization

First, the appropriate numbers of optimizers and convolution kernels were determined. Stochastic gradient descent (SGD) (Robbins and Monro, 1951) has been widely applied in the field of deep learning, while adaptive moment estimation (Kingma and Ba, 2014) offers better optimization performance. Adabound (Luo et al., 2019) dynamically crops the learning rate so that the algorithm is closer to Adam in the early stages of training and closer to SGD at the end. For CNNs, the receptive field and number of channels on the receptive field determine the

performance of the network. The convolution kernels considered in the experiments had the following structures: (16) 1-16-32-64-128-256-128-64-32-16-1; (32) 1-32-64-128-256-512-256-128-64-32-1. Four experimental groups were examined in the experiments: (1) Adam + (32), (2) Adabound + (32), (3) SGD + (32), and (4) Adabound + (16). The number of training epochs was set to 150, the batch size was set to 4, the weight decay was set to 5×10^{-8} , and the learning rate decreased by 0.1 after the 100th epoch. The experimental results are presented in **Table 1**. The loss curve is shown in **Figure 8**. In **Figure 8**, Adabound converges faster than the other optimizers. On the independent test dataset, the DCs of SGD, Adabound, and Adam are 0.8839, 0.8975, and 0.8922, respectively. Based on these results, Adabound and structure (32) were used in subsequent experiments.

Convolution kernels of different sizes have different receptive fields. The convolution kernel size of one encoder was kept the same as that in U-Net, while the convolution kernel size of the other encoder was modified as follows: (1) The 3×3 convolution of the amplified path was replaced by 5×5 convolution. (2) The two 3×3 convolutions were kept unchanged. (3) The 3×3 convolution of the amplified path was replaced by two 2×2 convolutions. Note that the padding is different when using 2×2 convolution. The experimental results presented in **Table 2** show that replacing a set of 3×3 convolutions with a set of 2×2 convolutions produces a better effect.

TABLE 2 | DC and JS for F-S-Net with different sizes of kernels and optimizers in the encoder-decoder for test dataset.

Sizes of convolution kernel	JS	DC
3x3-3x3	0.8172	0.8975
3x3-5x5	0.8226	0.9019
3x3-2x2(0101)	0.8234	0.9023
3x3-2x2(1010)	0.8226	0.9014

0101 and 1010 are the settings for the padding in each 2×2 convolution block.

It is necessary to modify the skip-connection to adapt to the inputs of the two encoders. An increase in skip-connection input would inevitably require feature fusion and dimensionality reduction. The order of 1×1 convolution and feature fusion may affect the performance of the network. Therefore, the four different structures shown in **Figure 9** were constructed.

Experiments were performed using the above four structures. The final experimental results are presented in **Table 3**, showing that better results are obtained by the skip-connection and dimension reduction of the two paths, respectively.

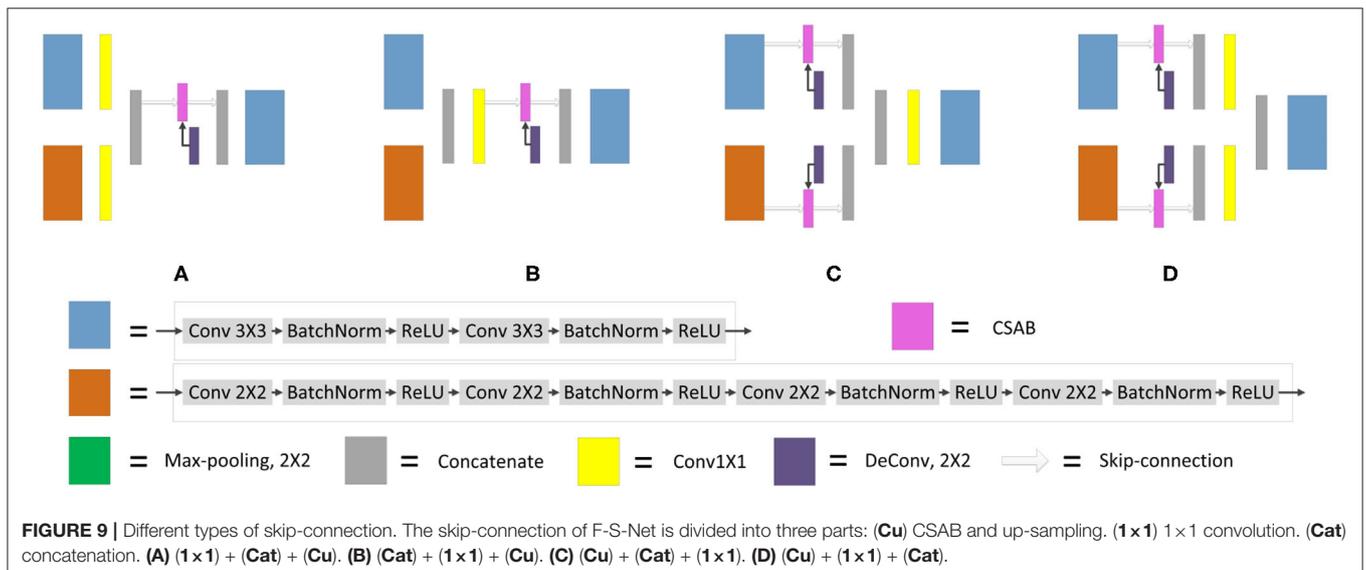
4.5. Ablation Analysis of Proposed Methods

The experimental results of the proposed structures with non-fusion and fusion were compared. It is clear that the improved structure and combination of modules are effective in enhancing the glioma segmentation results. The hyperparameters were set according to the previous optimization experiment. The training and testing samples for the experiment were taken from the glioma dataset. The fusion results in **Figure 10** clearly represent the overall area of the tumor, which makes the image features more obvious. The glioma can be accurately segmented and the network captures the specific outline and edge details of the lesion area in the image. **Table 4** presents the experimental results from using the proposed architecture.

DECSAU-Net is the segmentation sub-network in F-S-Net. When the proposed modules are removed, the network

TABLE 3 | DC and JS for F-S-Net with different types of skip-connection.

Skip-connection type	JS	DC
a	0.8234	0.9023
b	0.8019	0.8883
c	0.8109	0.8947
d	0.8280	0.9052



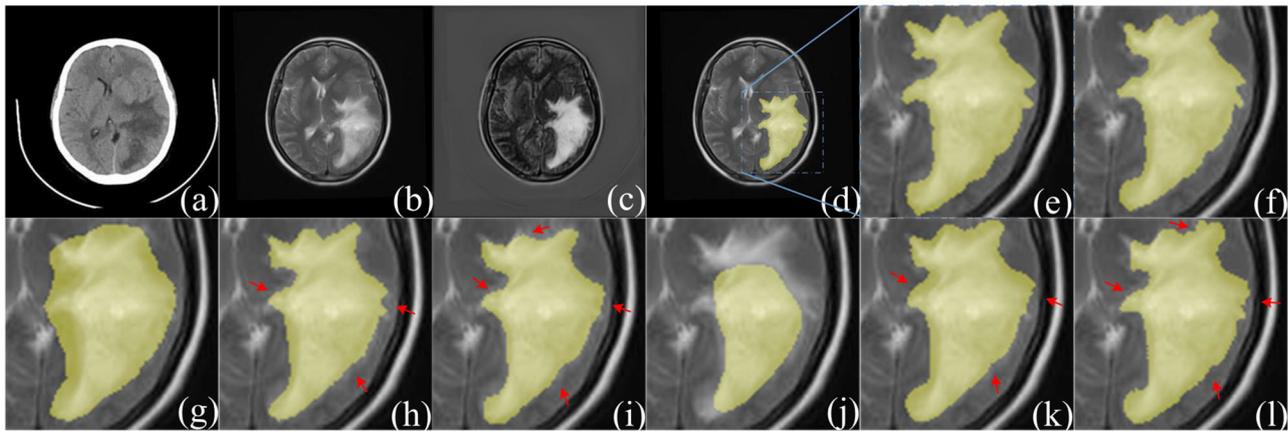


FIGURE 10 | Comparison of segmentation results between F-S-Net and other networks. **(a,b)** source images before fusion, **(c)** fusion result. Compared with **(a,b)**, the features in **(c)** are more obvious. The ground-truth glioma segmentation **(e)** is highlighted in **(d)**. Similarly, other model predictions are compared with those of F-S-Net **(f)**. **(g–l)** Are the results given by FCN8s, SegNet, DeeplabV3+, U-Net, R2U-Net, and AttU-Net, respectively. The missed dense predictions by other models are highlighted with red arrows.

TABLE 4 | Evaluation metric for ablation analysis of our methods with test dataset.

Model	ACC	PPV	JS	DC
U-Net (Ronneberger et al., 2015)	0.9916	0.8001	0.7656	0.8656
DECSAU-Net (Ours)	0.9938	0.8624	0.8193	0.8994
F-S-Net (Ours)	0.9943	0.9054	0.8280	0.9052

TABLE 5 | Evaluation metrics for different network architectures.

Model	ACC	PPV	JS	DC
FCN8s (Long et al., 2015)	0.9885	0.7714	0.6980	0.8197
SegNet (Badrinarayanan et al., 2017)	0.9931	0.8428	0.8039	0.8890
DeeplabV3+ (Chen et al., 2018a)	0.9931	0.8328	0.8066	0.8914
U-Net (Ronneberger et al., 2015)	0.9916	0.8001	0.7656	0.8656
R2U-Net (Alom et al., 2018)	0.9932	0.8472	0.8040	0.8905
AttU-Net (Oktay et al., 2018)	0.9934	0.8586	0.8087	0.8932
DECSAU-Net (Ours)	0.9938	0.8624	0.8193	0.8994
F-S-Net (Ours)	0.9943	0.9054	0.8280	0.9052

architecture is the same as the standard U-Net. Comparing the network models with and without DES and CSAB, it can be seen that the inclusion of DES and CSAB results in better performance. The PPV of DECSAU-Net is about 0.0623 higher than that of U-Net. The JS and DC values are about 0.0537 and 0.0338 higher, respectively. A comparison with U-Net shows that DES and CSAB improve the results of U-Net.

The results achieved with non-fusion and fusion approaches are now compared. The DC of the fused image is about 0.0058 higher than that of the image before fusion. The PPV of glioma segmentation after fusion is also higher at 0.9054. The difference in JS values shows that the result obtained after fusion is more similar to the ground truth. In general, the higher DC and

JS values demonstrate that the segmentation is more accurate after fusion.

4.6. Comparison With Other Methods

Table 5 compares the performance of different network architectures with that of the proposed F-S-Net after normalizing and enhancing the glioma data on the same test dataset. **Figure 10** shows the glioma segmentation results, which can be used to compare F-S-Net with other networks.

Several medical image segmentation architectures (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Alom et al., 2018; Chen et al., 2018a; Oktay et al., 2018) are outperformed by F-S-Net in both evaluations. The results in **Table 5** indicate that F-S-Net is more effective for performing accurate glioma segmentation. Compared with other network architectures, our method is more conducive to the segmentation of lesions as it maps multimodal medical images into the same semantic space. The advantage of F-S-Net is that the fusion of multimodal images makes the semantic information more conspicuous, and DES and CSAB allow the network to achieve a better segmentation effect.

5. DISCUSSION

Segmenting gliomas directly from CT or MRI images is a challenging task. In addition, the blurred edges of adjacent bones, blood vessels, or surgical packaging materials greatly increase the difficulty of segmentation.

Currently, most researchers directly input multimodal images into a network for learning. To the best of our knowledge, there are few reports on the segmentation of gliomas based on multimodal medical image fusion. To bridge this gap, F-S-Net has been proposed based on medical image fusion technology. Fusion and segmentation sub-networks are cascaded for end-to-end training, and two

new structures, DES and CSAB, are proposed based on the structural characteristics of U-Net. The basic idea of F-S-Net is to use fusion technology to produce images with more semantic information for the segmentation network, so as to obtain better segmentation results. DES and CSAB extract more detailed features and force the network to focus on the lesion area. Our work builds on existing techniques, such as CT and MRI image fusion. Medical image fusion techniques are not specifically designed for the segmentation task, but can provide images with richer semantic information for segmentation.

The most important innovation described in this paper is the ability to perform the task of glioma segmentation using image fusion. In the field of medical image analysis, better performance is often achieved by combining different technologies. The results in **Table 4** demonstrate the effectiveness of DES and CSAB, while those in **Tables 4, 5** demonstrate the improvement offered by using fusion technology for segmentation. Our network has the following advantages. First, image fusion can enrich the information available by integrating information between multimodal medical images. This method improves the quality of the image and facilitates the segmentation task. Second, the convolution kernels of different sizes in DES allow the network to obtain richer features. This helps to focus attention on the area of interest, and then obtains a better segmentation effect. Third, CSAB makes the network focus on the lesion area by applying different attention weights to the features. Our method not only integrates the complementary information from different modalities, but also extracts more detailed features. The experimental results show that F-S-Net outperforms several existing methods.

In summary, our proposed method will be helpful in allowing clinicians to diagnose and treat gliomas. More detailed segmentation results provide doctors with more complete boundary information of the tumor, and can better guide the resulting operations. In addition, better segmentation contributes to the reconstruction of the image data, which can provide more information for future monitoring and treatment planning. Our method overcomes the problem of incomplete semantic information and achieves good performance. The combination of segmentation and other medical imaging technologies will be explored in the future. This may improve clinical guidance in the diagnosis and treatment of glioma patients.

REFERENCES

- Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on U-net (R2u-net) for medical image segmentation. *arXiv [Preprint]*. arXiv:1802.06955. doi: 10.1109/NAECON.2018.8556686
- Altaf, F., Islam, S. M., Akhtar, N., and Janjua, N. K. (2019). Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access* 7, 99540–99572. doi: 10.1109/ACCESS.2019.2929365
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 6077–6086. doi: 10.1109/CVPR.2018.00636
- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

6. CONCLUSION

Glioma segmentation is a challenging and significant task in medical image segmentation. Based on medical image fusion technology, a cascade network was proposed to automatically segment gliomas from CT and MRI images. Our network obtained a DC of 0.9052 on the test dataset. Experimental results show that the combination of image fusion and image segmentation is effective. Our model provides a new method and a new idea for glioma segmentation based on deep learning, and is beneficial to the clinical diagnosis and treatment of patients. The proposed network is not only applicable to the segmentation of gliomas, but could also be easily applied to other medical image segmentation tasks.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because, datasets are open in the future. Requests to access the datasets should be directed to Deyun Zhang, zhangdeyun2016@163.com.

AUTHOR CONTRIBUTIONS

RS and JL: conceptualization. RS, CC, and MY: data curation. RS: methodology, project administration, visualization, and writing (original draft). RS, JL, and DZ: validation and writing (review and editing). All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was supported by two research grants: (1) Science and Technology Project grant from Anhui Province (Grant Nos. 1508085QH184 and 201904a07020098). (2) Fundamental Research Fund for the Central Universities (Grant No. WK 9110000032).

ACKNOWLEDGMENTS

The authors are grateful for the glioma dataset provided by Yijishan Hospital of Wannan Medical College. Thanks are also due for the ground truth glioma dataset provided by the First Affiliated Hospital of University of Science and Technology of China.

- Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., et al. (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. *Cancer J. Clin.* 69, 127–157. doi: 10.3322/caac.21552
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018a). “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 801–818.
- Chen, W., Liu, B., Peng, S., Sun, J., and Qiao, X. (2018b). “S3D-Unet: separable 3D U-net for brain tumor segmentation,” in *International MICCAI Brainlesion Workshop* (Granada: Springer), 358–368. doi: 10.1007/978-3-030-11726-9_32
- Cheng, J., Liu, J., Liu, L., Pan, Y., and Wang, J. (2019). “Multi-level glioma segmentation using 3D U-net combined attention mechanism with atrous convolution,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (San Diego, CA), 1031–1036. doi: 10.1109/BIBM47256.2019.8983092
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z
- Fan, F., Huang, Y., Wang, L., Xiong, X., Jiang, Z., Zhang, Z., et al. (2019). A semantic-based medical image fusion approach. *arXiv [Preprint]*. arXiv:1906.00225. Available online at: <https://arxiv.org/abs/1906.00225>
- Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv [Preprint]*. arXiv:1704.06857. doi: 10.1016/j.asoc.2018.05.018
- Hossain, M. Z., Soheli, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Comput. Surveys* 51, 1–36. doi: 10.1145/3295748
- Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for QSAR prediction. *IEEE J. Biomed. Health Inform.* 24, 3020–3028. doi: 10.1109/JBHI.2020.2977009
- Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). “No new-net,” in *International MICCAI Brainlesion Workshop* (Granada: Springer), 234–244.
- Jiang, Z., Ding, C., Liu, M., and Tao, D. (2019). “Two-stage cascaded U-net: 1st place solution to brats challenge 2019 segmentation task,” in *BrainLes@MICCAI* (Shenzhen). doi: 10.1007/978-3-030-46640-4_22
- Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2017a). “Ensembles of multiple models and architectures for robust brain tumour segmentation,” in *International MICCAI Brainlesion Workshop* (Quebec City, QC: Springer), 450–462. doi: 10.1007/978-3-319-75238-9_38
- Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017b). 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
- Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]*. arXiv:1412.6980. Available online at: <https://arxiv.org/abs/1412.6980>
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems* (Lake Tahoe), 1097–1105.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539
- Li, P., Wang, D., Wang, L., and Lu, H. (2018). Deep visual tracking: review and experimental comparison. *Pattern Recogn.* 76, 323–338. doi: 10.1016/j.patcog.2017.11.007
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005
- Liu, Y., Chen, X., Cheng, J., and Peng, H. (2017). “A medical image fusion method based on convolutional neural networks,” in *2017 20th International Conference on Information Fusion (Fusion)* (Xi’an), 1–7. doi: 10.23919/ICIF.2017.8009769
- Long, J., Shelhamer, E., and Darrell, T. (2015). “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440. doi: 10.1109/CVPR.2015.7298965
- Luo, L., Xiong, Y., Liu, Y., and Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. *arXiv [Preprint]*. arXiv:1902.09843. Available online at: <https://arxiv.org/abs/1902.09843>
- Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford University), 565–571. doi: 10.1109/3DV.2016.79
- Myronenko, A. (2018). “3D MRI brain tumor segmentation using autoencoder regularization,” in *International MICCAI Brainlesion Workshop* (Granada: Springer), 311–320. doi: 10.1007/978-3-030-11726-9_28
- Oktaç, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-net: Learning where to look for the pancreas. *arXiv [Preprint]*. arXiv:1804.03999. Available online at: <https://arxiv.org/abs/1804.03999>
- Ostrom, Q. T., Bauchet, L., Davis, F. G., Deltour, I., Fisher, J. L., Langer, C. E., et al. (2014). The epidemiology of glioma in adults: a “state of the science” review. *Neuro-oncology* 16, 896–913. doi: 10.1093/neuonc/nou087
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems* (Vancouver, BC), 8024–8035.
- Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 29, 2352–2449. doi: 10.1162/neco_a_00990
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems* (Montreal, QC), 91–99.
- Ristani, E., and Tomasi, C. (2018). “Features for multi-target multi-camera tracking and re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City), 6036–6046. doi: 10.1109/CVPR.2018.00632
- Robbins, H., and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* 22, 400–407. doi: 10.1214/aoms/1177729586
- Ronneberger, O., Fischer, P., and Brox, T. (2015). “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28
- Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). “Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks,” in *International MICCAI Brainlesion Workshop* (Quebec City, QC: Springer), 178–190.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861
- Xu, Y., Gong, M., Fu, H., Tao, D., Zhang, K., and Batmanghelich, K. (2018). “Multi-scale masked 3-D U-net for brain tumor segmentation,” in *International MICCAI Brainlesion Workshop* (Springer), 222–233. doi: 10.1007/978-3-030-11726-9_20
- Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2018). A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med. Image Anal.* 43, 98–111. doi: 10.1016/j.media.2017.10.002
- Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X. (2019). Object detection with deep learning: a review. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 3212–3232. doi: 10.1109/TNNLS.2018.2876865
- Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). “Unet++: A nested U-net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Granada: Springer), 3–11. doi: 10.1007/978-3-030-00889-5_1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Su, Liu, Zhang, Cheng and Ye. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.