Brain Tumor Segmentation Algorithm Based on Asymmetric Encoder and Multimodal Cross-Collaboration

Pengyue Zhang[1], Qiaomei Ma[2]*
Software School, North University of China, Taiyuan 030051, China[1, 2]
Shanxi Medical Imaging Artificial Intelligence Engineering Technology Research Center
(Central North University), Taiyuan 030051, China[2]

***Abstract*—To address the challenges of insufficient multimodal information fusion and insufficient long-range dependencies features extraction for brain tumor segmentation, this paper propose a novel network based on asymmetric encoder and multimodal cross-collaboration. The network employs an asymmetric encoder-decoder architecture. Firstly, the invert ConvNext split convolution (ICSC) block is used in the local refinement encoder and improved SwinTransformer with DscMLP enhancements (DscSwinTransformer) module is used in global associative encoder. The local and long-range dependencies of each stage of two parallel encoders can be well extracted by hybrid fusion. Moreover, this paper adds a multimodal cross-collaboration (MCC) module at the beginning of the two encoders to fully exploit the complementary information between modalities and reduce the reliance on a single modality during model training. Coordinate Attention (CA) is used in the bridge part of the encoder and decoder to capture important spatial location information. Then, the depthwise separable convolution (DscConv) module is used in the decoder branch to reduce the computation while maintaining good feature extraction ability. Finally, this paper uses a hybrid loss function of BCE, Dice and L2 loss to mitigate the problem of class datas imbalance. Experimental results show that our model achieves Dice coefficients of 0.897, 0.905 and 0.824 in the whole, core and enhanced tumor regions, respectively. These results show that the performance of our proposed method outperforms in comparison with several existing methods in core and enhanced tumor regions.**

***Keywords*—Brain tumor; multimodal cross-collaboration; asymmetric encoder; coordinate attention**

I. Introduction

A brain tumor is a mass or cluster of aberrant cells in the brain that impairs brain tissue function. Brain tumors are primarily classified as malignant or benign [1]. Malignant brain tumors are one of the most severe cancers at present, posing an increasing threat to human health. Brain tumors are classified as I-IV by the World Health Organization. Because the higher the grade of the brain tumor, the shorter the patient's survival time [2-3], early detection and treatment of brain tumors is critical. However, because the shape, size, location, and border of MRI images from various brain tumor patients vary, it is difficult to properly segment the brain tumor area. Manual segmentation of brain tumors by doctors is very time-consuming and inconsistent among different doctors for the same patient, while automatic segmentation techniques based on brain tumor MRI images can automatically locate and segment the shape, position and boundary of the brain tumor area, thus assisting doctors in diagnosing patients' conditions and alleviating their workload. Therefore, the research of brain tumor segmentation algorithm has significant scientific value and clinical relevance for efficient diagnosis of brain tumors.

At present, most segmentation networks do not properly use multi-modal complementary information. This study proposes a multi-modal cross-coordination feature fusion module, which reduces the feature dependence on a single mode and obtains rich context information of different modal complementary information. In order to obtain long-range dependencies information while extracting local feature information, this paper uses dual encoders to obtain spatial and coordinate attention information at different stages. In this paper, the mixed loss function is further designed to alleviate the problem of class imbalance in brain tumor data sets, so that the model can effectively segment different types of brain tumor regions.

II. Related Work

With the advancement of deep learning technologies, convoluted neural networks have emerged as the primary way for diagnosing brain tumor locations. U-net [4] is a typical segmentation network based on the encoder-decoder structure. Later, the Unet-type structure was further developed, such as Unet++ [5] with nested and densely connected structures, DenseUnet [7] that combines DenseNet [6] network and U-net, and Vnet [8] structure for volumetric segmentation. Convolutional neural networks can capture local features, but they have difficulty in modeling explicit long-range dependencies from the global feature space.

However, Local and global features are essential for dense prediction tasks. Vision Transformer [9] leverages self-attention mechanism to model long-range information, enabling CNN hybrid Transformer to fuse and extract local and distant features effectively. In this regard, TransBTS [10] network is proposed, which incorporates Transformer into the 3D CNN encoder-decoder architecture for the first time, enhancing global feature extraction. TransBTSV2 [11] further improves TransBTS by redesigning the Transformer module

and introducing deformable bottleneck module to capture shape-sensitive local features. SwinBTS [12] structure employs 3D SwinTransformer as both encoder and decoder of the network to extract global information from feature maps efficiently, using convolution operation for upsampling and downsampling. Unetr [13] network connects the Transformer encoder to the decoder with different resolutions through skip connections, capturing global multi-scale information more effectively.

Moreover, different brain tumor regions have large scale differences, and the receptive field of ordinary convolution is not enough to extract rich contextual feature information. In this regard, Liu et al. [14] proposes a lightweight ADHDC-Net network that combines hierarchical convolution with different dilation rates and tumor region relation-guided attention; Chang et al. [15] proposes a dual-path and multi-scale attention fusion module that merges feature maps with different receptive fields for dense pixel prediction; Rehman et al. [16] designs SDS-MSA-Net, which extracts features from 3D and 2D inputs separately, and uses selective depth supervision to assist the output, accelerating the model convergence speed, but at the same time processing 3D and 2D resources increases the computational cost. The above improved structures enhance the extraction ability of global features and multi-scale attention features respectively, but most of the current networks are limited to simple concatenation fusion of multi-modal brain tumor data input level, which cannot fully utilize the complementary fusion information between different modalities. Therefore, Liu et al. [17] designs a two-stage network that performs pixel-level fusion and feature-level fusion of multi-modal images to achieve more fine-grained utilization of multi-modal information; Zhou et al. [18] proposes an attention feature fusion module that can fuse different modalities and selectively extract useful feature information, but the core of the above networks still needs to improve the segmentation accuracy of the enhanced tumor region.

To solve the aforementioned challenges, our study offers an asymmetric encoder and multimodal cross-collaboration brain tumor segmentation network (AEMCCNet). This paper's primary contributions are summarized as follows:

1) This paper proposes an asymmetric encoder-decoder structure, where parallel local refinement encoder and global associative encoder use redesigned invert ConvNext split convolution (ICSC) block and improved SwinTransformer [19] with DscMLP enhancements (DscSwinTransformer) module respectively, which can effectively capture the fusion information of local details and long-range dependencies features in three stages of the encoder.

2) To reduce the model's dependence on a single brain tumor modality during training, this paper proposes a multimodal cross-collaboration (MCC) module, which can fully utilize the complementary information between modalities.

3) To obtain more accurate segmentation results, this paper uses coordinate attention (CA) Module [20] in AEMCCNet, which encodes the channels along horizontal and vertical directions. This transformation can capture remote features along one spatial direction and preserve precise location information along another direction, which is very important for generating spatial detail selective information.

4) To tackle the class imbalance issue, this paper employs a hybrid loss function composed of binary cross entropy, Dice, and L_2, which enhances brain tumor segmentation accuracy even further.

III. Methodology

A. AEMCCNet Network

The overall architecture of the brain tumor segmentation network based on asymmetric encoder and multimodal cross-collaboration proposed within this study is seen in Fig. 1.

The network model is an asymmetric encoder-decoder structure, where T1 and T1ce modalities are the inputs of the local refinement encoder; T2 and Flair modalities are the inputs of the global associative encoder. Both the local refinement encoder and the global associative encoder first use MCC module designed in this paper to fully learn the cross-modal features and reduce the model's dependence on a single modality [21]. Then this paper uses the ICSC Block and DscSwinTransformer module designed in this paper respectively, and the parallel dual-stream encoders fuse with each other at each stage, increasing the link throughout low-level detail features and high-level semantic features. Moreover, CA module is applied to the fused feature maps along two dimensions of MRI images to aggregate features, model long-range dependencies and channel transformation, and enhances the extraction of useful information while suppressing the influence of invalid information on tumor segmentation performance. Finally, depthwise separable convolution (DsConv) [22] is used in the decoder module to acquire the semantic details of the fusion information obtained from asymmetric dual-stream encoder and low-level decoder modalities.

B. MCC Module

To reduce the dependence on a single brain tumor modality during the model training process, and to better utilize the complementarity between T1, T1ce modalities and T2, Flair modalities to cross-extract features, this paper designs a MCC module, as shown in Fig. 2.

First, modality A and modality B separately go through 7×7 channel-by-channel convolution (DwConv) to obtain rich context information of a single modality, and then cross-multiply with the features of another modality after 1×1 convolution to obtain y_{11} and y_{21}, respectively, as shown in Eq. (1) and Eq. (2), to extract recognizable features from one modality to assist in correcting another modality;

$$y_{11} = \mathrm{Dw7\times7}(B) \otimes \mathrm{Conv1\times1}(A) \tag{1}$$

$$y_{21} = \mathrm{Dw7\times7}(A) \otimes \mathrm{Conv1\times1}(B) \tag{2}$$

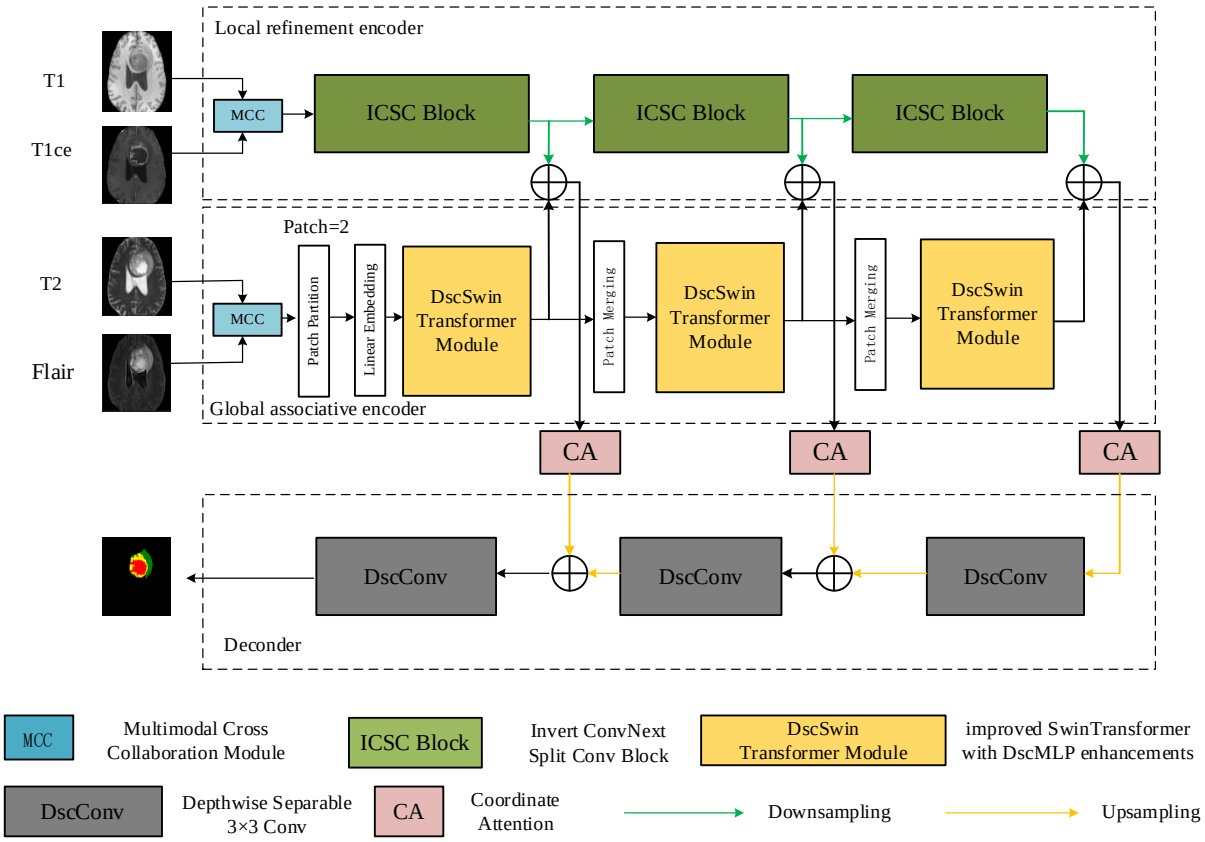


Fig. 1. Overall architecture of proposed AEMCCNet.

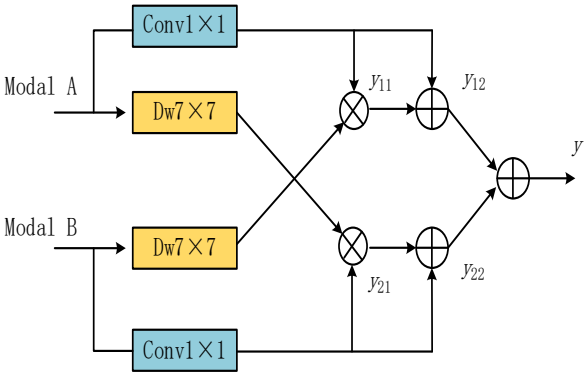


Fig. 2. Structure of a MCC module.

Then, the features of the same modality after 1×1 convolution are added and fused with the cross-fused features to generate the feature maps y_{12} and y_{22}, as stated in Eq. (3) and Eq. (4).

$$y_{12} = y_{11} + \text{Conv1}\times1(A) \tag{3}$$

$$y_{22} = y_{21} + \text{Conv1}\times1(B) \tag{4}$$

Finally, the branch modality features $y_{12}$and y_{22} are element-wise added and fused to generate the output feature map y of the module, as illustrated in Eq. (5).

$$y = y_{12} + y_{22} \tag{5}$$

C. ICSC Block

As shown in Fig. 3(a), MobileNetV2 [23] swaps the order of convolutional dimensionality increase and decrease in the Inverted Residuals structure, uses depthwise separable convolution to reduce the computational cost, and enhances the nonlinear expression ability of the network. As shown in Fig. 3(b), ConvNext Block [24] inherits the feature of wide convolutional dimension in the middle layer of Inverted Residul, and sets the depthwise separable convolution kernel size to 7×7, Padding=3. This paper draws on the advantages of these two modules and redesigns the ICSC Block of channel-split, as shown in Fig. 3 (c).

The input feature map of this module is $X \in R^{C\times H\times W}$,, where C is the number of channels. First, X is split into two $C/2$ branches X_{11} and X_{21} along the channel dimension. The left branch X_{11} uses 7×7 depthwise separable convolution to extract rich spatial context information, and then uses two 1×1 convolutions to increase and decrease the dimension respectively, obtaining the feature map X_1. The right branch X_{21} first uses 1×1 convolution to reduce the dimension, then uses 3×3 depthwise separable convolution to extract spatial rich information and increase the dimension, and then uses 1×1 convolution to reduce the dimension again, obtaining the feature map X_2. Then, the outputs of the two branches are concatenated and fused along the channel direction. Finally, the fused feature map is added with the original input feature map by identity connection to obtain the output feature map y of this module as shown in Eq. (6) to Eq. (8).

$$X_1 = Conv1(g(Conv1[L(Dw7(X_{11}))])) \quad (6)$$

$$X_2 = Conv1(Relu6(Dw3[B(Conv1(X_{21}))])) \quad (7)$$

$$Y = X + [X_1, X_2] \quad (8)$$

where, *DW7*() is a channel-wise convolution with a kernel size of 7×7, *Conv1*() is a convolution with a kernel size of 1×1, *L* is LN normalization operation, *B* is BN normalization operation, *g*() is gelu() activation function, *Relu6* is activation function, [·] is channel-wise concatenation and fusion.

D. DscSwinTransformer Module

Encoder-decoder structure based on CNN lacks the ability to capture long-range dependencies features, while lightweight SwinTransformer Block uses sliding window self-attention mechanism to capture global dependencies features information. In SwinTransformer Block [19], the multilayer perception (MLP) uses two fully connected layers for dimension transformation, but using fully connected layers in image segmentation causes partial segmentation information loss.

Inspired by the above content, this paper replaces the multilayer perception (MLP) in DscSwinTransformer Module with the designed depthwise separable perception (DscMLP), which can further refine the context information and improve the nonlinear transformation of features. As shown in Fig. 4, DscMLP takes the feature X after self-attention W-MSA, sliding window self-attention mechanism SW-MSA, and reshapes the shape of the feature map *X* first from [B, H×W, C] to X_1 in [B, C, H, W] dimensions, where B, C, H, and W are the batch size, the number of channels, the height, and the width, respectively, of the model training settings; Then it applies depthwise separable convolution and identity connection on X_1 in parallel respectively, and performs element-wise multiplication on the two-branch results; finally it reshapes the feature dimension to [B, H×W, C] dimension output feature map *Y*.

The DscSwinTransformer Module is used in the three stages of the global associative encoder, and the repetition number of SwinTransformer in each stage is 1; before the first stage the output feature map $y \in R^{C\times H\times W}$ of the multimodal cross-collaboration module is partitioned into *M* patches of size *P*×*P*,*P*=2 in the Patch Partition module, and each patch is reshaped into a one-dimensional vector $y_p \in R^{M\times(P\times P\times C)}$, then these patches are flattened along the channel direction and mapped to *D* dimensions by the Linear Projection module $E \in R^{(P\times P\times C)\times D}$, while adding a learnable position variable $E_{pos} \in R^{(P\times P\times C)\times D}$ to obtain the feature *z*, as shown in Eq. (9):

$$z = y_p E + E_{pos} \quad (9)$$

In the DscSwinTransformer Module, layer normalization (LN) is first used and residual connection is performed on W-MSA, SW-MSA, DscMLP, as shown in Fig. 4, the above process can be expressed as.

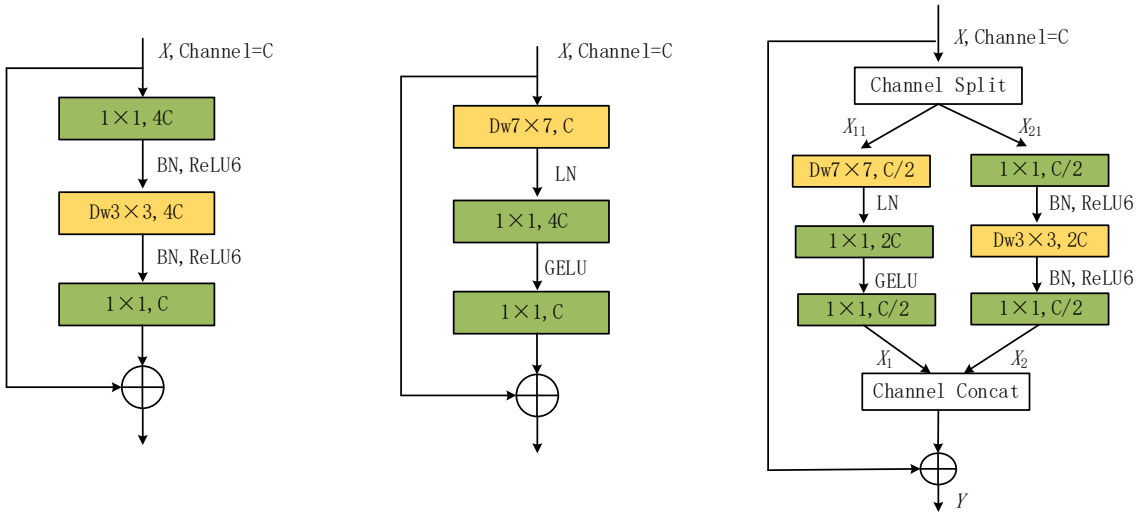
$$\hat{Z}^k = W - MSA(LN(Z^{k-1})) + Z^{k-1} \quad (10)$$

$$Z^k = DscMLP(LN(\hat{Z}^k)) + \hat{Z}^k \quad (11)$$

$$\hat{Z}^{k+1} = SW - MSA(LN(Z^k)) + Z^k \quad (12)$$

$$Z^{k+1} = DscMLP(LN(\hat{Z}^{k+1})) + \hat{Z}^{k+1} \quad (13)$$

To generate 2x downsampling, between the DscSwinTransformer modules use patch merging to increase dimensionality and decrease token numbers.



(a) InvertedResiduals Block (b) ConvNeXt Block (c) ICSC Block

Fig. 3. Comparison of Convolutional Blocks from left to right as (a) InvertedResidual Block, (b) ConvNeXt Block, (c) ICSC Block.

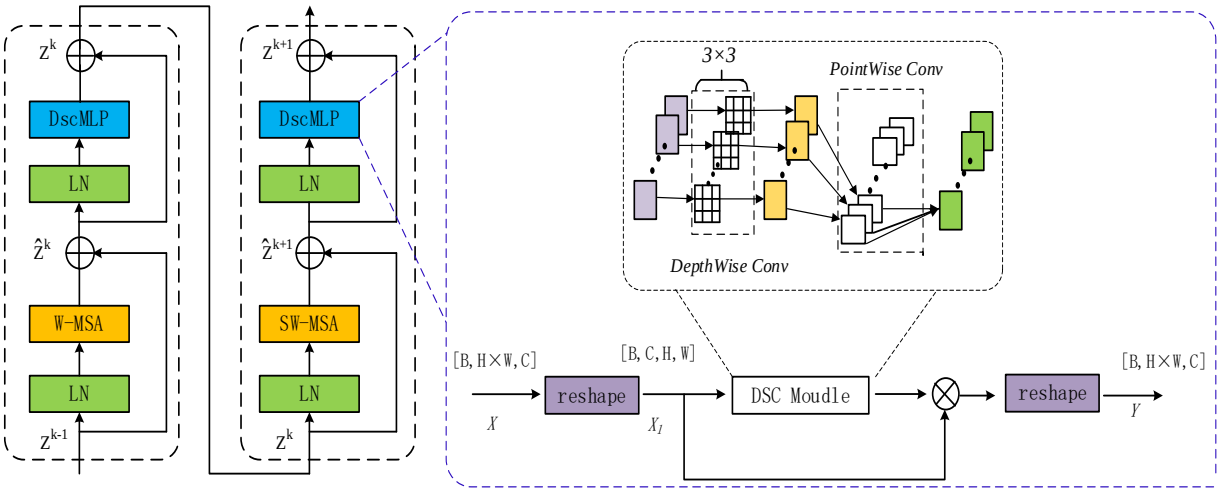


Fig. 4. Structure of DscSwinTransformer module.

E. CA Module

In deep network segmentation models, such as SE attention and CBAM attention, it has been proven that they can significantly enhance channel attention and spatial attention weights, and promote the model's segmentation performance, but they typically ignore positional details, which is vital for creating selective spatial features. Therefore, this paper introduces CA module [20], which embeds positional information into channel attention, as shown in Fig. 5.

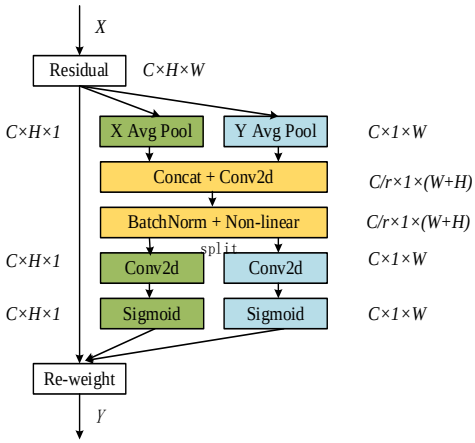


Fig. 5. Structure of CA module.

Coordinate information embedding alongside coordinated attention generation make up the two sections of the CA module. The coordinate details embedding implies encoding the input feature map *X* along both vertical and horizontal axes, respectively, using pooling kernels of shapes (H, 1) and (1, W) within the channels, to create two-dimensional feature maps that can capture distant features through one spatial direction and retain accurate positioning data along the other. Eq. (14) and Eq. (15) illustrate this:

$$Z_c^h(h)=\frac{1}{W}\sum_{0\le i\le W}X_c(h,i) \tag{14}$$

$$Z_c^w(w)=\frac{1}{H}\sum_{0\le j\le H}X_c(j,w) \tag{15}$$

where, $Z_c^h(h)$ is a result of the *C*th channel with height *h*, and $Z_c^w(w)$ is similarly.

The second transformation is the generation of CA module. First, $Z_c^h(h)$ and $Z_c^w(w)$ are concatenated, and then a 1×1 convolution function *F* and the feature mapping *f* of spatial data within multiple directions was extracted using an activation function that is nonlinear, as shown in Eq. (16).

$$f=\delta(F([Z^h,Z^w])) \tag{16}$$

Then *f* is decomposed into a pair of distinct tensors $f^h\in R^{C/r\times H}$ and $f^w\in R^{C/r\times W}$, and the convolution function F is used to transform them into tensors with the identical number of channels just like the input *X*, and the function of Sigmoid activating is utilizing to derive the attention weights g^h and g^w on two directions respectively. Ultimately, the module's initial feature map multiplies element by element with the two separate attention weights to yield the module's output *Y*, as indicated in Eq. (17) to Eq. (19).

$$g^h=\sigma(F(f^h)) \tag{17}$$

$$g^w=\sigma(F(f^w)) \tag{18}$$

$$Y(i,j)=X(i,j)\cdot g^h(i,j)\cdot g^w(i,j) \tag{19}$$

IV. Experimental Settings

A. Dataset

This paper uses the dataset from the Brain Tumor Segmentation (BraTS) competition in 2019, which contains 76 cases of low-grade glioma and 259 cases of high-grade glioma. This paper divides the training and testing data of the BraTS2019 dataset according to a ratio of 8:2. The

segmentation results are evaluated by the performance indicators of the whole tumor region (core tumor region and edema region), core tumor region (enhanced tumor region and necrosis region) and enhanced tumor region.

B. Data Preprocessing

Since training with 3D format images takes a long time and requires better GPU and more memory, this paper chooses to use 2D slices to train the proposed network. The size of the 3D data for each modality is 240×240×155. Since there is a lot of useless background information on the outer edge of the brain tumor data, which causes the problem of data class imbalance, this paper first crops the spatial size of the 3D data to 160×160×155 to eliminate the useless spatial background information, and then slices the data along the channel direction, transforming the 3D data into 155 slices of 160×160 2D slices to meet the needs of the model training in this paper.

Due to different imaging mechanisms, different modalities have different image contrast, so normalization is used to make the data intensity of different modalities balanced, which is conducive to the model using complementary information between different modalities. The normalization operation is shown in Eq. (20):

$$z = \frac{x - \bar{x}}{\lambda} \tag{20}$$

In the above equation, x is the cropped 2D sliced image, $\bar{x}$ is the mean value of the input image, λ is the standard deviation of the input image, and z is the normalized image.

As shown in Fig. 6, the images of four modalities and the ground truth label of a slice of a case after preprocessing in this paper are shown from left to right as (a) T1, (b) T1ce, (c)T2, (d) Flair and (e) Ground truth (GT) label. In the network model prediction image and the ground truth label, green represents edema tumor region, yellow represents enhanced tumor region and red represents necrosis and non-enhanced tumor region.

C. Experimental Environment Configuration

The software version is PyTorch 1.11.0 with Cuda 11.3, and the hardware environment consists of a 32 core CPU processor, 30GB RAM, and a GPU with NVIDIA RTX A5000, 24GB video memory. To update the model weights, this paper utilizes the Adaptive Moment Estimation (Adam) algorithm [25] as the optimizer; the detailed training experiment configuration is shown in Table I.

D. Evaluation Metrics

Four distinct assessment standards are employed to assess the segmentation accuracy of the model in this study to evaluate the how effective the suggested model algorithm. As described in Eq. (21), the Dice similarity coefficient is a value between 0 and 1. The closer to 1, the more similar the brain tumor segmentation result is to the manual label result, and the better the segmentation effect.

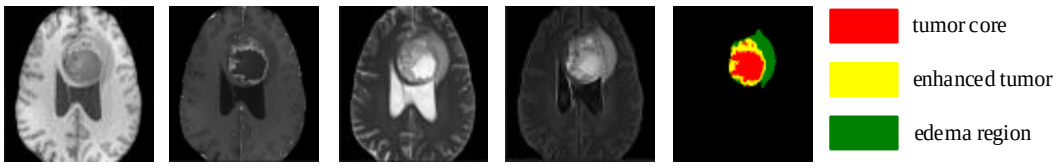
$$Dice = \frac{2TP}{2TP + FP + FN} \tag{21}$$

Sensitivity is a measure of the model's ability to predict positive pixels. Precision is used to measure the ability of the model to correctly predict pixels. As shown in Eq. (22) and Eq. (23).

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

Where TP stands for true positive pixels, FN represents for false negative pixels, FP means for false positive pixels, and TN indicates for true negative pixels.

$$Sensitivity = \frac{TP}{TP + FN} \tag{23}$$



(a) T1 (b) T1ce (c) T2 (d) Flair (e) GT

Fig. 6. Images after data preprocessing. from left to right as (a) T1, (b)T1ce, (c) T2, (d) Flair and (e) GT label.

TABLE I. EXPERIMENTAL CONFIGURATION

Configurations	Values
Software version	PyTorch11.0
GPU	NVIDIA RTX A5000
Optimizer	Adam
Initial learning rate	0.003
Momentum	0.09
Weight decay coefficient	0.00001
Batch szie	24
Tranning Epoches	300

To calculate the distance between the model segmentation border and the real label boundary, the Hausdorff distance (HD) has utilized, the higher the segmentation precision, the lower the Hausdorff distance. Eq. (24) shows the calculation formula.

$$Hausdorff = \max\left\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y)\right\} \quad (24)$$

Here y represents GT and x represents the predicted segmentation result, $d(x,y)$ is the Euclidean distance between x and y. In this paper, HD95 is used in the evaluation, which means taking the 95th percentile result.

E. Hybrid Loss Function

As indicated in Eq. (25), binary cross entropy (*BCE*) is often utilized as a loss function for performing segmentation operations for various medical data sets.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \quad (25)$$

Where N denotes the overall amount of output pixels, y_i means the la-bel value for the ith pixel, and p_i defines the model prediction value for the ith pixel. Since the BCE loss function assigns equal weights to foreground and background pixels, but there is a large difference in the proportion of foreground and background pixels in multimodal brain tumors, and foreground pixels account for only a minority, there is a problem of data class imbalance. This study applies the *Dice* loss function to the *BCE* loss function to overcome the problem of data class imbalance, as seen in Eq. (26):

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i}^{N} p_i y_i + \varepsilon}{\sum_{i}^{N} p_i + \sum_{i}^{N} y_i + \varepsilon} \quad (26)$$

The value given by the parameter has been set to 10^{-6} to ensure data stability. In addition, overfitting is prone to occur during the training of the model, so based on this L_2 loss is introduced to alleviate the overfitting problem during the training of the model, which is advantageous to network convergence. The L_2 loss function is shown in Eq. (27):

$$L_2 = \frac{1}{N}\Sigma_i^N \left(y_i - p_i\right)^2 \quad (27)$$

In summary, the hybrid loss function in this paper is shown in Eq. (28):

$$L = L_{\text{Dice}} + \alpha L_{\text{BCE}} + \beta L_2 \quad (28)$$

The approach of controlling hyperparameters is applied in this paper to evaluate the most effective settings. The variation range of hyperparameters α and β values is shown in Table II

First, β is set to 0 to test the best hyperparameter α. As shown in Fig. 7, the fluctuation range of α is between 0 and 1. When α is 0.5, the model in this work hits the peak point in the whole tumor (WT), core tumor (TC), and enhanced tumor (ET) regions, implying that its predictive ability is best at this time.

On this basis, the experimentation of the optimal superparameter β was continued, as shown in Fig. 8, where the fluctuation of β ranges from 0 to 0.1 spacing. The Dice coefficient of this paper's model on WT, TC and ET regions reaches 0.897, 0.905 & 0.824 when the finalized parameter α is 0.5 and β is 0.05, and the prediction performance of this paper reaches the best.

The best hyperparameters α=0.5 and β=0.05 are substituted into the hybrid loss function. When the model training iteration number is 295 rounds, as illustrated in Fig. 9, the model's training and validation loss values tend to be optimum.

F. Ablation Experiment

To evaluate the efficacy of the design along with addition of modules in this study, under the same experimental conditions of using the same loss function and parameters, this paper replaces the original 3×3 convolution module of the encoder-decoder structure with depthwise separable convolution, which serves as the Baseline structure of our model. This paper adds different modules to the Baseline structure, and Table III displays the outcomes. Where MCC stands MCC module, ICSC represents ICSC block CA indicates CA module and DscSwinT represents DscSwinTransformer Module. By incorporating the MCC module to the first layer of the encoder before entering the Baseline, the Dice values of the whole tumor, core tumor and enhanced tumor regions in the model increase by 0.6%, 0.9% and 1.5%, respectively. Based on Baseline, using the ICSC block, the performance of the model in the WT, TC and ET regions is further improved. Similarly, based on Baseline, using the DscSwinTransformer to obtain the accuracy indicators of the ET and TC regions are significantly improved. Based on Baseline, adding CA module, the Dice indicators of the WT and TC are significantly improved.

TABLE II. VARIATION RANGE OF HYPERPARAMETERS

Hyperparameters	Variation Range
α	Between 0 and 1
β	Between 0 and 0.1

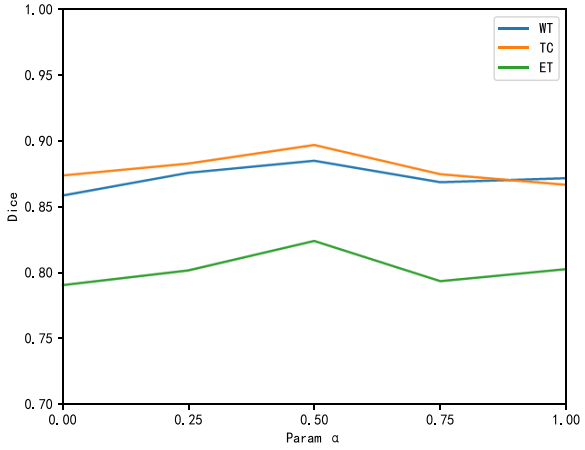


Fig. 7. Effect of hyperparameter α on model performance.

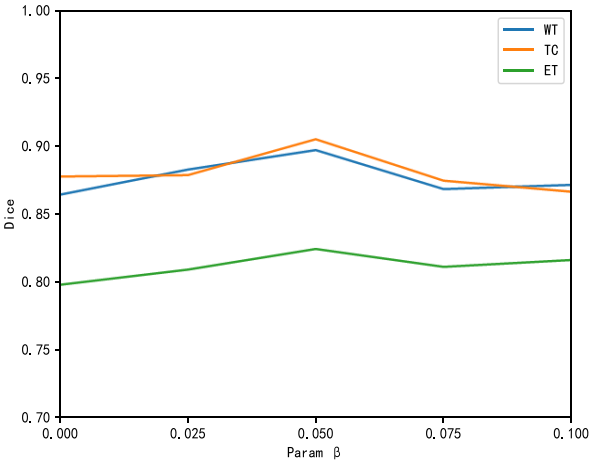


Fig. 8. Effect of hyperparameter β on model performance.

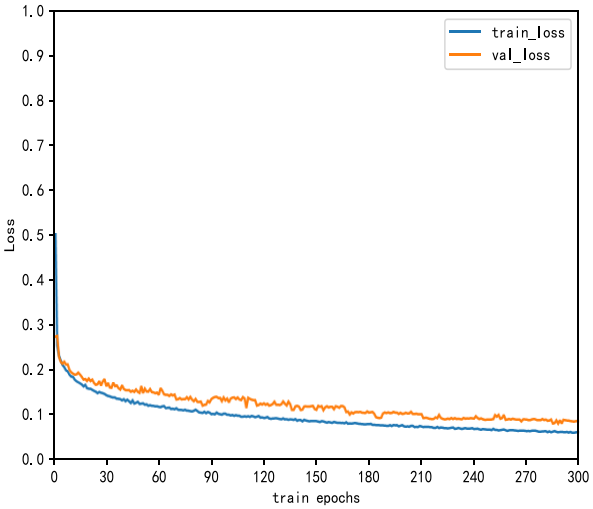


Fig. 9. Model training Loss variation.

Finally, this paper integrates the modules designed and used above into our structure. Compared with the Baseline structure, the Dice values of the WT, TC and ET increase by 5.1%, 5.3%, 6%, respectively. The hausdorff distance values of the three types of tumors are also the lowest values for ablation comparison experiments, proving the effectiveness of this method.

To test the effectiveness of adding different attentions to the bridging part between the asymmetric dual-stream encoder and decoder, this paper compares the CA module mechanism with SE [26] channel attention, CBAM [27] channel spatial attention, and ECA [28] efficient channel attention, respectively, as shown in Fig. 10 (a) and Fig. 10(b). After using the improved CA module in the decoder branch, the Dice similarity coefficient and HD95 of the model reach the best.

G. Experimental Results

To further verify the effectiveness of our method for multimodal brain tumor MR image segmentation, this paper uses part of the BraTS2019 dataset as the test set to compare with other advanced methods, and the results are shown in Table IV.

Compared with the classic 2D U-net and Unet++ networks, our model has a significant performance improvement in the whole, core and enhanced tumor regions. When compared with the advanced 3D models TransBTSV2, SwinBTS and MBANet, the 3D models have an advantage in segmenting the whole tumor due to their spatial continuity, but our 2D model uses DscSwinTransformer Module to strengthen the extraction of long-range dependencies features, making the Dice value of the whole tumor region close to the advanced 3D segmentation methods, and having the most optimal values in the overall evaluation indicators.

Finally, this paper segments some samples from the BraTS2019 test dataset and compare them with the input images and segmentation results as shown in Fig.11, where each row represents a patient case. U-net has over-segmentation phenomenon in the edema region of the second case, and also over-segments the core and enhanced tumor regions of the third case; The current advanced SwinBTS and TransBTSv2 networks have good segmentation effects on the edema region due to their spatial continuity, but they mis-segment part of the enhanced tumor region as necrotic tumor region in the second and third cases. Our proposed segmentation model improves by 5.7% and 2.2% respectively on the TC and ET tumor regions compared to the advanced TransBTSV2 network, showing that our method has better segmentation effects on the tumor core and enhanced regions.

TABLE III. MODULE ABLATION COMPARISON EXPERIMENTS

Baseline	MCC	ICSC	DscSwinT	CA	Dice			Hausdorff95(mm)		
					WT	TC	ET	WT	TC	ET
√					0.846	0.852	0.764	6.617	5.509	3.155
√	√				0.852	0.861	0.779	6.613	5.516	3.167
√		√			0.855	0.879	0.802	6.607	5.498	2.948
√			√		0.863	0.886	0.798	6.594	5.195	2.935
√				√	0.877	0.873	0.788	6.539	5.504	3.026
√	√	√	√	√	0.897	0.905	0.824	5.508	4.892	2.790

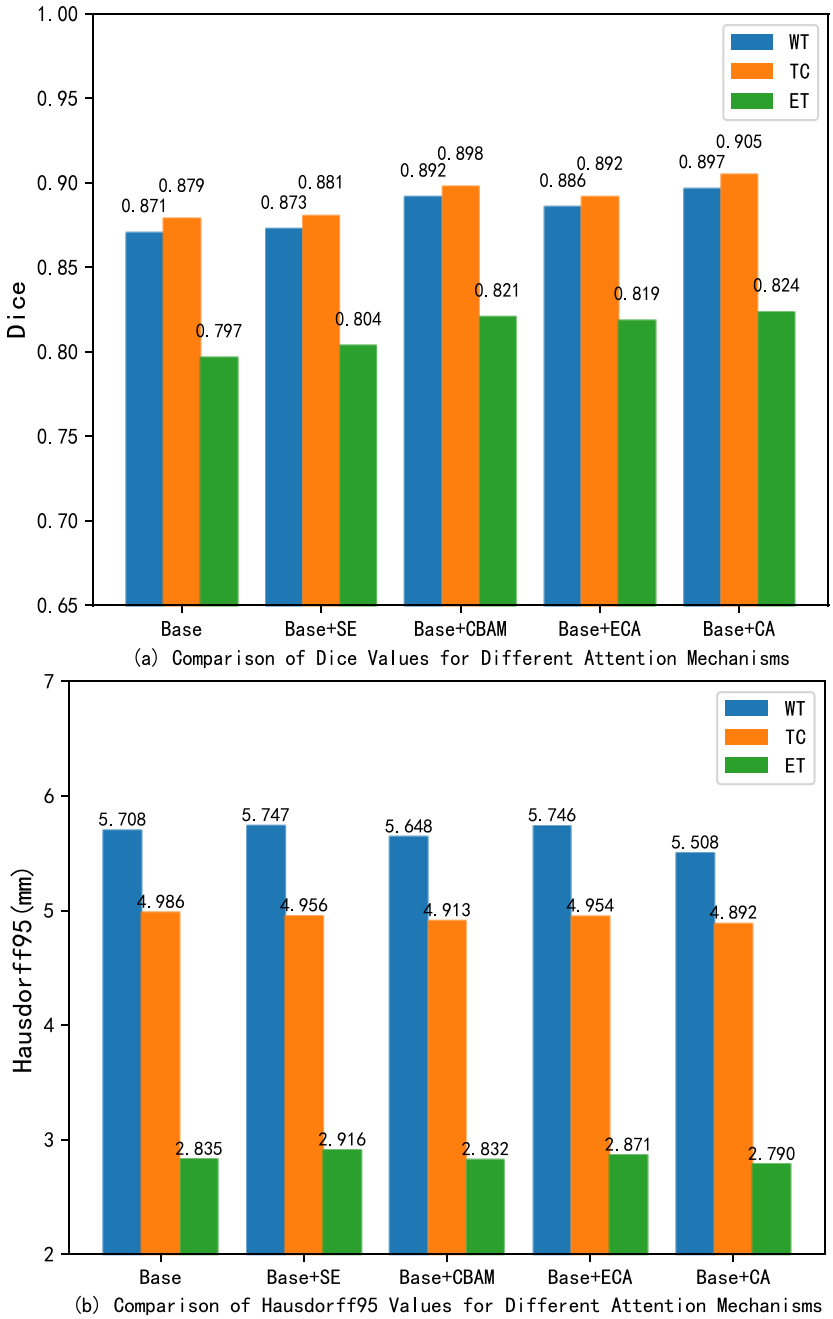
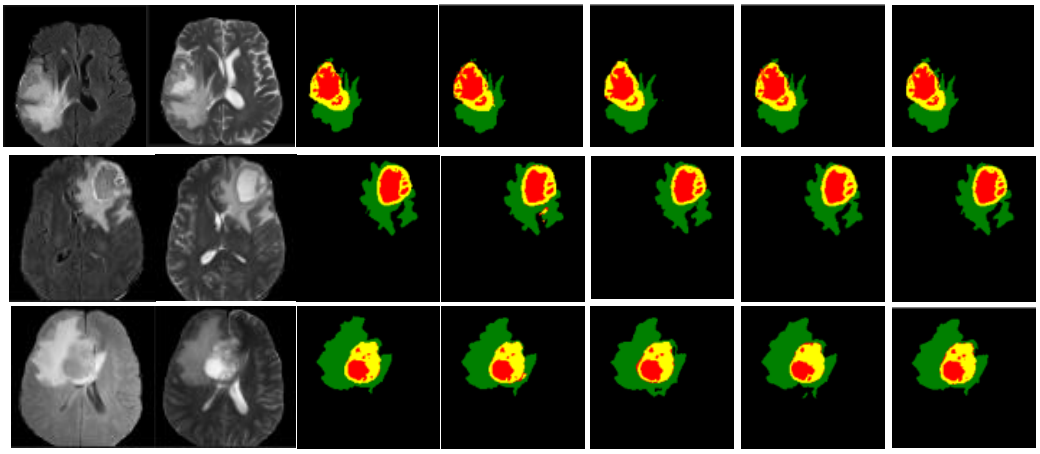


Fig. 10. Comparison of dice and hausdorff95 values with different attention modules added.

TABLE IV. COMPARATIVE EXPERIMENTS OF ADVANCED NETWORKS

Models	Dice↑			Precision↑			Sensitivity↑			Hausdorff95(mm) ↓		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
U-net [4]	0.834	0.823	0.769	0.871	0.898	0.800	0.856	0.908	0.813	6.648	6.596	4.062
Unet++ [5]	0.848	0.860	0.784	0.868	0.899	0.805	0.849	0.910	0.795	6.256	6.131	3.967
TransBTSv2 [11]	0.902	0.848	0.802	0.852	0.893	0.789	0.902	0.922	0.860	5.432	5.473	3.696
SwinBTS [12]	0.903	0.825	0.796	0.856	0.909	0.788	0.890	0.908	0.864	8.560	15.78	26.84
MBANet [29]	0.898	0.831	0.782	0.892	0.905	0.809	0.896	0.896	0.794	5.881	5.090	3.086
Ours	0.897	0.905	0.824	0.884	0.916	0.819	0.921	0.915	0.859	5.508	4.892	2.790



(a) Flair (b) T2 (c) GT (d) U-net (e) SwinBTS (f) TransBTSV2 (g) Ours

Fig. 11. Visual comparison of segmentation results: (a) Flair; (b) T2; (c) GT; (d) U-net; (e) SwinBTS; (f) TransBTSV2; (g) Ours.

H. Complexity Comparative Analysis

To better evaluate the performance of the model, we compare the number of parameters (Params) and the amount of computation (FLOPs) of the network model under the same input size, as shown in Table V. The quantity of parameters of our modelis reduced by about 75.45% compared to the CNN hybrid SwinTransformer structure SwinUnet, and the amount of computation is reduced by about 44.66%. Moreover, compared with the advanced 3D model SwinBTS, the number of parameters of our model is reduced by about 14.68%, and the amount of computation is reduced by about 70.09%. In summary, our model embeds depthwise separable convolution in each proposed module, and only uses one DscSwinTransformer Module in each stage of the global associative encoder. This substantially minimizes the number of parameters and the amount of computation required by the model, resulting in excellent segmentation performance for TC and ET regions.

TABLE V. COMPARATIVE ANALYSIS OF COMPLEXITY

Method	Params/M	FLOPs/G
U-net [4]	69.71	28.46
TransUnet [30]	96.07	48.34
TransBTS [10]	32.99	333.00
SwinBTS [12]	27.64	89.46
Unetr [13]	92.58	41.19
Ours	23.58	26.75

V. DISCUSSION

Due to insufficient hardware resources, this article only changed the number of rounds of model training on the basis of the same training parameters. The results showed that in the 295th round of model training, Dice, Precision, and HD95 in the model validation set were optimal. The sensitivity values in the TC and ET regions still had a gap compared to the optimal values of the network. This may be related to the removal of slices without lesion information during the process of slicing 3D images into 2D images in data preprocessing. Considering the segmentation performance presented by the four evaluation indicators, this article retains the 295th training model as the optimal model for the network to test the test set. If the GPU computing resources are sufficient, further increasing the training rounds can be considered to find the optimal value of the comprehensive evaluation indicators during the training process as the optimal model.

A lightweight AEMCCNet network model is proposed in this paper. It can be seen from subsections F and G of Section IV that compared with other network models, the proposed model has better segmentation effect in the core and enhanced tumor regions. In addition, the AEMCCNet network in Section IV, subsection H, reaches the optimal values of the number of parameters and the amount of computation. For the whole tumor area, 2D network cannot capture the information of adjacent sections of 3D brain tumor cases, but AEMCCNet is close to the segmentation effect of 3D network, and multiple adjacent sections can be combined for segmentation in the future.

VI. CONCLUSION

To address the problems of insufficient fusion of multimodal brain tumor information and inadequate extraction of long-range dependencies features, this paper adopts an asymmetric encoder-decoder structure, which incorporates the MCC module, ICSC block, DscSwinTransformer module, and CA module designed in this paper into the architecture, and offers an asymmetric encoder-based brain tumor segmentation algorithm with multimodal cross-collaboration. The MCC module can reduce the model's dependence on a single brain tumor modality during training and fully utilize the complementary information between modalities; the local refinement encoder branch uses the ICSC module to split channels and extract local detail features, enhancing the network's non-linear expression ability; the global associative encoder uses DscSwinTransformer module to strengthen the capture ability of long-range dependencies features; the bridge part between the asymmetric encoder and decoder uses the CA module to enhance the location weight of spatial detail

selective information; the decoder branch uses DscConv to maintain good feature extraction ability while reducing computation; During network training, a hybrid loss function is redesigned to handle the class imbalance and overfitting issues. The findings from the experiments reveal that the model's accurate segmentation of WT region is comparable to that of advanced 3D segmentation algorithms., while the Dice coefficient of TC and ET regions is better than other advanced models. At the same time, in the comparative experiment, it has the best values of other evaluation indicators, and uses DscConv throughout the model to minimize model parameters and calculations, but the extraction of edge detail information of enhanced tumor is still insufficient. Therefore, in future work, we will explore using an efficient encoder-decoder structure enhanced by edge operator attention or a low-parameter 2.5D network to further improve the segmentation accuracy of enhanced tumor region.

DATA AVAILABILITY

The datasets for this study can be found in the BraTS 2019 da taset available at: https://www.med.upenn. edu/cbica/brats2019/data.html.

ACKNOWLEDGMENT

We thank anonymous reviewers for valuable suggestions and comments. This work was supported by the Natural Science Foundation of Shanxi Province, China (No. 20210302123019).

REFERENCES

[1] Ostrom Q T, Cioffi G, Waite K, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. Neuro-oncology, 2021, 23(3): 1-105.

[2] Zhang R, Jia S, Adamu M J, et al. HMNet: Hierarchical Multi-Scale Brain Tumor Segmentation Network. Journal of Clinical Medicine, 2023, 12(2): 538.

[3] Wang J, Yu Z, Luan Z, et al. RDAU-Net: Based on a residual convolutional neural network with DFP and CBAM for brain tumor segmentation. Frontiers in Oncology, 2022,12:805263.

[4] Ronneberger O, Fischer P, Brox T.U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[5] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation. 4th Deep Learning in Medical Image Analysis Work, 2018: 3-11.

[6] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[7] Kaku A, Hegde C V, Huang J, et al. DARTS: DenseUnet-based automatic rapid tool for brain segmentation. arXiv preprint arXiv:1911.05567, 2019.

[8] Milletari F, Navab N, Ahmadi S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 4th International Conference on 3DVision. IEEE, 2016 :565-571.

[9] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020. arXiv preprint arXiv:2010.11929, 2010.

[10] Wang W, Chen C, Ding M, et al. Transbts: Multimodal brain tumor segmentation using transformer. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021: 109-119.

[11] Li J, Wang W, Chen C, et al. TransBTSV2: Towards Better and More Efficient Volumetric Segmentation of Medical Images. arXiv preprint arXiv:2201.12785, 2022.

[12] Jiang Y, Zhang Y, Lin X, et al. SwinBTS: A method for 3D multimodal brain tumor segmentation using swin transformer. Brain sciences, 2022, 12(6): 797.

[13] Hatamizadeh A, Tang Y, Nath V, et al. Unetr: Transformers for 3d medical image segmentation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 574-584.

[14] Liu H, Huo G, Li Q, et al. Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation. Expert Systems with Applications, 2023, 214: 119166.

[15] Chang Y, Zheng Z, Sun Y, et al. Dpafnet: A residual dual-path attention-fusion convolutional neural network for multimodal brain tumor segmentation. Biomedical Signal Processing and Control, 2023, 79: 104037.

[16] Rehman A, Usman M, Shahid A, et al. Selective Deeply Supervised Multi-Scale Attention Network for Brain Tumor Segmentation. Sensors, 2023, 23(4): 2346.

[17] Liu Y, Mu F, Shi Y, et al. Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. Frontiers in Neuroscience, 2022, 16:1000587.

[18] Zhou T. Modality-level cross-connection and attentional feature fusion based deep neural network for multi-modal brain tumor segmentation. Biomedical Signal Processing and Control, 2023, 81: 104524.

[19] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[20] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[21] Zhou Y, Liang X, Gu Y, et al. Multi-classifier interactive learning for ambiguous speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 695-705.

[22] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[23] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[24] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[25] Wu X, Huang F, Huang H. Fast stochastic recursive momentum methods for imbalanced data mining. 2022 IEEE International Conference on Data Mining (ICDM). IEEE, 2022: 578-587.

[26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computervision and pattern recognition. 2018: 7132-7141.

[27] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[28] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[29] Cao Y, Zhou W, Zang M, et al. MBANet: A 3D convolutional neural network with multi-branch attention for brain tumor segmentation from MRI images. Biomedical Signal Processing and Control, 2023, 80: 104296.

[30] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.