Received 5 April 2023, accepted 28 April 2023, date of publication 2 May 2023, date of current version 6 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272352

RESEARCH ARTICLE

TAGU-Net: Transformer Convolution Hybrid-Based U-Net With Attention Gate for Atypical Meningioma Segmentation

HONG HUANG [1], **PANPAN LIU** [2], **AND JIE LIU** [1]

[1] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, Shandong 264209, China
[2] Department of Neurosurgery, Weihai Municipal Hospital, Shandong University, Weihai, Shandong 264200, China

Corresponding authors: Panpan Liu (weihaislyy@163.com) and Jie Liu (lj2002@sdu.edu.cn)

ABSTRACT Meningioma is derived from the cap cells that reside on the arachnoid membrane. The atypical meninges of Grade II, a classification established by the World Health Organization, are included in one of the grades of meningioma. It has been discovered that early surgical resection significantly reduces the recurrence rate and mortality of tumors. Accurate segmentation of magnetic resonance images of brain tumors is crucial for diagnosing and treating atypical meningiomas. However, the traditional automatic segmentation framework heavily relies on convolution. The convolution-based segmentation network has limitations such as the size of the convolution kernels, a restricted receptive field, and a lack of spatial aggregation ability. To overcome these limitations, this paper presents a novel hybrid architecture named TAGU-Net, which combines Transformer and convolution based on U-Net with an attention gate. The TAGU-Net architecture extracts features of different resolution feature scales using convolutional neural network and Transformer. This approach effectively captures the image's long-distance dependency and global characteristics in the encoder stage, relying on the global self-attention mechanism of the Transformer. Additionally, the inductive bias of the convolution neural network is combined to enhance the local modeling information and improve the model's overall modeling ability. In the decoder phase, the attention gate is introduced to adaptively learn the skip connection information and up-sampling information in the network. This information is weighted and fused to highlight important features and suppress irrelevant features. To obtain better model training and avoid the vanishing gradient, deep supervision technology is used in the training process. Supplementary loss is added in some stages to supervise the training and achieve the best effect of atypical meningioma segmentation. The proposed method is evaluated on both the private atypical meningioma dataset and the publicly available BraTs2018 dataset.TAGU-Net has achieved Dice Scores of 97.67% and 97.62% and Jaccard index of 96.35% and 95.35% on the private atypical meningioma dataset and BraTs2018 dataset respectively, which is a state-of-the-art segmentation result beyond existing methods. According to the research results, the TAGU-Net model significantly improves atypical meningioma segmentation and can effectively assist doctors in processing MRI images.

INDEX TERMS U-Net, atypical meningioma segmentation, transformer, MRI, attention gate.

I. INTROUDUCTION

Intracranial meningiomas are extra-axial central nervous system tumors, frequently occur in the brain and spine [1], [2]. According to the World Health Organization (WHO), the lesions of meningiomas are classified as grade I (benign tumors), grade II (atypical), or grade III (anaplastic) [3], [4], and WHO grade II atypical meningiomas (AM) belong to one of the grades of meningiomas [5]. Intracranial meningiomas mostly present as low-grade (grade I) benign tumors, and high-grade (grade II or III) meningiomas account for 6% – 18% of all meningiomas [6]. Demonstrated that

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

resection can be curative for nearly 80% of benign tumors, but intracranial meningioma remains a dangerous disease [7]. However, high-grade meningiomas exhibit an increased risk of recurrence after treatment, exhibit aggressive behavior, and increase morbidity and decrease survival [8], [9], [10]. Numerous studies have shown that grade II and III meningiomas are recurrent, aggressive, and aggressive [11] and that grade III meningioma are considered the most aggressive, i.e., malignant. Therefore, the clinic is of great significance for diagnosing and segmenting grade II meningiomas, i.e., AM, especially as the tumor grows slowly and inhibits vital organs before progressing to malignancy. Early detection of AM holds significant value in the treatment of meningiomas and ultimately enhances patient survival rates.

The segmentation method based on traditional machine learning is not popular with the public because of its complexity, cumbersome operation, and low accuracy. Currently, the mainstream deep learning method still relies on the pure convolution architecture, and the pure Transformer and convolutional neural network (CNN) and Transformer hybrid architectures have their own defects. The traditional CNN segmentation network is limited by the size of the convolution kernel, which has the problems of limited receptive field and insufficient spatial aggregation ability [12]. While dilation convolution can increase the receptive field of CNN, it is not sufficient to overcome these problems [13]. Due to the lack of prior knowledge like CNN inductive bias (ie, locality and translation equivariance), the pure Transformer architecture requires a large amount of data to learn enough information, which is extremely difficult and particularly challenging on medical image data sets with few samples [14]. In the hybrid architecture of CNN and Transformer, Transformer typically operates on the feature map extracted by CNN [15]. Obviously, this approach leads to a significant loss of valuable information.

In this study, we propose a novel hybrid architecture that combines Transformer and convolution, based on U-Net with attention gate, to achieve automatic segmentation of atypical meninges. two types of Encoders are designed, namely ConvEncoder, and FormerEncoder. Different from the conventional hybrid architecture of Transformer and CNN, the proposed FormerEncoder does not model the feature maps extracted by CNN, but in the encoder stage, the two types of encoders extract the features of different resolution feature maps at different scales. ConvEncoder and FormerEncoder extract different information from different resolution feature maps at different scales, and the information obtained by the same Encoder at different scales and resolutions is also different, and the shallow features obtained in high resolution contain texture, contour and position information, while the deep features obtained in low resolution contain rich semantic information. In the encoder stage, the two types of encoders extracted the features of different resolution feature maps at different scales. At the same time, TAGU-Net fused the long-distance dependency and global features of images captured by FormerEncoder with the local features extracted by ConvEncoder to produce a more effective feature representation. Moreover, for FormerEncoder, it is a flexible and efficient encoder, which can replace the Former Encoder Block of FormerEncoder based on the characteristics of different tasks or different data sets, such as Swin-Transformer [16], PVT [17] or T2T-ViT [18]. In the decoder stage, we only use ConvDecoder to avoid the high complexity of the model. At the same time, we introduce the attention gatie mechanism to adaptively learn the skip connection information and the up-sampling information in the network, and carry out the weighted fusion of the two, highlight the important features and suppress the irrelevant features, realizing the feature reuse in the decoder stage. At the same time, in the training phase, we use deep supervision, and in some stages, we introduce an auxiliary loss function to carry out supervision training. The main contributions of this paper can be summarized as follows:

- A transformer convolution hybrid architecture based U-Net with attention gate is proposed for MRI segmentation and learning of atypical meningiomas. The results demonstrate that this framework surpasses state-of-the-art models in terms of performance.
- Use genetic algorithm-based adaptive histogram equalization to preprocess the original MRI image to enhance image details, thereby achieving a more precise segmentation.
- The FormerEncoder module is designed to capture global features at different scales and model the long-distance dependence of the image, and it is flexible and replaceable based on different data characteristics. In addition, the convolution features generated by ConvEncoder are fused to achieve a complementary structure.
- The Attention Gate module is introduced to adaptively learn feature information from different structures in the decoder branch, highlighting important features and suppressing irrelevant features.

The rest of this paper is organized as follows: The related work of atypical meningioma segmentation and segmentation network is presented in Section II, respectively. Section III introduces the data set used and provides a detailed description and the model framework and algorithm proposed in this paper. Subsequently, the experimental results are presented and analyzed in Section IV, including the performance comparison with other methods. Finally, Section V draws the main conclusions about the work introduced.

II. RELATED WORK

Clinically, for meningioma diagnosis and recognition divided into invasive and non-invasive methods, non-invasive medical imaging techniques such as computed tomography (CT) and magnetic resonance imaging (MRI) which is more favored in the diagnostic stage as brain tumor recognition tools, outperforming invasive methods such as tissue biopsy [19]. Among noninvasive medical imaging techniques, MRI is considered the most common technique for diagnosing meningiomas because it can provide detailed images and noninvasive properties of human tissues and organs.

Brain tumor segmentation is an essential step before applying any treatment, and the current standard method of brain tumor segmentation is manual and is based on expert experience. Experts must manually segment the MRI to delineate the target image. The surge in the number of patients can lead to a decrease in the quality of physicians' work, creating a situation of manual segmentation error. With the development of computer technology, computer-aided diagnosis (CAD) systems have been developed quickly and applied to segment tumors. A large number of studies have achieved great success in the fields of breast cancer [20], [21], brain tumors [22], [23] and other fields.

A. MACHINE LEARNING

For meningioma, the field of meningioma segmentation has developed rapidly. The main methods are divided into segmentation methods based on traditional machine learning and segmentation methods based on deep learning. There are two kinds of segmentation methods based on conventional machine learning, one is based on the unsupervised clustering method, and the other is to transform the segmentation problem into a pixel classification problem. Almahfud et al. used a combination of two K-means and Fuzzy C-means (FCM) grouping methods to detect brain tumors [24]. Benson et al. implemented an improved version of the fuzzy C-mean clustering and watershed algorithm. An effective way of selecting the initial centroid based on histogram calculation was proposed to improve the accuracy of clustering. In addition, a set-based tag detection method was proposed to avoid over-segmentation [25]. Saha and Hossain proposed a way to automatically classify brain images of MRI using K-means clustering, nonsubsampledcontourlet transform (NSCT), and support vector machine (SVM). Because NSCT has significant characteristics such as multiscale, multidirectional, and displacement invariance, K-means clustering and NSCT are used to segment brain images of MRI, which improves the efficiency and accuracy of segmentation [26]. Amin et al. used a fused eigenvector to apply a random forest (RF) classifier to classify between three sub-tumor regions using a mixture of gabor wavelet features (GWF), histograms of oriented gradient (HOG), local binary pattern (LBP), and segmentation based fractal texture analysis (SFTA) features [27]. Al-Dmour and Al-Ani proposed an efficient and fully automatic brain tissue segmentation algorithm based on clustering fusion technology. A Neural network simulates clustering and divides the target based on superpixels, three clustering algorithms, and a neural network [28]. Kaya et al. used principle component analysis (PCA) for multivariable data reduction and five standard PCA algorithms for target segmentation [29].

B. DEEP LEARNING

Among the segmentation methods based on deep learning, due to the excellent performance of the convolutional neural network in the field of image processing and computer vision, especially after the birth of AlexNet [30], CNN has ushered in a blowout of explosive development. At the same time, CNN architecture has become the leading choice in medical image segmentation. Kamnitsas et al. proposed a dual-channel, 11-layer deep three-dimensional convolutional neural network and designed an efficient and effective intensive training scheme while automatically adapting to the inherent class imbalance in the data, using the dual-channel architecture to combine local and more extensive context information [31]; Havaei et al. proposed a fully automatic brain tumor segmentation method based on deep neural network (DNN). By using dual-channel CNN architecture and cascade architecture, the system can more accurately model local label dependency by using local features and more global context features [32]; Díaz-Pernas et al. proposed a deep convolution neural network with multiscale methods. Inspired by the inherent multiscale operation of the human visual system (HVS), the input images are processed at three spatial scales along different processing paths. The multiscale processing strategy can effectively extract discriminatory texture features for different types of tumors [33]; Haq et al. proposed an integration and hybrid method based on a deep convolution neural network and machine learning classifier. By learning the feature map from the brain MRI image space to the tumor marker region through CNN, a faster region-based CNN was developed for tumor region localization, followed by the potential region network (RPN). Finally, the deep CNN and machine learning classifier were connected to achieve target segmentation [34]; Ding et al. proposed Stack Multi-Connection Simple Reduction Net (SMCSRNet) based on U-Net framework, which reduces the number of model parameters and adds bridging between stacked cascaded networks to improve information loss [35]; Maji et al. proposed an Attention Res-UNet (ARU-GD) with a guided decoder, which designs the loss function by guiding the decoder, and introduces the attention gate to focus on the activation of relevant information [36].

C. SEGMENTATION NETWORK

The complete convolution networks (FCNs) [37] proposed by Long et al. achieve the state-of-the-art (SOTA) of image segmentation and semantic segmentation under the premise of only using convolution; Ronneberger et al. proposed a symmetrical encoder-decoder structure of medical image segmentation network U-Net [12]. U-Net has played an excellent role in medical images with small data scale. Most future segmentation networks will continue to use the U-Net structure and make improvements; Ibtehaz et al. re-thought based on U-Net. Inspired by Inception [38], they replaced the traditional convolutional layer with a multi-resolution idea and introduced residual connection [39]. Instead of simply connecting the feature maps from the encoder to the decoder, they passed through the convolutional layer chain with residual connection and then combined with the decoder features to enhance the feature representation [40]; Influenced by Transformer [14], [41], a large number of studies have explored the feasibility of Transformer in the field of medical image segmentation. Hatamizadeh et al. used Transformer as an encoder to learn the sequence representation of the input quantity and effectively capture the global multi-scale information, and combined the information with the CNN decoder through the skip connection of different resolutions [42]; Wang et al. proposed a network based on Transformer's coder-decoder structure [43]. The 3D CNN is used to extract the spatial feature map to carefully transform the feature map of the global feature modeling of the input Transformer. At the same time, the decoder uses the features embedded in the Transformer and performs progressive up-sampling to predict the detailed segmentation map.

III. MATERIAL AND METHODS

A. PRIVATE DATASET

This study used the private atypical meningioma patient dataset from Weihai Municipal Hospital. In this dataset, researchers retrospectively collected pre-operative MRI scans of 203 subjects from 2010 to 2019. All subjects had the following MRI findings: (1) First operation of tumor resection in Weihai Municipal Hospital; (2) The grade of postoperative pathological diagnosis results was precise; (3) Preoperative high-quality cranial T2 weighted imaging (T2) and contrast-enhanced T1 weighted imaging (T1C) MRI; (4) Preoperative complete clinical data and information; (5) No history of surgery, gamma knife and other treatments; (5) No MRI sequence was incomplete (T2 / T1C) and imaging was free of artifacts.

Multimodal MRI delivers a great deal of information for segmentation and extraction of meningiomas, specifically for meningioma machine scans that provide hundreds of 2D imaged brain slices with high soft tissue contrast, the common MRI sequences are T1, T2, T1C, and fluid-attenuated inversion recovery (FLAIR). Each MRI sequence produces images with different tissue contrast, which has a different role in distinguishing tumors [44], [45]. T1 modality is usually used to process healthy tissue, T2 modality is more suitable for detecting the boundary of edematous regions, T1C modality highlights the tumor boundary, and flair modality is favorable for detecting edematous regions in cerebrospinal fluid.

In this study, we utilized T1C and T2 MRI sequences at the same time. Because different MRI sequences come from different signal information and belong to different modal information in a broad sense, the MRI sequences that use T1C and T2 simultaneously in this study belong to multimodal information fusion. In contrast, the joint multimodal information fusion usually constructs a multi-branch structure. Each mode has its flow; feature fusion is performed after feature extraction of different branches or streams. However, this method is challenging to obtain the characteristic relationship between different modes, and it is difficult to use the complementary information between the modal information. Moreover, although the mode represents different signals, it represents the same feature. To address this issue, we propose a novel approach that directly inputs the two modal information to effectively extract the relationship between different modal information.

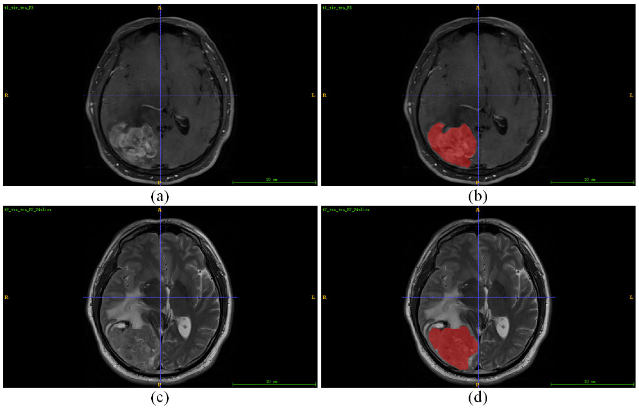


FIGURE 1. **Different modal MRI images and their corresponding segmentation results, (a) (b) are T1C modal slices and their segmentation results, and (c) (d) are T2 modal slices and their segmentation results.**

The resolution of most images in this dataset is 512 × 512, and the resolution of a few images is 432 × 512 and 496 × 512 in order to unify the resolution and facilitate image feeding into the model, we set all image resolutions to 512 × 512. Fig.1 shows the MRI images of different modes and their corresponding segmentation results.

B. PUBLIC DATASET

The BraTs dataset serves as a public benchmark for brain tumor segmentation, and for our study, we utilized the BraTs 2018 [44], [46] training dataset acquired from the official website to evaluate our proposed method. This dataset comprises two types of gliomas, high-grade glioma (HGG) typically classified as WHO grade III or IV, and low-grade glioma (LGG), typically classified as WHO grade I or II. Given that atypical meningiomas are only LGG, we focused our verification solely on LGG patients, totaling to 75 patients in the BraTs 2018 dataset. Each patient's MRI includes corresponding T1, T1C, T2, and FLAIR sequences, which led to a collection of 4845 slice images with each slice containing information from the four sequences. The size of each slice image was 160 × 160 × 4.

C. PREPROCESSING OF IMAGE DATA

Because medical images are very susceptible to noise, the quality of the obtained images could be lower, with noticeable noise and low contrast. However, the quality of the image has a significant impact on the subsequent diagnosis and segmentation. In low-quality images, the region of interest (ROI)

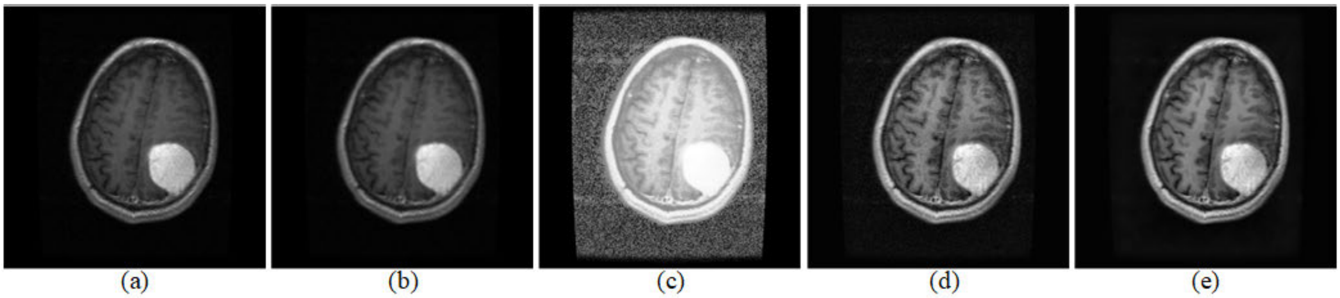


FIGURE 2. Different modal MRI images and their corresponding segmentation results, (a) original image, (b) after Gaussian filtering, (c) after AHE, (d) after CLAHE, and (e) after GAAHE.

may not be observed, resulting in abnormal diagnosis or segmentation. Therefore, it is necessary to de-noise and enhance the image's contrast. This preprocessing aims to solve the defects of MRI and generate the most precise and representative MRI possible to achieve the most accurate segmentation process. In this paper, we performed a series of preprocessing operations, and Gaussian filter denoising was used to remove general noise. Then we used a genetic algorithm-based adaptive histogram equalization (GAAHE) [47] to enhance the contrast of MRI.

1) GAUSSIAN FILTER

The gaussian filter is a smooth linear filter. Gaussian filter is used to smooth the image to remove noise. When calculating the gaussian smoothing result, the origin is the center point. Other points are weighted according to their positions on the standard distribution curve to obtain a weighted average value. The template used in this article is 5 × 5. The size of the Gaussian filter is publicly defined as follows:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

of which x and y indicates the size of the kernel filter, σ^2 is the variance of the Gaussian filter.

2) GENETIC ALGORITHM BASED ADAPTIVE HISTOGRAM EQUALIZATION

Adaptive histogram equalization (AHE) is commonly used to enhance contrast in medical images, but artifacts and noise amplification often occur in the actual process. It is also evident in Fig.2 that the artifacts and noise of MRI images after AHE are severe. Contrast limited adaptive histogram equalization (CLAHE) [48] is an improved approach of AHE, which suppresses the problem of AHE noise amplification by limiting the contrast. This paper uses the genetic algorithm-based adaptive histogram equalization method, which is also improved based on AHE. A new subdivision method is applied to the histogram through exposure threshold and optimal threshold to maintain brightness and reduce information loss. The threshold parameters are optimized using the concept of genetic algorithm. Then, modify each sub-histogram's probability density function (PDF) to improve the image quality. Fig.2 shows the comparison between the pre-processed image and the original image.

D. PROPOSED FRAMEWORK

This paper attempts to solve the segmentation problem of atypical meningioma MRI. The proposed framework is to design a hybrid model of transformer and convolution to adapt to the input of multimodal MRI sequences to achieve accurate segmentation. As shown in Fig.3, the framework is divided into three stages: preprocessing image data, TAGU-Net segmentation model, training, and model evaluation.

1) TAGU-NET NETWORK ARCHITECTURE

The TAGU-Net proposed by us is improved based on the classical segmentation network U-Net architecture, which consists of two parts, the encoder branch, and the decoder branch. The U-Net achieved precise positioning mainly by contracting and expanding paths. The encoder branch of the U-Net network is primarily composed of convolution and down-sampling operations, which are responsible for feature extraction. The decoder branch is used to restore the original resolution of the feature map. The connection between the two branches is mainly through a skip connection. The skip connection completes information fusion by splicing the underlying position information and deep semantic information.

In this paper, considering that U-Net belongs to a fully convolutional network and is limited by the local spatial information of convolution, we propose a novel hybrid architecture consisting of Transformer and convolution based on U-Net with attention gate, combining Transformer's global characteristics and long-distance dependence of the image, in which the attention gate acts on the skip connection and up-sampling. Fig.4 shows the overall architecture of the proposed TAGU-Net model.

The model's input is an MRI image, and the output is an AM mask image. In this model, we unified the size of the input MRI image and used the image resolution of 512 × 512 MRIs. All input images go through ConvEncoder and FormerEncoder, respectively, in the Encoder branch. After feature fusion, they are gradually down-sampled. After the final encoder, the size of the feature image has been reduced

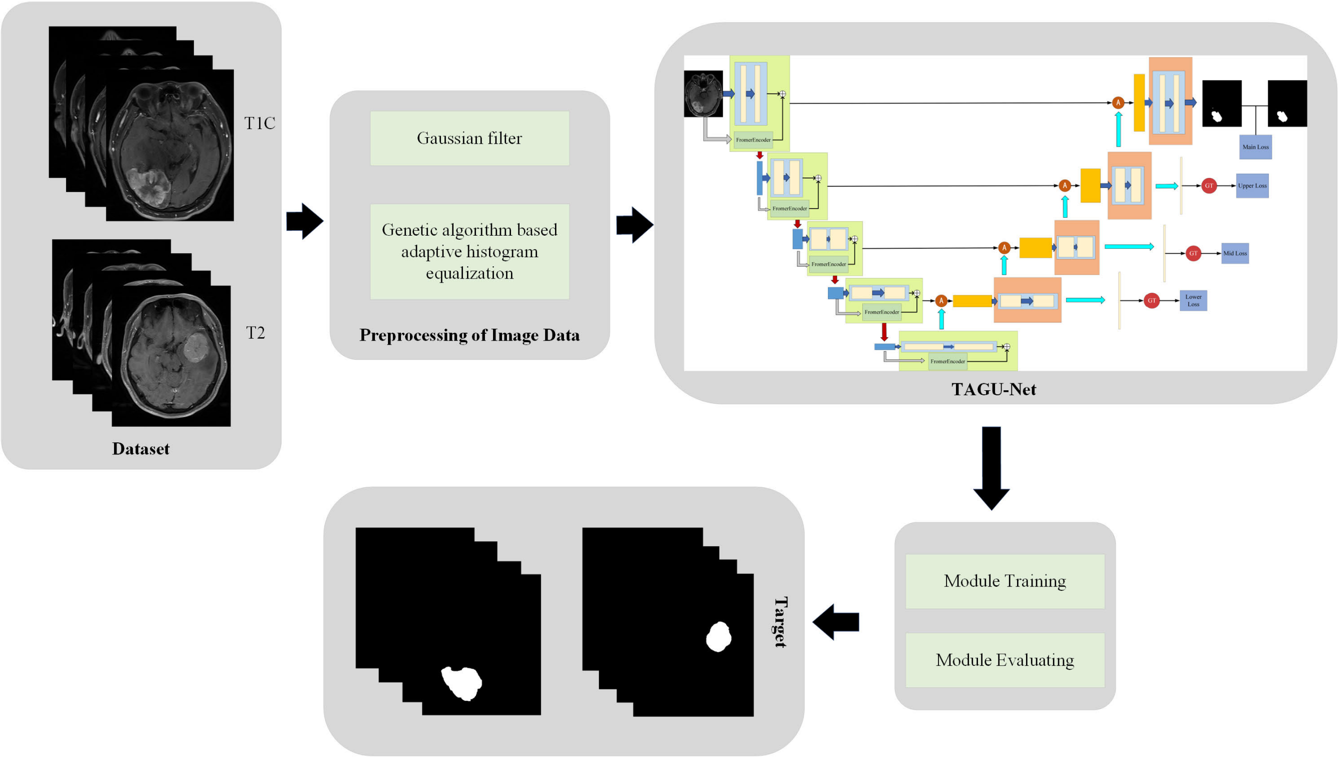

T1C
T2
Dataset
Gaussian filter
Genetic algorithm based adaptive histogram equalization
Preprocessing of Image Data
TAGU-Net
Module Training
Module Evaluating
Target

FIGURE 3. Block diagram for the proposed framework.

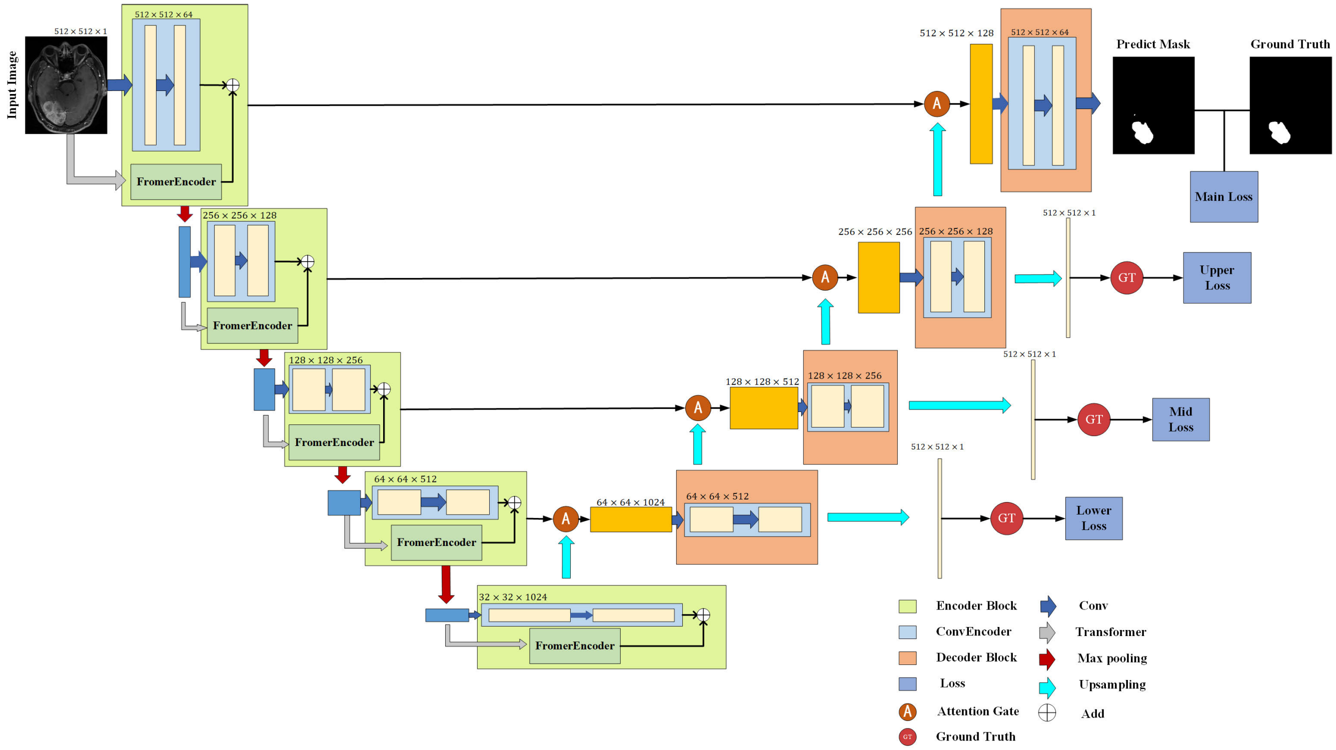


FIGURE 4. Structure diagram of TAGU-Net model, which mainly includes three modules: encoder branch, decoder branch, and deep supervison.

to the size of the original MRI image 1/16. The structure of ConvEncoder and FormerEncoder will be introduced later.

After passing the Encoder branch, the feature map enters the Decoder branch. The Decoder branch is mainly composed

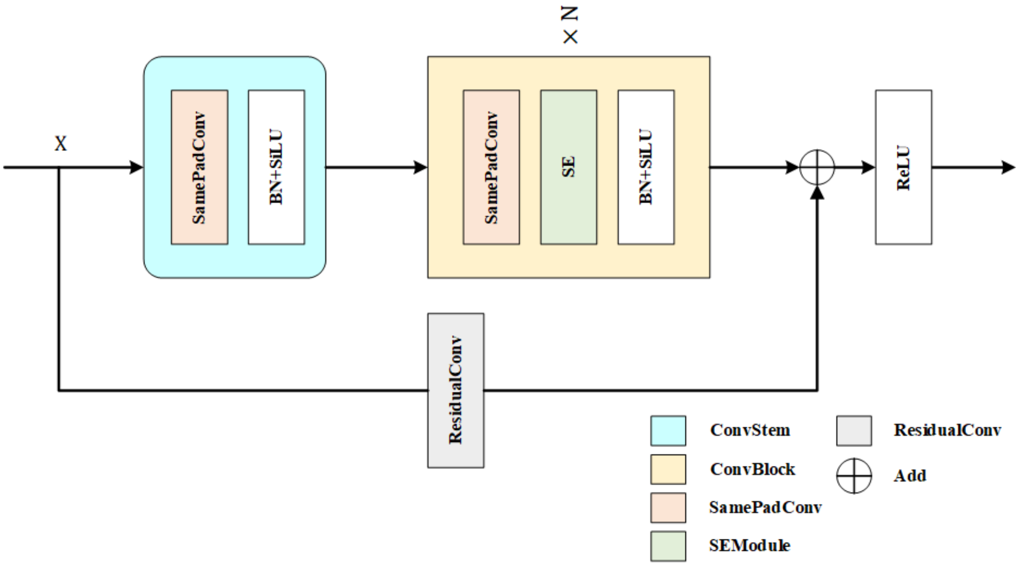


FIGURE 5. ConvEncoder structure diagram, which mainly includes three modules: convolution stem, convolution block, and residual convolution connection.

of ConvDecoder, attention gate, and deep supervision, and the ConvDecoder has the same structure as ConvEncoder. The attention Gate is primarily used for adaptive learning of skip connection information and up-sampling information, a weighted fusion of the two, highlighting important features and suppressing irrelevant features. The purpose of deep supervision is to train the network better and increase the training of the auxiliary loss supervision network.

2) ConvEncoder

ConvEncoder is used to extract the image's inductive bias and local feature information. The traditional ConvEncoder is full convolution, but the ConvEncoder in this paper is not full convolution. The main purpose is to learn the feature map weight better. We added the channel attention SE module [49]. After passing the SE module, ConvEncoder learned the correlation between channels and improved the weight of important channels in the subsequent feature fusion. ConvEncoder is the stack of convolution layers. The super parameter depth determines the number of convolution layers. To prevent ConvEncoder from following depth, the amount of calculation and parameters added needs to be more significant. We only add the SE module after the first layer of convolution. Meanwhile, because the size of MRI image at the time of input is 512 × 512 × 1, using the SE module has little effect. The SE module is not added in the first ConvEncoder. For the activation function in all volume layers $\sigma(x)$ using *SiLU* function, as shown below:

$$\sigma(x) = x * \frac{1}{1 + \mathrm{e}^{-\mathrm{x}}} \tag{2}$$

In addition, for each ConvEncoder, we have added the residual connection of the bottleneck structure to avoid vanishing gradient problem and network degradation. The equation is shown as follows:

$$X_0' = \sigma(BN(F(X, \{W_c\}))) \tag{3}$$

$$X_\ell' = \sigma\left(BN\left(SE\left(F\left(X_{\ell-1}', \{W_c\}\right)\right)\right)\right), \quad \ell = 1 \ldots N \tag{4}$$

$$X_N = X_N' + F(X, \{W_s\}) \tag{5}$$

where X is the input image $X \in R^{H \times W \times C}$, H, W is the resolution of the image, C is the number of channels, $F(*)$ indicates same-padding convolution operation, W_c is the weight of convolution, $BN(*)$ indicates batch normalization, $\sigma(*)$ indicates the activation function, $SE(*)$ indicates SE module, W_s is residual connection convolution weight, N represents the depth of ConvEncoder, Fig.5 shows the structure diagram of ConvEncoder.

3) FormerEncoder

Recently, Transformer has gradually become the primary means of natural language processing (NLP). At the same time, Transformer also shines brilliantly in computer vision (CV), and gradually becomes the basic component of a large number of CV. Transformer has also received much attention and research in medical image processing. Transformer's primary approach in CV is to split the input image into patches with different strategies, at the same time embedding the patch in high dimensions, and use the self-attention mechanism to model long-distance dependency. Transformers is immune to convolution imperfections. However, in the hybrid architecture based on Transformer and CNN, Transformer is usually used to model the feature map after CNN extracts features. It can be expected that such a method loses most of the image information, and the Transformer only models feature maps containing rich semantic information, and the representation of shallow features is missing. In this paper, we design a Transformer-based encoder

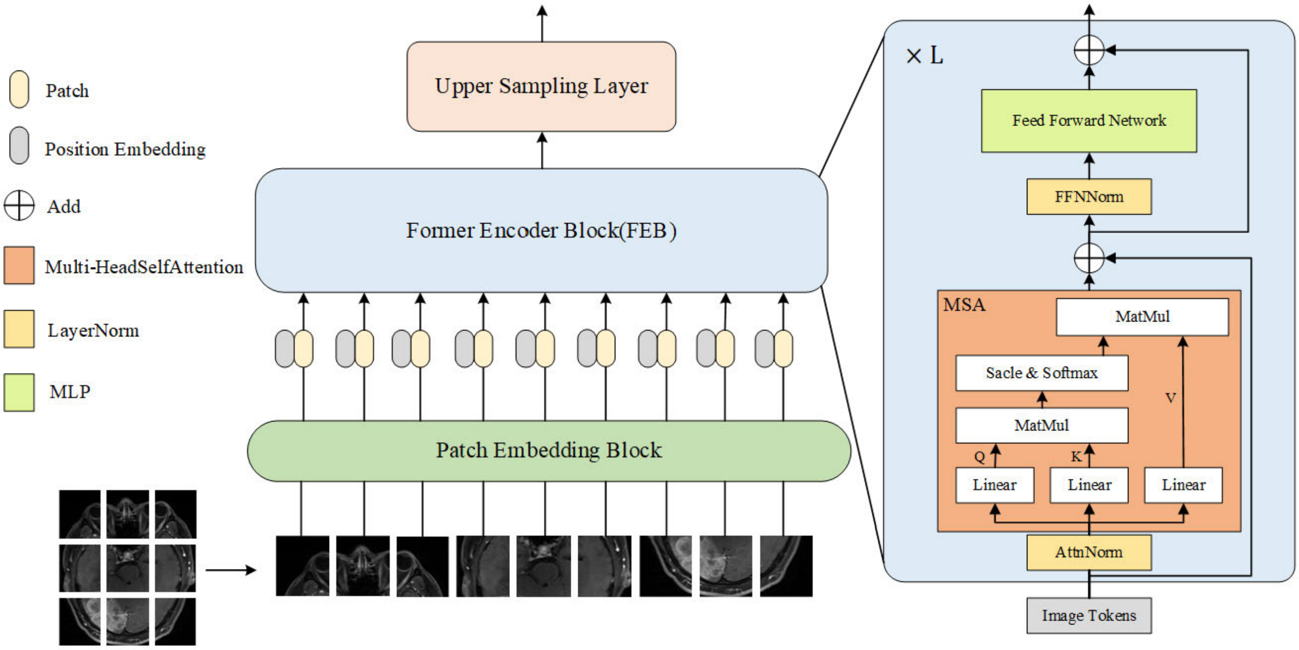


FIGURE 6. FormerEncoder structure diagram, which mainly includes three modules: patch embedding block, former encoder Block, and upper sampling layer.

called FormerEncoder. FormerEncoder will work with ConvEncoder to perform feature extraction on feature maps of different resolutions at different scales, combining different representations of convolution and Transformer and deep and shallow features will help the model perform better segmentation. Meanwhile, FormerEncoder can be flexibly replaced according to the task and data characteristics, such as Swin-Transformer [16], PVT [17] or T2T-ViT [18]. In this article, for the convenience of consideration, we only designed it based on the basic ViT.

FormerEncoder follows the classic ViT [14] architecture. FormerEncoder comprises three parts: Patch Embedding Block, Former Encoder Block (FEB), and Upper Sampling Layer (USL). Image in a FormerEncoder $X \in R^{H\times W\times C}$ enter the Patch Embedding Block and cut it into several non-overlapping patches $x_p \in R^{N\times P^2\times C}$, embedding the patch in high dimension, where P is the resolution of each patch, $N = HW/P^2$ is the number of patches generated. In FormerEncoder, the image is cut into patches from ordered spatial information to unordered sequence information. At this time, spatial position information is essential. To retain the spatial position information of the image, after all patches are embedded, we set a learnable position coding information position embedding $E_{pos} \in R^{N\times D}$ before the end of the Patch Embedding Block. Then, the embedded information and position embedding are fused and added, where D is the latent vector size set in the Patch Embedding Block. The equation definition of the Patch Embedding Block is as follows:

$$x_0 = \left[x_p^1 W_p; x_p^2 W_p; \cdots ; x_p^N W_p\right] + E_{\text{pos}}\ , \quad W_p \in R^{P^2\times C\times D} \tag{6}$$

FormerEncoder mainly includes FEB stack of, FEB it is divided into three parts: Multiheaded Self-Attention (MSA), Feed Forward Network (FFN) and LayerNorm (LN), MSA capture the long-distance dependency and global feature information of the image through self-attention, LN carry out normalization adjustment and finally pass FFN perform dimension transformation and mapping. The specific equation is defined as follows, where L is number of layers stacked for FEB, $d = D/H$ is the self-attention embedding dimension in the FEB and H is number of heads in MSA:

$$\begin{aligned} Q &= x_{\ell-1}W_q, K = x_{\ell-1}W_k, \\ V &= x_{\ell-1}W_v, W_q, W_k, W_v \in R^{D\times d} \end{aligned} \tag{7}$$

$$\text{Attention}(x_{\ell-1}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{8}$$

$$\begin{aligned} \text{MSA}(x_{\ell-1}) = \Big[&\text{Attention}^1(x_{\ell-1}); \text{Attention}^2(x_{\ell-1}); \\ &\cdots ; \text{Attention}^H(x_{\ell-1})\Big] \end{aligned} \tag{9}$$

$$x'_\ell = MSA(LN(x_{\ell-1})) + x_{\ell-1}, \quad \ell = 1 \ldots L \tag{10}$$

$$x_\ell = FFN\left(LN\left(x'_\ell\right)\right) + x'_\ell, \quad \ell = 1 \ldots L \tag{11}$$

$$y = USL\ (LN\ (x_L)) \tag{12}$$

The Upper Sampling Layer is the last component in FormerEncoder. Because the resolution of the image halves after Patch Embedding Block and FEB, the resolution does not decrease after ConvEncoder. To fuse the feature map obtained by the FormerEncoder with the feature map of the ConvEncoder, the resolution of the feature map needs to be restored. Fig.6 shows the structure diagram of the FormerEncoder.

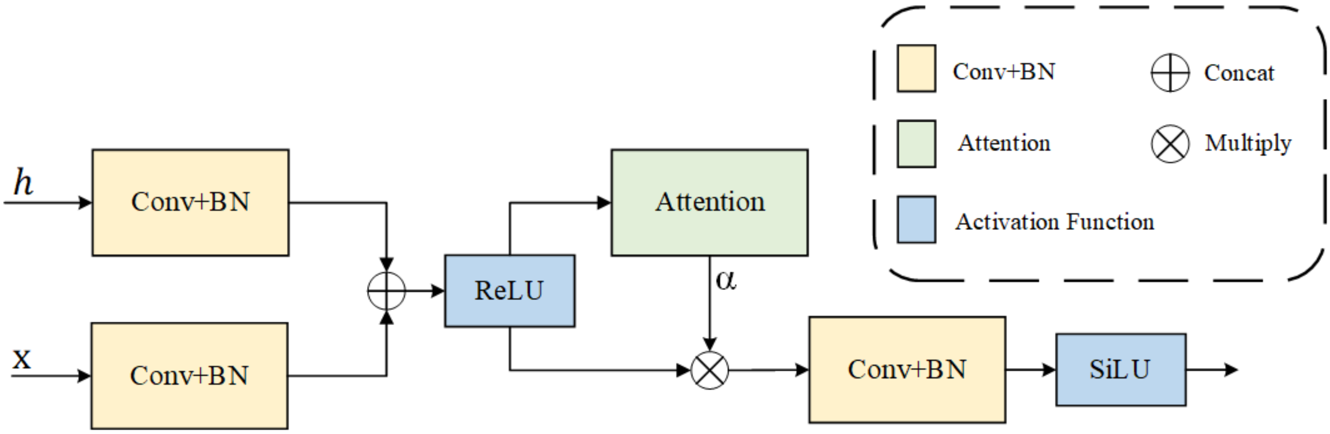


FIGURE 7. Schematic diagram of Attention gate.The feature map from encoder and decoder is processed in parallel, and the resulting gating signal controls the weight. The final feature map is obtained through the control of the gating signal.

4) ATTENTION GATE

The traditional U-Net structure only uses simple concatenation in skip-connection and up-sampling information fusion, and the more complex consideration is to use some nonlinear transformation to concatenate. However, these methods do not consider the correlation between skip-connection feature information and up-sampling feature information. In this paper, we propose an attention gate at this connection, which will consider both skip-connection feature information and up-sampling feature information. With this addition, the model can adaptively learn skip-connection feature information and up-sampling feature information and weigh the two. Highlight important features while suppressing irrelevant features. It can be seen from Fig.7 that the input of the attention gate is the skip connection feature information generated by the feature map of the encoder and the up-sampling feature information generated by the decoder of the upper layer. The skip connection feature information and the up-sampling feature information are operated in parallel, and finally, the concatenated fusion feature map is obtained. Some equations are defined as follows:

$$z' = ReLU([F(h, W_h); F(x_\ell, W_x)]), \ell = 1 \ldots H \quad (13)$$

$$\alpha = Sigmoid\left(BN\left(F\left(z', W_\alpha\right)\right)\right) \quad (14)$$

$$z_{\ell+1} = \sigma\left(BN\left(F\left(z' \cdot \alpha, W_x\right)\right)\right) \quad (15)$$

where h is the skip connection feature information generated by the feature map of the encoder, x_ℓ is up-sampling feature information generated by the upper decoder, H is the decoder depth, $F\,(*)$ indicates a convolution operation, α is the attention coefficient obtained.

5) DEEP SUPERVISION

Deep supervision [50] is one of the commonly employed to overcome the problems of vanishing gradients and slow convergence in neural networks. Its main idea is to add auxiliary classifiers to some hidden layers in the model as the network branch structure and supervise and train the backbone network. The most important thing about deep supervision is that it provides a method to judge the hidden layer feature map quality during the training process. In this study, we also use the deep supervision method to accelerate the convergence of the proposed network structure and supervise training. As seen in Fig.4, we added three groups of branch structures in the decoder branch. These three groups respectively perform depth supervising on feature maps of different resolutions and add auxiliary loss to calculate the corresponding loss of feature maps restored by the three groups of depth supervising during training, namely UpperLoss, MidLoss, and LowerLoss, which ultimately adds different weights to the main network loss MainLoss.

$$Loss = \alpha UpperLoss + \beta MidLoss + \gamma LowerLoss + \delta MainLoss \quad (16)$$

where α, β, γ, and δ is the weight coefficient corresponding to the loss, which determines the impact of the predicted loss on the whole loss at different scales, MainLoss will be given more weight.

6) LOSS FUNCTION

The most commonly used loss function in medical image segmentation is pixel-by-pixel cross entropy (CE). Image segmentation is the classification of each pixel. CE checks each pixel separately and makes the cross entropy of the predicted pixel value with ground truth one by one. The formula of CE is as follows:

$$\text{CELoss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i logp_i + (1-y_i)\, log(1-p_i)\right) \quad (17)$$

Among them, y_i is the real category of the input image pixels, p_i is the probability of prediction category 1, N is the number of all image pixels. Weighted cross entropy (WCE) improved CE by putting the weight before the loss of each corresponding class to alleviate the class imbalance. The formula is as follows:

$$\text{WCELoss} = -\frac{1}{N}\sum_{i=1}^{N}\left(wy_i logp_i + (1-y_i)\, log(1-p_i)\right) \tag{18}$$

Dice loss and IOU loss [51] is another function based on area loss, which aims to minimize the mismatch or maximize the overlapping area between the ground truth and the predicted segmentation. The formulas are as follows:

$$\text{DiceLoss} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} p_i} \tag{19}$$

$$\text{IOULoss} = 1 - \frac{\sum_{i=1}^{N} y_i p_i}{\sum_{i=1}^{N} (y_i + p_i - y_i p_i)} \tag{20}$$

In medical image segmentation, there are only one or two targets in an image, and the proportion of the target will be much smaller than the background. In essence, image segmentation is a classification problem, which causes the problem of class imbalance and severe imbalance between positive and negative sample scales. Focal loss [52] added a penalty item to solve this problem. Its basic idea is that the network will tend to predict only negative samples in the case of highly unbalanced categories. As a result, the prediction probability of negative samples p_i will be very high, and the return gradient is also huge. Adding $(1-p_i)^\gamma$ will reduce the loss of samples with high prediction probability, and increase the loss of models with low prediction probability, thus strengthening the attention to positive samples. The formula is defined as follows:

$$\text{FocalLoss} = -\frac{1}{N}\sum_{i=1}^{N}\Big(y_i\,(1-p_i)^\gamma\, logp_i + (1-y_i)\,p_i^\gamma\, log(1-p_i)\Big) \tag{21}$$

In this paper, we design a mixed loss function, which is the sum of Focal loss, Dice loss, and IOU loss. Its goal is to reduce the point-by-point cross-entropy of pixels through the maximum matching on the region. At the same time, because Focal loss is used, the problem of class inequality is solved to some extent, w_1, w_2, and w_3 is the weight coefficient of various losses, where *DiceLoss* will be given more weight.

$$\text{LossFunction} = w_1\text{FocalLoss}+w_2\text{DiceLoss}+w_3\text{IOULoss} \tag{22}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we use some evaluation metrics to evaluate the performance of our TAGU-Net model and the effectiveness of the experimental results. The model is mainly tested on the data set introduced in the second section, including training and testing. In the experiment process, we first discussed the comparison between the TAGU-Net model and the commonly SOTA segmentation model. In this section, we analyzed the performance of each model and the model we proposed. Then a group of ablation experiments is given to analyze the performance of some module designed of TAGU-Net proposed to confirm the superiority of our proposed methods in actual performance and excellent in AM segmentation.

A. IMPLEMENTATION DETAILS

The TAGU-Net model we proposed was implemented through Python 3.8.13 and Pytorch 1.12.1. All experiments were conducted in NVIDIA GTX 2080Ti GPU environment. In order to maximize the superiority of the proposed method and ensure the fairness of the experiment, all experiments use the same experimental settings and training strategies. The selection of some training configs and optimizers is as follows: the optimizer selects the Adam optimizer. The initial learning rate is set to 0.0001, $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $weight\_decay = 1e-5$ the learning rate adjustment strategy adopts cosine annealing, $T_{max} = 50$. Model parameters in 4 were updated in batches. The maximum number of epochs of training duration is set to 200. At the same time, we have normalized all image pixel values. The pixel value from $[0-255]$ adjusts to $[0-1]$ the image size is uniformly adjusted to 512×512. In terms of training strategy, to prevent over-fitting, we use the K-fold cross-validation training method, K is set to 5. Detailed hyperparameters see Table 1.

TABLE 1. Experimental configs and optimizer settings.

Training config	
weight init	Kaiming normal
optimizer	Adma
init learning rate	0.0001
optimizer momentum	$\beta 1 = 0.9, \beta 2 = 0.999$
weight decay	$1e-5$
batch size	4
patch size	16
training epochs	200
learning rate scheduler	Cosine Annealing
learning rate warm restart	50
K-fold	5

B. EVALUATION METRICS

To evaluate the performance of the proposed model, we adopted the following evaluation metrics commonly used for segmentation tasks: Dice score (Dice) and Jaccard index (Jac) are the two most essential segmentation indicators. The definition of metrics is as follows:

$$\text{Dice} = \frac{2 * TP}{2 * TP + FP + FN} \tag{23}$$

$$\text{Jac} = \frac{TP}{TP + FP + FN} \tag{24}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{25}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{26}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{27}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{28}$$

$$\text{Hausdorff Distance(HD)} = \max_{a \in A} \left\{ \min_{b \in B} \|a - b\| \right\} \tag{29}$$

The four essential items in the formula are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Hausdorff distance (HD) is a measure that describes the similarity between two sets of points. It defines the distance between two groups of points and is also commonly used for segmentation metrics. HD is sensitive to the boundary of segmentation, which is mainly used to measure the accuracy of boundary segmentation. In the experiment, we use 95% HD, the 95th percentile of HD. Compared with HD, this metric is slightly stable for small outliers.

C. PRIVATE DATASET EXPERIMENTAL RESULTS

Through private dataset experiments, we compared the performance of the proposed model with some SOTA models, and the results are shown in Table 2. It can be seen from the experimental results that the proposed TAGU-Net can obtain the highest Dice and Jac, which indicates that the TAGU-Net has higher performance than these SOTA models, and the prediction mask generated by TAGU-Net is highly consistent with the ground truth mask.

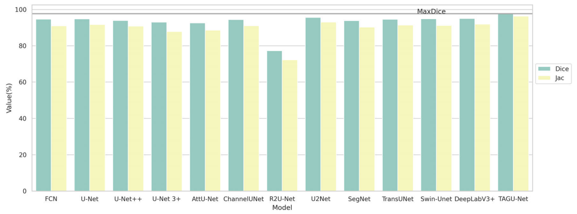


FIGURE 8. Comparison of dice and Jac histogram of each model. The right side is the Dice and Jac histogram of the proposed TAGU-Net.

FCN fused the characteristic images with different sampling coefficients through strip structure and full convolution and restored the resolution by the operation of up-pooling and transposed convolution, reaching the SOTA of pixel-level segmentation at that time; U-Net achieved better performance with a symmetric encoder-decoder structure and the skip connection between the encoding feature and the decoding feature; U-Net++ redesigns the skip connection based on U-Net so that the decoder can aggregate different scale features to achieve the effect of dense connection; U-Net 3+ proposed a full-scale skip connection, which combines low-level details from different scale feature maps with high-level semantics to maximize the use of full-scale feature maps and improve segmentation accuracy; AttU-Net introduces the attention mechanism into U-Net, and designs the attention gate in the skip connection. The soft-attention method gradually strengthens the weight of local ROI, inhibits the activation in unrelated regions, and reduces the redundant part of the skip. This method is similar but different from the attention gate proposed in this paper. The attention gate proposed in this paper aims to obtain the concatenate feature map of the skip connection and decoder features through the attention mechanism.The comparative results about the attention gate experiments are given in Table 4; ChannelUNet uses spatial channel-wise convolution, which can perform convolution operation along the feature map channel direction to extract the mapping relationship of spatial information between pixels, which is conducive to learning the mapping relationship between pixels in feature maps; R2U-Net applies recurrent neural network and residual network to U-Net, and designs recurrent residual layer to add features to better extract features; SegNet is a segmentation network based on FCN with encoder and decoder structure. In pooling operation, a Pooling Indices method is proposed to save pooled point source information; U2Net proposes a two-layer nested u-shaped structure, which turns the simple convolution structure in UNet into RUS (Residual U-blocks). RUS realizes the mixture of feature maps of different scales and different receptive fields through this u-shaped structure, which can capture more global information from different scales; TransUNet is an attempt to combine with Transformer. It uses the transformer's encoder structure on the encoder structure to enhance the representation of features, and the rest still follows the U-Net architecture; Swin-Unet is a pure transformer-based U-shaped architecture. The contextual features extracted based on Swin-transformer are upsampled by a decoder with a patch expanding layer, and the spatial resolution of the feature map is restored through skip connection and multi-scale feature fusion of the encoder, further segmentation prediction; DeepLabv3+ uses dilated convolution to solve the problem of the receptive field, and obtains multi-scale object information based on spatial pyramid pooling. Furthermore, it uses a fully-connected conditional random field to improve the ability of the model to capture structural information and solve the problem of fine segmentation.

To make a reliable comparison, we compared the results of these studies with our work. The TAGU-Net proposed by us has reached the highest level in important metrics, Dice, and Jac, surpassing other SOTA models. Fig.8 shows the performance of each model in the Dice score and Jac index. In terms of AM segmentation, the Dice of TAGU-Net is 97.67%, and the Jac is 96.35%. Except for TAGU-Net, the best performer is U2Net. Its Dice is 95.56%, and the Jac is 92.03%. In contrast, TAGU-Net absolute accuracy is 2.11% higher in Dice and 3.36% higher in Jac. In terms of relative accuracy, Dice is 2.21% higher and Jac is 4.69% higher. In the evaluation metrics of 95HD, DeepLabv3+ is 0.456, while TAGU-Net is 0.550, which lags behind DeepLabv3+ by a narrow margin and is also far higher than

TABLE 2. Comparison results between the proposed model and SOTA models. The bold black value indicates the highest score of all methods in the corresponding metric.

Method	Metrics						
	Dice (%)	Jac (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	95HD
FCN [37]	94.64	90.91	94.48	99.91	99.79	95.99	1.495
U-Net [12]	94.76	91.65	95.30	99.91	99.81	96.00	1.261
U-Net++ [53]	93.89	90.80	94.64	99.91	99.80	95.71	1.370
U-Net 3+ [54]	92.99	87.82	92.46	99.87	99.74	94.70	3.592
AttU-Net [55]	92.60	88.60	93.91	99.86	99.74	93.86	3.191
ChannelUNet [56]	93.26	90.64	94.21	99.89	99.83	94.55	1.561
R2U-Net [57]	77.26	72.24	83.94	99.65	99.31	83.97	25.79
U2Net [58]	95.56	92.03	95.43	99.93	99.85	95.94	0.668
SegNet [59]	93.81	90.25	94.85	99.89	99.78	94.90	1.821
TransUNet [15]	94.59	91.37	94.79	99.92	99.81	96.20	0.710
Swin-Unet [60]	94.85	91.18	95.26	99.91	99.81	95.49	0.805
DeepLabv3+ [13]	95.01	91.88	94.94	99.93	99.83	96.63	**0.456**
TAGU-Net	**97.67**	**96.35**	**97.76**	**99.96**	**99.92**	**98.51**	0.550

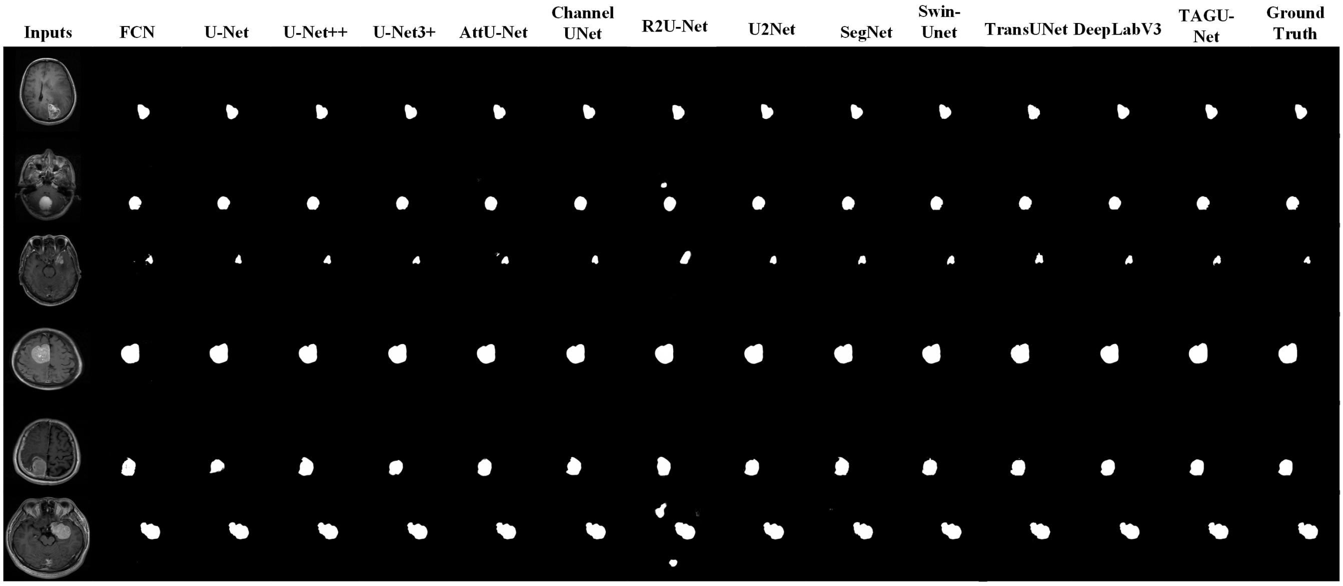


FIGURE 9. Comparison of prediction mask and ground truth of each model.

other models. In addition to the Dice and Jaccard metrics, Table 2 presents a comparison of the sensitivity, specificity, accuracy, and precision of our model with those of state-of-the-art models. Sensitivity refers to the ability of the method to detect tumors in MRI pixels, while specificity reports the ability to identify MRI pixels without tumors. Our proposed TAGU-Net model demonstrated a sensitivity value of 97.76% for atypical meningiomas, indicating its ability to accurately detect tumor-associated pixels in MRI. Similarly, the TAGU-Net model exhibited a specificity of 99.96% for atypical meningiomas, demonstrating a strong ability to distinguish tumor and non-tumor pixels. Finally, accuracy describes how well the model classified each pixel class (tumor/non-tumor class). Compared to state-of-the-art models, the proposed TAGU-Net model exhibits the highest pixel-wise recognition ability, achieving the highest values in various metrics. Our proposed model is generally superior to other models in AM segmentation.

At the same time, we use the above model to generate the predicted mask image and visually compare it with the ground truth. As shown in Fig.9, the first column on the left is the MRI image of the input model, the mask image

TABLE 3. Results of different modules ablation experiment. The bold black value indicates the highest score of all methods in the corresponding metric.

FormerEncoder	Attention Gate	Deep supervision	Dice (%)	Jac (%)
—	—	—	94.76	91.65
✔	—	—	96.58	94.78
—	✔	—	95.88	93.21
—	—	✔	95.21	92.33
✔	✔	—	97.12	95.69
—	✔	✔	95.74	93.14
✔	—	✔	96.82	94.61
✔	✔	✔	**97.67**	**96.35**

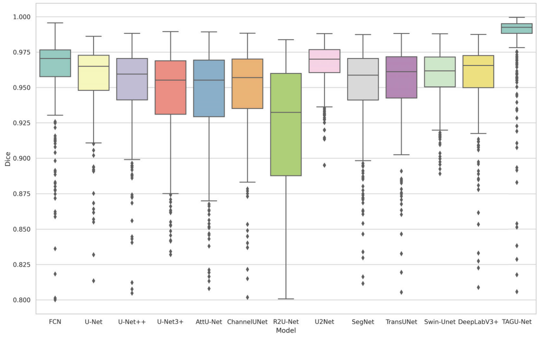


FIGURE 10. Box-plot of dice score of each model.

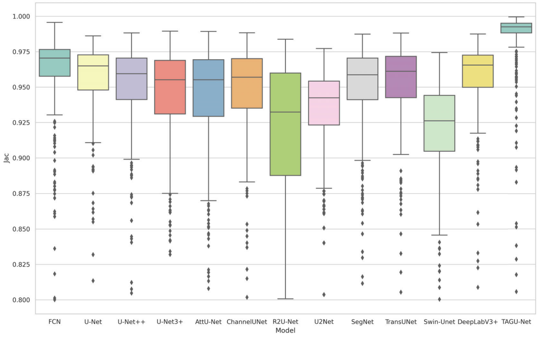


FIGURE 11. Box-plot of Jac index of each model.

generated by the model from left to right is the ground truth of the image, and the penultimate column is the mask image produced by the proposed model. It is evident from Fig.9 that the mask image generated by TAGU-Net is the closest to the ground truth. However, other models have different results in generating mask images due to their characteristics and generally have defects.

In addition, we have calculated the Dice distribution and Jac distribution of the output of each model, which is displayed in the form of a boxplot. Fig.10 and Fig.11 show the Dice and Jac boxplot of the proposed method and other SOTA models respectively. It can be seen from Fig.10 and Fig.11 that the box diagram of TAGU-Net is at the far right. Excluding outliers, the median, maximum and minimum values of the Dice and Jac of the proposed method are higher than those of other methods. It can be seen from the figure that the performance of the proposed TAGU-Net is much higher than that of other models.

D. ABLATION EXPERIMENT RESULTS

To evaluate the effectiveness of our proposed method, we have carried out many ablation experiments, mainly discussing the impact of FormerEncoder, Attention Gate, and Deep supervision on TAGU-Net. For the sake of simplicity, we only select the Dice and Jac, and the experimental results are shown in Table 3.

From the results of the ablation experiment, it can be seen that the FormerEncoder has the most significant impact, improving the performance by 1.89% in Dice, and 3.41% in Jac because it provides a global modeling capability for the model, and the input through the FormerEncoder is information from different scales. This multi-scale information will enable the model to obtain richer semantic information when fused. The role of the Attention Gate must be addressed. As can be seen from Table 3, the results obtained without the structure of the Attention Gate are generally lower, which also fully proves that the Attention Gate adaptively learns the skip connection feature information and up-sampling feature information in the decoder branch, effectively enhancing the activity of important information while suppressing the activity of irrelevant information. The impact of deep supervision is not particularly clear in Table 3. We find that the effects of deep supervision are not so fixed. In most cases, improving the model's performance is beneficial, and occasionally it does not work. Based on the experimental results and the mechanism of the deep supervision, we can speculate that in most cases, the deep supervision increases the loss during the model's training to prevent vanishing gradient, which makes the model better optimized and improve the performance of the model. In a small part of the time, deep supervision can only play a role if the model has converged or the model fitting ability is limited. The experimental results also show that the best results can be achieved simultaneously using FormerEncoder, Attention Gate, and Deep supervision, which

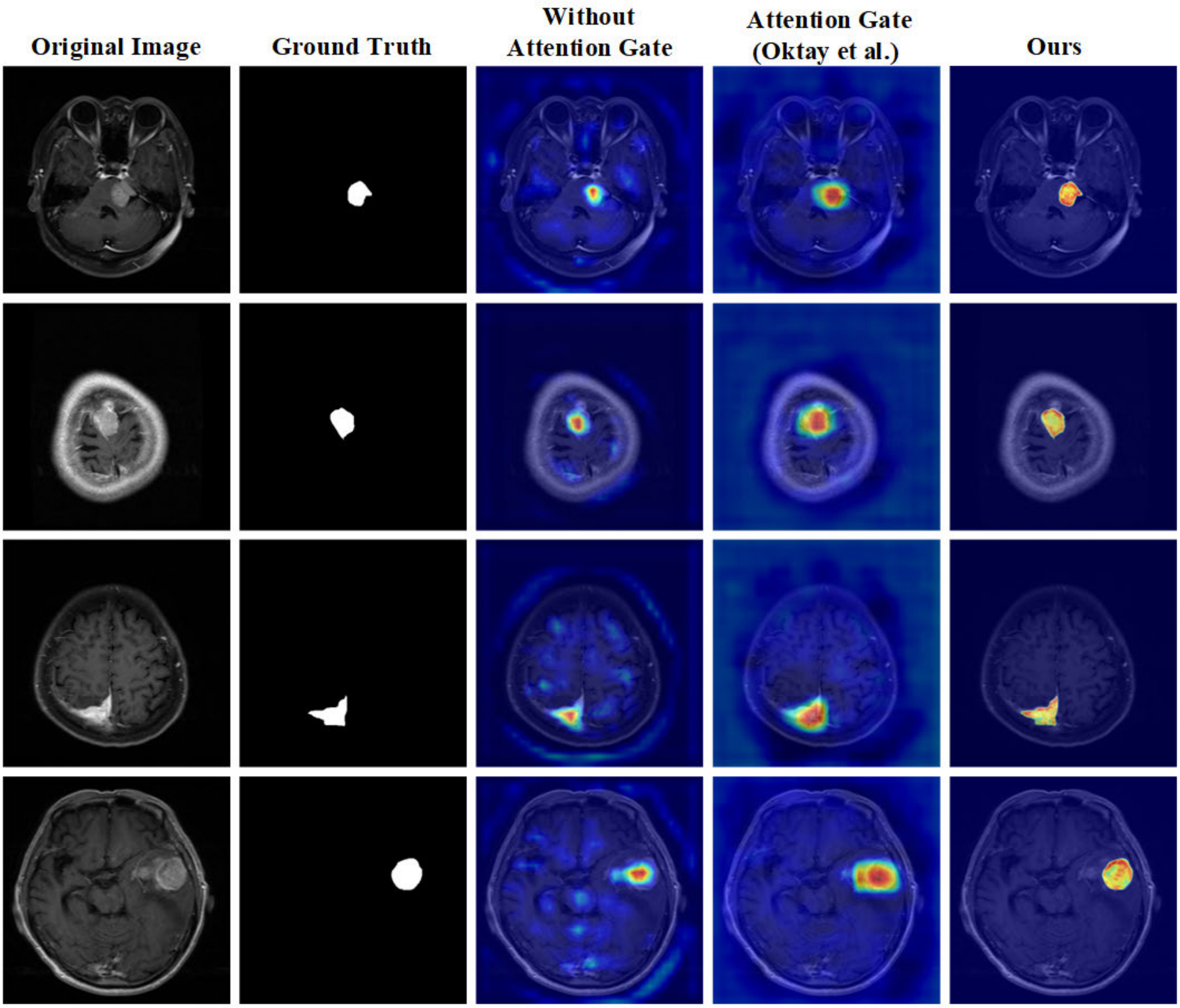


FIGURE 12. The first column represents the original image, and the second column shows the ground truth. The other columns are heatmaps about the corresponding methods. The deeper the red, the more attention the pixel gets.

TABLE 4. Comparison results between the proposed model and other attention gate modules. The bold black value indicates the highest score of all methods in the corresponding metric.

Module	Dice (%)	Jac (%)
Attention gate [55]	96.84	93.89
Ours	**97.67**	**96.35**

improves the performance by 3.07% in Dice, and 5.12% in Jac compared with backbone. Overall, each module is indispensable for achieving better performance.

In our experiments, we also perform an experimental comparison of loss functions to verify the effectiveness of the proposed hybrid loss function. As shown in Table 5, compared with WCELoss, our proposed hybrid loss improves 1.61% in Dice and 2.91% in Jac, achieving the best results, which verifies that our loss can facilitate model optimization. In Table 4, we compare the attention gate proposed by Oktay et al. with

TABLE 5. Results of different loss function ablation experiment. The bold black value indicates the highest score of all methods in the corresponding metric.

Loss function	Dice (%)	Jac (%)
WCELoss	96.12	93.62
DiceLoss	96.57	94.13
IOULoss	96.33	94.25
FoaclLoss	96.69	94.38
Ours	**97.67**	**96.35**

TABLE 6. Comparison BraTs 2018 dataset results between the proposed model and SOTA models. The bold black value indicates the highest score of all methods in the corresponding metric.

Method	Metrics						
	Dice (%)	Jac (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	95HD
FCN [37]	92.20	88.15	93.34	99.72	99.47	92.46	2.971
U-Net [12]	90.17	85.18	91.03	99.66	99.38	90.39	2.766
U-Net++ [53]	93.65	90.59	93.92	99.82	99.64	94.30	1.429
U-Net 3+ [54]	90.66	85.98	91.23	99.69	99.39	91.34	3.257
AttU-Net [55]	93.25	90.35	93.68	99.77	99.65	93.97	1.341
ChannelUNet [56]	92.60	88.60	93.91	99.86	99.74	93.86	3.191
U2Net [58]	90.52	85.13	92.75	99.62	99.42	89.46	2.165
SegNet [59]	92.55	88.26	93.27	99.76	99.56	92.84	1.431
TransUNet [15]	90.18	85.03	91.55	99.60	99.37	90.02	2.545
Swin-Unet [60]	93.64	90.32	93.13	99.82	99.74	94.17	1.720
DeepLabv3+ [13]	91.15	90.11	92.54	99.68	99.33	90.25	2.185
TAGU-Net	**97.62**	**95.35**	**97.84**	**99.91**	**99.83**	**97.40**	**0.624**

the attention gate of this paper. The results show that the proposed attention gate has a better effect. In Fig.12, the comparison of the heatmap generated by the Grad-CAM [61] based on different methods is shown. It can be clearly seen that the region of interest of the proposed method tends to coincide with the ground truth. It is worth emphasizing that the heatmap generated by TAGU-Net pays more attention to the meningioma boundary region, achieving better accurate segmentation.

E. BraTs 2018

The proposed architecture was compared with other state-of-the-art models used for semantic segmentation on the BraTs 2018 dataset as shown in Table 6. The metrics of the TAGU-Net model, namely the Dice and Jac, demonstrate that our proposed model surpasses all other state-of-the-art models in LGG segmentation. Specifically, the Dice score is 97.62% and the Jaccard index is 95.35%, indicating superior performance compared to other models. Similar to the experimental results presented in Table 2, TAGU-Net performs equally well on the public benchmark dataset BraTs 2018. It is evident that our model generally outperforms other models in terms of low-grade glioma segmentation. Therefore, the proposed framework has been demonstrated to be capable of accurately distinguishing tumor tissue from other brain tissues (normal and pathological) while precisely following tumor tissue boundaries.

V. CONCLUSION

In the task of atypical meningioma segmentation, the shape and size of atypical meningioma are irregular, and the boundary is not apparent, especially in MRI images with a lot of noise. Therefore, how to accurately segment atypical meningioma accurately is very significant,arduous and challenging. In this study, we used GAAHE to improve the quality of MRI images. At the same time, experimental verification is carried out under the TAGU-Net framework we propose. The proposed TAGU-Net is a hybrid architecture of convolution and transformer. It combines ConvEncoder and FormerEncoder in the encoder branch and introduces Attention Gate in the decoder branch. ConvEncoder and FormerEncoder extract different information from feature maps of different resolutions at different scales, effectively reducing the small drawbacks of the limited receptive field in convolution, while aggregating information from different encoders at various scales. At the same time, FormerEncoder can well capture global features with its unique properties, and the long-distance dependency of the image is modeled to retain fine details. Moreover, it is flexible and replaceable based on different tasks and data characteristics. Furthermore, the Attention Gate adaptively learns the skip connection information and up-sampling information at the decoder stage, highlights the essential features, and suppresses the irrelevant features when fusing the two features. In addition, we have built three sets of losses and one main loss at different scales through the in-depth monitoring technology to help the model learn and train better. TAGU-Net can effectively extract features from MRI images and fuse features of different scales, and achieve accurate segmentation of atypical meningiomas through these proposed modules. We have conducted rigorous experimentation on both a private atypical meningioma dataset and the publicly available BraTs 2018 benchmark dataset. Our proposed methodology has been found to achieve state-of-the-art atypical meningioma segmentation. In comparison to other models, our model has exhibited superior segmentation results, boasting higher levels of accuracy and precision with Dice and Jaccard coefficients of 97.67% and 96.35%, respectively, in the private dataset, and 97.62% and 95.35%, respectively, in the BraTs 2018 dataset.

ACKNOWLEDGMENT

The authors are grateful to the anonymous referee, who made valuable suggestions to help improve the article. Special thanks to Weihai Municipal Hospital for providing the MRI dataset.

REFERENCES

[1] Q. T. Ostrom, H. Gittleman, P. Farah, A. Ondracek, Y. Chen, Y. Wolinsky, N. E. Stroup, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006–2010," *Neuro-Oncol.*, vol. 15, no. 2, pp. 1–56, Nov. 2013.

[2] A. Kunimatsu, N. Kunimatsu, K. Kamiya, M. Katsura, H. Mori, and K. Ohtomo, "Variants of meningiomas: A review of imaging findings and clinical features," *Japanese J. Radiol.*, vol. 34, no. 7, pp. 459–469, Jul. 2016.

[3] D. N. Louis, A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, "The 2016 World Health Organization classification of tumors of the central nervous system: A summary," *Acta Neuropathol.*, vol. 131, no. 6, pp. 803–820, Jun. 2016.

[4] A. Perry, S. L. Stafford, B. W. Scheithauer, V. J. Suman, and C. M. Lohse, "Meningioma grading: An analysis of histologic parameters," *Amer. J. Surgical Pathol.*, vol. 21, no. 12, pp. 1455–1465, Dec. 1997.

[5] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, and P. Kleihues, "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*, vol. 114, no. 2, pp. 97–109, Aug. 2016.

[6] M. Cho, J.-D. Joo, I. A. Kim, J. H. Han, C. W. Oh, and C.-Y. Kim, "The role of adjuvant treatment in patients with high-grade meningioma," *J. Korean Neurosurgical Soc.*, vol. 60, no. 5, pp. 527–533, Sep. 2017.

[7] M. Preusser, P. K. Brastianos, and C. Mawrin, "Advances in meningioma genetics: Novel therapeutic opportunities," *Nature Rev. Neurol.*, vol. 14, no. 2, pp. 106–115, Feb. 2018.

[8] M. J. Riemenschneider, A. Perry, and G. Reifenberger, "Histological classification and molecular genetics of meningiomas," *Lancet Neurol.*, vol. 5, no. 12, pp. 1045–1054, Dec. 2006.

[9] S. Schob, C. Frydrychowicz, M. Gawlitza, L. Bure, M. Preuß, K.-T. Hoffmann, and A. Surov, "Signal intensities in preoperative MRI do not reflect proliferative activity in meningioma," *Transl. Oncol.*, vol. 9, no. 4, pp. 274–279, Aug. 2016.

[10] M. Nowosielski, N. Galldiks, S. Iglseder, P. Kickingereder, A. von Deimling, M. Bendszus, W. Wick, and F. Sahm, "Diagnostic challenges in meningioma," *Neuro-Oncol.*, vol. 19, no. 12, pp. 1588–1598, Nov. 2017.

[11] K. Violaris, V. Katsarides, M. Karakyriou, and P. Sakellariou, "Surgical outcome of treating grades II and III meningiomas: A report of 32 cases," *Neurosci. J.*, vol. 2013, pp. 1–4, Nov. 2013.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE ICCV*, Oct. 2021, pp. 10012–10022.

[17] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 568–578.

[18] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 558–567.

[19] G. Minchev, G. Kronreif, W. Ptacek, C. Dorfer, A. Micko, S. Maschke, F. G. Legnani, G. Widhalm, E. Knosp, and S. Wolfsberger, "A novel robot-guided minimally invasive technique for brain tumor biopsies," *J. Neurosurg.*, vol. 132, no. 1, pp. 150–158, Jan. 2020.

[20] S. R. Kebede, T. G. Debelee, F. Schwenker, and D. Yohannes, "Classifier based breast cancer segmentation," *J. Biomimetics, Biomater. Biomed. Eng.*, vol. 47, pp. 41–61, Nov. 2020.

[21] X. Yu, Q. Zhou, S. Wang, and Y. Zhang, "A systematic survey of deep learning in breast cancer," *Int. J. Intell. Syst.*, vol. 37, no. 1, pp. 152–216, Jan. 2022.

[22] A. Tiwari, S. Srivastava, and M. Pant, "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019," *Pattern Recognit. Lett.*, vol. 131, pp. 244–260, Mar. 2020.

[23] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for brain tumor segmentation," in *Proc. 6th Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2021, pp. 118–132.

[24] M. A. Almahfud, R. Setyawan, C. A. Sari, D. R. I. M. Setiadi, and E. H. Rachmawanto, "An effective MRI brain image segmentation using joint clustering (K-means and fuzzy C-means)," in *Proc. Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRITI)*, Nov. 2018, pp. 11–16.

[25] C. C. Benson, V. Deepa, V. L. Lajish, and K. Rajamani, "Brain tumor segmentation from MR brain images using improved fuzzy c-means clustering and watershed algorithm," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 187–192.

[26] C. Saha and M. F. Hossain, "MRI brain tumor images classification using K-means clustering, NSCT and SVM," in *Proc. 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr., Comput. Electron. (UPCON)*, Oct. 2017, pp. 329–333.

[27] J. Amin, M. Sharif, M. Raza, and M. Yasmin, "Detection of brain tumor based on features fusion and machine learning," *J. Ambient Intell. Humanized Comput.*, vol. 10, pp. 1–17, Nov. 2018.

[28] H. Al-Dmour and A. Al-Ani, "A clustering fusion technique for MR brain tissue segmentation," *Neurocomputing*, vol. 275, pp. 546–559, Jan. 2018.

[29] I. E. Kaya, A. A. Pehlivanl, E. G. Sekizkardeş, and T. Ibrikci, "PCA based clustering for brain tumor segmentation of T1w MRI images," *Comput. Methods Programs Biomed.*, vol. 140, pp. 19–28, Mar. 2017.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[31] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2016.

[32] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[33] F. J. Díaz-Pernas, M. Martínez-Zarzuela, M. Antón-Rodríguez, and D. González-Ortega, "A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network," *Healthcare*, vol. 9, no. 2, p. 153, 2021.

[34] E. U. Haq, H. Jianjun, X. Huarong, K. Li, and L. Weng, "A hybrid approach based on deep CNN and machine learning classifiers for the tumor segmentation and classification in brain MRI," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–18, Aug. 2022.

[35] Y. Ding, F. Chen, Y. Zhao, Z. Wu, C. Zhang, and D. Wu, "A stacked multi-connection simple reducing net for brain tumor segmentation," *IEEE Access*, vol. 7, pp. 104011–104024, 2019.

[36] D. Maji, P. Sigedar, and M. Singh, "Attention Res-UNet with guided decoder for semantic segmentation of brain tumors," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103077.

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–14.

[42] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 574–584.

[43] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. 24th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 109–119.

[44] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[45] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic Features," *Sci. Data*, vol. 4, no. 1, pp. 1–13, 2017.

[46] S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.

[47] U. K. Acharya and S. Kumar, "Genetic algorithm based adaptive histogram equalization (GAAHE) technique for medical image enhancement," *Optik*, vol. 230, Jan. 2021, Art. no. 166273.

[48] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI signal Process. Syst. Signal, Image Video Technol.*, vol. 38, no. 1, pp. 35–44, 2004.

[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[50] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[51] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2016, pp. 234–244.

[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[53] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[54] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[55] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[56] Y. Chen, K. Wang, X. Liao, Y. Qian, and P. A. Heng, "Channel-UNet: A spatial channel-wise convolutional neural network for liver and tumors segmentation," *Frontiers Genet.*, vol. 10, p. 1110, Nov. 2019.

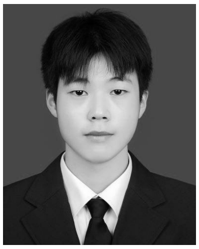
[57] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[58] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U^2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Jan. 2020, Art. no. 107404.

[59] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Oct. 2016.

[60] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2023, pp. 205–218.

[61] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.



HONG HUANG was born in Kaifeng, Henan, China, in 2002. He is currently pursuing the B.S. degree in computer science with Shandong University, China. His research interests include deep learning and computer vision.

PANPAN LIU received the M.S. degree from Zhejiang University. He is currently pursuing the D.S. degree with Tiantan Hospital. His current research interests include the diagnosis of meningiomas, gliomas, and pituitary tumors.

JIE LIU received the M.S. degree from Shandong University, Weihai. He is currently a Research Associate with the Department of Computer Science, Shandong University. His current research interests include astronomical data mining and machine learning algorithms.