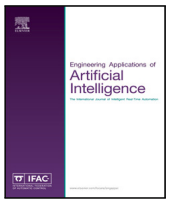
Engineering Applications of Artificial Intelligence 171 (2026) 114299

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Research paper

# A lightweight brain tumor segmentation network based on cross-modality feature fusion

Yawen Fan *, Chenziyi Huang, Xiang Wang, Chaoyuan Wang, Quan Zhou, Jianxin Chen

*National Engineering Research Center of Communications and Networking, Nanjing University of Posts & Telecommunications, Nanjing, 210003, China*

ARTICLE INFO

*Keywords:*
Brain tumor segmentation
Multi-modality
Cross-modality feature fusion
Attention mechanism
Lightweight

ABSTRACT

In clinical diagnosis and treatment, the segmentation of brain tumors using multi-modality Magnetic Resonance Imaging (MRI) is crucial. Effectively leveraging information from different modalities is challenging, as they have varying sensitivities to tumor regions. To address this, we propose a four-branch encoder structure that incorporates distinct attention mechanisms for each modality, enabling the learning of discriminative features. Furthermore, we introduce a novel lightweight Cross-Modal Feature Fusion (CMFF) module to enhance feature representation. Additionally, we reduce the number of convolutional layers to prevent overfitting. Experimental results on the Brain Tumor Segmentation (BraTS) 2021 Challenge demonstrate that our framework achieves superior segmentation performance with a significantly reduced parameter count. Specifically, our method obtains average Dice scores of 93.12% for whole tumor, 89.50% for tumor core, and 85.92% for enhancing tumor, using only 3.66 M parameters. These results highlight the model's strong performance and efficiency compared to existing baseline and state-of-the-art methods.

## 1. Introduction

Brain tumors are a life-threatening medical condition caused by the cancerous growth of cells in the brain. Therefore, accurate brain tumor segmentation (BTS) is crucial for improving brain tumor diagnosis and prognosis evaluation. This process involves recognizing and segmenting tumor regions from healthy tissues using medical imaging technologies (Biratu et al., 2021; Litjens et al., 2017; Liu et al., 2023; Ranjbarzadeh et al., 2023; Zhu et al., 2023). Nowadays, magnetic resonance imaging (MRI) (Di Ieva et al., 2021; Mohammed et al., 2023) has become a routine examination method due to its high resolution, strong soft tissue contrast, and noninvasive nature. MRI typically includes four modalities: T1, T1C, T2, and T2Flair, as shown in Fig. 1. The T1 image is useful for observing anatomical structures, although it may not clearly display lesions; T1C involves injecting contrast agents into the bloodstream before MRI, making areas of active blood flow more apparent in the imaging, which is important for enhancing tumor detection; The T2 sequence displays lesions, allowing for the judgment of the entire tumor, while T2Flair, a fluid-attenuated inversion recovery (FLAIR) sequence, is brighter with larger water content and can be utilized to identify peritumoral edema areas. When evaluating brain cancers, radiologists typically combine data from all four modalities, with the T1C sequence often providing a higher diagnostic yield for brain tumor cores. Segmenting brain tumors could be greatly aided by these clinical insights.

Compared with common image semantic segmentation, the MRI-based tumor segmentation task faces three main challenges. Firstly, medical image datasets are often limited in size, which poses a challenge for effectively training deep neural networks. Secondly, MRI is not only multi-modal but also 3D volumetric, requiring consideration of the correlation among different modalities and 3D spatial information. Finally, due to the diverse shape and size variations of brain tumors, accurately localizing and segmenting them is a complex and difficult task.

In recent years, numerous deep neural networks have been applied to brain tumor segmentation, demonstrating remarkable success (Jyothi and Singh, 2023; Rehman et al., 2023; Allah et al., 2023; Chukwujindu et al., 2024; Bougourzi and Hadid, 2025). In particular, fully convolutional neural(FCN) (Zhao et al., 2018) networks have attracted much attention due to their ability to achieve pixel-level semantic segmentation (Long et al., 2015). Based on FCN, Ronneberger et al. (2015) proposed U-Net, which is particularly suitable for medical image segmentation tasks with low data requirements and has become a mainstream algorithm in the field of brain tumor segmentation (Zhu et al., 2023). However, these 2D models do not leverage 3D spatial information. Therefore, 3D fully convolutional neural networks, such as 3D UNet (Ahmad et al., 2021) and nnU-Net (Isensee et al., 2021) have gained popularity in volumetric brain tumor segmentation due

* Correspondence to: No.66, new model road, Nanjing, zip code: 210003, China.
*E-mail address:* ywfan@njupt.edu.cn (Y. Fan).

https://doi.org/10.1016/j.engappai.2026.114299
Received 31 August 2024; Received in revised form 23 July 2025; Accepted 22 February 2026

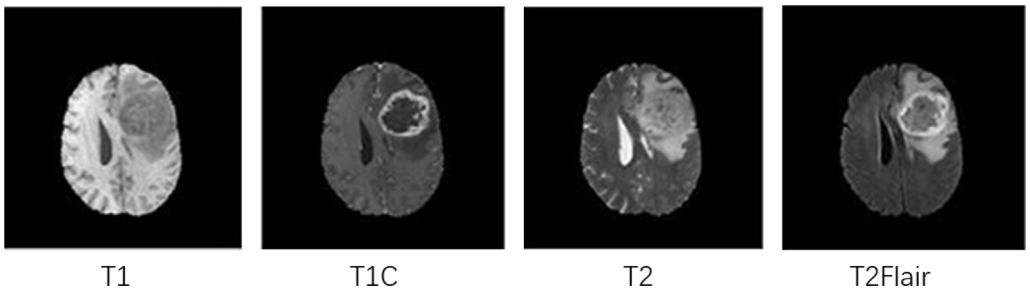



**Fig. 1.** Samples of four modalities.

to their ability to automatically learn high-dimensional feature representations from volumetric MRI (Dolz et al., 2017). Additionally, methods based on the Vision Transformer(ViT) architecture, known for its global modeling capability through multi-head self-attention mechanisms, have also been applied to brain tumor segmentation, including TransUnet (Chen et al., 2021; Zhu et al., 2024a), TransBTS (Wenxuan et al., 2021), Transsea (Liu et al., 2024) and GH-UNet (Wang et al., 2025). However, although various optimizations have been proposed, they still have high complexity and computational costs, as well as data requirements.

From the perspective of multi-modality fusion, most existing methods directly concatenate modalities together and feed them into models without considering the significant differences among them (Ahmad et al., 2021; Isensee et al., 2019; Wenxuan et al., 2021; Fang and Wang, 2022; Ali et al., 2022). Unlike these early fusion strategies, feature fusion methods utilize many convolutional branches corresponding to modalities or groups, and then fuse the learned features using different strategies (Syazwany et al., 2021; Tseng et al., 2017; Zhao et al., 2022; Menze et al., 2010). For example, Ding et al. (2021) proposed a Region-aware Fusion network to fuse available modalities features. Zhuang et al. (2022) designed a cross-modality feature interaction module to fuse multi-modal features in the encoding stage. However, the methods have complex architectures and are not easily adaptable to specific modality requirements, particularly due to difficulties in distinguishing multimodal feature representations.

To overcome the aforementioned limitations, this research aims to propose a novel 3D lightweight network based on cross-modality feature fusion to segment brain tumors from multi-modality MRI. The U-shaped structure with a multi-branch encoder is adopted to learn modal-specific features, where different attention mechanisms are used according to the modal characteristics. To effectively and efficiently fuse multi-modality features, we design a cross-modality feature fusion block. Overall, the main contributions of this study are three-fold:

- We propose a lightweight 3D multi-modality brain tumor segmentation model by considering the distinctiveness of different modalities and their correlations. Our model adopts a four-branch encoder structure, allowing the incorporation of modal-specific attention mechanisms based on clinical knowledge. This design ensures adaptability to handle cases where modalities are missing or unavailable.
- A novel cross-modality feature fusion block is designed based on lightweight attention mechanisms. Specifically, features from different modalities are fused at each layer and then feed to decoder.
- We conduct experiments on the BraTS2021 benchmark. The experimental results demonstrate that the proposed framework achieves superior or comparable accuracy compared to the state-of-the-art while maintaining fewer parameters.

The rest of this paper is organized as follows. In Section 2, we review research related to the proposed method. In Section 3, we provide an overview of the proposed network structure and loss function. Section 4 presents implementation details, evaluation metrics, experimental results, and visualization analysis. In Section 5, we conclude the paper by summarizing the main contributions, discussing its limitations, and outlining directions for future work.

## 2. Related work

In this section, we briefly review existing methods from three key aspects: segmentation model architectures, cross-modality fusion methods, and attention mechanisms.

### 2.1. Brain tumor segmentation network

Brain tumor segmentation has benefited greatly from recent advances in deep learning, especially through CNNs, hybrid models, and efficient sequence modeling frameworks. This section reviews representative approaches from these three angles, with a focus on their structural designs and practical strengths (see Table 1).

Convolutional neural networks (CNNs) have become the backbone of medical image segmentation, particularly in brain tumor analysis using 3D MRI scans. By leveraging the volumetric nature of MRI data, 3D CNNs effectively preserve spatial continuity and anatomical structure across slices. Urban et al. (2014) pioneered a multi-modal brain tumor segmentation model by replacing traditional 2D convolutional kernels with 3D kernels, enabling richer contextual learning from full volumetric input. Building on this foundation, numerous models have been proposed to enhance the expressive power and generalization ability of 3D CNNs. Notable examples include Attention U-Net (Oktay et al., 2018), which incorporates spatial attention to emphasize lesion regions; UNet++ (Zhou et al., 2018), which enhances feature fusion through nested and dense skip connections; and nnU-Net (Isensee et al., 2021), an automated pipeline that adapts architectures and training strategies to specific datasets. Despite these advancements, conventional CNNs are inherently limited in capturing long-range dependencies, which constrains their effectiveness in modeling complex anatomical structures.

Therefore, in medical image segmentation, several Transformer-based architectures have been proposed to enhance global contextual understanding. TransBTS (Wenxuan et al., 2021) combines CNN-based encoders with Transformer bottlenecks, enabling effective integration of local and global features in 3D medical images. TransUNet (Chen et al., 2021) embeds Transformer layers within a CNN backbone to enrich semantic representation, while Swin-Unet (Hatamizadeh et al., 2021) adopts hierarchical window-based attention for efficient context modeling. Despite their promising performance, Transformer-based models often suffer from high computational and memory costs, limiting their scalability to high-resolution volumetric data and real-time clinical applications.

Recently, the Mamba model (Gu and Dao, 2023) introduces a new class of efficient sequence models with linear-time complexity and low memory usage. Its medical extensions, such as U-mamba (Ma et al., 2024) and LightM-UNet (Liao et al., 2024), adapt this framework for 3D

**Table 1**
Comparison of representative models based on different architectural paradigms.

| Type | Networks | Advantage | Limitation |
|---|---|---|---|
| CNN based | UNet3D (Ahmad et al., 2021), nnU-Net (Isensee et al., 2021), UNet++(Zhou et al., 2018), Attention U-Net (Oktay et al., 2018) | Strong locality modeling, low computational cost, and ease of training. | Limited global context modeling; reduced performance in complex multi-modal tasks. |
| Transformer based | TransBTS (Wenxuan et al., 2021), TransUNet (Chen et al., 2021), Swin-UNETR (Hatamizadeh et al., 2021) | Better global dependency modeling, enhanced performance on complex anatomical structures. | High computational cost, large parameter size, and slow convergence. |
| Mamba based | U-mamba (Ma et al., 2024), LightM-Unet (Liao et al., 2024) | Efficient long sequence modeling with lower latency than transformers. | Still in early stages in medical imaging; requires further validation and tuning. |

segmentation by embedding Mamba modules into U-Net-style architectures. These models aim to capture global context more efficiently than traditional transformers. However, challenges remain due to the limited size of medical datasets and the computational demands of processing high-resolution 3D volumes.

Consequently, CNNs continue to serve as the primary architecture for 3D MRI brain tumor segmentation. To achieve a balance between segmentation performance and computational efficiency, transformer components and other advanced modules are frequently embedded within CNN-based encoder–decoder frameworks, rather than used as standalone replacements.

### 2.2. Cross-modality fusion

In brain tumor segmentation, different modalities can provide distinct structural and functional information, which they can be complementary to each other. Therefore, cross-modality feature fusion of MRI has become an important topic for brain tumor segmentation. The goal of fusion is to take advantages of different MRI modalities to enhance the visibility and segmentation performance of tumor regions. According to the level of information, cross-modality fusion strategies can be divided into three levels. Pixel-level fusion (Badrinarayanan et al., 2015) methods typically stack images from different modalities in the channel dimension to form a single multi-channel input. These inputs are then processed by convolutional neural networks or other models for joint training, resulting in fused feature maps used for segmentation tasks. However, these methods assume a simple linear relationship among the modalities, and ignore the correlations between different modalities (Chen et al., 2014; Wang et al., 2018).

The feature-level fusion technique uses each or a group of modal images as input for training independent segmentation models. Subsequently, the learned feature representations from each separate network are combined within network layers. Finally, the fused output is transmitted to the decision layer, which determines the final segmentation results. It is thought that the feature-level fusion approach can effectively take advantage of the intricate interaction between these modalities (Chen et al., 2019b). Syazwany et al. (2021) incorporated a bi-directional feature pyramid network into their four encoders- one decoder architecture to realize cross-modality feature fusion. In Mo et al. (2020), the authors proposed a cross-modality convolution to fuse information among these modalities. A modality-aware module was created by Zhang et al. (2021) to provide more effective information sharing between various modalities.

Decision-level fusion involves integrating the results obtained from individual modalities or models at the decision-making stage. Its aim is to enhance the overall accuracy and reliability of the final decision by leveraging the complementary information provided by different modalities. For example, after training a single network for each modality, Nie et al. (2016) combined the features of all the upper layers of the networks. Experimental results showed that the suggested model outperformed earlier techniques in terms of accuracy. Kamnitsas et al. (2018)utilized the majority voting technique to assign each voxel the labels corresponding to the majority of the individual networks.

Recent work (Ullah et al., 2023; Yang et al., 2023; Zhu et al., 2024b) has effectively demonstrated that, for medical image segmentation tasks, the use of the middle-level fusion method, that is feature-level fusion, performs better than input-level fusion. While compared with decision-level fusion, feature-level fusion can better utilize the correlations between various modalities at different scales. Therefore, we will adopt a feature-level fusion mechanism in our work.

### 2.3. Attention mechanisms

The attention mechanism, borrowed from human perception and visual cognition, has become a standard tool in computer vision tasks, as demonstrated in Syazwany et al. (2021), Vaswani et al. (2017), Havaei et al. (2017), Jang and Cho (2024), Alwadee et al. (2025). In these tasks, the main goal of attention mechanism is to generate channel or spatial weight maps using the feature representations learned by neural networks. To present a fused attention mechanism integrating spatial and channel attention, Woo et al. (2018) created a convolutional block attention module(CBAM) that is commonly employed in classification or segmentation networks. Additionally, to help the network focus on the regions of interest(ROIs), Oktay et al. (2018) designed a type of Attention U-Net gate in the decoder, which can improve the segmentation performance. In conclusion, the attention mechanism is frequently used to emphasize ROIs and suppress unimportant information.

In addition to feature representation, attention mechanisms have also been adopted to realize multi-modality feature fusion (Mo et al., 2020; Liu et al., 2022; Zhou et al., 2020). Mo et al. (2020) used the attention mechanism for feature fusion after dividing the various modalities into primary and auxiliary modalities. Zhou et al. (2020) created a fusion module based on the attention mechanism after extracting the data from each of the four brain tumor modalities independently. To address variations among multiple modalities for a specific segmentation task, Liu et al. (2022) presented a multimodal feature refinement module with attention-based modality selection feature fusion. For multi-modality fusion, a dual-branch hybrid encoder incorporating a Modality-Correlated Cross-Attention block (MCCA) (Lin et al., 2023) is created.

Given the powerful capability of the attention mechanism in feature extraction, we will employ it for both specific modal feature representation and cross-modality feature fusion.

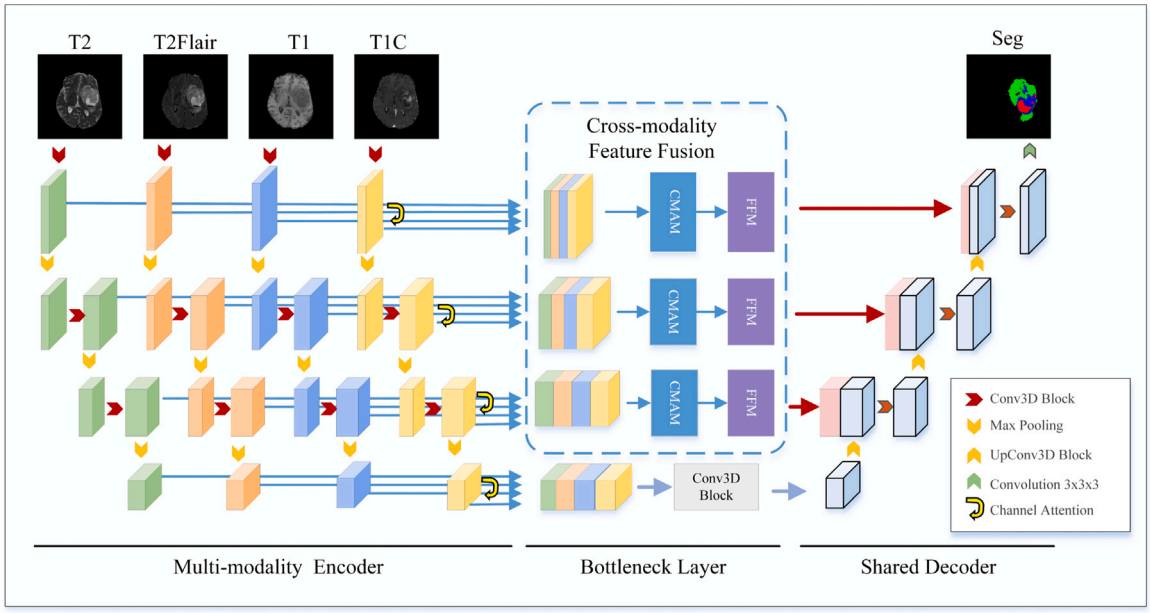



**Fig. 2.** Overall model structure. The model adopts a U-shaped structure with a four-branch encoder, a cross-modality feature fusion module, and a shared decoder.

## 3. Method

The proposed segmentation model for brain tumors is shown in Fig. 2. It is built upon a 3D U-shaped structure, which is an end-to-end full-convolution encoder–decoder network. A feature-level fusion strategy is employed to thoroughly explore and utilize the information within and across four modalities of MRI sequences. To extract independent features from different modalities, we adopt a four-branch encoder structure with three layers of convolution. Features from these modalities are concatenated along the channel dimension and connected to the decoder via skip links. The proposed Cross-Modality feature fusion (CMFF) module is embedded in the skip connection to adaptively capture interactions between modalities. In the bottleneck layer, high-level feature maps from all four modalities are directly concatenated after max-pooling. In the shared decoder, the fused encoder features, enriched with detail information from the skip connections, are merged with up-sampled features to reconstruct the segmentation results.

### 3.1. Multi-modality encoder

In MRI-based brain tumor segmentation tasks, the tumor region is usually divided into three parts: enhancing tumor (ET), tumor core(TC), and whole tumor(WT). TC includes enhancing tumor, necrosis, and non-enhancing tumor (NET), while WT consists of TC and edema(ED). Based on expert knowledge, it was observed that different modalities have varying effects on these tumor regions. For instance, experimental results (Tseng et al., 2017) indicate that T2 and T2Flair modalities have better segmentation effects on WT. Therefore, simply stacking and feeding multi-modal MRI sequences into the model may not effectively exploit the significance of each modality for tumor segmentation. To address this issue, our model adopts a non-interacting four-branch encoder structure in the feature extraction process. This approach aims to generate more discriminant features, while also ensuring high adaptability to deal with the modality missing problem.

The multi-modality images are denoted by $[\mathbf{X}_{T1}, \mathbf{X}_{T1C}, \mathbf{X}_{T2}, \mathbf{X}_{T2F}]$ and their corresponding encoders are denoted by $[\mathrm{E}_{T1}, \mathrm{E}_{T1C}, \mathrm{E}_{T2}, \mathrm{E}_{T2F}]$. Instead of having four convolutional layers as in the classical 3D U-Net, each encoder consists of only three 3D convolutional layers followed by max-pooling to generate feature maps with four-level resolutions. The purpose is to create a lightweight model and prevent overfitting. In each layer, the ConvBlock (Zhao et al., 2022) can be expressed as:.

$$\mathrm{ConvBlock}(\mathbf{F}) = \sigma(BN(\mathrm{Conv3D}^{3\times3\times3}(\mathbf{F}))). \tag{1}$$

Based on clinical knowledge in radiology, the T1C sequence is suggested to be more effective in diagnosing brain tumors compared to other modalities. This modality offers improved contrast between healthy and abnormal tissues, aiding in the accurate identification and assessment of brain tumors. In semantic segmentation tasks, the advantages of the T1C sequence become more apparent. Its high contrast and clear delineation of tumor boundaries make it particularly well-suited for accurately segmenting the tumor core and areas of enhancement. Therefore, the T1C sequence plays a crucial role in guiding therapeutic decision-making and monitoring the progression of brain tumors. To further enhance the effectiveness of the T1C modal encoder, we introduce a lightweight channel attention module (CAM) in $E_{T1C}$, as shown in Fig. 3. (W, H, and D stand for the three spatial dimensions of width, height, and depth, respectively. C stands for the number of channels.)

The CAM block can be formulated by Eq. (2)

$$\mathrm{M}_{\mathrm{CAM}}\left(\mathbf{F}_{\mathrm{in}}^{l}\right) = \mathrm{Sigmoid}\left(\mathrm{MLP}\left(\mathrm{AvgP}\left(\mathbf{F}_{\mathrm{in}}^{l}\right) + \mathrm{MLP}\left(\mathrm{MaxP}\left(\mathbf{F}_{\mathrm{in}}^{l}\right)\right)\right)\right). \tag{2}$$

The MLP used in the CAM consists of two fully connected layers: the first layer reduces the channel dimension from $C$ to $C/r$ (where r=16 in our experiments) with a ReLU activation, and the second layer projects it back from $C/r$ to $C$ without any activation.

Then the procedure of each layer in is as Eq. (3)

$$\begin{cases} \hat{\mathbf{F}}_{T1C}^{l} = \mathrm{ConvBlock}\left(\mathrm{Max\,P}\left(\mathbf{F}_{T1C}^{l-1}\right)\right) \\ \mathbf{F}_{T1C}^{l} = \mathrm{M}_{\mathrm{CAM}}\left(\hat{\mathbf{F}}_{T1C}^{l}\right) \otimes \hat{\mathbf{F}}_{T1C}^{l} + \hat{\mathbf{F}}_{T1C}^{l} \end{cases}, \tag{3}$$

where $l$ denotes the convolutional block layer level and $\otimes$ stands for multiplication of elements. $\mathbf{F}_{T1C}^{l-1}$ and $\mathbf{F}_{T1C}^{l}$ represent the output feature maps of layer $l-1$ and layer $l$ of $\mathrm{E}_{T1C}$, respectively.

### 3.2. Bottleneck layer

Different from the classical U-net-like structure, we keep the encoder lightweight by employing only three convolutional layers. Consequently, the multimodal features obtained from the third convolutional

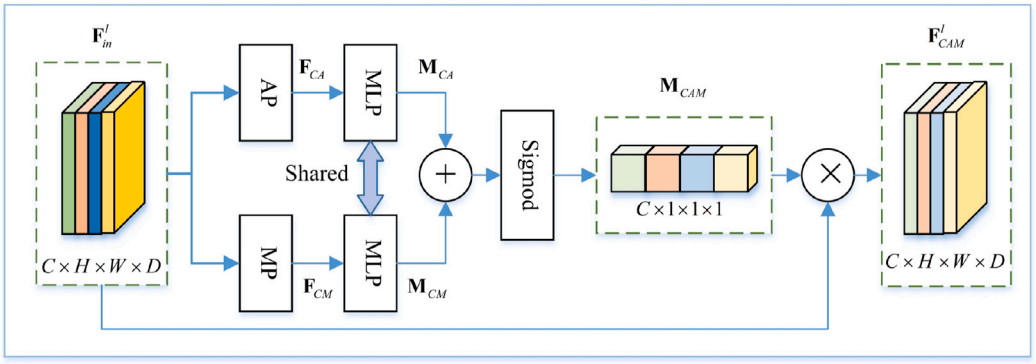



**Fig. 3.** Channel Attention Module.

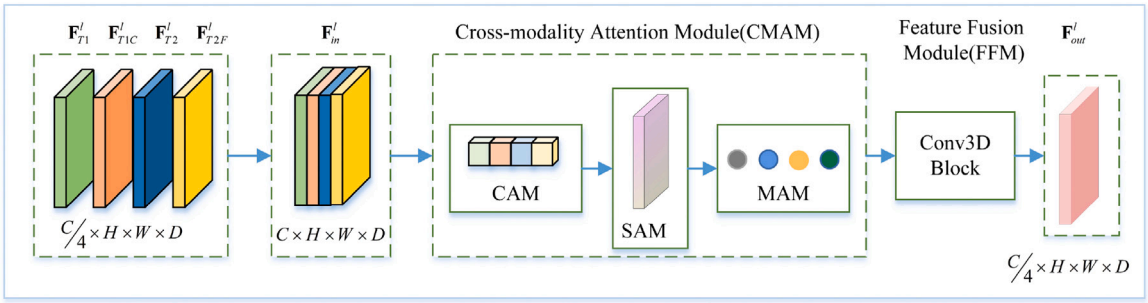



**Fig. 4.** The proposed Cross-modality Feature Fusion Module.

layer are directly concatenated after max-pooling. The process within the bottleneck layer is defined as follows.

$$\begin{cases} \hat{\mathbf{F}}^4_{T1} = \mathrm{MaxP}\left(\mathbf{F}^3_{T1}\right) \\ \hat{\mathbf{F}}^4_{T1C} = \mathrm{MaxP}\left(\mathbf{F}^3_{T1C}\right) \\ \hat{\mathbf{F}}^4_{T2} = \mathrm{MaxP}\left(\mathbf{F}^3_{T2}\right) \\ \hat{\mathbf{F}}^4_{T2F} = \mathrm{MaxP}\left(\mathbf{F}^3_{T2F}\right) \end{cases}, \tag{4}$$

$$\mathbf{F}_{\text{brige}} = \mathrm{ConvBlock}\left(\ \mathrm{Concat}\ \left(\hat{\mathbf{F}}^4_{T1}, \hat{\mathbf{F}}^4_{T1C}, \hat{\mathbf{F}}^4_{T2}, \hat{\mathbf{F}}^4_{T2F}\right)\right), \tag{5}$$

where Concat(·) denotes the concatenation operation.

### 3.3. Cross-modality feature fusion module

Fusing features from different modalities is crucial for multi-modality MRI sequences. We designed a fusion module based on attention mechanisms to highlight the data that significantly contribute to brain tumor segmentation across various channels, spatial dimensions, and modalities. The proposed module for cross-modality feature fusion (CMFF), as illustrated in Fig. 4, consists of three stages: feature concatenation, attention, and feature fusion.

First, the concatenation of each output from the convolutional block of each modality is done as follows.

$$\mathbf{F}^l_{\text{in}} = \mathrm{Concat}\left(\mathbf{F}^l_{T1}, \mathbf{F}^l_{T1C}, \mathbf{F}^l_{T2}, \mathbf{F}^l_{T2F}\right). \tag{6}$$

And then, this concatenated tensor serves as the input for the cross-modality attention module(CMAM), allowing the fusion of features across channel, spatial and modalities. CBAM (Convolutional Block Attention Module) (Woo et al., 2018) is a lightweight architecture that uses a combination of channel and spatial attention to improve the performance of convolutional neural networks. Inspired by the CBAM, we propose a novel lightweight attention module that incorporates a modality attention module (MAM) into CBAM. This MAM assigns varying weights to different modalities across modal fusion sections. Let $\mathrm{M}_{CAM}, \mathrm{M}_{SAM}$ and $\mathrm{M}_{MAM}$ represent the channel attention module(CAM), spatial attention module(SAM) and modality attention model(MAM), respectively. The overall attention process can be summarized as follows,

$$\begin{cases} \mathbf{F}^l_{CAM} = \mathrm{M}_{CAM}\left(\mathbf{F}^l_{in}\right) \otimes \mathbf{F}^l_{in} \\ \mathbf{F}^l_{SAM} = \mathrm{M}_{SAM}\left(\mathbf{F}^l_{CAM}\right) \otimes \mathbf{F}^l_{CAM} \\ \mathbf{F}^l_{MAM} = \mathrm{M}_{MAM}\left(\mathbf{F}^l_{SAM}\right) \oplus \mathbf{F}^l_{SAM} \end{cases}, \tag{7}$$

where ⊗ stands for multiplication of elements, and ⊕ represents element-wise addition. At last, the final cross-modality fused feature is calculated as follow:

$$\mathbf{F}^l_{\text{out}} = \mathrm{ConvBlock}\left(\mathbf{F}^l_{\text{MAM}}\right). \tag{8}$$

The attention values are broadcast in accordance with the multiplication process: modality attention values are broadcast along the spatial and channel dimension. The computing procedure for each attention module are explained in the following. In practical applications, the feature fusion module in each layer can adopt different configurations of attention mechanisms, which provides great flexibility.

#### 3.3.1. Channel attention module(CAM)

We employ the same channel attention module as used in the T1C encoder, with the only difference being that the input has changed from single T1C modality to multi-modality features. CAM block is formulated by Eq. (9)

$$\mathbf{F}^l_{CAM} = \mathrm{M}_{CAM}\left(\mathbf{F}^l_{in}\right) \otimes \mathbf{F}^l_{in}. \tag{9}$$

#### 3.3.2. Spatial attention module (SAM)

Spatial attention allows model to concentrate on particular areas of an input image. The fundamental idea behind spatial attention is to enhance the performance of the model by using a gating mechanism

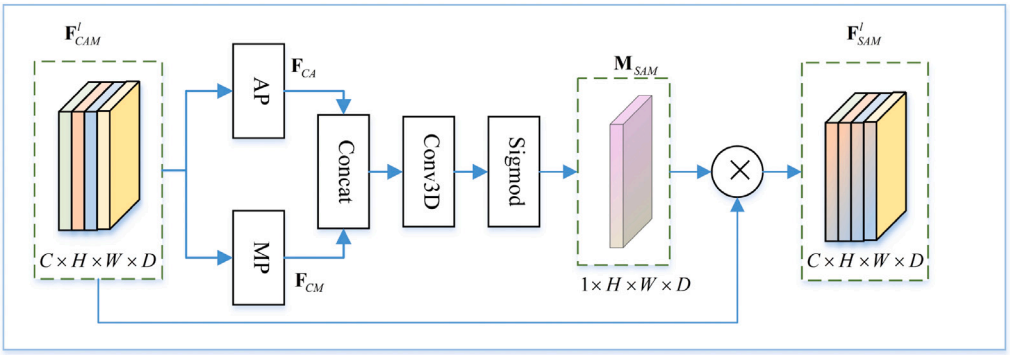



**Fig. 5.** Spatial Attention Module.

that can pick out or concentrate on particular areas of the input picture. SAM block is demonstrated in Fig. 5 and formulated by Eq. (10)

$$\begin{aligned} \mathrm{M}_{SAM}\left(\mathbf{F}^l_{CAM}\right) &= \mathrm{Sigmoid}\left(\mathrm{Conv3D}^{7\times7\times7}\left(\mathrm{AvgP}\left(\mathbf{F}^l_{CAM}\right)\right.\right. \\ &\quad \left.\left. + \mathrm{MaxP}\left(\mathbf{F}^l_{CAM}\right)\right)\right) \\ \mathbf{F}^l_{SAM} &= \mathrm{M}_{SAM}\left(\mathbf{F}^l_{CAM}\right) \otimes \mathbf{F}^l_{CAM}. \end{aligned} \tag{10}$$

### 3.3.3. Modality attention module (MAM)

After the contacted modal feature passes through the CAM and SAM attention modules, the feature map $\mathbf{F}^l_{SAM} \in \mathbf{R}^{C\times H\times W\times D}$ is generated. The Modality Attention Module(MAM) takes $\mathbf{F}^l_{SAM} \in \mathbf{R}^{C\times H\times W\times D}$ as input and splits it into four feature maps $[\hat{\mathbf{F}}^l_{T1}, \hat{\mathbf{F}}^l_{T1C}, \hat{\mathbf{F}}^l_{T2}, \hat{\mathbf{F}}^l_{T2F}]$ corresponding to four modalities, respectively, with dimensions of $[C/4, H, W, D]$. Then, each modality feature tensor passes through an adaptive average pooling layer with a stride of 1, which reduces spatial dimensions from $H\times W\times D$ to $1\times1\times1$. Further, the channel dimensions are compressed from $C/4\times H\times W\times D$ to $1\times1\times1\times1$ by using global average pooling(GAP), which directly obtains the scalar weights of each modal. Then, the fusion feature map is generated by multiplying the feature map of each modality$[\hat{\mathbf{F}}^l_{\mathbf{T1}}, \hat{\mathbf{F}}^l_{T1C}, \hat{\mathbf{F}}^l_{T2}, \hat{\mathbf{F}}^l_{T2F}]$ with the corresponding weight $[\alpha_{T1}, \alpha_{T1C}, \alpha_{T2}, \alpha_{T2F}]$. Finally, the weighted modal feature maps are concatenated and the resulting feature map has dimensions of $C\times H\times W\times D$. MAM block is demonstrated in Fig. 6 and formulated by Eqs. (11) and (12). Modality attention can make the fusion process pay more attention to the importance of different modalities.

$$\begin{cases} \tilde{\mathbf{F}}^l_{T1} = \alpha_{T1} \otimes \hat{\mathbf{F}}^l_{T1} \\ \tilde{\mathbf{F}}^l_{T1C} = \alpha_{T1C} \otimes \hat{\mathbf{F}}^l_{T1C} \\ \tilde{\mathbf{F}}^l_{T2} = \alpha_{T2} \otimes \hat{\mathbf{F}}^l_{T2} \\ \tilde{\mathbf{F}}^l_{T2F} = \alpha_{T2F} \otimes \hat{\mathbf{F}}^l_{T2F} \end{cases}. \tag{11}$$

The output of the MAM block is

$$\mathbf{F}^l_{MAM} = \mathrm{Concat}\left(\tilde{\mathbf{F}}^l_{T1}, \tilde{\mathbf{F}}^l_{T1C}, \tilde{\mathbf{F}}^l_{T2}, \tilde{\mathbf{F}}^l_{T2F}\right) + \mathbf{F}^l_{SAM}. \tag{12}$$

### 3.4. Shared decoder

As symmetrical to the encoder, the decoder also consists of three convolution layers. Each basic module includes two 3D convolutions, followed by a batch normalization layer and an activation layer. After transposed convolution, the outputs of the feature fusion block are concatenated with the feature maps of corresponding decoder layers. At last, the feature map is passed through a 3D convolution layer to obtain the final segmentation result, which can be formulated as follows:

$$\begin{cases} \mathbf{F}^3_D = \mathrm{ConvBlock}\left(\mathrm{Concat}\left(\mathbf{F}^3_{\mathrm{out}}, \mathrm{TranConv}\left(\mathbf{F}_{\mathrm{bridge}}\right)\right)\right) \\ \mathbf{F}^2_D = \mathrm{ConvBlock}\left(\mathrm{Concat}\left(\mathbf{F}^2_{\mathrm{out}}, \mathrm{TranConv}\left(\mathbf{F}^3_D\right)\right)\right) \\ \mathbf{F}^1_D = \mathrm{ConvBlock}\left(\mathrm{Concat}\left(\mathbf{F}^1_{\mathrm{out}}, \mathrm{TranConv}\left(\mathbf{F}^2_D\right)\right)\right) \\ \mathbf{F}_{\mathrm{seg}} = \mathrm{Conv}^{3\times3\times3}\left(\mathbf{F}^1_D\right) \end{cases}, \tag{13}$$

where $F_{seg}$ is the segmentation map for the brain tumor.

### 3.5. Loss function

The loss function used in this research is the combination of dice and cross entropy loss:

$$L_{\mathrm{total}} = \alpha L_{\mathrm{dice}} + (1-\alpha)L_{CE}, \tag{14}$$

where $\alpha$ represents the weight of Dice loss.

Because the dice loss ignores a considerable number of background pixels when computing the intersection ratio, it could solves the problem of imbalanced positive and negative samples. $L_{dice}$ is calculated as:

$$L_{\mathrm{dice}} = 1 - \frac{2\sum_k^K \sum_{i\in N} p_i^k g_i^k}{\sum_k^K \sum_{i\in N} p_i^k + \sum_k^K \sum_{i\in N} g_i^k}. \tag{15}$$

The cross-entropy loss $L_{CE}$ is often used to solve the classification problem of multiple labels, which is calculated as:

$$L_{CE} = -\sum_i^N \sum_{k=1}^K g_i^k \log\left(p_i^k\right), \tag{16}$$

where $K$ represents the number of tumor regions, $N$ represents the number of voxels. $g_i^k$ denotes the ground truth, and $p_i^k$ denotes the probability prediction for class $k$ of the $i$ th voxel. In the experiments, we set $\alpha = 0.5$

## 4. Experimental analysis

### 4.1. Dataset

The experiment was conducted on the publicly available BraTs2021 dataset from MICCAI, which consists of 1251 patients. The dataset was divided into three groups, and the ratio of training set, validation set and test set was 8:1:1. The dataset includes 3D scans with images of size 155 $x$ 240 $x$ 240 pixels and contains four modalities of MRI images (T1, T1C, T2, T2Flair), as illustrated in Fig. 1. The tumor region is segmented into three distinct sub-regions, denoted by the colors green, blue, and red. Specifically, the green region corresponds to edema (ED), the blue region signifies the enhancing tumor (ET), and the red region

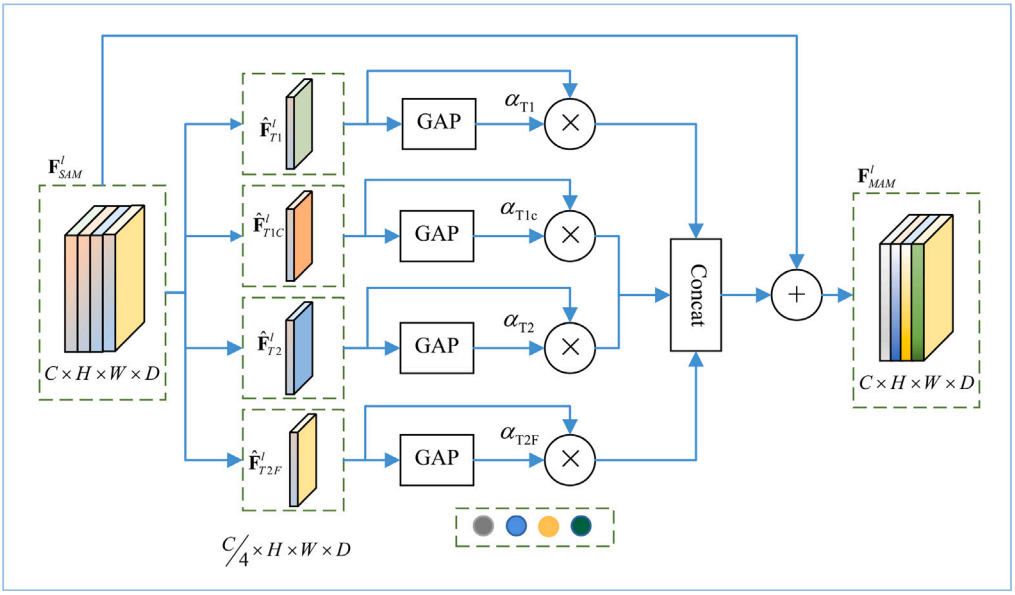



**Fig. 6.** Modality Attention Module.

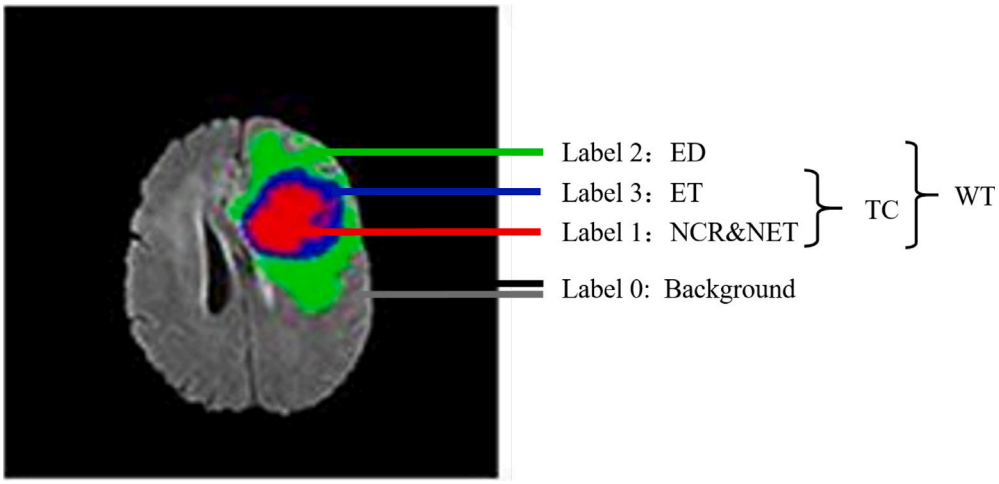



**Fig. 7.** Illustration of tumor regions.

represents the non-enhancing tumor (NCR&NET). To comprehensively assess the segmentation accuracy, the tumor is further classified into three categories: ET, tumor core (TC) including both ET and NCR&NET, and the whole tumor (WT) including all three sub-regions, as shown in Fig. 7.

### 4.2. Data processing and training strategy

In this paper, the public dataset is randomly divided into three sets, with 0.8 used for training set, 0.1 used for validation, and 0.1 used for testing. Since the contrast of the four modalities in the dataset varied, each modality was normalized using the Z-score method. The data were then averaged and normalized to have an average of 0 and a standard deviation of 1.

All of the experiments were carried out with the NVIDIA GeForce RTX 3090 GPU and Adam (Ding et al., 2021) as the optimizer on the Pytorch 1.10 platform. During training, each input sequence was randomly cropped into a size of 128 × 160 × 160, and random rotation(90°, 180° and 270°), mirror flip, and strength offset were performed with a probability of 0.01. The number of training epochs is set 60, with a batch size of 1. During the training process, a weight decay of 5e-4 was employed, and the nnUnet (Woo et al., 2018) learning rate strategy was adopted. Initially, a learning rate of 4e-3 was set, which gradually decayed to 2e-3 prior to termination.

### 4.3. Evaluation metric

(1) Dice Coefficient(Dice): a measure of set similarity. It offers insights into how closely the segmented regions of two samples align when compared.

$$\text{Dice} = \frac{2 * |V_P \cap V_G|}{|V_P| + |V_G|}, \tag{17}$$

where $V_G$ and $V_P$ represent the point sets of ground truth and prediction segmentation, respectively.

(2) Hausdorff Distance (HD): a measure of the similarity between two sets of points, which is defined as

$$H(V_P, V_G) = \max\left(h(V_P, V_G), h(V_G, V_P)\right), \tag{18}$$

where $\|\cdot\|$ is the distance metric between point sets $V_P$ and $V_G$ (e.g., L2 or Euclidean distance).

$$h(V_\rho, V_G) = \max_{\rho \in V_r}\left\{\min_{g \in V} \|p - g\|\right\}. \tag{19}$$

**Table 2**
Intra-modality Channel attention modules are added to different modality.

| Method | Dice(%) ↑ | | | |
|---|---|---|---|---|
| | ET | TC | WT | Avg |
| Base | 83.43 | 86.91 | 90.34 | 86.89 |
| Base+T1-CAM | 83.35 | 86.24 | 90.88 | 86.82 |
| Base+T1C-CAM | **84.76** | **87.95** | 90.97 | **87.89** |
| Base+T2-CAM | 83.51 | 85.82 | **91.45** | 86.92 |
| Base+Flair-CAM | 82.83 | 85.58 | 91.85 | 86.75 |
| Base+(T1C+F)-CAM | 84.15 | 87.26 | 91.22 | 87.54 |

$$h\left(V_G, V_\rho\right) = \max_{g \in V_G} \left\{ \min_{\rho \in V_r} \|g - p\| \right\}. \tag{20}$$

When it comes to evaluating medical image segmentation, the Hausdorff distance exhibits greater sensitivity toward the segmented boundary, whereas the Dice coefficient demonstrates a stronger focus on the internal filling of the mask.

(3) Intersection over Union (IoU): the spatial overlap between the predicted and ground truth masks is quantified as

$$IoU = \frac{\left|V_P \cap V_G\right|}{\left|V_P \cup V_G\right|}. \tag{21}$$

(4) Recall (Re): It reflects the sensitivity of the model, which is defined as

$$\mathrm{Re} = \frac{TP}{TP + FN}. \tag{22}$$

(5) Accuracy (Acc): It indicates the model's overall performance in accurately identifying both positive and negative pixels, which is defined as

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{23}$$

where TP, FP, and FN refer to the true positive, false positive, and false negative, respectively.

### 4.4. Ablation experiment

We conducted several ablation experiments to evaluate the effectiveness of the main components including the proposed T1C-CAM, different configurations of cross-modality feature fusion block, and the number of convolution layers.

#### 4.4.1. Effectiveness of intra-modality attention in encoder

In this subsection, we evaluated the performance of the Channel Attention Module (CAM) across different modalities in the encoders. The backbone of the proposed network without any attention module is used as the baseline method, and then add CAM to different modality encoders to validate their effects. From Table 2, it is obvious that the T1-CAM method did not show improvement compared to the baseline and even exhibited a slight disadvantage. However, employing T1C-CAM increases the Dice scores for ET, TC, and WT by 1.33%, 1.04% and 0.63%, respectively. Similar results were obtained by incorporating T2-CAM and Flair-CAM, with notable increases in Dice scores for WT by 1.11%, and 1.51%, respectively, but without improvement for ET and TC. Furthermore, when channel attentions were applied simultaneously to T1C and Flair modalities, improvements in Dice scores were observed for all ET, TC, and WT regions. These experimental results are consistent with the prior knowledge that T1C and Flair are more conducive to segmenting tumor core and edema regions, respectively, resulting in noticeable improvements in segmentation performance. Consequently, considering the overall performance and model lightweight, channel attention was exclusively integrated into the encoding part of T1C.

**Table 3**
Cross-modality attention modules are added to the feature fusion block. The red, blue, and teal colors represent the top-1, top-2, and top-3 performance respectively.

| Method | Dice (%) ↑ | | | |
|---|---|---|---|---|
| | ET | TC | WT | Avg |
| Base | 84.77 | 87.96 | 90.92 | 87.88 |
| Base+CAM+SAM | 85.63 | 88.53 | 91.25 | 88.47 |
| Base+CAM+SAM+MAM | 84.77 | 88.52 | 91.66 | 88.32 |
| Base+MAM+CAM+SAM | 84.83 | 88.64 | 91.65 | 88.37 |
| Base+MAM(L2)+CAM+SAM | 85.86 | 89.47 | 91.53 | 88.95 |
| Base+CAM+SAM+MAM(L2) | 85.97 | 89.98 | 91.64 | 89.20 |

#### 4.4.2. Effectiveness of cross-modality attention in feature fusion block

To evaluate the effectiveness of various attention component configurations within the cross-modality feature fusion (CMFF) block, an ablation study was conducted. For this experiment, we use the selected model (Base+T1C-CAM) in the last subsection as the base method. Subsequently, we introduce attention components at various positions to evaluate their effects.

- Base+CAM+SAM: Just adding channel attention module (CAM) and spatial attention module (SAM) to the feature fusion block in each layer.
- Base+CAM+SAM+MA: Further adding modality attention module (MAM) to the feature fusion block in each layer.
- Base+MAM+CAM+SAM: Changing the position of modality attention module (MAM).
- Base+MAM(L2)+CAM+SAM: The modality attention module is incorporated only in the middle layer(L2).
- Base+CAM+SAM+MAM(L2): Changing the position of modality attention module(L2).

Table 3 presents the objective comparison of these methods. It is obvious that each method has achieved some improvements in segmentation. For example, by adding the channel and spatial attention at each layer, the Dice scores of three tumor regions are increased obviously. Moreover, by adding the modality attention module, the segmentation performance of WT regions has reached its peak (91.66%). Among them, the baseline model has the lowest segmentation accuracy, while the last method achieves the best results with Dice scores for ET, TC, and WT, increasing by 1.2%, 2.02%, and 0.72%, respectively.

Overall, the ablation experimental results show the effectiveness of the attention mechanism in cross-modality feature fusion, which can guide the model to pay more attention to the information that is more useful for segmentation, whether it is channel and spatial attention or modality attention. Additionally, the results also indicate that introducing modality attention (i.e., a scalar for each modality) during feature fusion at intermediate layers leads to the best performance. This is because the middle-layer features not only contain detailed low-level information, but also contain rich semantic information. Therefore, the last method is adopted in our final model.

#### 4.4.3. Holistic ablation experiment

To further verify the effectiveness of each technical novelty in the proposed model, an ablation study is conducted in this subsection. (1) is the baseline model without T1C-CAM and CMAM. (2) is the baseline model with T1C-CAM. (3) is the proposed model with a 4-layer convolutional architecture. (4) is our final model.

As evident from Table 4, the integration of attention mechanisms in both feature detection and fusion stages leads to a substantial improvement in Dice scores. However, when comparing models (3) and (4), it is observed that incorporating more convolutional layers does not result in an improvement in segmentation performance. Conversely, there is a slight decrease in performance, especially for the TC. This indicates that for medical image segmentation with limited training samples, an excessive number of convolutional layers may lead to overfitting.

**Table 4**
Ablation studies on layers of convolution, T1-CAM and CMAM.

| Ablation | | | | Dice(%) ↑ | | |
|---|---|---|---|---|---|---|
| Model | Conv Layers | T1-CAM | CMAM | ET | TC | WT |
| (1) | 3 | | | 83.43 | 86.91 | 90.34 |
| (2) | 3 | ✓ | | 84.77 | 87.96 | 90.92 |
| (3) | 4 | ✓ | ✓ | 85.19 | 88.34 | 91.20 |
| (4) | 3 | ✓ | ✓ | **85.97** | **89.98** | **91.64** |

**Table 5**
Ablation experiments on modality missing.

| Modality | | | | Dice (%) ↑ | | | HD95 (mm) ↓ | | |
|---|---|---|---|---|---|---|---|---|---|
| T1C | T1 | T2 | FLAIR | ET | TC | WT | ET | TC | WT |
| ✓ | ✓ | ✓ | ✓ | 85.78 | 89.02 | 91.04 | 2.42 | 2.56 | 2.41 |
| ✓ | × | ✓ | ✓ | 82.74 | 85.79 | 89.79 | 3.71 | 3.66 | 2.94 |
| ✓ | ✓ | × | ✓ | 82.53 | 85.24 | 85.10 | 5.60 | 5.66 | 9.99 |
| ✓ | ✓ | ✓ | × | 82.08 | 79.64 | 71.04 | 5.19 | 15.84 | 16.81 |

#### 4.4.4. Ablation experiments on modality missing

Table 5 presents the results of an ablation study designed to evaluate the model's robustness to missing imaging modalities during inference. All models were trained using the complete set of four modalities (T1, T1C, T2, and FLAIR), while testing was conducted under conditions where one modality was selectively removed.

When all modalities are available during testing, the model achieves its best performance across all metrics . Removing T1 during testing causes only a slight performance drop, indicating that T1 contributes less to final segmentation than other modalities. In contrast, removing T2 or T2Flair leads to a more substantial degradation, especially in the WT region (Dice drops to 85.10% and 71.04%, respectively). This suggests that T2 and FLAIR provide critical information for delineating the full tumor extent. In terms of HD95, missing FLAIR results in the largest spatial prediction error, particularly for WT and TC (15.84 mm and 16.81 mm, respectively), highlighting its importance in spatially accurate tumor boundary estimation.

Overall, the results indicate that the model exhibits a certain degree of generalization under partial modality absence; however, the FLAIR sequence is essential for achieving reliable whole tumor segmentation, particularly in terms of boundary accuracy. These findings emphasize the necessity of incorporating modality-specific robustness into multi-modal segmentation models, especially for real-world clinical deployment where certain modalities may be unavailable.

### 4.5. Comparative experiment

In this subsection, we compare our proposed model with six other methods, including UNet3D (Ahmad et al., 2021), Attention Unet (Oktay et al., 2018), nnU-Net (Isensee et al., 2019), UNet++ (Zhou et al., 2018), TransBTS (Wenxuan et al., 2021), and LightMUnet (Liao et al., 2024). All of these models were re-implemented based on the released codes. The experiments were conducted on the same computer hardware with identical learning rates and dataset assignments. The quantitative and qualitative results were obtained without any post-processing. Due to memory limitations, the input size for TransBTS is set as $128 \times 128 \times 128$.

#### 4.5.1. Quantitative comparison results

As shown in Tables 6 and 7, the proposed model obtains more competitive performance compared with other models. Among all models, our model obtains the best results in average Dice scores of 85.92% for enhancing tumor, 89.50% for tumor core, and 93.12% for whole tumor. In particular, the proposed model outperforms the others on all metrics for ET and TC, with the only exception being a slightly lower recall than the LightM-Unet. Additionally, compared to TransBTS, which combines Transformer and U-Net, our model outperforms it on almost all metrics, just except for the sensitivity of WT. This highlights the superior performance of our model. However, these results may be affected by the resolution of the input image. Furthermore, compared with the latest model LightM-Unet, our proposed model also has comparable performance.

#### 4.5.2. Qualitative comparison results

In order to provide a qualitative comparison of the models' performance, we visualized some sample segmentation results and their corresponding ground truths in Fig. 8. It can be observed that the segmentation results generated by our proposed model are closer to the ground truth when compared to the other models. Particularly, the proposed model obtains more accurate green regions.

We also demonstrate the comparison of the 3D volumetric segmentation results with the ground truth in Fig. 9. As can be seen, the 3D tumors generated by our model is very close to the ground truth.

#### 4.5.3. Comparison of model parameters

Table 8 provides a comparison of the parameter counts for classical Unet series models and hybrid models. TransBTS has the largest parameter count at 32.99M, reflecting the computational burden often associated with Transformer based models. In contrast, classical convolutional models such as UNet3D (3.68M) and UNet++ (6.87M) demonstrate much lower parameter complexity, making them more lightweight and efficient. Our model with 4 convolutional layers has 8.04M parameters, lower than Attention UNet and TransBTS, while offering a good balance between capacity and efficiency. The 3-layer version reduces the size to 3.66M, comparable to UNet3D, and is the most compact among the compared methods. Notably, LightM-UNet, a recent lightweight model, has 6.15M parameters, which is still higher than our 3-layer design. This suggests that our method achieves competitive or even superior parameter efficiency while supporting scalable design based on application needs.

#### 4.5.4. Discussion

The model proposed in this study is more accurate in the detailed segmentation of the red necrotic area and better in enhancing the blue tumor area compared to other models. Due to the modality attention assigning different weights to different modalities in a certain proportion, according to the analysis of intermediate data in the experimental process, the attention matrix output by the modal attention layer assigns more weight to the T1ce sequence, which is approximately 1.5 times more than other modalities. The T1C sequences have more obvious characteristic information for the segmentation of ET and TC, and the details of the red and blue parts of the segment are more accurate.

## 5. Conclusion

In this paper, we present a novel 3D lightweight network for brain tumor region segmentation from multi-modality MRI. To learn the modal-specific features, the clinical knowledge-inspired multi-branch encoder is adopted. By simply introducing channel attention into the T1C modality, obvious improvement can be obtained. To realize the modality fusion efficiently and effectively, we propose the cross-modality feature fusion (CMFF) module based on a lightweight attention mechanism in channel, spatial, and modality dimensions. To further make the 3D model light, only three layers of convolution are used. Compared with other SOTA methods on BraTS21, our method works better or has comparable performance in brain tumor segmentation tasks.

Despite the promising results, our method still has several limitations. The current framework does not explicitly enforce the learning of causally stable features, which may limit its robustness under distribution shifts. All experiments are conducted on public datasets, which may not fully reflect the variability and noise in real-world clinical

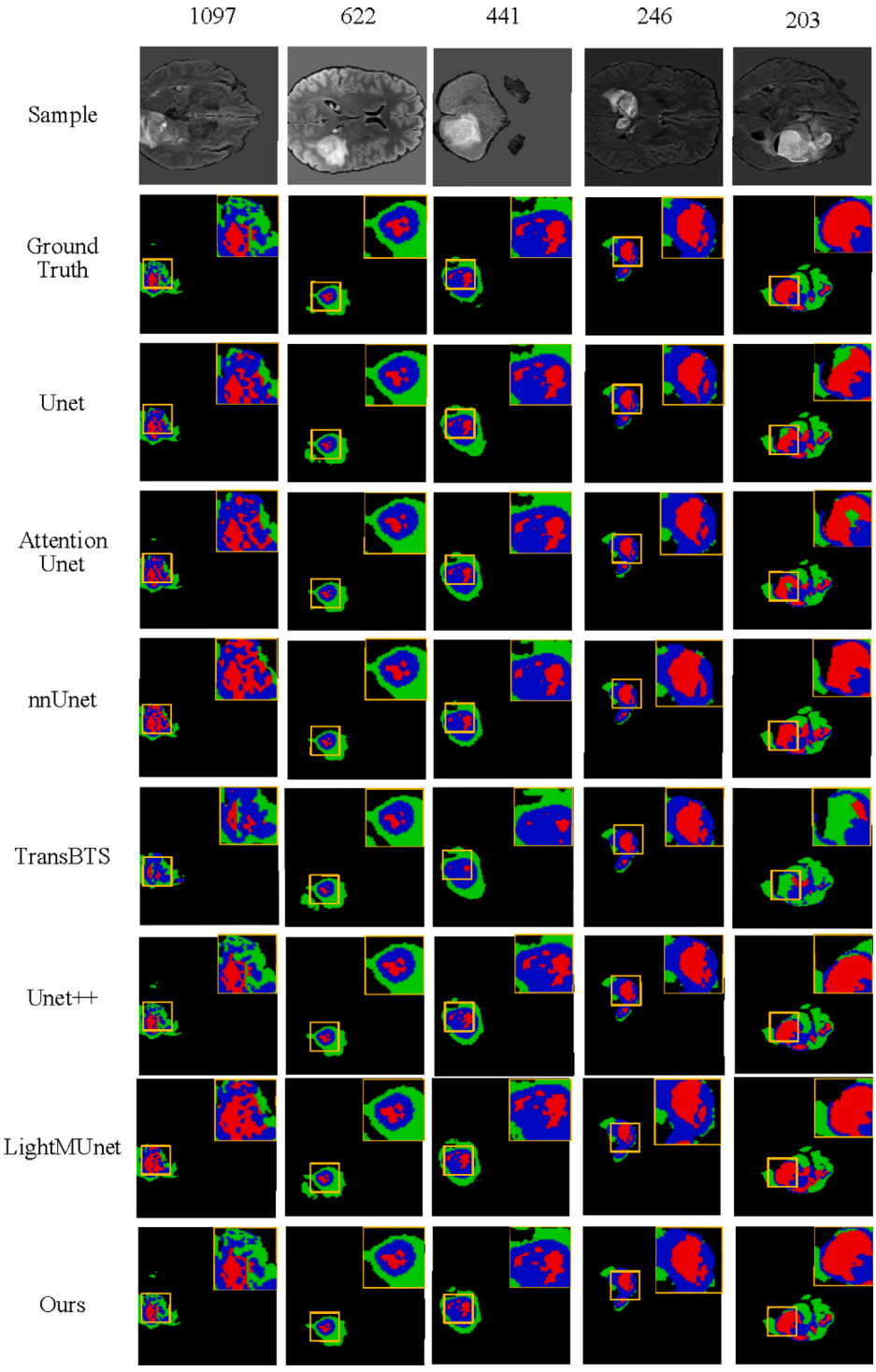



**Fig. 8.** Visualization of quantitative results on BraTS21, where yellow box used to observe the segmentation details, and the colors green, blue, and red correspond to ED, ET and NCR/NET regions, respectively.

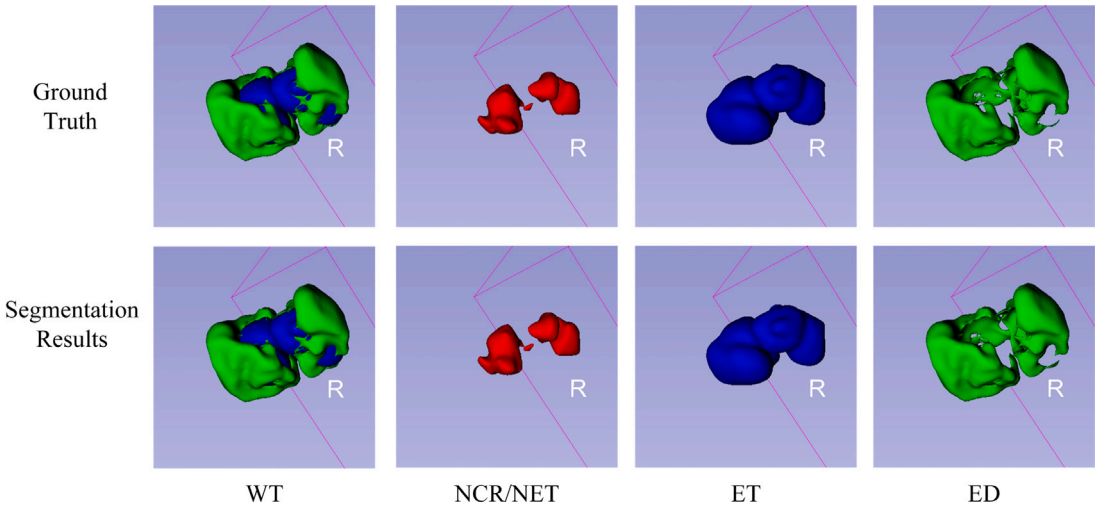



**Fig. 9.** 3D visualization of our model's segmentation results. The top row is shows the ground truth. The bottom row shows our segmentation results.

**Table 6**
Performance comparison of different segmentation models on the BraTS21 dataset using Dice (%) and HD95 (mm) across three tumor subregions.

| Model | Dice (%) ↑ | | | HD95 (mm) ↓ | | |
|---|---|---|---|---|---|---|
| | ET | TC | WT | ET | TC | WT |
| UNet3D | 82.32 ± 0.17 | 86.16 ± 0.23 | 90.16 ± 0.34 | 3.20 ± 0.16 | 2.64 ± 0.19 | 2.14 ± 0.19 |
| Attention-Unet | 83.38 ± 0.48 | 86.47 ± 0.66 | 90.82 ± 0.27 | 2.78 ± 0.45 | 2.63 ± 0.28 | 2.51 ± 0.22 |
| nnU-Net | 83.66 ± 0.23 | 87.37 ± 0.25 | 92.16 ± 0.11 | 3.60 ± 0.19 | 2.69 ± 0.11 | 3.84 ± 0.14 |
| UNet++ | 80.05 ± 1.49 | 83.19 ± 2.16 | 87.33 ± 2.20 | 2.89 ± 0.50 | 2.84 ± 0.44 | 3.53 ± 0.81 |
| TransBTS | 78.50 ± 0.54 | 81.71 ± 0.54 | 89.45 ± 0.46 | 6.07 ± 0.03 | 8.53 ± 0.19 | 5.42 ± 0.26 |
| LightM-Unet | 85.78 ± 0.07 | 89.33 ± 0.39 | 92.39 ± 0.51 | 2.71 ± 0.45 | 2.24 ± 0.76 | 2.42 ± 0.40 |
| **Ours** | **85.92 ± 0.18** | **89.50 ± 0.43** | **93.12 ± 2.66** | **2.47 ± 0.73** | **2.59 ± 0.61** | **2.64 ± 0.17** |

**Table 7**
Performance comparison of different segmentation models on the BraTS21 dataset using IoU (%), Sensitivity (%) and Accuracy (%) across three tumor subregions.

| Model | IoU (%) ↑ | | | Sensitivity (%) ↑ | | | ACC (%) ↑ |
|---|---|---|---|---|---|---|---|
| | ET | TC | WT | ET | TC | WT | |
| UNet3D | 72.15 ± 0.12 | 80.38 ± 0.20 | 84.18 ± 0.19 | 74.57 ± 1.25 | 86.37 ± 1.02 | 90.29 ± 0.97 | 96.23 ± 0.45 |
| Attention-UNet | 75.05 ± 0.16 | 80.48 ± 0.78 | 85.11 ± 0.49 | 74.10 ± 0.05 | 84.61 ± 1.20 | 89.54 ± 0.52 | 97.23 ± 0.22 |
| nnU-Net | 70.16 ± 0.06 | 82.53 ± 0.47 | 85.99 ± 0.38 | 74.15 ± 1.13 | 86.60 ± 0.88 | 90.52 ± 0.65 | 96.89 ± 0.25 |
| UNet++ | 72.35 ± 1.04 | 77.58 ± 1.56 | 81.88 ± 1.98 | 74.46 ± 3.58 | 80.72 ± 1.92 | 85.07 ± 2.69 | 96.96 ± 0.66 |
| TransBTS | 69.98 ± 0.39 | 79.25 ± 0.14 | 81.79 ± 0.40 | 72.83 ± 1.45 | 84.89 ± 1.12 | 91.59 ± 0.85 | 95.73 ± 0.73 |
| LightM-Unet | 76.74 ± 0.89 | 83.32 ± 0.63 | 87.16 ± 0.80 | 76.76 ± 1.89 | 90.11 ± 4.84 | 91.25 ± 1.81 | 97.38 ± 0.43 |
| Ours | **77.99 ± 0.24** | **83.58 ± 0.08** | **85.22 ± 0.22** | **79.34 ± 0.24** | **87.11 ± 2.68** | **89.64 ± 1.13** | **97.23 ± 0.28** |

**Table 8**
Model parameters.

| Model | Parameters |
|---|---|
| UNet3D | 3.68 M |
| Attention Unet | 14.75 M |
| Unet++ | 6.87 M |
| TransBTS | 32.99 M |
| LightM-Unet | 6.15 M |
| Ours (4 Conv Layers) | 8.04 M |
| Ours (3 Conv Layers) | 3.66 M |

environments. In future work, we plan to design a causality-driven learning framework that facilitates the extraction of modality-invariant and causally stable features, with the goal of enhancing robustness against modality variability and dataset bias (Qu et al., 2024). Additionally, we aim to develop adaptive modules to handle incomplete modality inputs (Yang et al., 2023; Chen et al., 2019a). Importantly, we will validate our method on multi-center clinical data collected from real hospital settings, to better evaluate its generalization capability and practical applicability in real-world brain tumor segmentation tasks.

## CRediT authorship contribution statement

**Yawen Fan:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Chenziyi Huang:** Writing – original draft. **Xiang Wang:** Data curation. **Chaoyuan Wang:** Methodology. **Quan Zhou:** Methodology. **Jianxin Chen:** Software, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yawen Fan, Chenziyi Huang, Chaoyuan Wang and Jianxin Chen hold a granted patent (CN116740513B) concerning the methodology presented in this study. This potential competing interest has been disclosed in accordance with journal policy. The remaining authors declare no competing interests.

## Acknowledgment

This research is partly supported by the National Natural Science Foundation of China (62476139).

## Data availability

Data will be made available on request.

## References

Ahmad, P., Qamar, S., Shen, L., Rizvi, S.Q.A., Ali, A., Chetty, G., 2021. Ms unet: Multi-scale 3d unet for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. Springer, pp. 30–41.

Ali, S., Li, J., Pei, Y., Khurram, R., Rehman, K.U., Mahmood, T., 2022. A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal MR image. Arch. Comput. Methods Eng. 29 (7), 4871–4896.

Allah, A.M.G., Sarhan, A.M., Elshennawy, N.M., 2023. Edge U-Net: Brain tumor segmentation using MRI based on deep U-net model with boundary information. Expert Syst. Appl. 213, 118833.

Alwadee, E.J., Sun, X., Qin, Y., Langbein, F.C., 2025. LATUP-net: A lightweight 3D attention U-net with parallel convolutions for brain tumor segmentation. Comput. Biol. Med. 184, 109353.

Badrinarayanan, V., Handa, A., Cipolla, R., (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv 2015. arXiv preprint arXiv:1505.07293.

Biratu, E.S., Schwenker, F., Ayano, Y.M., Debelee, T.G., 2021. A survey of brain tumor segmentation and classification algorithms. J. Imaging 7 (9), 179.

Bougourzi, F., Hadid, A., 2025. Recent advances in medical imaging segmentation: A survey. arXiv preprint arXiv:2505.09274.

Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.-A., 2019a. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 447–456.

Chen, C., Liu, X., Ding, M., Zheng, J., Li, J., 2019b. 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. Springer, pp. 184–192.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Chukwujindu, E., Faiz, H., Sara, A.-D., Faiz, K., De Sequeira, A., 2024. Role of artificial intelligence in brain tumour imaging. Eur. J. Radiol. 176, 111509.

Di Ieva, A., Russo, C., Liu, S., Jian, A., Bai, M.Y., Qian, Y., Magnussen, J.S., 2021. Application of deep learning for automatic segmentation of brain tumors on magnetic resonance imaging: a heuristic approach in the clinical scenario. Neuroradiology 63, 1253–1262.

Ding, Y., Yu, X., Yang, Y., 2021. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3975–3984.

Dolz, J., Ayed, I.B., Yuan, J., Desrosiers, C., 2017. HyperDense-net: A hyper-densely connected CNN for multi-modal image semantic segmentation. arXiv preprint arXiv:1710.05956.

Fang, L., Wang, X., 2022. Brain tumor segmentation based on the dual-path network of multi-modal MRI images. Pattern Recognit. 124, 108434.

Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. Springer, pp. 272–284.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Isensee, F., Petersen, J., Kohl, S.A., Jäger, P.F., Maier-Hein, K.H., 2019. nnu-net: Breaking the spell on successful medical image segmentation. 1, (1–8), p. 2, arXiv preprint arXiv:1904.08128.

Jang, G.-B., Cho, S.-B., 2024. Multi-instance attention network for anomaly detection from multivariate time series. Cybern. Syst. 55 (6), 1417–1440.

Jyothi, P., Singh, A.R., 2023. Deep learning models and traditional automated techniques for brain tumor segmentation in MRI: a review. Artif. Intell. Rev. 56 (4), 2923–2969.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2018. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3. Springer, pp. 450–462.

Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L., 2024. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. arXiv preprint arXiv:2403.05246.

Lin, J., Lin, J., Lu, C., Chen, H., Lin, H., Zhao, B., Shi, Z., Qiu, B., Pan, X., Xu, Z., et al., 2023. CKD-TransBTS: clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. IEEE Trans. Med. Imaging 42 (8), 2451–2461.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, Y., Ma, Y., Zhu, Z., Cheng, J., Chen, X., 2024. TransSea: Hybrid CNN–transformer with semantic awareness for 3-D brain tumor segmentation. IEEE Trans. Instrum. Meas. 73, 16–31.

Liu, Y., Mu, F., Shi, Y., Cheng, J., Li, C., Chen, X., 2022. Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. Front. Neurosci. 16, 1000587.

Liu, Z., Tong, L., Chen, L., Jiang, Z., Zhou, F., Zhang, Q., Zhang, X., Jin, Y., Zhou, H., 2023. Deep learning based brain tumor segmentation: a survey. Complex Intell. Syst. 9 (1), 1001–1026.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Ma, J., Li, F., Wang, B., 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.

Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., Golland, P., 2010. A generative model for brain tumor segmentation in multi-modal images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part II 13. Springer, pp. 151–159.

Mo, S., Cai, M., Lin, L., Tong, R., Chen, Q., Wang, F., Hu, H., Iwamoto, Y., Han, X.-H., Chen, Y.-W., 2020. Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. Springer, pp. 429–438.

Mohammed, Y.M., El Garouani, S., Jellouli, I., 2023. A survey of methods for brain tumor segmentation-based MRI images. J. Comput. Des. Eng. 10 (1), 266–293.

Nie, D., Wang, L., Gao, Y., Shen, D., 2016. Fully convolutional networks for multi-modality isointense infant brain image segmentation. In: 2016 IEEE 13Th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1342–1345.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Qu, J., Xiao, X., Wei, X., Qian, X., 2024. A causality-inspired generalized model for automated pancreatic cancer diagnosis. Med. Image Anal. 94, 103154.

Ranjbarzadeh, R., Caputo, A., Tirkolaee, E.B., Ghoushchi, S.J., Bendechache, M., 2023. Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. Comput. Biol. Med. 152, 106405.

Rehman, M.U., Ryu, J., Nizami, I.F., Chong, K.T., 2023. RAAGR2-Net: A brain tumor segmentation network using parallel processing of multiple spatial frames. Comput. Biol. Med. 152, 106426.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Syazwany, N.S., Nam, J.-H., Lee, S.-C., 2021. MM-BiFPN: multi-modality fusion network with bi-FPN for MRI brain tumor segmentation. IEEE Access 9, 160708–160720.

Tseng, K.-L., Lin, Y.-L., Hsu, W., Huang, C.-Y., 2017. Joint sequence learning and cross-modality convolution for 3D biomedical segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6393–6400.

Ullah, F., Nadeem, M., Abrar, M., Al-Razgan, M., Alfakih, T., Amin, F., Salam, A., 2023. Brain tumor segmentation from MRI images using handcrafted convolutional neural network. Diagnostics 13 (16), 2650.

Urban, G., Bendszus, M., Hamprecht, F., Kleesiek, J., et al., 2014. Multi-modal brain tumor segmentation using deep convolutional neural networks. MICCAI BraTS (Brain Tumor Segmentation) Chall. Proc. Win. Contrib. 31–35.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, S., Li, G., Gao, M., Zhuo, L., Liu, M., Ma, Z., Zhao, W., Fu, X., 2025. GH-unet: group-wise hybrid convolution-VIT for robust medical image segmentation. Npj Digit. Med. 8 (1), 426.

Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2018. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3. Springer, pp. 178–190.

Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L., 2021. Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 109–119.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam:Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Yang, H., Zhou, T., Zhou, Y., Zhang, Y., Fu, H., 2023. Flexible fusion network for multi-modal brain tumor segmentation. IEEE J. Biomed. Health Informatics 27 (7), 3349–3359.

Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., He, Z., 2021. Modality-aware mutual learning for multi-modal medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 589–599.

Zhao, C., Liu, K., Chen, W., Pei, Z., Feng, Y., 2022. Multi-Modality Brain Tumor Segmentation Network Based on Collaborative Feature Fusion. In: 2022 IEEE 17th Conference on Industrial Electronics and Applications. ICIEA, IEEE, pp. 1122–1127.

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., Fan, Y., 2018. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med. Image Anal. 43, 98–111.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis. Springer, pp. 3–11.

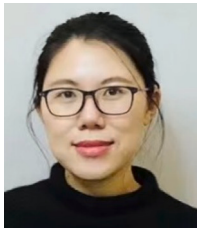
Zhou, T., Ruan, S., Guo, Y., Canu, S., 2020. A multi-modality fusion network based on attention mechanism for brain tumor segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 377–380.

Zhu, Z., He, X., Qi, G., Li, Y., Cong, B., Liu, Y., 2023. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. Inf. Fusion 91, 376–387.

Zhu, Z., Sun, M., Qi, G., Li, Y., Gao, X., Liu, Y., 2024a. Sparse dynamic volume TransUNet with multi-level edge fusion for brain tumor segmentation. Comput. Biol. Med. 172, 108284.

Zhu, Z., Yu, K., Qi, G., Cong, B., Li, Y., Li, Z., Gao, X., 2024b. Lightweight medical image segmentation network with multi-scale feature-guided fusion. Comput. Biol. Med. 182, 109204.

Zhuang, Y., Liu, H., Song, E., Hung, C.-C., 2022. A 3D cross-modality feature interaction network with volumetric feature alignment for brain tumor and tissue segmentation. IEEE J. Biomed. Health Inform. 27 (1), 75–86.



**Yawen Fan** received the B.S. and M.S. degree form Hohai Universit, Nanjing, China, in 2003 and 2005, respectively, and the Ph.D. degree with major on Electronics Engineering from Shanghai Jiao Tong University, Shanghai, China in 2014. From December 2018 to December 2019, she was a visiting scholar at Arizona State University, Temp. AZ, USA. She is now an assistant professor at Nanjing University of Posts and Telecommunications, Nanjing, P. R. China. Her research interests include computer vision, deep learning, and causality analysis.

**Chenziyi Huang** received his bachelor's degree from the School of Hubei University in 2021. He is currently a master's student at the School of Communication and Information Science and Technology, Nanjing University of Posts and Telecommunications, with research interests in medical image analysis, artificial intelligence, and deep learning.

**Xiang Wang** received his bachelor's degree from the School of Xiamen University of Technology in 2023. He is currently a master's student at the School of Communication and Information Science and Technology, Nanjing University of Posts and Telecommunications, with research interests in medical image analysis, artificial intelligence, and deep learning.

**Chaoyuan Wang** received his bachelor's degree from the School of Electronic and Information Engineering of North China Institute of Science and Technology in 2021. He is currently a master's student at the School of Communication and Information Science and Technology, Nanjing University of Posts and Telecommunications, with research interests in medical image analysis, artificial intelligence, and deep learning.

**Quan Zhou** received the B.S. degree in electronics and information engineering from the China University of Geosciences, Wuhan, China, in 2002, and the M.S and Ph.D. degrees in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, in 2006 and 2013, respectively. He is currently a Full Professor with the National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include deep learning, pattern recognition, and computer vision.

**Jianxin Chen** was born on February 1973. He received Ph.D. degree with major on Electronics Engineering from Shanghai Jiaotong University in 2007. After that, he worked in the computer college of Nanjing University of Posts and Telecommunications. From May 2008 to July of 2009, he worked as a postdoctoral in IPP Hurray Research Group, Portugal. Now he is an associate professor in the information and telecommunication engineering school of Nanjing University of Posts and Telecommunications. Mr. Chen's research interests include radiomics, cyber–physical system, wearable computing, etc.