Original Article

CMAF-Net: a cross-modal attention fusion-based deep neural network for incomplete multi-modal brain tumor segmentation

Kangkang Sun[1,2,3,4], Jiangyi Ding[2,3,4], Qixuan Li[2,3,4], Wei Chen[1,2,3,4], Heng Zhang[2,3,4], Jiawei Sun[2,3,4], Zhuqing Jiao[1*], Xinye Ni[2,3,4*]

[1]School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, China; [2]Department of Radiotherapy, The Affiliated of Changzhou No. 2 People's Hospital of Nanjing Medical University, Changzhou, China; [3]Jiangsu Province Engineering Research Center of Medical Physics, Changzhou, China; [4]Center of Medical Physics, Nanjing Medical University, Changzhou, China

Contributions: (I) Conception and design: K Sun, X Ni, Z Jiao; (II) Administrative support: Z Jiao, X Ni; (III) Provision of study materials or patients: K Sun, X Ni; (IV) Collection and assembly of data: J Ding, Q Li, W Chen; (V) Data analysis and interpretation: K Sun, H Zhang, J Sun; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to this work.

Correspondence to: Xinye Ni, PhD. Department of Radiotherapy, The Affiliated of Changzhou No. 2 People's Hospital of Nanjing Medical University, 68 Gehu Middle Road, Wujin District, Changzhou 213164, China; Jiangsu Province Engineering Research Center of Medical Physics, Changzhou, China; Center of Medical Physics, Nanjing Medical University, Changzhou, China. Email: nxy@njmu.edu.cn; Zhuqing Jiao, PhD. School of Computer Science and Artificial Intelligence, Changzhou University, 21 Gehu Middle Road, Wujin District, Changzhou 213164, China. Email: jzq@cczu.edu.cn.

Background: The information between multimodal magnetic resonance imaging (MRI) is complementary. Combining multiple modalities for brain tumor image segmentation can improve segmentation accuracy, which has great significance for disease diagnosis and treatment. However, different degrees of missing modality data often occur in clinical practice, which may lead to serious performance degradation or even failure of brain tumor segmentation methods relying on full-modality sequences to complete the segmentation task. To solve the above problems, this study aimed to design a new deep learning network for incomplete multimodal brain tumor segmentation.

Methods: We propose a novel cross-modal attention fusion-based deep neural network (CMAF-Net) for incomplete multimodal brain tumor segmentation, which is based on a three-dimensional (3D) U-Net architecture with encoding and decoding structure, a 3D Swin block, and a cross-modal attention fusion (CMAF) block. A convolutional encoder is initially used to extract the specific features from different modalities, and an effective 3D Swin block is constructed to model the long-range dependencies to obtain richer information for brain tumor segmentation. Then, a cross-attention based CMAF module is proposed that can deal with different missing modality situations by fusing features between different modalities to learn the shared representations of the tumor regions. Finally, the fused latent representation is decoded to obtain the final segmentation result. Additionally, channel attention module (CAM) and spatial attention module (SAM) are incorporated into the network to further improve the robustness of the model; the CAM to help focus on important feature channels, and the SAM to learn the importance of different spatial regions.

Results: Evaluation experiments on the widely-used BraTS 2018 and BraTS 2020 datasets demonstrated the effectiveness of the proposed CMAF-Net which achieved average Dice scores of 87.9%, 81.8%, and 64.3%, as well as Hausdorff distances of 4.21, 5.35, and 4.02 for whole tumor, tumor core, and enhancing tumor on the BraTS 2020 dataset, respectively, outperforming several state-of-the-art segmentation methods in missing modalities situations.

Conclusions: The experimental results show that the proposed CMAF-Net can achieve accurate brain

tumor segmentation in the case of missing modalities with promising application potential.

Keywords: Brain tumor segmentation; magnetic resonance imaging (MRI); missing modalities; multimodal fusion; cross-modal attention fusion (CMAF)

Submitted Jan 03, 2024. Accepted for publication Apr 19, 2024. Published online Jun 27, 2024.
doi: 10.21037/qims-24-9
View this article at: https://dx.doi.org/10.21037/qims-24-9

Introduction

A brain tumor is an abnormal growth of brain cells, which is considered one of the most prevalent and deadly diseases (1). Relevant data show that brain tumors account for approximately 1.6% of the incidence and 2.5% of the mortality of all tumors, posing a great threat to human health (2). Accurate segmentation of brain tumors provides important information to assess disease progression and develop treatment plans. Magnetic resonance imaging (MRI) is a high-performance imaging technology that causes little damage to the human body. It is extensively used to examine brain tumors because it can provide high-resolution images of the anatomical structure of soft tissues (3). Multimodal MRI is an imaging sequence under different imaging parameters that preserves the structural features of brain diseases from multiple perspectives. In the multimodal MRI of brain tumors, commonly used MRI sequences include fluid-attenuated inversion recovery (FLAIR), T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), and T2-weighted (T2) (4). Different modalities can describe different kinds of pathological features of brain tumor structures, and multimodal image information can effectively complement one another. *Figure 1* shows a typical example of multimodal MRI for brain tumor segmentation in the BraTS 2020 dataset (5). From left to right, the MRI images of the 4 modalities and the corresponding tumor region labels are shown. As shown in *Figure 1*, different modality sequences highlight different features, and limitations exist in characterizing these features with only a single-modality MRI. Combining multiple modality images can provide comprehensive information for analyzing different subregions of brain tumors and improving the accuracy of diagnosis and segmentation (6). However, manual segmentation of different subregions of brain tumors from multimodal MRI usually requires substantial time and effort from expert radiologists (7), and the results are usually subjective. Therefore, designing an automatic multimodal MRI brain tumor segmentation algorithm is needed to improve the accuracy and efficiency of clinical diagnosis (8).

For the auxiliary diagnosis technology of medical images, some studies have mainly focused on the classification

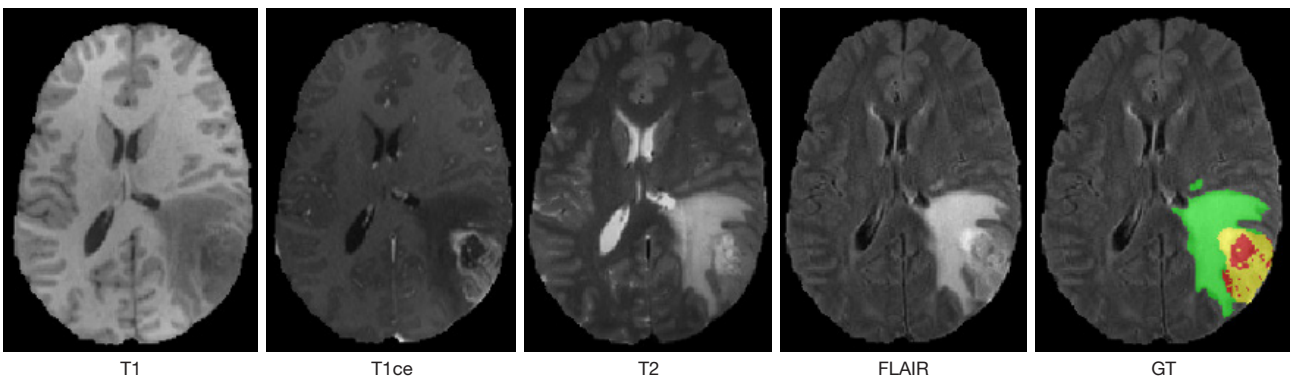

Figure 1 Example of data from the BraTS 2020 dataset. From left to right, the images represent 4 MRI modalities: T1, T1ce, T2, FLAIR, and the ground truth labels. In the ground truth image, the green region represents edema; the yellow region represents enhanced tumor; and the red region represents necrotic and non-enhancing tumor. T1, T1-weighted; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; FLAIR, fluid-attenuated inversion recovery; GT, ground truth; MRI, magnetic resonance imaging.

of brain tumors. For example, Badjie *et al.* (9) proposed a new model combining convolutional neural network (CNN) and AlexNet to enhance brain tumor detection and classification in MR images. Khan *et al.* (10) proposed a brain tumor detection and classification framework based on saliency map and deep learning feature optimization, and designed an improved dragonfly optimization algorithm for optimal feature selection, but their framework is limited by incorrect tumor segmentation. Later, Kurdi *et al.* (11) used the Harris Hawks optimized convolutional neural network (HHOCNN) method to classify brain tumors, which further improved the overall tumor recognition accuracy. The focus of these works was still mainly on the task of brain tumor classification, but the current image segmentation research is also becoming more and more important, such as the realization of CNN-based joint segmentation and classification to improve the disease detection accuracy of benchmark and clinical images (12), or the use of deep learning methods to segment brain tumor images.

Deep learning is widely used in various computer vision problems because of its strong feature extraction ability and high adaptability. CNNs show good performance in brain tumor segmentation. Mainstream CNN-based networks with an encoder–decoder structure such as 3D U-Net (13), Attention U-Net (14), and V-net (15) provide promising results for 2-dimensional (2D) and 3-dimensional (3D) brain tumor segmentation tasks. However, CNN-based methods also have limitations. The localized nature of convolution limits the model when dealing with long-range dependencies, resulting in insufficient ability to learn global features and remote features, which are essential for accurately segmenting tumors of different types and sizes. To address this issue, Transformer (16) has been rapidly applied to brain tumor MRI image segmentation tasks with good performance by establishing connections between token features and employing a self-attention mechanism to be able to better capture global contextual information. Compared with the traditional CNN, the Transformer model solves the local inductive bias and improves the ability to deal with non-local interactions by introducing a self-attention mechanism. Existing Transformer-based brain tumor segmentation methods can compensate for the problem of limited receptive field of traditional CNN, but the computational cost is high (17,18). To reduce the computational complexity of Transformer, Swin Transformer (19) utilizes window-based self-attention to reduce parameters and computation, as well as a shifted-window mechanism for global dependency modeling. Inspired by this, Hatamizadeh *et al.* (20) proposed Swin UNETR (UNEt TRansformers), a new architecture for the semantic segmentation of brain tumors by using multimodal MRI images. It uses Swin Transformer as an encoder and a CNN-based decoder. Among the techniques using artificial neural networks, there are also various preprocessing techniques such as data augmentation, transfer learning, and test time augmentation that can improve the performance of classifiers with limited training data. Transfer learning can accelerate the CNN training process by using pre-trained models. Deep CNN combined with transfer learning techniques can achieve intelligent recognition and classification of brain tumors, which helps to reduce the work intensity of doctors (21). Data augmentation processes can solve problems with limited datasets and improve classification performance, and are often incorporated into model training (22). The combined use of these techniques can bring more possibilities and development opportunities for research and practice in areas such as medical image analysis. In addition, multimodal joint learning can help the model to mine the relationship between modal data and establish complementary links between modalities (23). This strategy is widely used in the field of medical image analysis (24). Compared with using unimodal MRI, segmentation using multimodal MRI can fully utilize the complementary information among different modalities and significantly improve the segmentation performance (25). To date, numerous multimodal methods for automated brain tumor segmentation have been proposed (26-28). Among them, a common way to use multimodal MRI images is to directly concatenate them in the channel dimension as inputs early in the network, but it does not take full advantage of the complementary relationship among different modal images (26). To address this problem, recent works have improved the performance and expressiveness of the model by using high-level features extracted from each modality. They use a hierarchical fusion strategy in the middle layers of the network (27,28). Xing *et al.* (27) proposed a nested Modality Aware Feature Aggregation module for multimodal fusion, which utilizes a nested Transformer to establish intra- and inter-modal long-range correlations for a more efficient feature representation. Considering the significant differences between modalities, TranSiam (29) uses 2 parallel CNNs to extract features specific to each modality, and fuses the features of different modalities through a locally-aware aggregation block to establish correlations between the features of different modalities. However, it focuses on

generating shared representations by fusing multi-modal data, without considering modality-specific features. F^2Net (30) has an obvious advantage, that is, it retains modal-specific features while exploring complementary information, thereby improving segmentation performance. The methods discussed above have a better segmentation effect when the multimodal MRI data are complete. However, considering factors such as scanning cost and data quality, not all patients have complete multimodal MRI data. As a result, the problem of missing modalities arises, and segmentation performance is reduced. Some existing brain tumor segmentation methods (31,32) can use only data with complete modalities to complete the segmentation task, which cannot be directly applied to modality missing scenarios in clinical practice. Therefore, the development a robust multimodal brain tumor segmentation method is urgently needed to solve the missing modality problem.

Current mainstream segmentation methods for dealing with missing modalities can be roughly divided into 2 kinds. One relatively straightforward approach is to train a dedicated network for each possible combination of MRI modalities (8), which is a complex and inefficient training process despite better segmentation performance. Another more typical solution is to synthesize the missing modalities by generative modeling and then using the complete modalities for segmentation (33,34). For example, Zhan *et al.* (33) proposed Multi-Scale Gate Mergence Generative Adversarial Network (MGM-GAN), an end-to-end multimodal MRI synthesis architecture, based on conditional generation adversarial networks. The model extracts valuable features from multiple MRI modalities and is used to synthesize missing or corrupted images in the clinic. However, this approach requires training the generative model for each missing modality, which not only introduces additional computational cost and manual intermediate steps, but also the generative network training faces an instability problem that may lead to unsatisfactory synthesis results. Ultimately, the segmentation results are affected. The third solution to incomplete modality is knowledge distillation (35,36). Li *et al.* (36) proposed a Deeply-supervised knowledGE tranSfer neTwork (DIGEST), constructing a knowledge transfer learning framework that enables a student model to learn modality-sharing semantic information from an instructor model pre-trained with complete multimodal MRI data. It can achieve accurate brain tumor segmentation in different modality-absent scenarios. However, the performance and robustness of the student model may be affected by the teacher model. Similarly, D^2-Net (37) segments brain tumors with missing modalities and identifies the relationship between the state and the tumor region by aligning the characteristics of the disentangled binary teacher network with the overall student network. Meanwhile, Qiu *et al.* (38) proposed a Category Aware Group Self-Support Learning framework for incomplete multi-modal brain tumor segmentation. Unlike previous distillation algorithms, the framework does not use a larger network as the teacher, but a self-supporting group of students as the teacher, and no new models or parameters are introduced. The first 'dedicated' method performs well in segmentation effect; however, they are resource-intensive. Since there may be various missing modes, these methods usually need to train multiple networks, which requires substantial time and computer resources. The synthetic network training in the second method requires certain computing and storage resources, and its performance is affected by the quality of the synthetic image of the missing mode. The third method based on knowledge distillation needs to train a series of student models for each missing modality, which leads to huge space and time costs. At the same time, this method also has the risk of inaccurate information caused by direct transmission of knowledge between modalities. Although knowledge distillation can mine useful knowledge from the teacher network to guide the feature learning of the student network, it is still challenging to enable the student model to obtain key feature representations from all modalities. Recent approaches have attempted to create a shared feature space to retrieve missing information and establish a unified model for all missing modality cases (39-41). Specifically, Havaei *et al.* (40) proposed Hetero-Modal Image Segmentation (HeMIS). It trains a feature extractor for each modality and computes first-order moments and second-order moments of the available features to fuse multimodal information and deal with missing modalities. After the HeMIS network, Dorent *et al.* (41) proposed a Hetero-Modal Variational Encoder-Decoder (U-HVED), which extends the Multi-Modal Variational Auto-Encoder and has been shown to be robust to missing modes. However, these 2 methods may not be able to aggregate features efficiently and perform poorly when critical modes are missing. Zhang *et al.* (42) proposed a Transformer-based incomplete multimodal learning approach for brain tumor segmentation, aiming to fuse the features of all modalities into more comprehensive features, but it ignores the use of the correlation between different modalities. Region-aware Fusion Network for Incomplete Multi-modal Brain Tumor

Segmentation (RFNet) (43) achieves area-aware fusion through a tumor area-aware module, but fails to establish cross-modal long-range dependencies. Furthermore, Konwer *et al.* (44) proposed a new training strategy to solve the missing problem in brain tumor segmentation under limited full-modal supervision through metal learning and modal combination as a meta-task. These previous works are cleverly designed and robust under missing modality conditions, but these methods still need to rely on complete modalities during training. Unlike previous methods, our approach is effective and resource efficient because it does not require training specific models for individual missing modality cases. Meanwhile, our framework simulates only the missing modalities by randomly "dropping out" the imaging modalities throughout the training process without changing the model structure.

To overcome the above limitations, the present study aimed to adopt a more efficient incomplete multimodal learning strategy to learn modality-invariant shared-feature representations by fusing any number of available modalities in the latent space. A unified model was then established for all possible missing-modality scenarios. We propose a deep CNN [cross-modal attention fusion-based deep neural network (CMAF-Net)] consisting of a 3D U-Net (13) and a cross-modal attention fusion (CMAF) module for brain tumor segmentation with missing modalities. Our network can automatically capture complementary information from incomplete multimodal data and learn brain tumor shared feature representations to handle the problem of missing modalities in real-world scenarios. In our design, we used 4 independent coding branches for each modality, where the convolutional block and 3D Swin block (45) learn local features and global dependencies, respectively. Then, we designed a CMAF module, which can effectively utilize the complementary information between different modalities to learn modality-shared potential representations. Finally, the modality-invariant shared features obtained after the fusion stage were upsampled to obtain the final segmentation results. Extensive experimental results on BraTS 2018 and 2020 (5) demonstrated that CMAF-Net exhibited a superior performance compared to other popular networks in brain tumor segmentation under various missing modality settings. The main contributions of this paper are summarized as follows:

- We developed a multi-encoder and single-decoder brain tumor segmentation network called CMAF-Net to solve the incomplete multimodal brain tumor segmentation problem. Particularly, we designed a new CMAF module that can adequately model the correlation between multimodal data and handle the case of missing modalities by fusing multimodal features to learn modality-invariant shared representations.
- To fully utilize the respective advantages of Convolution and Transformer for better segmentation performance, we proposed a shift-window based 3D Swin block to capture long-distance contextual interactions while maintaining low computational complexity.
- In order to segment brain tumors more accurately, we introduced a spatial and channel self-attention module in the decoding stage to gradually aggregate multimodal and multi-level features. With these 2 modules, the network can selectively emphasize informational features while suppressing less useful features.

The rest of this paper is organized as follows. The 'Methods' section elaborates the overall framework of the proposed CMAF-Net and the detailed structure of each module and the 'Experiments and results' section describes the experimental setup and details, as well as provides the experimental results and analysis. The 'Discussion' section presents the discussion of study implications and offers future research directions. In the last section, the conclusion is presented.

Methods

The proposed network architecture is first introduced. Each component including the 3D Swin block and the CMAF module is then described in detail. Finally, details of the loss function and hyperparameters used in the training phase are discussed. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Network architecture

To overcome the challenges of missing modalities in clinical practice, we proposed a brain tumor segmentation algorithm based on CMAF-Net. Our CMAF-Net is based on a multi-encoder 3D U-Net (13) architecture using incomplete multimodal MRI data with $M=\{FLAIR, T2, T1, T1ce\}$. It comprises an encoder stage, a fusion stage, and a decoder stage, as shown in *Figure 2*. In the encoder phase, CMAF-Net takes multimodal MRI images as input. First, the available MRI modalities are fed into each independent

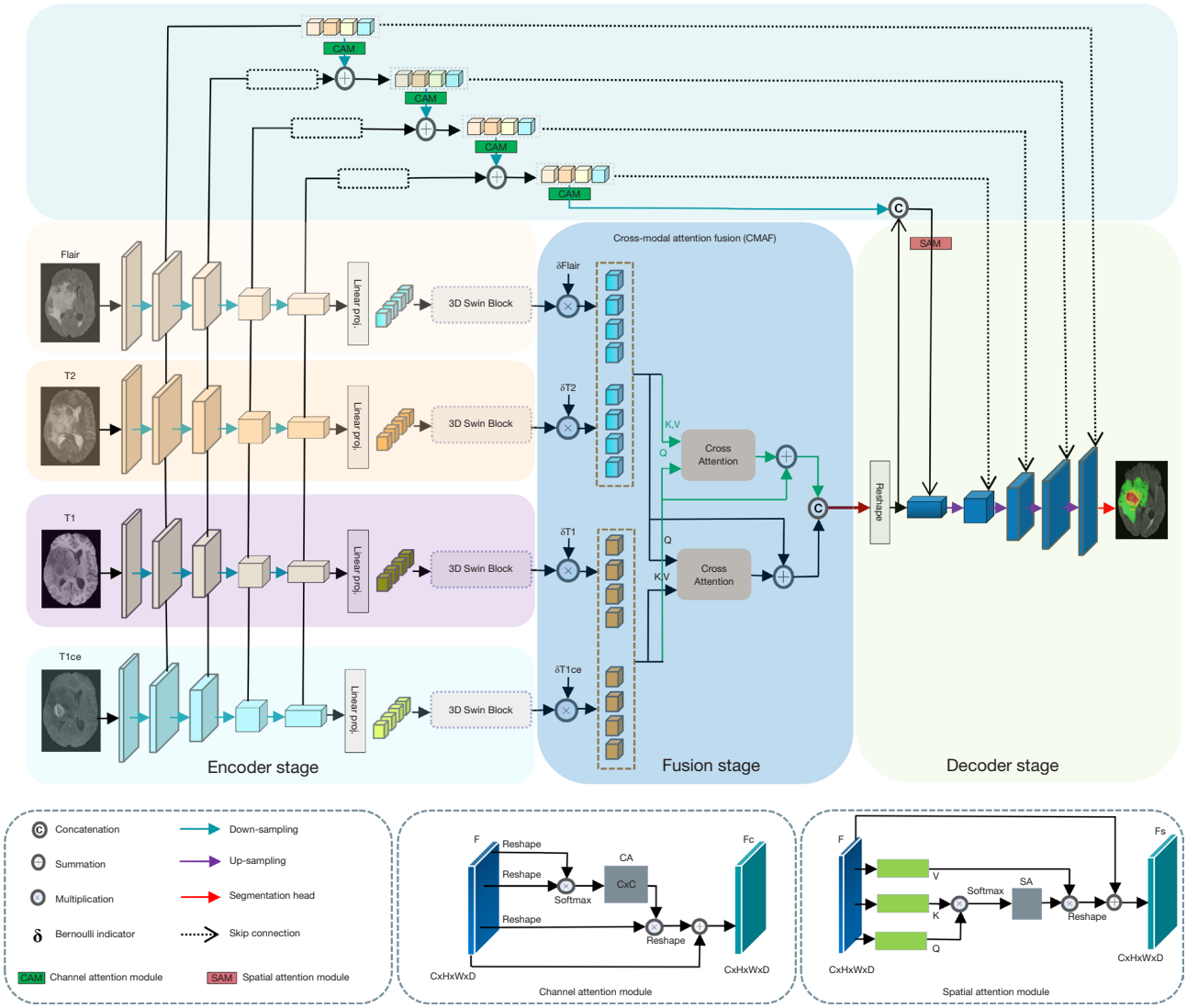

Figure 2 An overview of our proposed network architecture, including encoder stage, fusion stage, and decoder stage. 3D Swin block is used to capture global information. The CMAF block is used to fuse multimodal features. CAM, channel attention module; SAM, spatial attention module; CMAF, cross-modal attention fusion; T2, T2-weighted; T1, T1-weighted; T1ce, contrast-enhanced T1-weighted; FLAIR, fluid-attenuated inversion recovery; 3D, three-dimensional; F, input feature map; C, channels of the feature map; H, height of feature map; W, width of feature map; D, depth of feature map; CA, channel attention; Fc, feature map weighted by channel attention; Fs, feature map weighted by spatial attention; SA, spatial attention; V, value; K, key; Q, query.

encoder to extract the features of different modalities. Then, the local feature maps of each modality generated by the convolutional encoder are then fed into the constructed 3D Swin block to model the long-range correlation. In the multimodal fusion stage, we proposed a new multimodal fusion module called CMAF module. It efficiently fuses features among different modalities to learn a modality-invariant shared representation corresponding with the tumor region and cope with the situations of missing modalities. Finally, the decoder progressively aggregates multimodal and multilevel features by using spatial and channel self-attention modules and then upsamples the fused features to the original input resolution to make pixel-level predictions.

Encoder stage

In the encoder stage, the hybrid encoder can efficiently bridge the convolutional encoder and the 3D Swin block to extract local and global features within a specific modality, respectively. Specifically, given a multimodal 3D MRI image $x=\{x_m\}_{m=1}^{M}$, $x\in\Re^{H\times W\times D\times 1}$, where M is the number of modalities, D is the number of slices, H and W are the height and width of the inputs, respectively. A convolutional encoder is used to extract rich local 3D contextual features, and then 3D Swin blocks are constructed to model the long-range dependencies of the extracted feature maps.

Convolutional encoder

The convolution encoder is stacked by 3×3×3 convolution blocks and downsampled by using a convolution with a step size of 2. Specifically, we construct a five-stage encoder, each stage comprises two convolution blocks through the GroupNorm + ReLu + 3D Conv structure. The encoder gradually encodes each modal image into a low-resolution feature map with local context $F_m^{L-local}\in\Re^{C\times\frac{D}{16}\times\frac{H}{16}\times\frac{W}{16}}$, $m\in[1,M]$, where $\left(\frac{D}{16},\frac{H}{16},\frac{D}{16}\right)$ is 1/16 of the input spatial resolution H, W, and depth dimension D; M is the number of modal images. The channel dimension C and the number of stages L in the encoder are set to 128 and 5, respectively.

3D Swin block

Traditional CNNs can accurately segment brain tumors in local MRI images, but this method may not perform well for challenges such as artifacts that may exist in brain tumor MRI images. For dense prediction tasks such as brain tumor segmentation, it is crucial to consider both local and global information. CNNs excel in extracting local information, but they have limited capability in extracting global information. Relying solely on local information may not effectively solve some of the complex problems mentioned above. Several studies have confirmed that incorporating global information can enhance the model's ability to capture correlations among different parts of the image and help improve the segmentation effect (17,46). CNN and Transformers are 2 mainstream architectures. CNN is proficient at extracting local features through convolutional operations, but its limited receptive field hinders its ability to capture global representations. Based on the shift window mechanism, Swin Transformer retains the powerful global information extraction ability of the Transformer. It offers the potential to address the aforementioned issues by efficiently processing information that deserves attention through the self-attention mechanism. Therefore, to address the challenge posed by the convolutional encoder's inability to capture long-range dependencies within each modality, we proposed a 3D Swin block inspired by the module design in the Swin Transformer (19). This block incorporates a multi-head self-attention (MSA) mechanism to facilitate effective long-range context modeling, thereby enhancing the model's capacity to comprehend and process global information. The main difference is that we extend the Swin Transformer block to 3D structures. An important factor limiting the Transformer (16) structure's application in medical image tasks is the consumption of excessive computational resources and storage space when an entire image is converted into a sequence for self-attention computation. Consequently, inefficiency and performance degradation ensue. Conversely, Swin Transformer (19) is a method of calculating self-attention within a local window based on a shifted window, which greatly reduces the number of parameters while showing better feature-learning ability. The structure of the 3D Swin block is shown in *Figure 3*. It consists of 2 basic units, and each subunit comprises a LayerNorm layer, an attention module, a LayerNorm layer, and an MLP module. The first subunit uses the 3D windowed multi-head self-attention (3D W-MSA) module, whereas the second one uses the 3D shifted window multi-head self-attention mechanism (3D SW-MSA) module.

The 3D Swin block processes embeddings in a sequence-to-sequence manner based on an attention mechanism. The linear embedding layer is responsible for converting the local feature map $F_m^{L-local}$ generated by the convolutional encoder into non-overlapping patches. Each patch is considered a "token". Subsequently, the patches are mapped to C (C is set to 512) dimensional vectors. For efficient modeling, we propose to compute self-attention within a local window. By arranging the windows, the input feature maps are evenly divided in a non-overlapping manner. Assuming that each window contains $M\times M\times M$ patches, the standard multi-head self-attention computation is performed separately for each window. In the computation of W-MSA, information exchange is lacking due to the non-overlapping windows being unconnected to one another, which limits its modeling capability. To solve this problem, cross-window connectivity is introduced to transfer information among neighboring windows in SW-MSA while maintaining

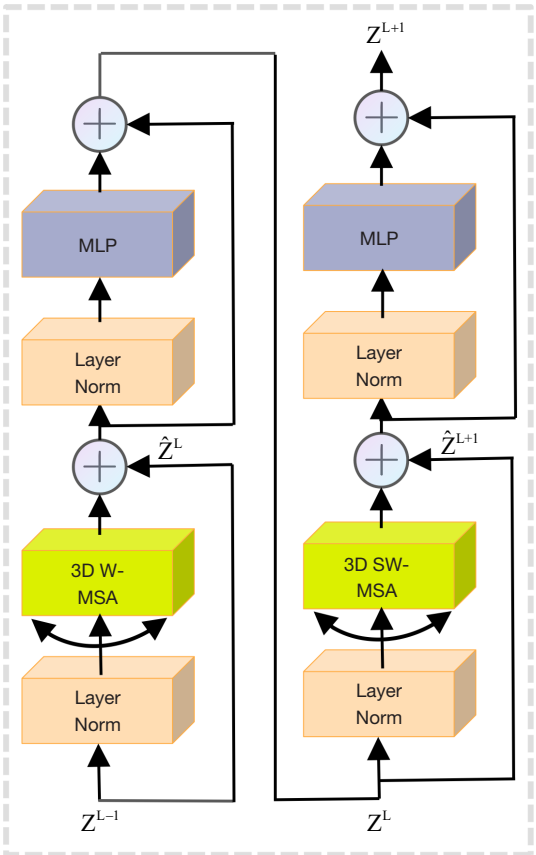

Figure 3 The structure of 3D Swin block. MLP, multi-layer perceptron; 3D W-MSA, 3D windowed multi-head self-attention; 3D SW-MSA, 3D shifted window multi-head self-attention mechanism; 3D, three-dimensional.

effective computation in non-overlapping windows. In SW-MSA, the $8\times8\times8$ feature map is uniformly partitioned into $2\times2\times2$ windows of size $4\times4\times4(M=4)$. First, MSA calculation is performed on each window. Second, the shift window mechanism is used to realize the communication between windows to achieve the effect of global modeling. Through the above operations, information exchange between windows can be realized. We used W-MSA and SW-MSA alternately to achieve cross-window connection. Under the shift structure, the entire calculation process of 3D Swin Block is as follows:

$$\hat{Z}^{L}=3DW-MSA\left(LN\left(Z^{L-1}\right)\right)+Z^{L-1} \quad [1]$$

$$Z^{L}=MLP\left(LN\left(\hat{Z}^{L}\right)\right)+\hat{Z}^{L} \quad [2]$$

$$\hat{Z}^{L+1}=3DSW-MSA\left(LN\left(Z^{L}\right)\right)+Z^{L} \quad [3]$$

$$Z^{L+1}=MLP\left(LN\left(\hat{Z}^{L+1}\right)\right)+\hat{Z}^{L+1} \quad [4]$$

where L denotes the block layer; $\hat{Z}^{L}$ and $\hat{Z}^{L+1}$ denote the output features of the (S)W-MSA module and the MLP module, respectively. As described above, 3D Swin block can effectively extract global and long-range dependencies through the self-attention mechanism and generates the feature maps F_m^{global} with global context within each modality through 3D Swin blocks. The W-MSA module and the SW-MSA module primarily comprise the multi-head self-attention mechanism and the trainable relative position encoding. The specific formula is shown in Eq. [5]:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{T}}{\sqrt{d_k}}+B\right)V \quad [5]$$

where Q, K, and V are the query, key, and value, respectively; d_k is the dimension of the key; and B is the learnable relative position encoding.

Fusion stage

According to the clinical knowledge of radiologists diagnosing brain tumors from multiple MRI modalities, a strong structural correlation is known to exist between imaging modalities (47). The T1 and T1ce modalities are usually paired to detect the tumor core region, and the FLAIR and T2 modalities are often interpreted jointly. Therefore, a relatively strong correlation may exist between these 2 pairs of modalities. In the fusion stage, we consider T1, T1ce as 1 modality (T1-T1ce) and FLAIR, T2 as the other modality (FLAIR-T2) because the difference between them is very small (48). We cascade the embeddings from modality-specific encoders into input multimodal tokens by combining them 2-by-2 based on the correlation between the modalities, defined respectively as follows:

$$FLAIR\text{-}T2^{token}=\left[\delta_{Flair}F_{Flair}^{global},\delta_{T2}F_{T2}^{global}\right] \quad [6]$$

$$T1\text{-}T1ce^{token}=\left[\delta_{T1ce}F_{T1ce}^{global},\delta_{T1}F_{T1}^{global}\right] \quad [7]$$

where $(\cdot,\cdot)$ indicates the concatenation operation. During training, we randomly set δ_m to 0 to simulate the missing modality. In case of missing modalities, the token for the missing modalities is replaced by a zero vector.

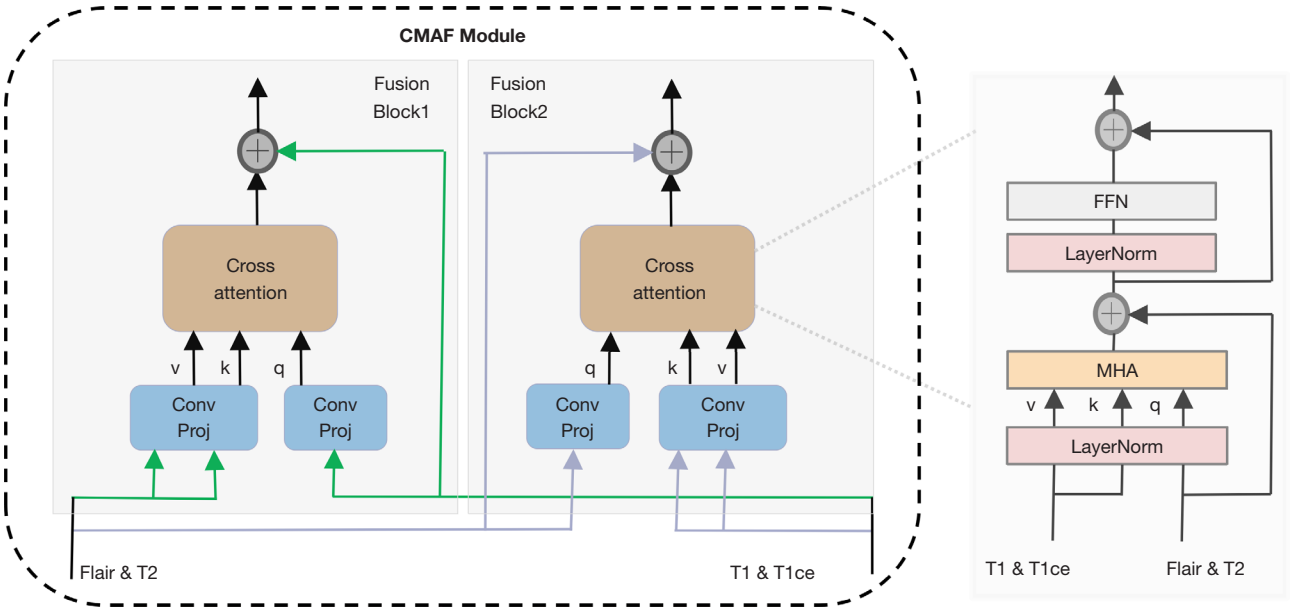

Figure 4 The structure of CMAF Module. It consists of fusion block1 and fusion block2. CMAF, cross-modal attention fusion; V, value; K, key; Q, query; Conv Proj, convolutional projection; FLAIR, fluid-attenuated inversion recovery; T2, T2-weighted; T1, T1-weighted; T1ce, contrast-enhanced T1-weighted; FFN, feed forward network; MHA, multi-head attention.

Multimodal data, which offer a more comprehensive understanding of information than any single modality alone, often encounter the challenge of incomplete modalities. Therefore, the effective integration of these multimodal data has become a crucial concern, particularly in scenarios where modalities are incomplete. Multimodal data fusion can make full use of the complementary information among multiple modal data to improve the segmentation accuracy. Accordingly, we propose a new cross-attention-based fusion mechanism, called CMAF module, which can fully model the correlation between multimodal data and handle the case of missing modalities by fusing multimodal features to learn modality-invariant shared representations. The motivation behind the CMAF module is to mine the complementary information between multimodal data to stably establish long-range dependency across modalities while learning from incomplete modalities more flexibly. The CMAF module is described in detail below.

CMAF module

The CMAF module first receives the features of 2 different modalities and subsequently fuses the features of the 2 modalities by cross-attention. As shown in *Figure 4*, the CMAF module comprises 2 fusion blocks. Fusion block 1 fuses the features of T1-T1ce into FLAIR-T2. Fusion block 2 fuses FLAIR-T2 to T1-T1ce. In the interactive fusion of available modality information, we use the principle of cross-attention (49) to realize the information interaction among different modalities. We have used fusion block 1 as an example for illustration. We initially use linear transformation to map the features of T1-T1ce to Q (query) vectors followed by the features of FLAIR and T2 to K (key) and V (value) vectors. The input triple $(\text{Query}, \text{Key}, \text{Value})$ of cross-attention is calculated as follows:

$$\begin{aligned} Q &= T1\text{-}T1ce^{token} W_q, \\ K &= FLAIR\text{-}T2^{token} W_k, \\ V &= FLAIR\text{-}T2^{token} W_v \end{aligned} \quad [8]$$

where W_q, W_k, and W_v are the weight matrix. For Q from $T1\text{-}T1ce$, it fuses cross-modal information by attention weighting with K and V from $FLAIR\text{-}T2$ while retaining $T1\text{-}T1ce$ modality-specific information through residual connections, and vice versa. In other words, cross-attention highlights similar features between the 2 modalities by establishing correlations between them. Considering that these features can be extracted from both modality data, they should have high confidence (29) and are important for

representing correlation and shared information between modalities. Given a token sequence from 2 different modality images, the cross-attention (CA) is denoted as follows:

$$CA\left(T1\text{-}T1ce^{token}, FLAIR\text{-}T2^{token}\right) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad [9]$$

$$y_1 = CA\left(T1\text{-}T1ce^{token}, FLAIR\text{-}T2^{token}\right) + T1\text{-}T1ce^{token} \quad [10]$$

where $FLAIR\text{-}T2^{token}$ denotes the features of $FLAIR$ and $T2$, $T1\text{-}T1ce^{token}$ denotes the features of $T1$ and $T1ce$, y_1 denotes the output of fusion block 1, and $CA\left(T1\text{-}T1ce^{token}, FLAIR\text{-}T2^{token}\right)$ denotes cross-attention.

Decoder stage

In the decoder stage, we first cascade the outputs y_1 and y_2 from the two fusion blocks of the fusion stage. This process can be expressed as follows:

$$F^{global} = Concat\left(y_1, y_2\right) \quad [11]$$

where y_1 and y_2 denote the fused deep features of CMAF module output, respectively, and $Concat(\cdot)$ represents the concatenation of channel dimensions.

To enhance the accuracy and robustness of the proposed model, we developed 2 independent self-attention modules during the decoding stage: the channel attention module (CAM) and the spatial attention module (SAM). These modules were designed to specifically address channel characteristics and spatial location dependencies, respectively. CAM and SAM can strengthen inter-channel relationships and obtain a more comprehensive contextual representation. For example, the literature (50) introduced an innovative attention mechanism that seamlessly integrates features from 2 different perspectives: modal and spatial, for the purpose of multimodal fusion. This innovative fusion approach has been shown to be highly effective across a range of tasks.

The decoder adopts a structure symmetric to the encoder, and the function of the convolutional decoder is to perform an upsampling operation on the fused feature representations to recover the spatial resolution of the feature map for final pixel-level classification. The output sequence F^{global} of the fusion stage is reshaped into a high-level 4D feature map corresponding with the size before flattening. To obtain spatially detailed features that enrich the available modalities, we progressively cascade the feature maps obtained at each level of the encoder for different modalities and use skip connection to fuse the encoder and decoder features for finer segmentation. To obtain a more comprehensive representation of shared features and to fully utilize the information between multiple modalities, a CAM is used to integrate low-level multimodal features. Each channel map is regarded as a modality-specific feature map, with features from different modalities interacting and correlating with each other, which means that the network can better understand the correlations between different modalities, thus improving the representation of shared feature maps for tumor regions. To model the interaction of multilevel features, low-level shared features from the CAM are downsampled with higher-level multimodal features cascaded to the decoder for brain tumor segmentation. Meanwhile, the multimodal fusion features F^{global} from the fusion stage are cascaded with the previous level of shared features learned by the channel self-attention module. The SAM is then used to model the long-range spatial relationships of the higher-level features.

Loss function

The performance and robustness of the model can be improved by using a joint loss function. Cross-entropy loss function is a basic loss function for medical image segmentation. However, the cross-entropy loss function does not easily deal with category imbalance. Dice loss is used to evaluate the overlap between the predictions and the ground truth, which can reduce the effect of category imbalance. Therefore, for end-to-end training of the entire network, we use a loss that is the sum of the cross-entropy L_{cross} and the dice loss function L_{dice}. This loss can be formulated as follows:

$$L = L_{dice} + L_{cross} \quad [12]$$

$$L_{dice} = 1 - 2\frac{\sum_{i=1}^{C}\sum_{j=1}^{N} p_{ij} g_{ij} + \varepsilon}{\sum_{i=1}^{C}\sum_{j=1}^{N}\left(P_{ij} + g_{ij}\right) + \varepsilon} \quad [13]$$

$$L_{cross} = -\sum_x g_l(x)\log\left(p_l(x)\right) \quad [14]$$

where N is the number of samples, C is the number of segmentation classes, P_{ij} is the predicted probability that pixel i belongs to tumor class j, g_{ij} is the actual probability that pixel i belongs to tumor class j, ε is a small constant to avoid being divided by 0, $p_l(x)$ is the estimated probability that pixel x belongs to class l, and $g_l(x)$ is the

Table 1 Hyperparameters of the proposed CMAF-Net model

Hyperparameters	Values
Optimizer	AdamW
Epochs	800
Batchsize	1
Initial learning rate	1e−4
Weight decay	1e−5

CMAF-Net, cross-modal attention fusion based deep neural network.

truth label of pixel x.

Hyperparameters selection

This subsection describes the hyperparameter settings chosen to produce efficient results. We selected the hyperparameters for model training after extensive experimentation. *Table 1* lists the hyperparameters used for the training of the CMAF-Net model, where the learning rate is tuned smaller to fit the other hyperparameters. AdamW (27) can help the model converge faster. A batch size of 1 reduces the demand for computational resources.

We chose AdamW (27) as the optimizer to achieve the best loss reduction during the training process. This optimization technique has an adaptive learning rate and weight decay mechanism, which can improve the model performance by automatically adjusting the learning rate and controlling the model complexity according to the different conditions of the parameters during the training process. As opposed to other optimization techniques (e.g., Sgdm or RMSprop), we choose AdamW because of its easy implementation, efficient memory utilization and faster learning rate. Recently, AdamW has shown excellent performance in deep learning applications such as medical image analysis (51).

Results

We briefly describe herein the dataset and evaluation metrics used for the experiments and then present a description of the implementation details. We demonstrate the effectiveness of our proposed CMAF-Net method by comparing it with other state-of-the-art methods on the Brain Tumor Segmentation Challenge (BraTS) datasets. Ablation experiments are performed to evaluate the effectiveness of each component of our CMAF-Net.

Datasets and evaluation metric

Datasets

We evaluated our CMAF-Net on the publicly available brain tumor segmentation datasets BraTS 2018 and BraTS 2020. BraTS 2018 and 2020 contain 285 and 369 annotated brain tumor samples, respectively. Each case contains 4 aligned modalities including FLAIR, T1, T1ce, and T2, as well as expert manually segmented labeled images. The ground-truth labels provided by BraTS include enhanced tumor (ET), edema (ED), necrosis, and non-enhancing tumor core (NCR/NET) regions. Each modality is 240×240×155 in size and has been skull stripped, co-registered, and resampled to a resolution of 1 mm^3. To provide a fair comparison, we divided the 285 case samples in the BraTS 2018 dataset into 190 cases as the training set and the other 95 cases as the validation set using the same split list as in (41), and we conducted experiments using 3-fold cross-validation. Regarding the BRATS 2020 dataset containing 369 training samples, we divided the dataset according to the ratio 220:74:75 (training/validation/testing) in (36). We compared our method with other methods on 2 datasets. In addition, as BraTS 2020 has a larger dataset size compared with BraTS 2018, our ablation experiments were mainly conducted on this dataset in this study.

Evaluation metrics

In our experiments, we used Dice similarity coefficient (DSC) and Hausdorff distance, which are commonly used in brain tumor segmentation research, to evaluate the segmentation results. The DSC showed the similarity between the prediction results and the ground truth. A higher value of DSC corresponded with a better segmentation results. The calculation formula of DSC is as follows:

$$DSC = \frac{2|P \cap T|}{|P| + |T|} \quad [15]$$

where P is the prediction result of the model, and T is the ground-truth label.

The Hausdorff distance metric is a measure of boundary similarity to evaluate the distance between model prediction region and ground truth region. The lower the value of Hausdorff distance, the better the predicted mask. Hausdorff distance is defined as:

$$Hausdorff\ distance = \max\left\{\sup_{p\in\partial P} d_{\min}(g,p), \sup_{g\in\partial G} d_{\min}(p,g)\right\} \quad [16]$$

where ∂P and ∂G denote the tumor boundary point sets of the network prediction and the ground truth, respectively,

and $d_{\min}(g,p)$ indicates the minimum Euclidean distances between voxel g and voxels in a set P.

Implementation details

Our CMAF-Net was implemented on the PyTorch framework by using an Nvidia RTX 3090 (Nvidia, Santa Clara, CA, USA). We used 3D U-Net (13) as the backbone network for CMAF-Net. During model training, the AdamW optimizer (27) was used for optimization with an initial learning rate of 0.0001. The input image size was 128×128×128 voxels, the batch size was set to 1, and the number of training times was 800 epochs. A simple data enhancement strategy including random rotation, cropping, and random flipping was used. Following (42), we utilized a modality missing mask to discard modality and simulate various missing modalities scenarios. During the training process, we randomly set some modalities as zero matrices to simulate the case of missing modalities (41). In the testing phase, for the case of missing modalities, we set the missing modalities as zero vectors.

Comparison with state-of-the-art methods on incomplete MRI data

We performed a series of comparative experiments on the BraTS 2018 and BraTS 2020 datasets to demonstrate the effectiveness of our proposed method. In Section (Comparison on BRATS 2020), we compare our method with other state-of-the-art brain tumor segmentation methods with missing modalities. In Section (Comparison on BRATS 2018), we compare our method with state-of-the-art full-modality methods.

Comparison on BRATS 2020

To evaluate the effectiveness of our proposed approach, we compared our proposed CMAF-Net with several SOTA models. These models include 2 representative models U-HeMIS (40) and U-HVED (41) that learn shared feature representations in the latent space, DIGEST (36), D^2-Net (37), and GSS (38) that are based on teacher-student distillation framework, RFNet (43) based on CNN, and mmFormer (42) based on Transformer. For a fair comparison, we used the same data split (36) and reference the results directly. At the same time, we reproduced the results of mmFormer (42) on our data split by running the author's code. *Table 2* shows the quantitative results of all methods on the BraTS 2020 dataset for 15 different missing MRI modalities. *Table 2* shows that the average results of our method in most missing modality cases outperform those of the previous methods. Our method achieves the highest values for the average DSC metrics in the 3 nested tumor regions, namely, 87.9%, 81.8%, and 64.3%, respectively. Notably, in the difficult-to-segment tumor core region, our method outperformed other methods on most subsets of missing modalities. This further supported the effectiveness of CMAF-Net for richer cross-modal information learning, which is more conducive to dealing with cases with missing modalities. Our CMAF-Net performed better in almost all cases where 2 or 3 modalities were missing. Moreover, on tumor segmentation with only 1 available modality, Dice was improved by 12.1% on average. This finding suggests that the proposed CMAF-Net can successfully learn potential shared feature representations for brain tumor segmentation, showing the effectiveness and robustness of the proposed model. *Figure 5* provides a visual representation in the form of a bar graph, comparing the segmentation results predicted by our network and other methods on the BraTS 2020 dataset. It is illustrated that our method achieved outstanding segmentation performance on this dataset. For qualitative analysis, *Figure 6* compares the segmentation results of our method with existing ones (40), (41), and (42) for a different number of missing modalities from the same patient. We can see that our method produced more accurate segmentation results in most of the cases and especially performs more clearly with stronger generalization ability when only 1 modality was available.

To evaluate the robustness of our model to missing modalities, we randomly selected several cases from the BraTS 2020 dataset. The segmentation results in the case of missing modalities are shown in *Figure* 7. With increased number of missing modalities, the segmentation results progressively deteriorated; however, the decline was not sudden and sharp, so we can still obtain a good segmentation result. Furthermore, even with only 1 FLAIR modality available, our CMAF-Net was able to segment the brain tumor appropriately.

Comparison on BRATS 2018

To further evaluate the proposed method, we also compared the segmentation performance with the state-of-the-art methods using 3-fold cross-validation on the BraTS 2018 dataset. The comparison results are shown in *Table 3*. From *Table 3*, we can see that our proposed model obtained the best average DSC in all 3 tumor regions compared to other methods. Our method outperformed

Table 2 Comparison with different methods such as U-HeMIS (40), U-HVED (41), DIGEST (36), D^2-Net (37), mmFormer (42), RFNet (43), and GSS (38) on the BraTS 2020 dataset

Methods	Type: FLAIR/T1/T1c/T2															AVG
	○/○/○/●	○/○/●/○	○/●/○/○	●/○/○/○	○/○/●/●	○/●/●/○	●/●/○/○	○/●/○/●	●/○/○/●	●/○/●/○	●/●/●/○	●/●/○/●	●/○/●/●	○/●/●/●	●/●/●/●	
WT																
U-HeMIS	76.7	65.2	58.8	79.6	81.1	69.8	83.2	79.2	85.4	85.2	86.4	86.9	87.9	83.1	88.5	79.8
U-HVED	80.9	65.4	83.5*	83.5	83.5	69.7	84.3	81.7	87.6	86.5	86.7	88.2	89.2	84.0	89.2	82.9
DIGEST	78.6	74.4	68.6	84.2	85.5	77.6	85.8	84.3	87.8	88.6	88.5	88.8	89.7	86.0	90.2	83.9
D^2-Net	77.8	65.5	46.6	81.3	82.0	76.5	83.8	80.4	86.2	85.0	88.1	89.0	89.2	82.5	89.4	80.2
mmFormer	85.8	78.1	79.2	85.8	87.5	80.7	88.1	87.4	89.9	89.3	89.7	89.8	90.2	88.0	90.1	86.6
RFNet	86.1	78.5	79.7	86.3	87.9	81.4	89.0	88.2	89.7	89.9	90.0	90.0	90.4	88.2	90.3	87.0
GSS	86.6	78.8	80.3	86.5	88.1	82.0	89.5	87.9	90.0	89.7	90.1	90.2	90.0	88.5	90.4	87.2
Ours	86.9*	81.0*	81.4	88.3*	88.4*	83.1*	89.9*	88.3*	90.2*	90.1*	90.4*	90.4*	90.8*	88.8*	90.9*	87.9*
TC																
U-HeMIS	50.7	68.3	38.9	51.6	74.6	70.8	55.9	52.7	59.6	73.4	74.6	60.7	76.8	76.1	77.1	64.1
U-HVED	57.9	70.7	53.8	53.8	79.5	73.2	55.7	59.2	62.7	76.7	77.0	63.4	80.1	79.7	80.2	68.2
DIGEST	60.1	83.3	54.5	60.4	87.1	84.1	65.0	63.8	67.8	86.1	85.6	67.5	87.2	86.1	87.0*	75.0
D^2-Net	58.8	71.4	43.2	55.5	82.6	79.3	58.6	62.7	68.5	79.7	79.2	65.0	82.2	83.1	81.6	70.1
mmFormer	74.2	86.2	69.6	68.0	87.5	86.3	73.4	75.8	75.9	86.3	86.4	76.2	87.1	87.0	86.5	80.4
RFNet	74.0	86.5	69.8	68.3	87.6*	86.2	73.5	75.6	76.4	86.7	86.6	76.4	87.4*	87.2	86.9	80.6
GSS	74.8	85.8	71.6	71.4	86.5	86.1	75.0	75.4	76.6	86.4	86.8	77.3	87.0	86.7	85.7	80.9
Ours	76.0*	86.6*	72.1*	75.4*	87.4	87.0*	76.3*	76.2*	77.1*	87.5*	87.2*	77.6*	86.7	87.6*	86.8	81.8*
ET																
U-HeMIS	21.2	66.2	13.1	26.5	71.8	68.2	28.4	25.3	20.8	71.6	74.1	33.0	72.9	72.5	73.4	49.3
U-HVED	31.0	66.3	22.5	22.5	72.6	68.1	20.4	30.5	34.6	72.6	72.7	32.7	73.7	72.5	73.6	51.1
DIGEST	39.1	78.3*	35.1	39.8	81.0	79.6	42.9	42.4	44.6	78.7	77.7	45.5	80.3*	82.3*	81.2*	61.9
D^2-Net	20.8	72.5	19.4	21.2	76.7	72.2	25.6	27.5	29.1	70.2	73.3	31.7	70.8	71.6	72.8	50.4
mmFormer	44.9	75.9	40.4	42.2	77.2	77.2	45.3	47.2	49.4	77.0	77.0	49.9	76.6	78.5	76.5	62.3
RFNet	45.0	76.3	41.2	42.4	78.0	77.5	46.1	47.8	48.7	77.6	77.2	50.0	76.8	78.7	76.9	62.7
GSS	45.7	77.0	42.3	42.8	79.3	77.6	47.3	48.9	48.5	77.8	77.4	50.2	77.1	78.5	78.3	63.2
Ours	46.9*	77.6	44.0*	44.9*	81.2*	79.8*	48.1*	49.8*	50.8*	78.8*	77.8*	51.3*	77.6	78.1	77.8	64.3*

The results reported in the table are the DSC for the different missing sceneries. The higher the DSC value, the better the results. ○ and ● indicate the missing and available modalities, respectively. *, the best results. AVG denotes the average results on one target region across all the situations. FLAIR, fluid attenuated inversion recovery; T1, T1-weighted; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; WT, whole tumor; TC, tumor core; ET, enhancing tumor; DSC, dice similarity coefficient.

the current state-of-the-art method, GSS (38), with average DSC increases of 1.6%, 0.5%, and 0.6% in the WT, TC, and ET regions, respectively. Moreover, our proposed method outperformed the compared methods in most of the modal combinations on BRATS 2018, indicating the effectiveness of our proposed method. Since our goal was to also help with segmentation in the absence of modalities, we also compared the results of different numbers of modalities as input with state-of-the-art methods, namely U-HeMIS (40), U-HVED (41), DIGEST (36), D^2-Net (37), mmFormer (42), RFNet (43), and GSS (38). The results are shown in *Table 4*. It can be seen that the proposed method

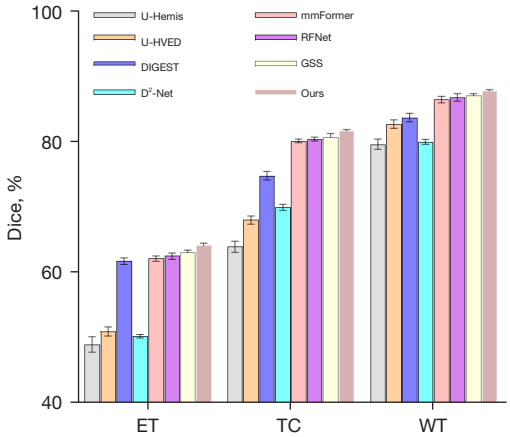

Figure 5 Bar graph of the DSC of comparison experiments on the BraTS 2020 dataset. Error bars show standard error. DSC, dice similarity coefficient; ET, enhancing tumor; TC, tumor core; WT, whole tumor.

is robust and superior to up-to-date approaches GSS (38), regardless of using 1, 2, 3, or 4 modalities as input. With the increase of the number of missing modalities, the improvement of CMAF-Net was more impressive. These results demonstrate the effectiveness of CMAF-Net in incomplete brain tumor segmentation. In *Figure 8*, our method is compared with other methods on the BraTS 2018 dataset using boxplots of DSC obtained from the segmentation results of 15 missing modality cases and histograms generated from average results. The figure illustrates the characterization of the distribution of the segmentation results of the different methods for various modality missing cases.

In addition, we also compare the Hausdorff Distance of our proposed method with other methods on both datasets, as shown in *Table 5*. Compared to other methods, CMAF-Net achieves optimal performance on Hausdorff Distance. The results of the Hausdorff Distance evaluation indicate that each region's boundary is segmented more accurately.

Comparison with the state-of-the-art methods on full modalities

To explore the performance of our model on full modalities, we also compared it with the recent state-of-the-art 3D U-Net (13), Attention U-Net (14), TransBTS (17), F^2Net (30), and Swin-UNETR (20) on multimodal brain tumor segmentation with full MRI modalities. The hyperparameters of all experiments are almost identical. We reproduced the results by using the official codebase. The results of the different methods are compared in *Table 6*. Our method achieved the best results in terms of average Dice scores and average Hausdorff distance for all tumor regions segmented. These results show that our multimodal representation of CMAF-Net learning is robust to missing modalities and effective for full modalities.

Ablation study

In this section, by adding different components individually to the baseline model, we conducted a comprehensive ablation experiment to evaluate the effectiveness of each component in CMAF-Net. The baseline was the proposed method without using the 3D Swin module, the CMAF module, and the hybrid attention module. We present in *Table 7* the quantitative results of the average DSC of the 3 variants with CMAF-Net over all possible subsets of the input modalities. When the 3D Swin module was applied to the baseline, the DSCs of ET, WT, and TC increased by 1.8%, 0.7%, and 0.7%, respectively. The above results show that the proposed 3D Swin block can significantly improve the segmentation performance, which is attributed to the fact that the 3D Swin block can model long-range dependencies and learn global and detailed features. On the basis of the above architecture, the CMAF block was further introduced into the network. The results showed that the DSC is improved in all 3 tumor regions. This finding can be explained by the fact that CMAF further improves the feature learning ability of the model by fully integrating the complementary information among different modalities and establishing long-range dependencies among them. Additionally, a complete segmentation network model was constructed by adding hybrid attention modules (SAM and CAM). The segmentation accuracy was found to further improve, resulting in a 1.1% increase in the average DSC. We effectively demonstrated that SAM can learn the importance of different spatial regions, thereby enhancing the feature representation. CAM can increase the network's attention to key features to further improve the results. As a result, the proposed CMAF-Net shows significant improvement in DSC for all 3 types of tumor regions compared with the baseline, showing that each module contributes to the performance improvement. In order to verify the contribution of each term in our proposed joint loss function, we compared the performance differences of CMAF-Net using different loss functions. It was found that the loss function of CMAF-Net removed any item

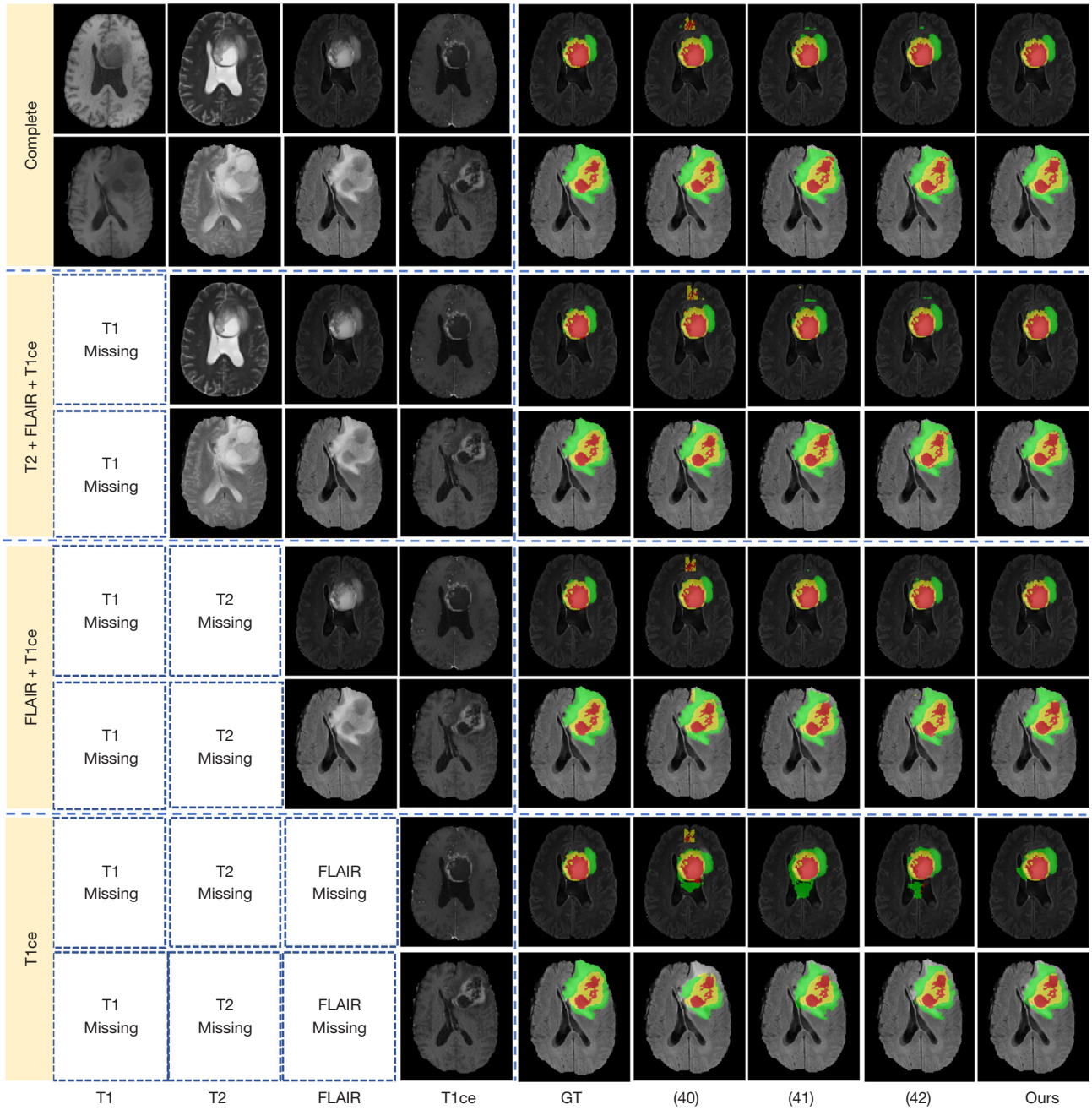

Figure 6 Comparison of segmentation results on four cases of missing modalities: complete modalities; FLAIR, T1ce, T2; FLAIR, T1ce; T1ce. From the left to right are 4 MRI modalities: T1, T2, FLAIR, and T1ce; the fifth column presents the ground truth of 2 patients, the sixth to eighth columns show the results of the state-of-the-art methods, and the rightmost column shows our segmentation results. T2, T2-weighted; FLAIR, fluid-attenuated inversion recovery; T1ce, contrast-enhanced T1-weighted; T1, T1-weighted; GT, ground truth; MRI, magnetic resonance imaging.

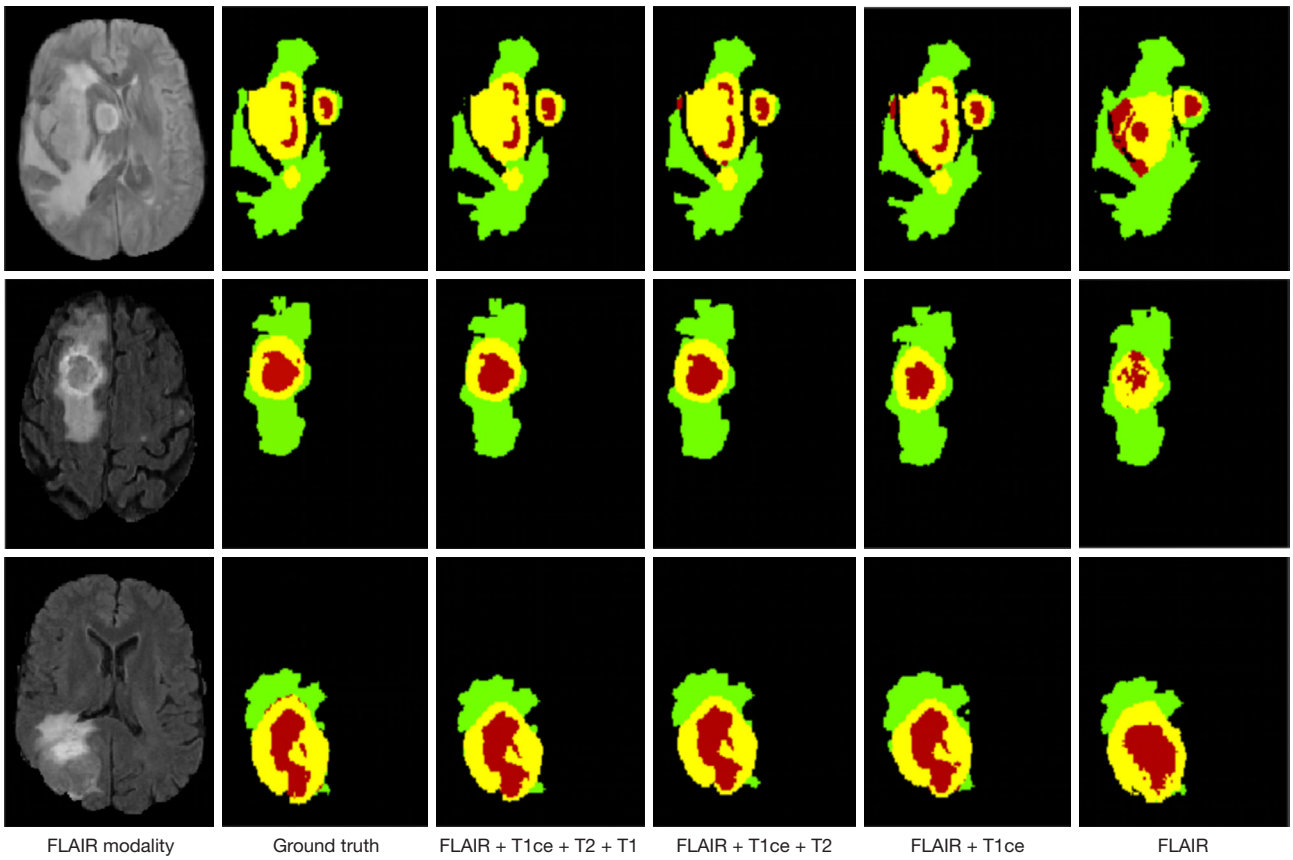

Figure 7 Examples of the segmentation results of CMAF-Net with various available modalities. FLAIR, fluid-attenuated inversion recovery; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; T1, T1-weighted; CMAF-Net, cross-modal attention fusion based deep neural network.

will reduce the segmentation performance of the model. The quantitative evaluation shown in *Table 8* verifies the effectiveness of the joint loss function proposed in this paper.

In order to demonstrate the effectiveness of each module in the proposed model on the segmentation results, *Figure 9* shows the visualization of the segmentation results we obtained using different models under the four available modal combinations. Compared with the baseline model, it is clearly shown that the segmentation performance can be improved by adding the proposed modules, making the segmentation results of our CMAF-Net closer to the ground truth. To show the distribution characteristics of the segmentation results of the 4 methods, *Figure 10* shows the ablation experimental results of different modules in 75 test samples, and the DSC are represented by boxplots and a bar graph. It can be seen from the figure that the proposed CMAF-Net presents fewer outliers, indicating that it has high stability and consistency.

Discussion

Brain tumor segmentation is an important step in the clinical diagnosis and development of treatment plans. It can provide a reference for assisted diagnosis and surgical planning. The brain tumor segmentation of multimodal MRI images using deep learning methods has great potential and excellent performance. However, in the actual clinical data acquisition, due to various reasons such as equipment failure, data corruption, and human error, the final collected data often have different degrees of missing modality. Consequently, performance degradation of existing multimodal MRI brain tumor segmentation methods occurs. To solve the above problems, we propose in this work a brain tumor segmentation network CMAF-Net based on CMAF by using incomplete multimodal MRI data.

Table 3 Comparison with different methods such as U-HeMIS (40), U-HVED (41), DIGEST (36), D^2-Net (37), mmFormer (42), RFNet (43) and GSS (38) on the BraTS 2018 dataset

Methods	Type: FLAIR/T1/T1c/T2															AVG
	○/○/○/●	○/○/●/○	○/●/○/○	●/○/○/○	○/○/●/●	○/●/●/○	●/●/○/○	○/●/○/●	●/○/○/●	●/○/●/○	●/●/●/○	●/●/○/●	●/○/●/●	○/●/●/●	●/●/●/●	
WT																
U-HeMIS	80.9	61.5	57.6	52.4	82.4	68.4	64.6	82.4	82.9	68.9	72.3	83.4	83.8	83.9	84.7	74.0
U-HVED	79.8	53.6	49.5	84.3	81.3	64.2	85.7	81.5	87.5	85.9	86.7	88.0	88.0	82.3	88.4	79.1
DIGEST	80.4	77.2	61.6	84.5	83.6	70.8	86.9	83.3	88.8	86.4	87.0	88.7	89.4	83.4	89.9	82.8
D^2-Net	76.3	42.8	15.5	84.2	84.1	62.1	87.3	80.1	87.9	87.5	87.7	88.4	88.8	80.9	88.8	76.2
mmFormer	84.5	74.1	72.0	86.2	86.3	77.0	87.1	85.6	89.0	87.9	88.5	89.3	89.8	86.7	90.0	84.9
RFNet	85.1	73.6	74.8	85.8	85.6	77.5	87.4	86.7	89.2	88.7	88.9	89.4	90.0	86.9	90.1	85.3
GSS	85.7	75.8	75.4	86.7	86.9	78.5	88.1	87.0	89.3	89.2	89.4	89.7	90.2	87.7	90.2	86.0
Ours	86.9*	81.8*	80.9*	87.9*	88.2*	83.3*	89.2*	87.9*	89.6*	89.6*	89.8*	89.8*	90.4*	88.5*	90.4*	87.6*
TC																
U-HeMIS	57.2	65.2	37.3	26.0	76.6	72.4	41.1	60.9	57.6	71.4	76.0	60.3	77.5	78.9	79.4	62.5
U-HVED	54.6	59.5	33.9	57.9	73.9	67.5	61.1	56.2	62.7	75.0	77.0	63.1	76.7	75.2	77.7	64.8
DIGEST	61.3	78.7	60.6	62.5	81.4	75.8	69.6	64.2	71.1	79.4	78.3	65.5	83.5	82.6	81.8	73.1
D^2-Net	56.7	65.1	16.8	47.3	80.3	78.2	61.6	63.2	62.6	80.8	80.9	63.7	80.7	79.0	80.1	66.5
mmFormer	70.5	82.1	68.7	68.3	83.0	82.1	72.4	73.4	72.5	83.1	82.9	73.8	83.0	82.8	82.6	77.4
RFNet	70.7	82.4	69.2	68.9	83.2	82.7	73.1	73.8	73.7	82.7	82.7	74.6	83.2	83.3	83.1	77.8
GSS	71.2	82.8	70.4	73.6*	82.8	83.0	74.0	74.4	74.1	83.0	82.9	75.7	83.6	83.5	83.4	78.6
Ours	73.5*	83.1*	71.5*	72.5	83.5*	83.2*	75.2*	75.3*	75.2*	83.2*	83.1*	76.2*	83.9*	83.7*	83.6*	79.1*
ET																
U-HeMIS	25.6	62.0	10.1	11.7	67.8	66.2	10.7	32.3	30.2	66.1	68.5	31.0	68.7	69.9*	70.2*	46.0
U-HVED	22.8	57.6	8.6	23.8	67.8	61.1	27.9	24.2	32.3	68.3	68.6	32.3	68.9	67.7	69.0	46.8
DIGEST	32.3	62.2	33.7	32.0	68.4	68.0	38.6	40.4	39.7	67.9	68.0	37.5	68.1	69.5	67.8	52.9
D^2-Net	16.0	66.3	8.1	8.1	68.5	68.0	9.5	16.5	17.4	64.8	65.7	19.4	66.4	68.3	68.4	42.1
mmFormer	36.0	65.8	35.2	32.9	67.1	67.0	37.5	41.3	39.8	66.8	67.0	40.8	67.9	67.2	68.0	53.4
RFNet	36.5	66.4	35.9	33.5	67.6	67.4	38.8	40.7	40.3	67.1	67.6	41.5	68.4	67.6	67.8	53.8
GSS	36.8	67.0	36.4	37.9*	68.0	67.7	40.9*	41.1	40.6	67.4	68.0	44.3*	68.7	68.2	68.1	54.7
Ours	37.6*	67.8*	37.2*	37.5	68.7*	68.4*	40.5	41.5*	41.8*	68.5*	68.8*	44.0	69.1*	69.0	68.5	55.3*

The results reported in the table are the DSC for the different missing sceneries. The higher the DSC value, the better the results. ○ and ● indicate the missing and available modalities, respectively. *, the best results. AVG denotes the average results on one target region across all the situations. FLAIR, fluid attenuated inversion recovery; T1, T1-weighted; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; WT, whole tumor; TC, tumor core; ET, enhancing tumor; DSC, dice similarity coefficient.

First, we proposed a 3D Swin block based on a shift window to capture long-range context information. Long-range context information is very important in medical image segmentation. When segmenting tumor tissues that are far away from the tumor core region and have no obvious tumor features, CNN-based models can lead to insufficient modeling of global information due to the limitation of receptive fields (52). Thus, the segmentation results are affected. Given that the convolutional operation in CNN can capture only local information, the Transformer model can better capture long-range dependencies through the self-attention mechanism. Compared with the traditional

Table 4 Improvements of CMAF-Net upon U-HeMIS (40), U-HVED (41), DIGEST (36), D^2-Net (37), mmFormer (42), RFNet (43), and GSS (38) with different numbers of input modalities evaluated by DSC (%)

Methods	Numbers of input modalities			
	1	2	3	4
WT				
U-HeMIS	+85.1	+78.2	+35.1	+5.7
U-HVED	+70.3	+41.7	+13.5	+2.0
DIGEST	+34.7	+28.0	+10.0	+0.5
D^2-Net	+118.7	+38.8	+12.7	+1.6
mmFormer	+20.7	+14.9	+4.2	+0.4
RFNet	+18.2	+12.7	+3.3	+0.3
GSS	+13.9	+8.8	+1.5	+0.2
TC				
U-HeMIS	+114.9	+95.6	+34.2	+4.2
U-HVED	+94.7	+79.2	+34.9	+5.9
DIGEST	+37.5	+34.1	+17.0	+1.8
D^2-Net	+114.7	+48.9	+22.6	+3.5
mmFormer	+11.0	+9.1	+4.4	+1.0
RFNet	+9.4	+6.8	+3.1	+0.5
GSS	+2.6	+4.3	+1.2	+0.2
ET				
U-HeMIS	+70.7	+56.1	+12.8	-1.7
U-HVED	+67.3	+47.8	+11.2	-0.5
DIGEST	+19.9	+6.4	+7.8	+0.7
D^2-Net	+81.6	+84.7	+31.1	+0.1
mmFormer	+10.2	+9.9	+8.0	+0.5
RFNet	+7.8	+7.5	+5.8	+0.7
GSS	+2.0	+3.7	+1.7	+0.4

CMAF-Net, cross-modal attention fusion-based deep neural network; 3D, three-dimensional; DSC, dice similarity coefficient; WT, whole tumor; TC, tumor core; ET, enhancing tumor.

Transformer, the 3D Swin block introduces a shift-window mechanism to achieve window-to-window communication. It reduces the amount of computation and introduces a locality prior, so it can be better applied to the segmentation task. Brain tumors are usually highly localized (53) and spatially variable, requiring the modeling of features at different scales and different locations. Therefore, applying 3D Swin block based on shift window to the brain tumor segmentation task is reasonable. As shown in *Table* 7, the results of ablation experiments demonstrated that our proposed 3D Swin block can effectively improve the segmentation performance and prove its effectiveness. Second, we designed a cross-attention based CMAF module to deal with the missing modality cases by fusing features from different modalities to learn shared potential representations. It is able to adaptively learn the correlations between modalities and thus better capture complementary information. Specifically, the CMAF module can effectively utilize the complementarities between different modalities during training. Even if some modalities are missing, the network can still synthesize the information of other available modalities to infer the features of the missing modalities, maintaining accuracy and stability. Finally, the channel and spatial self-attention modules are used in the decoding stage to gradually aggregate multimodal and multilevel features and better focus on the key information in the feature map, thereby further improving the accuracy of segmentation. For the task of brain tumor image segmentation, the use of the attention mechanism extracts richer semantic information and pays more attention to the information of small-area brain tumors. Consequently, the segmentation effect of brain tumors is improved (54).

Existing methods for incomplete multimodal brain tumor segmentation mainly differ in their fusion strategies. A common approach is to calculate the mean and variance of each available modality and fuse the corresponding features with equal importance. However, this method may not efficiently aggregate the features of the missing modalities, potentially compromising the effectiveness of fusion. Another approach involves directly integrating modal features through convolution. However, the fusion strategy based on convolution may not be adequate for integrating global information due to the localized nature of convolution operations and the inherent bias of weight sharing. This limitation inevitably hinders the modeling of remote dependencies. In contrast, the CMAF module utilizes a cross-attention mechanism that is able to capture correlations between different modalities and interactively propagate information in a dynamic and learnable manner. In this way, global information and remote dependencies can be captured effectively. Therefore, introducing the CMAF module to supplement the global information is a simple solution. In the field of computer vision, multimodal fusion is important for joint modeling and understanding

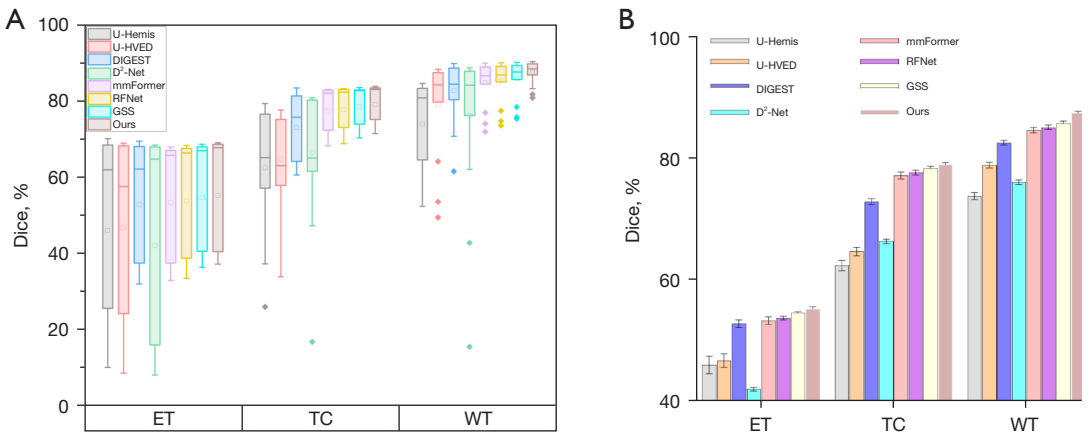

Figure 8 Boxplots and bar graph of the DSC of the comparison experiment results. (A) The results for 15 cases of missing modalities. (B) The average results of different methods and ours. Error bars show standard error. ET, enhancing tumor; TC, tumor core; WT, whole tumor; DSC, dice similarity coefficient.

Table 5 Comparison of different methods in terms of average Hausdorff distance on BraTS 2018 and BraTS 2020 datasets

Dataset	Method	Average Hausdorff distance		
		ET	TC	WT
BraTS 2018	U-HeMIS	15.38	17.62	16.22
	U-HVED	15.11	16.63	14.87
	DIGEST	8.93	10.11	10.06
	D^2-Net	11.74	13.67	13.54
	mmFormer	7.32	7.54	6.82
	RFNet	7.09	7.30	6.63
	GSS	6.54	6.87	5.91
	Ours	5.95*	6.59*	4.38*
BraTS 2020	U-HeMIS	15.40	18.37	20.32
	U-HVED	13.86	16.65	21.00
	DIGEST	8.72	12.70	13.05
	D^2-Net	11.48	8.64	5.67
	mmFormer	8.09	5.94	5.01
	RFNet	7.62	6.15	4.82
	GSS	6.73	5.79	4.65
	Ours	4.02*	5.35*	4.21*

*, the best results. Average Hausdorff distance denotes the average results on 1 target region across all the situations. ET, enhancing tumor; TC, tumor core; WT, whole tumor.

between images and texts. By integrating image and text data from multiple modalities, better image description generation, visual question and answer, and other tasks can be realized (55). Our CMAF-Net is flexible and scalable in processing multimodal data and solving segmentation tasks, where the multimodal fusion module can be easily adapted to other multimodal network architectures and research areas. This makes our approach promising for a wide range of applications in various fields. In addition, the complete training and inference process of our CMAF-Net is conducted in an end-to-end manner, and the method is generalized to handle any number of modality data, which can help segmentation even if some modalities are missing.

Extensive segmentation results were compared with other methods on incomplete and complete MRI data from the BraTS 2020 dataset. The results showed that our method solves the problem of missing modalities in brain tumor segmentation and obtains the feature information in multimodal data, thereby performing effective feature fusion to segment brain tumors more accurately. This finding fully demonstrates the effectiveness of CMAF-Net and the advantages of utilizing multimodal data fusion. Currently, the use of this method can help to segment brain tumors when 1 or more modalities are missing and obtain better segmentation results, which has great value in clinical practice. In actual clinical practice, doctors need to use image data for tumor localization and evaluation to develop

Table 6 Comparison of full modality performance of different methods

Method	DSC (%) (mean ± SD)				Hausdorff distance			
	WT	TC	ET	Average	WT	TC	ET	Average
3D U-Net (13)	87.3±0.321	80.7±0.351	73.8±0.439	80.6	5.68	5.19	7.24	6.04
Attention U-Net (14)	87.8±0.265	83.4±0.323	74.8±0.404	82.0	5.37	5.49	7.07	5.98
TransBTS (17)	88.3±0.200	85.8±0.252	77.5±0.321	83.9	5.60	8.13	6.21	6.65
F^2Net (30)	91.2±0.252	85.4±0.264	77.0±0.350	84.5	3.73	6.02	4.93	4.89
Swin-UNETR (20)	91.4*±0.208	86.2±0.305	76.9±0.360	84.8	3.61*	4.48	8.39	5.49
Our	90.9±0.200	86.8*±0.231	77.8*±0.265	85.2*	3.65	4.45*	6.16*	4.75*

*, the best results. Average indicates the average results of the 3 target regions. Higher DSC indicate better results, while lower Hausdorff distance indicate better results. DSC, dice similarity coefficient; SD, standard deviation; WT, whole tumor; TC, tumor core; ET, enhancing tumor; 3D, three-dimensional.

Table 7 Ablation study of critical components of CMAF-Net

Method	Average DSC (%)			
	WT	TC	ET	Average
Baseline	85.8	77.9	60.3	74.7
Baseline + 3D Swin block	86.5	78.6	62.1	75.7
Baseline + 3D Swin block + CMAF	87.1	80.4	63.3	76.9
Baseline + 3D Swin block + CMAF + CAM	87.5	81.1	64.0	77.5
CMAF-Net	87.9*	81.8*	64.3*	78.0*

*, the best results. CMAF-Net, cross-modal attention fusion-based deep neural network; DSC, dice similarity coefficient; WT, whole tumor; TC, tumor core; ET, enhancing tumor; 3D, three-dimensional; CMAF, cross-modal attention fusion; CAM, channel attention module.

Table 8 Performance comparison of different loss functions in CMAF-Net

Loss function	Average DSC (%)			
	WT	TC	ET	Average
L_{cross}	86.8	81.4	63.8	77.3
L_{dice}	87.4	81.1	64.0	77.5
L	87.9*	81.8*	64.3*	78.0*

*, the best results. CMAF-Net, cross-modal attention fusion-based deep neural network; DSC, dice similarity coefficient; WT, whole tumor; TC, tumor core; ET, enhancing tumor; L, loss function; cross, the cross-entropy.

treatment plans (56). If the tumor segmentation model can provide reliable results despite missing modalities, it will become an important auxiliary tool for doctors and help them understand the patient's condition more accurately to support clinical decision making. In addition, a tumor segmentation model that handles missing modalities can streamline clinical workflows and reduce the extra time and effort consumed by physicians dealing with missing data. By integrating the model into medical imaging devices or diagnostic software, physicians can obtain accurate tumor segmentation results more quickly, thereby saving diagnostic time and improving diagnostic accuracy. This integration can increase productivity by allowing physicians to focus more on diagnosis and treatment planning rather than spending a significant amount of time in the process of segmenting tumors. With the model's accurate tumor segmentation results, doctors are able to develop personalized treatment plans with a clearer understanding of the tumor's size, shape, location, and relationship to surrounding tissues. This helps to optimize surgical and radiotherapy protocols, monitor tumor growth, assess treatment efficacy, and adjust treatment strategies for

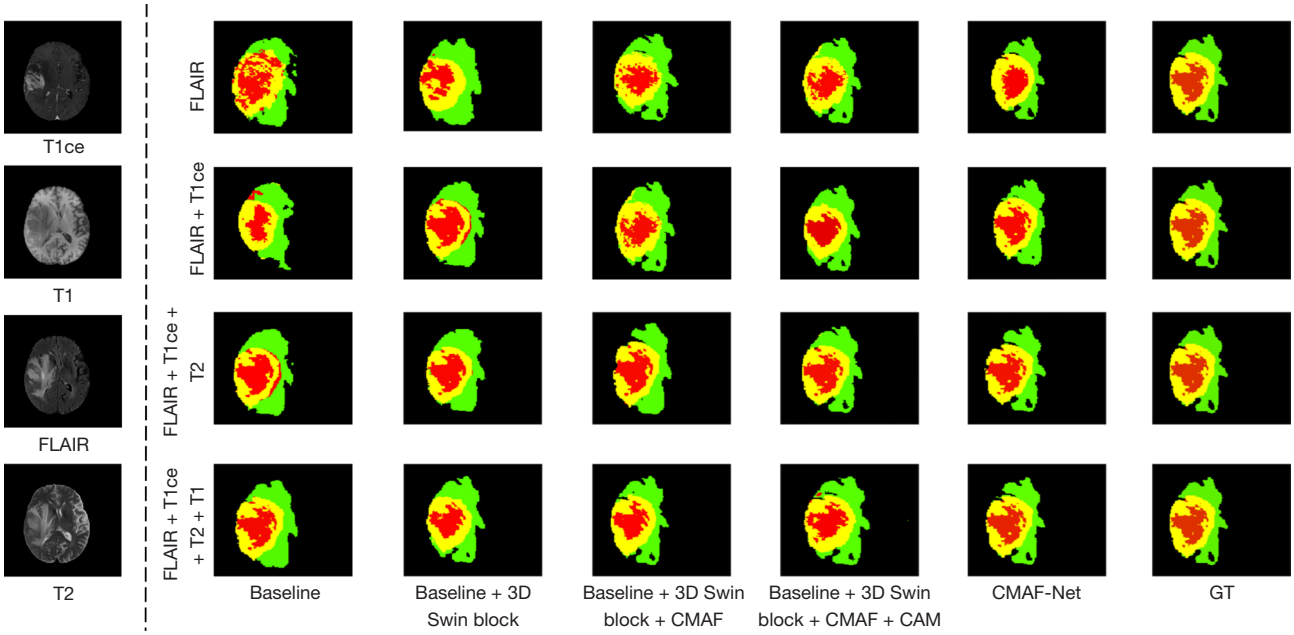

Figure 9 Visual comparison of the effects of different components of CMAF-Net in the ablation study. T1ce, contrast-enhanced T1-weighted; T1, T1-weighted; FLAIR, fluid attenuated inversion recovery; T2, T2-weighted; 3D, three-dimensional; CMAF, cross-modal attention fusion; CAM, channel attention module; CMAF-Net, cross-modal attention fusion-based deep neural network; GT, ground truth.

better clinical outcomes. Advances in artificial intelligence, especially deep learning, have enabled the development of automated brain tumor segmentation techniques that overcome manual segmentation methods reliant on heavy labor and subjective operator judgment. Beyond its impact on brain tumor segmentation, AI has broader implications in medical imaging. For example, AI has the potential to help radiologists to identify and prioritize patients with the most severe or complex cases (57). Additionally, compared to the traditional imaging workflow that heavily relies on human labor, AI enables more safe, accurate, and efficient imaging solutions (58).

However, the limitation of this method is the inability to recover missing modalities, and accurate clinical diagnosis and treatment rely on complete and accurate data. In future work, we will try to explore multi-task learning to complement the missing modalities and thus provide richer image information for brain tumor diagnosis. We also plan to extend our proposed CMAF-Net to the problem of segmentation of completely different modalities [e.g., computed tomography (CT) and MRI] and explore more ways to further improve the proposed method's performance. At the same time, we also realize that due to the complexity of the model, CMAF-Net has specific limitations of large memory and time resource usage during training. Although the introduction of new modules can bring higher characterization capabilities to the model, it also bears the limitations of more parameters, difficult optimization, and the need for more data volume. Therefore, we will further optimize the model in the future while reducing memory consumption and making it more lightweight.

Conclusions

We have proposed an effective CMAF-based 3D segmentation model CMAF-Net for incomplete multimodal brain tumor segmentation. It fully combines the advantages of CNN and Transformer structures to improve the ability of extracting local features and global context information. We also designed a cross attention-based multimodal feature fusion module that models long-range correlations between different modalities and learns modality-invariant shared feature representations. Ablation experiments validated the effectiveness of key components in CMAF-Net. The experimental results on the BraTS 2018 and BraTS 2020 datasets showed that the proposed method achieves excellent performance and robustness in dealing with multimodal brain tumor segmentation with missing

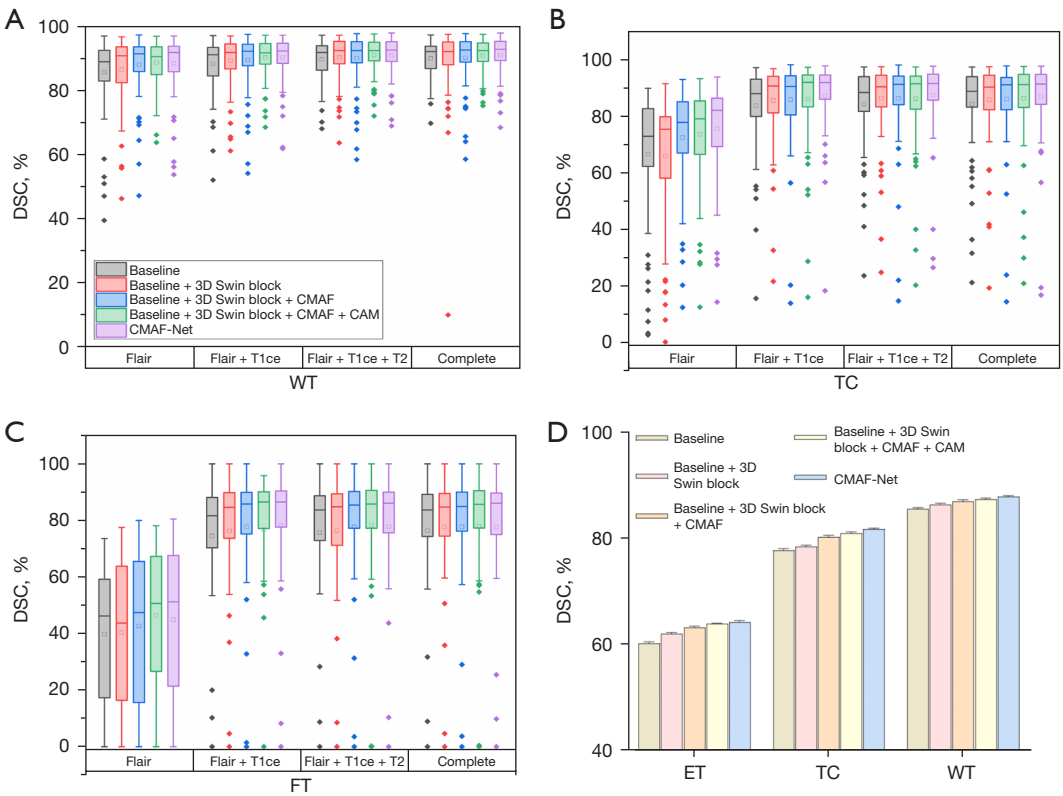

Figure 10 Boxplots and bar graph of the DSC of ablation experiments for different modules. (A-C) The DSC of WT, TC, and ET, respectively. (D) The bar chart comparison result of different components. 3D, three-dimensional; CMAF, cross-modal attention fusion; CAM, channel attention module; CMAF-Net, cross-modal attention fusion based deep neural network; FLAIR, fluid attenuated inversion recovery; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; WT, whole tumor; TC, tumor core; ET, enhancing tumor; DSC, dice similarity coefficient.

modalities. In the future, we will continue to explore how to improve our multimodal model so that it can be applied to real-world clinical scenarios and improve the performance of brain tumor segmentation.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (No. 62371243, to X.N.), the General Program of Natural Foundation of Jiangsu Province (No. BK20231190, to X.N.), the General Program of Jiangsu Provincial Health Commission (No. M2020006, to X.N.), the Jiangsu Provincial Medical Key Discipline Cultivation Unit of Oncology Therapeutics (Radiotherapy) (No. JSDW202237, to X.N.), the Jiangsu Provincial Key Research and Development Program Social Development Project (No. BE2022720, to X.N.), and the Changzhou Social Development Project (No. CE20235063, to X.N.).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-9/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as

revised in 2013).

References

1. Krishna PR, Prasad V, Battula TK. Optimization empowered hierarchical residual VGGNet19 network for multi-class brain tumour classification. Multimed Tools Appl 2023;82:16691-716.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
3. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. Inf Fusion 2023;91:376-87.
4. Rehman MU, Ryu J, Nizami IF, Chong KT. RAAGR2-Net: A brain tumor segmentation network using parallel processing of multiple spatial frames. Comput Biol Med 2023;152:106426.
5. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 2015;34:1993-2024.
6. Diao Y, Li F, Li Z. Joint learning-based feature reconstruction and enhanced network for incomplete multi-modal brain tumor segmentation. Comput Biol Med 2023;163:107234.
7. Peng Y, Sun J. The multimodal MRI brain tumor segmentation based on AD-Net. Biomed Signal Process Control 2023;80:104336.
8. Wang Y, Zhang Y, Liu Y, Lin Z, Tian J, Zhong C, Shi Z, Fan J, He Z. ACN: Adversarial co-training network for brain tumor segmentation with missing modalities. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, Essert C. editors. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. Springer, 2021:410-20.
9. Badjie B, Ülker ED. A Deep Transfer Learning Based Architecture for Brain Tumor Classification Using MR Images. Inf Technol Control 2022;51:332-44.
10. Khan MA, Khan A, Alhaisoni M, Alqahtani A, Alsubai S, Alharbi M, Malik NA, Damaševičius R. Multimodal brain tumor detection and classification using deep saliency map and improved dragonfly optimization algorithm. Int J Imaging Syst Technol 2023;33:572-87.
11. Kurdi SZ, Ali MH, Jaber MM, Saba T, Rehman A, Damaševičius R. Brain Tumor Classification Using Meta-Heuristic Optimized Convolutional Neural Networks. J Pers Med 2023;13:181.
12. Rajinikanth V, Kadry S, Nam Y. Convolutional-Neural-Network Assisted Segmentation and SVM Classification of Brain Tumor in Clinical MRI Slices. Inf Technol Control 2021;50:342-56.
13. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W. editors. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer, 2016:424-32.
14. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D. Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999, 2018.
15. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV) 2016:565-71.
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems 2017;30:6000-10.
17. Wang W, Chen C, Ding M, Li J, Yu H, Zha S. Transbts: Multimodal brain tumor segmentation using transformer. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, Essert C. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, Springer, 2021:109-19.
18. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: Transformers make

strong encoders for medical image segmentation. arXiv:2102.04306, 2021.
19. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021:10012-22.
20. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi A, Bakas S. editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021. Lecture Notes in Computer Science, Springer, 2021:272-84.
21. Lamrani D, Cherradi B, El Gannour OE, Bouqentar MA, Bahatti L. Brain tumor detection using mri images and convolutional neural network. Int J Adv Comput Sci Appl 2022;13:452-60.
22. Asif S, Zhao M, Tang F, Zhu Y. An enhanced deep learning method for multi-class brain tumor classification using deep transfer learning. Multimed Tools Appl 2023;82:31709-36.
23. Acar E, Levin-Schwartz Y, Calhoun VD, Adali T. ACMTF for fusion of multi-modal neuroimaging data and identification of biomarkers. 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 2017:643-7.
24. Rahaman MA, Chen J, Fu Z, Lewis N, Iraji A, Calhoun VD. Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness. Annu Int Conf IEEE Eng Med Biol Soc 2021;2021:3267-72.
25. Nie D, Wang L, Gao Y, Shen D. Fully convolutional networks for multi-modality isointense infant brain image segmentation. Proc IEEE Int Symp Biomed Imaging 2016;2016:1342-5.
26. Peiris H, Hayat M, Chen Z, Egan G, Harandi M. A robust volumetric transformer for accurate 3D tumor segmentation. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science, Springer, 2022:162-72.
27. Xing Z, Yu L, Wan L, Han T, Zhu L. NestedFormer: Nested modality-aware transformer for brain tumor segmentation. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science, vol 13435. Springer, 2022:140-50.
28. Zhang Y, Yang J, Tian J, Shi Z, Zhong C, Zhang Y, He Z. Modality-aware mutual learning for multi-modal medical image segmentation. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, Essert C. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, 2021:589-99.
29. Li, X, Ma, S, Tang, J, Guo, F. TranSiam: Aggregating multi-modal visual features with locality for medical image segmentation. Expert Syst Appl 2024;237:121574.
30. Yang H, Zhou T, Zhou Y, Zhang Y, Fu H. Flexible Fusion Network for Multi-Modal Brain Tumor Segmentation. IEEE J Biomed Health Inform 2023;27:3349-59.
31. Zhou C, Ding C, Wang X, Lu Z, Tao D. One-pass Multi-task Networks with Cross-task Guided Attention for Brain Tumor Segmentation. IEEE Trans Image Process 2020. [Epub ahead of print]. doi: 10.1109/TIP.2020.2973510.
32. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T. editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science, Springer, 2019:311-20.
33. Zhan B, Li D, Wu X, Zhou J, Wang Y. Multi-Modal MRI Image Synthesis via GAN With Multi-Scale Gate Mergence. IEEE J Biomed Health Inform 2022;26:17-26.
34. Lee D, Moon WJ, Ye JC. Assessing the importance of magnetic resonance contrasts using collaborative generative adversarial networks. Nat Mach Intell 2020;2:34-42.
35. Hu M, Maillard M, Zhang Y, Ciceri T, Barbera GL, Bloch I, Gori P. Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Maria A. Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. Springer, 2020:772-81.
36. Li H, Li C, Huang W, Zheng X, Xi Y, Wang S. DIGEST: Deeply supervIsed knowledGE tranSfer neTwork learning for brain tumor segmentation with incomplete multi-modal MRI scans. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 2023:1-4.
37. Yang Q, Guo X, Chen Z, Woo PYM, Yuan Y. D(2)-Net: Dual Disentanglement Network for Brain Tumor

Segmentation With Missing Modalities. IEEE Trans Med Imaging 2022;41:2953-64.
38. Qiu Y, Chen D, Yao H, Xu Y, Wang Z. Scratch Each Other's Back: Incomplete Multi-Modal Brain Tumor Segmentation via Category Aware Group Self-Support Learning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023:21317-26.
39. Zhou T, Canu S, Vera P, Ruan S. Latent Correlation Representation Learning for Brain Tumor Segmentation With Missing MRI Modalities. IEEE Trans Image Process 2021;30:4263-74.
40. Havaei M, Guizard N, Chapados N, Bengio Y. Hemis: Hetero-modal image segmentation. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W. editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Lecture Notes in Computer Science, Springer, 2016:469-77.
41. Dorent R, Joutard S, Modat M, Ourselin S, Vercauteren T. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Tap PT, Khan A. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science, Springer, 2019:74-82.
42. Zhang Y, He N, Yang J, Li Y, Wei D, Huang Y, Zhang Y, He Z, Zheng Y. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science, Springer, 2022:107-17.
43. Ding Y, Yu X, Yang Y. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021:3975-84.
44. Konwer A, Hu X, Bae J, Xu X, Chen C, Prasanna P. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023:21415-25.
45. Jiang Y, Zhang Y, Lin X, Dong J, Cheng T, Liang J. SwinBTS: A Method for 3D Multimodal Brain Tumor Segmentation Using Swin Transformer. Brain Sci 2022;12:797.
46. Ma J, He J, Yang X. Learning Geodesic Active Contours for Embedding Object Global Information in Segmentation CNNs. IEEE Trans Med Imaging 2021;40:93-104.
47. Lin J, Lin J, Lu C, Chen H, Lin H, Zhao B, Shi Z, Qiu B, Pan X, Xu Z, Huang B, Liang C, Han G, Liu Z, Han C. CKD-TransBTS: Clinical Knowledge-Driven Hybrid Transformer With Modality-Correlated Cross-Attention for Brain Tumor Segmentation. IEEE Trans Med Imaging 2023;42:2451-61.
48. Zhou T. Modality-level cross-connection and attentional feature fusion based deep neural network for multi-modal brain tumor segmentation. Biomed Signal Process Control 2023;81:104524.
49. Lin H, Cheng X, Wu X, Yang F, Shen D, Wang Z, Song Q, Yuan W. Cat: Cross attention in vision transformer. 2022 IEEE International Conference on Multimedia and Expo (ICME), 2022:1-6.
50. Rahaman MA, Garg Y, Iraji A, Fu Z, Chen J, Calhoun V. Two-dimensional attentive fusion for multi-modal learning of neuroimaging and genomics data. 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an, China, 2022:1-6.
51. Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, Nath V, Hatamizadeh A. Self-supervised pre-training of swin transformers for 3d medical image analysis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022:20730-40.
52. Xu Y, Yang Y, Zhang L. DeMT: Deformable mixer transformer for multi-task learning of dense prediction. Proceedings of the AAAI Conference on Artificial Intelligence, 2023:3072-80.
53. Liang J, Yang C, Zeng L. 3D PSwinBTS: An efficient transformer-based Unet using 3D parallel shifted windows for brain tumor segmentation. Digit Signal Process 2022;131:103784.
54. Tian W, Li D, Lv M, Huang P. Axial Attention Convolutional Neural Network for Brain Tumor Segmentation with Multi-Modality MRI Scans. Brain Sci 2022;13:12.
55. Tan Y, Li C, Qin J, Xue Y, Xiang X. Medical Image Description Based on Multimodal Auxiliary Signals and Transformer. Int J Intell Syst 2024. doi: 10.1155/2024/6680546.
56. Ottom MA, Rahman HA, Dinov ID. Znet: Deep Learning Approach for 2D MRI Brain Tumor Segmentation. IEEE J Transl Eng Health Med 2022;10:1800508.
57. Alexander A, Jiang A, Ferreira C, Zurkiya D. An Intelligent Future for Medical Imaging: A Market Outlook

on Artificial Intelligence for Medical Imaging. J Am Coll Radiol 2020;17:165-70.
58. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, He K, Shi Y, Shen D. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. IEEE Rev Biomed Eng 2021;14:4-15.

Cite this article as: Sun K, Ding J, Li Q, Chen W, Zhang H, Sun J, Jiao Z, Ni X. CMAF-Net: a cross-modal attention fusion-based deep neural network for incomplete multi-modal brain tumor segmentation. Quant Imaging Med Surg 2024;14(7):4579-4604. doi: 10.21037/qims-24-9