*electronics*

*Article*

# MMAformer: Multiscale Modality-Aware Transformer for Medical Image Segmentation

**Hao Ding [1], Xiangfen Zhang [1,*], Wenhao Lu [2], Feiniu Yuan [1,*] and Haixia Luo [1]**

[1] College of Information, Mechanical and Electrical Engineering, Shanghai Normal University (SHNU), Shanghai 201418, China; 1000448120@smail.shnu.edu.cn (H.D.); 1000467742@smail.shnu.edu (H.L.)

[2] College of Computing and Data Science, Nanyang Technological University (NTU), Nanyang Avenue, Singapore 639798, Singapore; wenhao.lu@ntu.edu.sg

* Correspondence: xiangfen@shnu.edu.cn (X.Z.); yfn@ustc.edu (F.Y.)

**Abstract:** The segmentation of medical images, particularly for brain tumors, is essential for clinical diagnosis and treatment planning. In this study, we proposed MMAformer, a Multiscale Modality-Aware Transformer model, which is designed for segmenting brain tumors by utilizing multimodality magnetic resonance imaging (MRI). Complementary information between different sequences helps the model delineate tumor boundaries and distinguish different tumor tissues. To enable the model to acquire the complementary information between related sequences, MMAformer employs a multi-stage encoder, which uses a cross-modal downsampling (CMD) block for learning and integrating the complementary information between sequences at different scales. In order to effectively fuse the various information extracted by the encoder, the Multimodal Gated Aggregation (MGA) block combines the dual attention mechanism and multi-gated clustering to effectively fuse the spatial, channel, and modal features of different MRI sequences. In the comparison experiments on the BraTS2020 and BraTS2021 datasets, the average Dice score of MMAformer reached 86.3% and 91.53%, respectively, indicating that MMAformer surpasses the current state-of-the-art approaches. MMAformer's innovative architecture, which effectively captures and integrates multimodal information at various scales, offers a promising solution for tackling complex medical image segmentation challenges.

**Keywords:** brain tumor segmentation; cross-modal downsampling; multimodal gated aggregation; multimodality; multiscale; Transformer

check for updates

**Citation:** Ding, H.; Zhang, X.; Lu, W.; Yuan, F.; Luo, H. MMAformer: Multiscale Modality-Aware Transformer for Medical Image Segmentation. *Electronics* **2024**, *13*, 4636. https://doi.org/10.3390/electronics13234636

Academic Editors: Jungpil Shin, Hoang D. Le and Md. Al Mehedi Hasan

Received: 30 October 2024
Revised: 16 November 2024
Accepted: 21 November 2024
Published: 25 November 2024

## 1. Introduction

An abnormal tissue growth in or near the brain or surrounding regions, known as a brain tumor, can be either benign or malignant [1]. Malignant gliomas represent the most frequent and aggressive type of primary brain tumors, posing significant harm to patients [2]. The automated diagnosis of these tumors can greatly improve the prognosis of patients by enabling earlier and more accurate detection. This article focuses on the challenges and advancements in the automated diagnosis of malignant glioma. Multimodality magnetic resonance imaging (MRI) analyzes the brain by providing rich multimodality information and the commonly used MRI sequences. Referring to Figure 1a, this includes T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2-fluid attenuated inversion recovered (T2-flair) images, which are generated by these different sequences that are usually complementary in imaging, and they are useful in distinguishing between the enhancing tumor, peritumoral edema and tumor core in each of the three specific objects [3–5]. These different tumors are shown in Figure 1b. For clinicians, these automatically segmented tumors with multiple regions are very helpful in clinical diagnosis and therapy.

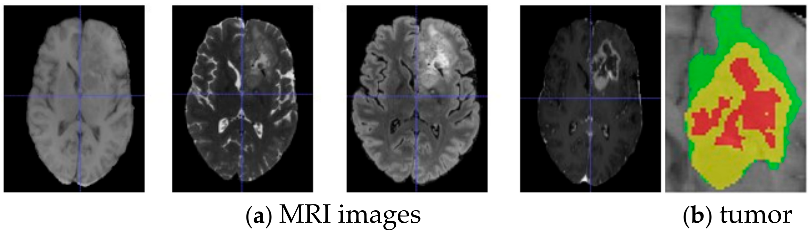


**Figure 1.** Multimodal MRIs. Adapted from [6]. (**a**) T1, T2, T2-flair, T1Gd; (**b**) red: tumor core; yellow: enhancing tumor; green: peritumoral edema.

In recent years, deep learning-based image segmentation has become a widespread task in medical image analysis, which is backed by numerous studies [7]. Currently, convolutional neural networks (CNNs) have made significant advancements in the field of image analysis and processing [8], and brain tumor segmentation has become a particularly active area of research.

In terms of creating segmentation models, the currently used segmentation models can be categorized into 2D and 3D models. The earliest 2D segmentation model is the Fully Convolutional Network (FCN) proposed by Long et al. [9], which is capable of directly outputting pixel-level labels. However, due to excessive feature information loss and insufficient feature utilization, the network suffers from several issues, including the tendency to produce nulls and discontinuous segmentation results. Subsequently, the VGG model [10] builds on FCN with a deeper architecture with smaller convolutional filters to achieve better segmentation. However, the increased depth also led to higher computational costs and memory usage, which are limitations in practical applications. Notably, the introduction of the U-Net [11] model marked the emergence of a new network architecture that consists of downsampling and upsampling phases, which are commonly referred to as the encoder and decoder. In the downsampling stage, the process involves extracting progressively smaller-sized features from an image, reducing its size through a series of convolutional and pooling layers, and increasing the number of image channels to obtain low-resolution, high-level features. In the upsampling stage, the process reverses by enlarging the image through specialized convolutions, decreasing the number of channels, and finally restoring the original size of the target image. U-Net has been proven to achieve outstanding success in fields with small datasets, such as medical imaging [12]. However, its performance can be limited by its reliance on symmetric architecture, which may not fully capture complex spatial hierarchies, especially in 3D medical imaging.

Brain tumor images are obtained by MRI and CT scanning. In 3D MRI and CT images, slicing 3D images into 2D images will lose a lot of spatial information, while 3D models help to preserve this spatial connection. Numerous 3D models inspired by the U-Net model have been developed [13–15]. For instance, the V-Net model [16] adapts the 2D U-Net into a 3D segmentation model, significantly improving accuracy. To enhance the V-Net architecture, which originally featured just a single downsampling path, Guan et al. [17] introduced the 3D AGSE-VNet. A combined segmentation model based on V-Net that integrates the SE module and AG module, the 3D AGSE-VNet cleverly utilizes the channel relationship to enhance the useful information in the channel and suppress the useless information in the channel by using the global information. However, the local receptive fields of CNNs limit their capacity to model the global context, and U-Net may have difficulty in effectively capturing details far from the region in the image when processing more complex medical images.

Therefore, Vaswani et al. [18] introduced the Vision Transformer (ViT) module [19] into various fields of computer vision as an attention-based mechanism. This module addresses the shortcomings of CNNs by providing superior capabilities to capture and model long-range features. Following the introduction of the ViT module, the CNN–Transformer combination models [20–23] were quickly developed. In the field of brain tumor image segmentation, Wang et al. [23] proposed the TransBTS model, which is a multimodal brain tumor image segmentation model that integrates the attention mechanism of the Transformer with CNNs. This model employs CNNs to create an encoder–decoder architecture, utilizes 3D CNNs to capture spatial and depth information, and leverages the ViT module to model long-range features, achieving high-resolution segmentation results. However, the encoder–decoder model that combines the Transformer and CNN modules faces some challenges [24], such as excessive computational demands, requiring more training data, and longer training times. To mitigate these computational requirements, Yu et al. [25] used only a simple nonparametric pooling operator as a token mixer; such a streamlined token mixer also delivers excellent performance and warrants further research on low-computation and efficient pooling.

In multimodal magnetic resonance imaging, earlier multimodal MRI segmentation approaches typically conduct modality integration by combining input multimodal MRIs at the beginning [26,27] or middle of the network [28,29], which limits the exploration of nonlinear dependencies between modalities. We use a more refined fusion method that allows the model to progressively map the modality of interest by progressively fusing feature maps at different resolutions. In this work, as shown in Figure 2, we propose a framework named Multiscale Modality-Aware Transformer (MMAformer). The key contributions of this work are outlined as follows:

- In this paper, we proposed a four-stage Transformer framework designed for processing parallel multimodal medical images by using the cross-modality downsampling (CMD) module. CMD modules are used to select modality types of interest to the model at different scales, enhancing the modality awareness of the Transformer network for multimodal medical images;
- We designed a multimodality gated aggregation block that combines a dual-attention mechanism with multi-gated clustering, efficiently enhancing and integrating spatial, channel, and modal features across different imaging modalities;
- We conducted sufficient experiments on several datasets to validate the segmentation accuracy of the proposed model. For example, we increased the Dice score from 90.66% to 91.53% on the BraTS2021 dataset compared to the previous method, which is an improvement of about 1%. The convergence stability of the model is illustrated by more experiments in the experimental section.

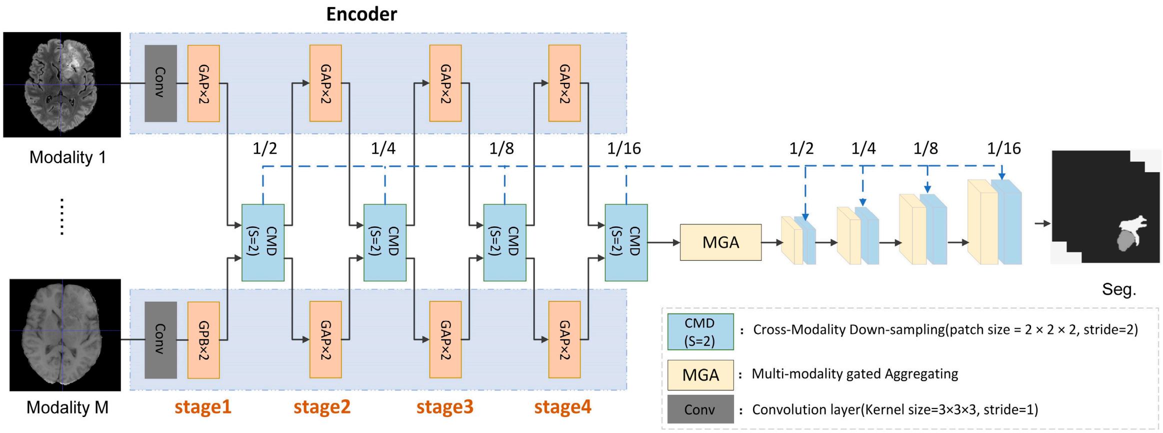


**Figure 2.** An overview of MMAformer. The four-stage CMD module is applied at different scales and the MGA is responsible for fusing multimodal information.

## 2. Related Work

### *2.1. CNN-Based Segmentation Networks*

Convolutional neural networks (CNNs) are extensively employed for image segmentation in the medical image domain [12]. Key CNN models include an architecture of an encoder, a decoder, and skip connections with notable examples such as 3D-UNet [30], SegResNet [31], and nnUNet [32]. Traditional 2D models like U-Net often face challenges in medical image segmentation due to the limited receptive fields of convolution and the lack of 3D spatial information, especially the contextual information of neighboring cuts. To fully leverage the three-dimensional spatial information of images, contemporary approaches process the entire image as input rather than individual slices. 3D-UNet enhances the traditional U-Net by implementing fully 3D convolutional operations on volumetric data, capturing the spatial context more effectively by processing entire image volumes rather than individual slices, which improves accuracy in cases where complex structures in 3D medical images need to be precisely delineated. The nnUNet stands out by automatically adapting its architecture and training protocols to various medical imaging datasets. It optimizes parameters like input resolution, batch size, and learning rate based on the dataset's specific characteristics, delivering high performance across diverse tasks without manual tuning. In medical image processing, the flexibility and adaptability of nnUNet greatly extends its range of applications.

### *2.2. Transformers-Based Segmentation Networks*

Transformer-based models effectively capture long-range dependencies and complex spatial interactions, enhancing the capabilities of traditional convolutional neural networks. TransUNet [20] integrates the strengths of U-Net and Transformers, using CNNs to capture detailed spatial information at high resolutions and Transformers to identify global dependencies across the data. UNetFormer [33] merges the U-Net architecture with a Transformer module to significantly improve feature extraction capabilities, particularly in capturing the context of large image regions, leading to better segmentation accuracy and robustness. SwinUNETR [27] incorporates the Swin Transformer within the U-Net framework and adopts a window shifting scheme, which reduces computational complexity while efficiently processing high-resolution medical images, making it ideal for precise 3D medical imaging. HyperDense-Net [34] combines dense connection layers with the relationship modeling capabilities of the Transformer, ensuring extensive feature learning and robust information flow, which enhances segmentation efficiency for complex anatomical structures. MmFormer [29] utilizes a multi-head self-attention mechanism to integrate features from different imaging modalities, providing insights into cross-modal dynamics essential for tasks such as multimodal brain tumor segmentation.

However, these networks simply stack multimodal data and do not fully exploit the mutual information between different modalities. Moreover, these methods focus only on extracting multimodal information from the bottleneck and do not fully utilize modal information at multiple scales. In contrast, we propose an end-to-end framework that can retrieve multimodal information from various scales and fuse the spatial, channel, and intermodal information simultaneously at the middle layer.

## 3. Methods

In this study, the preparation of the experimental methodology and setup took about 6 months, including designing the model, building the data preprocessing pipeline, and fine tuning the experimental parameters. The subsequent data collection and experiment execution phases took about 3 months. Finally, it took about 1 month to obtain the results and perform the comparative analysis.

We found that relying on a single imaging modality is inadequate for accurately delineating tumors and the surrounding edema. To achieve more precise predictions, this paper utilized four MRI modalities as parallel inputs, leveraging the complementary nature of multimodality image data. For instance, T2 and T2-FLAIR modalities, which employ water suppression techniques, effectively identify regions of high-moisture content such as edema to improve the discrimination of different regions. Additionally, the T1ce modality, which provides contrast enhancement in vascular regions, supplements the segmentation of the tumor core, where vascular structures are more prominent.

As shown in Figure 2, our method is used for the parallel processing of multimodal imaging data and employs a four-stage structural encoder. Each stage includes a global averaging pooling block and a cross-modal downsampling block that enables the model to search and learn modalities of interest. In addition, a Multimodal Gated Aggregation block is designed to combine the dual-attention mechanism with multiple gated clustering to efficiently integrate and enhance the spatial, channel, and modal features of multimodal images.

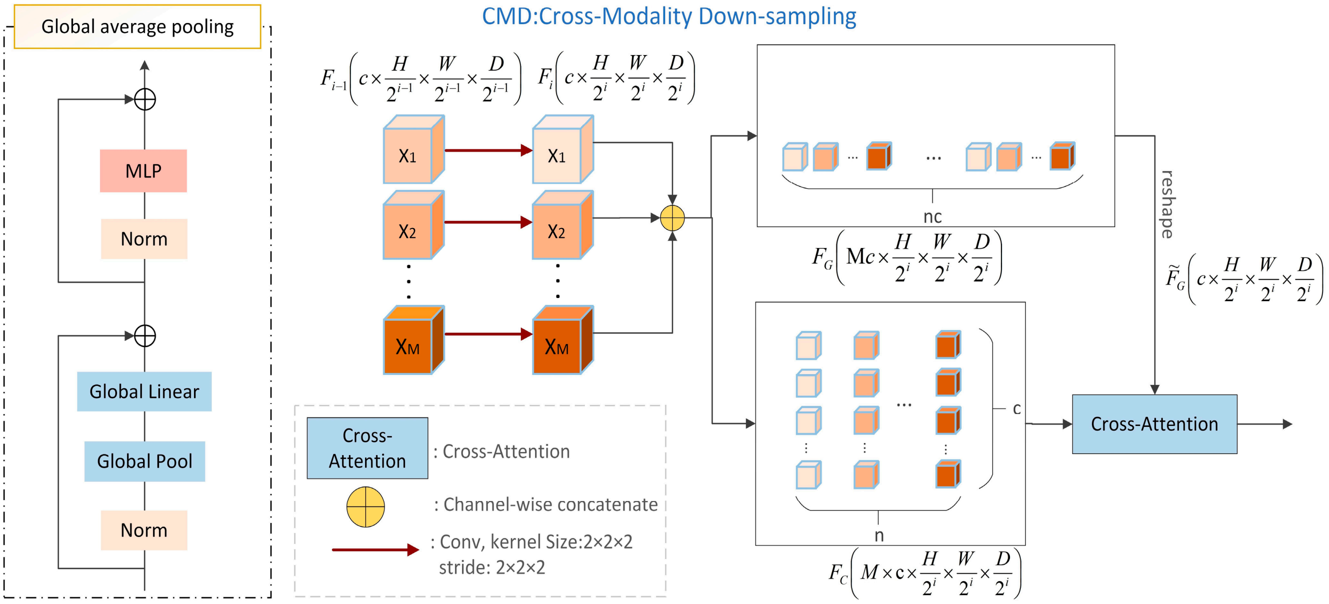
### 3.1. Dataset and Pre-Processing

Dataset: The brain tumor segmentation challenge (BraTS Chanllenge) is the oldest of all the competitions of the Medical Image Computing and Computer-Assisted Intervention Society and has been running for 10 years. It is one of the most popular competitions in the field of medical image processing. For evaluation, we used their provided BraTS2020 and BraTS2021 datasets for brain tumor segmentation. These datasets provide diverse multimodal MRI images, including T1-weighted, T1-weighted post-contrast, T2-weighted, and T2-FLAIR sequences, facilitating comprehensive assessment across different tumor types and imaging modalities.

Pre-processing: Initially, we partitioned the BraTS20 [5] and BraTS21 [6] segmentation datasets roughly into training, testing, and validation sets in a ratio of 60:20:20. Subsequently, we employed MONAI [35] for the pre-processing process. The pre-processing steps for the training process included the normalization of 3D MRI images, random cropping of patches with a resolution of 128 × 128 × 128 from multimodal 3D MRI images, and data augmentation. The augmentation process involved random per-channel intensity shifts between $-0.1$ and 0.1 and random scaling of intensity between 0.9 and 1.1. Additionally, random axis mirror flips were applied with a probability of 0.5 for all major axes.

### 3.2. Encoder

To enhance the utilization of independent information from each modality, this paper adopts a parallel processing approach for multimodality data, as shown in Figure 3. From modality 1 to modality M, the input images for each modality are generated using different magnetic resonance imaging techniques. The input image $X_m \in R^{H \times W \times D}$, where $m \in (1, M)$ and M = 4, with W, H, and D representing the width, height, and depth of the input image, respectively.



Global average pooling
MLP
Norm
Global Linear
Global Pool
Norm
CMD:Cross-Modality Down-sampling
$F_{i-1}\left(c\times\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times\frac{D}{2^{i-1}}\right)$
$F_i\left(c\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}\right)$
X1
X2
XM
X1
X2
XM
nc
$F_G\left(Mc\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}\right)$
reshape
$\widetilde{F}_G\left(c\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}\right)$
c
n
$F_C\left(M\times c\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}\right)$
Cross-Attention
Cross-Attention
: Cross-Attention
: Channel-wise concatenate
: Conv, kernel Size:2×2×2 stride: 2×2×2

(**a**) (**b**)

**Figure 3.** Encoder. (**a**) global average pooling; (**b**) cross-modality downsampling.

The encoder is divided into five stages. Stage 0 consists of a convolutional layer with a kernel size of 1 × 1 × 1, transforming the input image $X_m$ into $F_{0,m} \in R^{C\times H\times W\times D}$, C is the number of feature channels in the first stage, H, W, D is the feature map size in the first stage, where C = 16. Stages 1 to 4 each comprise two global average pooling (GAP) blocks and one cross-modal downsampling (CMD) block, progressively encoding each modal image into multi-level feature $F_{i,m} \in R^{c\times h\times w\times d}$, where i denotes the stage number ranging from 1 to 4, c is the number of characteristic channels in stage i, and (h, w, d) is the feature map size in stage i, with $(h, w, d) = \left(\frac{H}{2^i}, \frac{W}{2^i}, \frac{D}{2^i}\right)$, and $c = C \times 2^i$.

Global Average Pooling

Inspired by prior research, this paper employs the Transformer architecture, which is known for its superior global feature learning over traditional Convolutional Neural Networks (CNNs). It is proposed to replace the multi-head self-attention mechanism in Transformers with global average pooling, aiming to reduce computational complexity while maintaining effectiveness. To balance computational load and feature extraction capabilities, as shown in Figure 3a, this paper incorporated a Global Average Poolformer (GAP) block for encoding across different modalities. As illustrated, a GAP block consists of a learnable global average pooling and a Multi-Layer Perceptron (MLP) block. The computation is as follows [25]:

$$\begin{gathered} Y = GAP(LN(X))W + X \\ Z = MLP(LN(Y)) + Y \end{gathered} \tag{1}$$

where X, Z are inputs and outputs, respectively, GAP is global average pooling, LN(∗) is layer normalize, and W is a learnable parameter in the linear layer.

### 3.3. Cross-Modality Downsampling

At the end of each stage, cross-modality downsampling (CMD) is utilized to select features from modalities of interest. As shown in Figure 3b, M input features $F_{i-1}$ are processed through a convolutional layer with a kernel size of 2 and stride of 2, producing M features $F_i \in R^{c\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}}$. Initially, the features of the different modalities are concatenated along the channel dimension, resulting in $F_G \in R^{Mc\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}}$, which is then reshaped to $\widetilde{F}_G \in R^{c\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}}$. Subsequently, a new modality dimension is concatenated, forming $F_C \in R^{M\times c\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}}$. Finally, to calculate the cross-attention between $\widetilde{F}_G$ and $F_C$, the query (Q), key (K) and value (V) in the attention formula are computed as follows [18]:

$$Q = \widetilde{F}_G W_q, \quad K = F_C W_k, \quad V = F_C W_v \tag{2}$$

where $W_q$, $W_k$, and $W_v$ represent the weight matrices for the query, key, and value. Cross-attention is computed between $\widetilde{F}_G$ and $F_C$, while the input $F_{i,m}$ for the next stage comes from the cross-attention operation [18]:

$$F_{i,m} = CA\left(\widetilde{F}_G, F_C\right) = \widetilde{F}_G + SoftMax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

where CA denotes cross-attention and d is the dimensionality of Q, K, V. Using the CMD module, the model is able to prioritize the learning of modal features of interest, thus enhancing the integration of information across different modalities.

### 3.4. Multimodality Gated Aggregating

The encoder progressively encodes each modal image into advanced features, $F_{i,m} \in R^{c\times h\times w\times d}$, where m ranges from 1 to M, and i ranges from 1 to 4. At the highest-level features, the number of layers i= 4, and $(\mathrm{h}, \mathrm{w}, \mathrm{d}) = \left(\frac{\mathrm{H}}{16}, \frac{\mathrm{W}}{16}, \frac{\mathrm{D}}{16}\right)$ represents a reduction to 1/16 of the original dimensions of depth D, width W, and height H in the input space. The channel dimension c = 128, respectively.

Given the highest-level features $\mathrm{F}_{4,\mathrm{m}} \in \mathrm{R}^{128\times\frac{\mathrm{H}}{16}\times\frac{\mathrm{W}}{16}\times\frac{\mathrm{D}}{16}}$, $\mathrm{m} \in (1, \mathrm{M})$, a new dimension is introduced in these features, namely the modality dimension. The MGA module utilizes a dual-attention block [36] and multi-gated clustering for end-to-end operations to integrate spatial, channel, and modal information.

As shown in Figure 4, the first part employs the dual-attention block. The input feature map is fed into two feature extraction blocks, and after passing through the position feature extraction block and channel feature extraction block, respectively, the image integrating specific position and channel information is output.

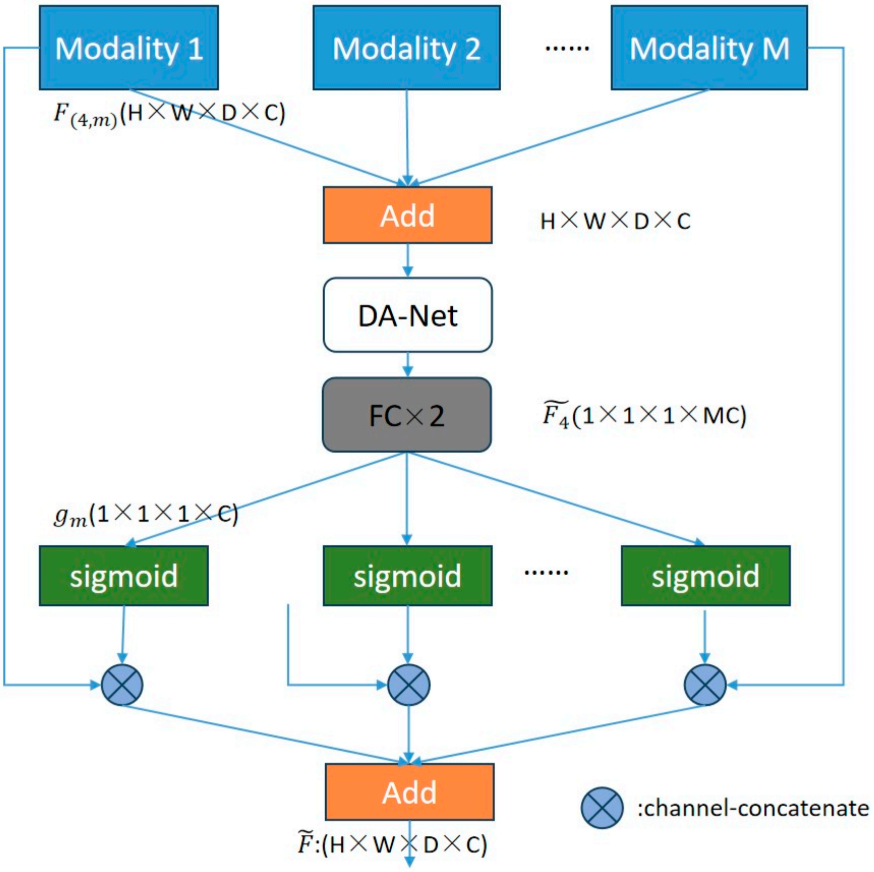


**Figure 4.** Multimodality gated aggregating block.

In the second part, multi-gated clustering further computes the global relationships between different modalities and facilitates the fusion of intermodal information. This paper introduces a modality-sensitive gating strategy, which uses sigmoid gating to model intermodal information.

### 3.4.1. Cluster Spatial and Channel Information

Initially, given the highest-level features $\mathrm{F}_{4,\mathrm{m}} \in \mathrm{R}^{128\times\frac{\mathrm{H}}{16}\times\frac{\mathrm{W}}{16}\times\frac{\mathrm{D}}{16}}$, $\mathrm{m} \in (1, \mathrm{M})$, as shown in Figure 4, the high-level feature maps $\mathrm{F}_{4,\mathrm{m}}$ of M modalities are summed across the spatial and channel dimensions to obtain $\mathrm{F}_4 \in \mathrm{R}^{128\times\frac{\mathrm{H}}{16}\times\frac{\mathrm{W}}{16}\times\frac{\mathrm{D}}{16}}$. Subsequently, $\mathrm{F}_4$ is input into the dual-attention (DA) block to capture its global feature dependencies in both the spatial and channel dimensions, resulting in $\widetilde{\mathrm{F}_4}$. The details are computed as follows:

$$\widetilde{F_4} = \mathrm{DA}\left(\sum_{m=1}^{4} F_{4,m}\right) \tag{4}$$

where $\sum_{m=1}^{4}(*)$ is a pixel summing operation, m refers to the number of modalities (there are four modalities), and $\mathrm{DA}(*)$ is the dual-attention block. This component is used to cluster spatial and channel information.

3.4.2. Cluster Intermodal Information

Given tokens $\widetilde{F_4}$, fully connected layers and convolutional layers are used here to perform the operations of channel squeezing and expanding to obtain the point attention vector in modal dimension with a shape of (1, 1, 1, C), and C = 128. After passing through the gating function, correlations of different modal dimensions are included in $g_m$, which is calculated as follows:

$$g_m = \sigma\left(\mathrm{Conv}\left(\mathrm{FC}\left(\widetilde{F_4}\right)\right)\right) \tag{5}$$

where $\mathrm{FC}(*)$ is the fully connected layer, $\mathrm{Conv}(*)$ is a three-dimensional convolutional layer, and $\sigma(*)$ is the sigmoid function.

Then, $\widetilde{F} \in R^{128\times\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}}$ is the output of MGA, which simultaneously integrates spatial, channel and intermodal information. It can be computed as follows:

$$\widetilde{F} = \sum_{m=1}^{4} F_{4,m} \odot g_m \tag{6}$$

where $\odot$ is the element-wise multiplication. This approach improves the characterization of features and is important for dealing with complex multimodal data.

### *3.5. Loss Function*

In this study, we employed a combination of Dice Loss [37] and Cross-Entropy Loss [38] to optimize our model. Each of these loss functions has its own strengths, and by combining them, we can enhance the overall performance and robustness of the model.

Dice Loss. Unlike the traditional Dice coefficient, Dice Loss calculates the loss using continuous probability values, making it more suitable for model training. The formula is as follows [37]:

$$\text{Dice loss} = 1 - \frac{2\sum_i^N p_i g_i + \in}{\sum_i^N p_i + \sum_i^N g_i + \in} \tag{7}$$

where N is the number of samples, $p_i$ represents the predicted values, $g_i$ represents the ground truth values, and $\in$ is an infinitesimal number. Since the Dice Loss method focuses on the relative overlap between the predicted and true regions, the Dice Loss method performs well in dealing with category imbalance.

Cross-Entropy Loss. Cross-Entropy Loss is a function used to measure the difference between two probability distributions. In medical image segmentation tasks, the cross-entropy loss can be used in segmentation tasks to calculate the difference between the predicted boundary and the true boundary. The formula is as follows [38]:

$$\text{Cross} - \text{Entropy Loss} = -\sum_i^N g_i \log(p_i) \tag{8}$$

where $g_i$ is the true label of the sample (usually 0 or 1), and $p_i$ is the predicted probability value. Cross-Entropy Loss optimizes the model by maximizing the log probability of the true labels, making the predictions more accurate.

Combination of Dice Loss and Cross-Entropy Loss. Both pixel and boundary accuracy are important in segmentation tasks. By combining Dice Loss and Cross-Entropy Loss, we can leverage the strengths of both to refine model training. The combined loss function is as shown below:

$$\text{Total Loss} = \alpha \times \text{Dice Loss} + (1 - \alpha) \times \text{Cross} - \text{Entropy Loss} \tag{9}$$

where $\alpha$ represents the hyperparameters that adjust the relative weights of the two loss functions. By properly tuning these weights, we can better balance the needs of segmentation and classification tasks.

By using this combined loss function, our model is more effective in dealing with class imbalance and distinguishing between classes.

### *3.6. Evaluation Metrics*

We used the Dice score and 95% Hausdorff distance (HD95) as quantitative metrics for comparison. The Dice score quantifies the degree of pixel overlap between predicted and ground truth segmentations, and the HD95 metric quantifies the maximum boundary discrepancy, providing a holistic evaluation of segmentation accuracy.

Together, the Dice score and HD95 provide complementary insights into the performance of our segmentation model. The Dice score quantifies the overall pixel overlap, making it a valuable metric for assessing the model's accuracy in capturing relevant structures, while HD95 offers a more detailed evaluation of boundary precision, which is critical in medical image segmentation tasks where the exact delineation of structures is paramount. These metrics allow for a holistic evaluation of the segmentation quality, ensuring both accurate structure representation and precise boundary delineation.

## **4. Experimental Setup**

### *4.1. Implementation Details*

We deployed the model's training on an NVIDIA GTX 3090 GPU, and it uses the PyTorch framework. The loss function employed was a combination of Dice Loss and Cross-Entropy Loss, where $\alpha$ is 0.5. The optimizer used the AdamW [39] with a weight decay of $10^{-4}$. The learning rate was empirically set to $10^{-4}$. The batch size per GPU was set to 1. All models were trained for 500 epochs with a linear warmup and a cosine annealing learning rate scheduler. The best model checkpoints from the validation set were used for inference. Inference was conducted using a sliding window approach with a 0.5 overlap for neighboring voxels.

### *4.2. Experiments*

In our comparative analysis, we compared our network with representative segmentation methods in the field in recent years to evaluate the results. For fairness, we used publicly available implementations of these methods and re-trained their networks to obtain the best segmentation results. The detailed experiments are as follows.

#### 4.2.1. Experiment 1—Comparison in the BraTS2020 Dataset

The training environment is consistent across the methods compared with up to 600 epochs on the BraTS2020 dataset. The segmentation task focuses on delineating the whole tumor (WT), enhanced tumor (ET), and tumor core (TC) regions. We used a random selection method to divide the BraTS2020 dataset into a training set (220), a validation set (80), and a test set (69). The evaluation metrics used in the laboratory were Dice score and HD95.

#### 4.2.2. Experiment 2—Comparison in the BraTS2021 Dataset

The training environment is consistent across the methods compared with up to 500 epochs on the BraTS2021 dataset. The segmentation task focuses on delineating the whole tumor (WT), enhanced tumor (ET), and tumor core (TC) regions. We used a random selection method to divide BraTS2021 into a training set (750), a validation set (250), and a test set (251). The evaluation metrics used in the laboratory were Dice score.

#### 4.2.3. Experiment 3—Ablation Study on BraTS2021

We also conducted controlled trials of the proposed GAP, CMD, and MGA modules, and we conducted ablation experiments on the three modules in three phases: training, validation, and testing. We conducted ablation experiments on the training stage, the validation stage and the test stage, respectively. We also plotted the ablation experiment training process and evaluated the metrics using Dice scores.

#### 4.2.4. Experiment 4—Comparison of Convergence Stability of Training Models

We chose two representative methods from the last two years in medical image segmentation to compare with our method, and used the Dice score as a metric to evaluate the training stability of the three methods.

## 5. Results and Discussion

In Experiment 1, we compare our method with seven other methods (including three CNN-based methods and four CNN-combined Transformer methods). They are 3D-UNet [13], SegResNet [31], nnUNet [32] and SwinUNet (2D) [22], TransBTS [23], UNETR [26] and SwinUNETR [27]. As shown in Table 1, the Dice and HD95 scores of each model for segmenting the three different tumor regions on the BraTS2020 dataset are listed along with their average scores on this dataset. Our model complexity was moderate with 18.76 million parameters and 104.6 GFLOPs. MMAformer achieved the highest Dice scores on TC and ET, the lowest HD95 scores on WT and TC, the second-best Dice score on WT. Our method achieved excellent quantitative results in both average Dice metric and average HD95 metric with mean values of 86.3 and 4.774, respectively.

**Table 1.** Quantitative comparison on BraTS2020 dataset. Evaluation of indicator use Dice score values (Dice) and Hausdorff distance values (HD95). Whole tumor (WT), enhancing tumor (ET), and tumor core (TC). Bold represents the best indicator. Down arrows indicate that lower scores are better and up arrows indicate that higher scores are better. Bolding indicates optimal scores.

| Methods | Parm | FLOPs | WT | | TC | | ET | | Ave | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (M) ↓ | (G) ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ |
| 3D-UNet 122 | 5.75 | 1449.59 | 88.2 | 5.113 | 83.0 | 6.604 | 78.2 | 6.715 | 83.1 | 6.144 |
| SegResNet 31 | 18.79 | 185.23 | 90.3 | 4.578 | 84.5 | 5.667 | 79.6 | 7.064 | 84.8 | 5.763 |
| nnUNet 32 | 5.75 | 1449.59 | 90.7 | 6.94 | 84.8 | 5.069 | 81.4 | 5.851 | 85.6 | 5.953 |
| SwinUNet (2D) 21 | 27.17 | 357.49 | 87.2 | 6.752 | 80.9 | 8.071 | 74.4 | 10.644 | 80.8 | 8.489 |
| TransBTS 22 | 32.99 | 333 | 91.0 | 4.141 | 85.5 | 5.894 | 79.1 | **5.463** | 85.2 | 5.166 |
| UNETR 25 | 92.58 | 41.19 | 89.9 | 4.314 | 84.2 | 5.843 | 78.8 | 5.598 | 84.3 | 5.251 |
| SwinUNETR 26 | 62.5 | 295 | **92.0** | 4.907 | 85.3 | 7.218 | 80.5 | 11.419 | 85.9 | 7.848 |
| MMAformer (ours) | 18.76 | 104.6 | 91.3 | **3.873** | **86.6** | **4.559** | **81.1** | 5.890 | **86.3** | **4.774** |

In Experiment 2, we also compared our network to seven SOTA methods from recent years. They are DynUnet (2021) [32], SegtransVAE (2022) [40], SwinUNETR (2022) [27], PSwinBTS (2022) [41] and CKD-TransBTS (2023) [42]. As shown in Table 2, the Dice scores of each model for segmenting the three different tumor regions in the BraTS2021 dataset are listed, along with their average scores on this dataset. Our model achieved the highest Dice score on TC and ET and the second-best Dice score on WT. Notably, MMAformer achieved the best overall performance in the average Dice score with a mean value of 91.53. On smaller tumor regions, the Dice values of WT and ET surpass the previous models, suggesting that the ability of MMAformer to segment smaller targets is enhanced by integrating modality information.

The automatic segmentation of small targets in medical images is affected by the limitations of individual MRI images themselves. As can be seen from Experiment 1 and Experiment 2, our model in the fusion of different modalities of MRIs, each modality's MRI features complement each other, obtaining 1.3 points and 0.6 points of enhancement on BraTS2020 and 2 points and 0.3 points of enhancement on BraTS2021 for the segmentation of small targets ET and TC, respectively.

**Table 2.** Compared to SOTA on BraTS2021. Evaluation of indicator using Dice score values. Whole tumor (WT), enhancing tumor (ET), and tumor core (TC). Bold represents the best indicator.

| Methods | WT | TC | ET | Ave | Year |
|---|---|---|---|---|---|
| DynUnet 32 | 92.88 | 89.71 | 85.81 | 89.46 | 2021 |
| SegtransVAE 40 | 92.54 | 89.99 | 86.22 | 89.58 | 2022 |
| SwinUNETR 26 | 92.73 | 89.98 | 86.81 | 89.84 | 2022 |
| PSwinBTS 41 | **93.62** | 90.43 | 88.25 | 90.76 | 2022 |
| CKD-TransBTS 42 | 93.33 | 90.16 | 88.50 | 90.66 | 2023 |
| MMAformer (ours) | 93.58 | **92.21** | **88.78** | **91.53** | ours |

As shown in Figure 5, we performed the visualization on BraTS2020. A comparison with other methods shows that our method is more accurate in segmenting the small target enhancing tumor. This is due to the fact that our model allows for complementary multimodal information in the encoder, and segmenting small targets is more advantageous.

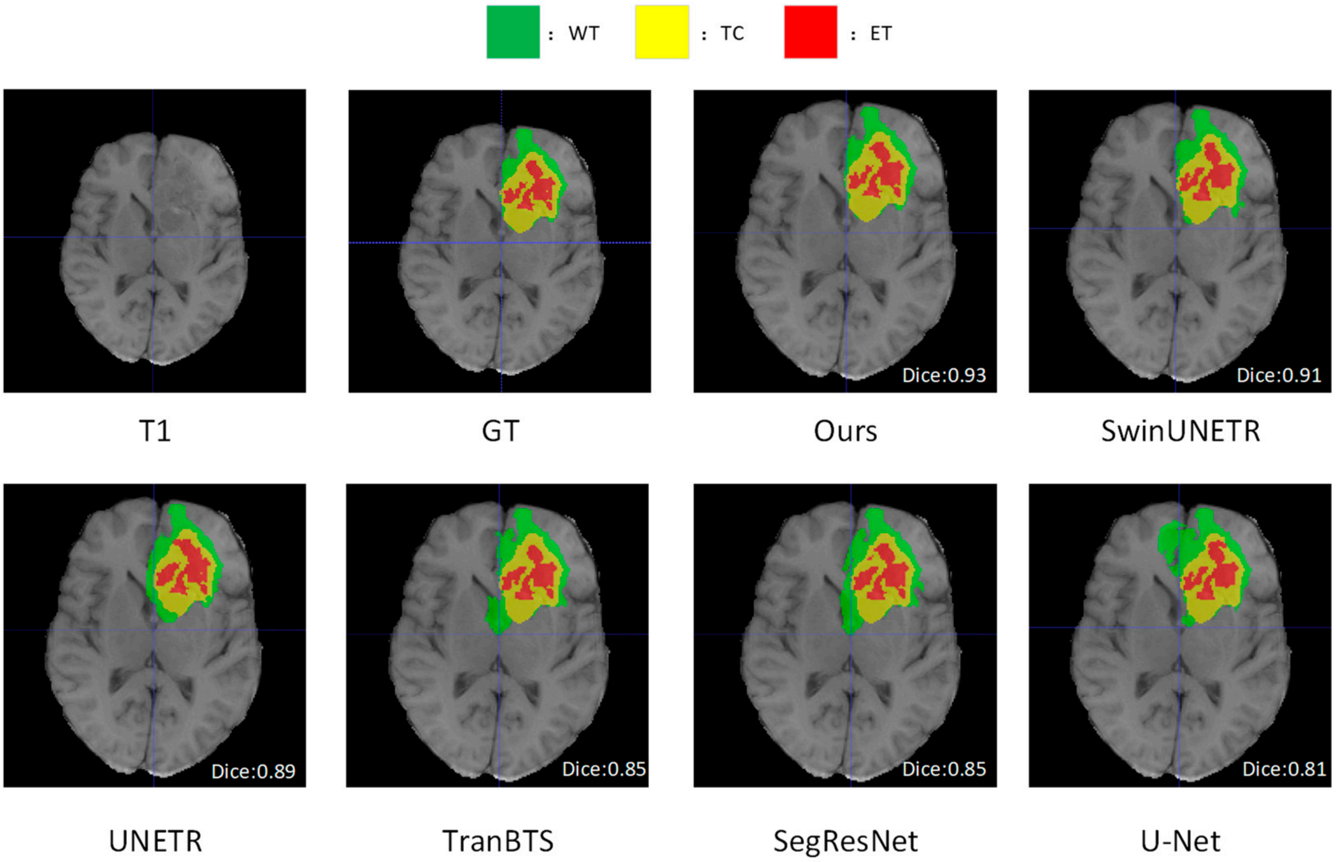


**Figure 5.** Comparison with other methods on BraTS2020.

In Experiment 3, we also conducted controlled trials of the proposed GAP, CMD, and MGA modules. As shown in Table 3, we conducted ablation experiments on the three modules in three stages: training, validation, and testing. In the U-Net method, we deploy multiple U-shaped coding layers in parallel as encoders and then input different modal images in parallel to extract features and perform feature fusion by concatenate. The Baseline method is the replacement of the encoded portion with GAP; the Baseline + CMD method has CMD blocks added; Baseline + CMD + MGA has an MGA module added, which is the MMAformer. We use the Dice score as an evaluation metric, and this ablation experiment demonstrates that the addition of GAP, CMD, and MGA modules can improve model performance as a whole. As shown in Figure 6, analyzing three curves using Dice score as an evaluation metric, the CMD and MGD modules improve the predictive accuracy and stability of our model. Compared to the Baseline method and Baseline + CMD method, our CMD blocks can help the model extract intermodal relationships between features of different scales. Compared to the Baseline + CMD method and Baseline + CMD + MGA, our MGA module can simultaneously fuse spatial, channel and modal information during feature fusion.

**Table 3.** Ablation study at different stages on BraTS2021. Evaluation of indicator using Dice score values. Different stages mean training stage, validation stage and testing stage.

| Methods | Training | Validation | Testing |
|---|---|---|---|
| U-Net | 89.68 | 89.83 | 89.10 |
| Baseline | 90.23 | 90.48 | 90.35 |
| Baseline + CMD | 90.66 | 90.88 | 90.80 |
| Baseline + CMD + MGA | 91.56 | 91.60 | 91.53 |

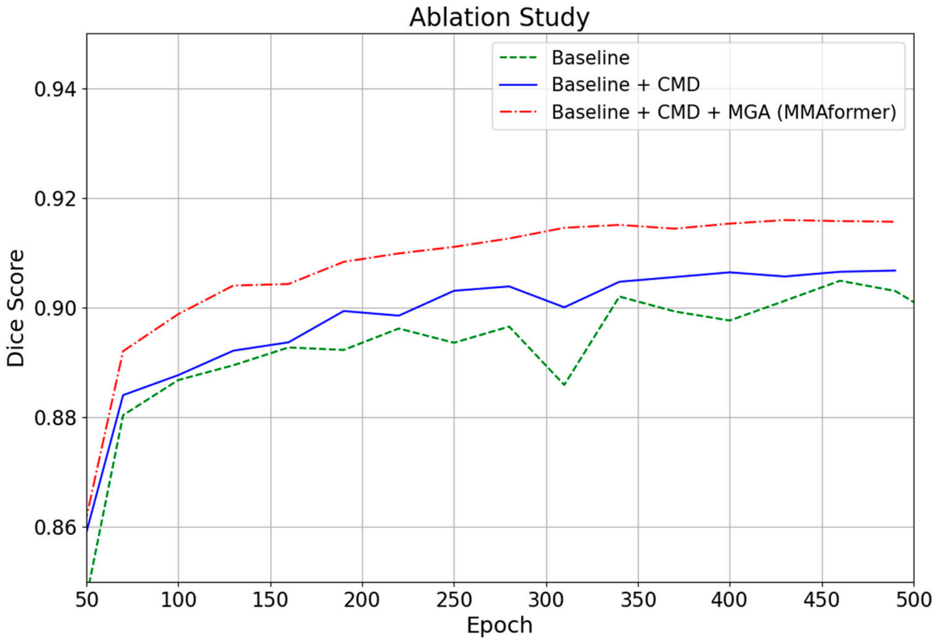


**Figure 6.** Ablation study. The red solid dashed line represents baseline, the yellow solid line represents baseline +CMD, and the green dashed line represents MMAformer. Horizontal axis is epochs, vertical axis is Avg Dice values.

In Experiment 4, we trained SwinUNETR, CKD-TransBTS and MMAformer on the Brats2021 dataset using the same hyperparameters. As shown in Figure 7, the red color represents MMAformer, which has faster convergence, higher Dice scores, and better overall stability than the other two models.

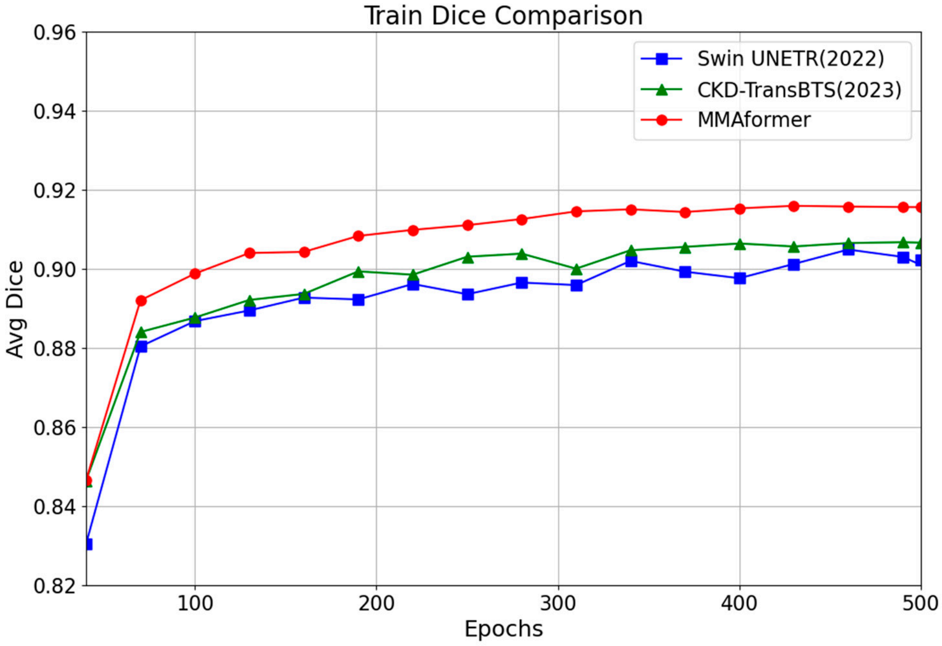


**Figure 7.** Comparison of convergence stability of training models. The three lines represent Swin-UNTER (2022), CKD-TransBTS (2023), and MMAformer (ours). Horizontal axis is epochs, vertical axis is Avg Dice values.

*Limitation*

Our model complexity was moderate with 18.76 million parameters and 104.6 GFLOPs. Our model achieves lightweight architecture to reduce computational complexity, but the CMD module, which uses cross-attention to fuse multimodal information, incurs a high computational cost. Currently, we handle up to four modalities on a 3090 Ti GPU, but processing more modalities or increasing patch resolution may result in longer processing times. Additionally, the current model should be extended with a more comprehensive training dataset to improve generalization and facilitate its application to a wider range of medical imaging tasks. Finally, the model should be optimized for compatibility with other imaging modalities, such as CT and ultrasound, to enable the real-time segmentation of a broader range of medical images.

## 6. Conclusions

In this study, we proposed MMAformer, which is a Multiscale Modality-Aware Transformer model. We stack the GAP blocks in the encoder to save a lot of computation. The encoder extracts the features of M modalities in parallel. Cross-modality downsampling is used to fuse the intermodal information of multiscale features during downsampling, enriching the integration of information across different modalities. High-level features are subsequently fused with spatial, channel, and modal features through the Multimodality Gated Aggregation module. Through these modules, the network can gradually learn features of interest from an early stage, efficiently extracting and fully mixing features of different patterns. The validity of MMAFormer is verified on the BraTS2020 and BraTS2021 datasets. Our framework is independent of the number of modalities and data types; it can extend to any other multimodal medical data. In future explorations, we will further explore the potential of jump connectivity to enable the back-and-forth exchange of multimodal information.

**Author Contributions:** Conceptualization, H.D. and X.Z.; Methodology, H.D.; Software, H.D.; Validation, H.D., X.Z. and W.L.; Formal Analysis, H.D.; Investigation, H.D.; Resources, H.D.; Data Curation, H.L.; Writing—Original Draft Preparation, H.D.; Writing—Review and Editing, X.Z. and

W.L.; Visualization, H.D.; Supervision, X.Z.; Project Administration, X.Z.; Funding Acquisition, X.Z. and F.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (Grant Nos. 61862029 and 62171285) and the General Research Fund of Shanghai Normal University (Grant Nos. KF2021100 and Sk201220).

**Data Availability Statement:** The BraTS2020 dataset and BraTS2021 dataset used in the experiment can be downloaded at https://www.kaggle.com/datasets/awsaf49/brats2020-training-data (accessed on 10 December 2023), and https://www.med.upenn.edu/cbica/brats2021/#Data2 (accessed on 4 March 2024).

**Acknowledgments:** The authors would like to thank all contributors and reviewers for their insightful comments and suggestions that helped improve this manuscript.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Balwant, M.K. A Review on Convolutional Neural Networks for Brain Tumor Segmentation: Methods, Datasets, Libraries, and Future Directions. *IRBM* **2022**, *43*, 521–537. [CrossRef]
2. Ostrom, Q.T.; Patil, N.; Cioffi, G.; Waite, K.; Kruchko, C.; Barnholtz-Sloan, J.S. Cbtrus statistical report: Primary brain and central nervous system tumors diag-nosed in the united states in 2013–2017. *Neuro-Oncology* **2020**, *22* (Suppl. 1), iv1–iv96. [CrossRef] [PubMed]
3. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117. [CrossRef] [PubMed]
4. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overallsurvival prediction in the brats challenge. *arXiv* **2018**. [CrossRef]
5. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef]
6. Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F.C.; Pati, S.; et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv* **2021**. [CrossRef]
7. Xie, H.; Yang, D.; Sun, N.; Chen, Z.; Zhang, Y. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognit.* **2019**, *85*, 109–119. [CrossRef]
8. Pak, M.; Kim, S. A review of deep learning in image recognition. In Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), Kuta Bali, Indonesia, 8–10 August 2017; pp. 1–3. [CrossRef]
9. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2015**, *39*, 3431–3440. [CrossRef]
10. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–14. [CrossRef]
11. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
12. Du, G.; Cao, X.; Liang, J.; Chen, X.; Zhan, Y. Medical Image Segmentation based on U-Net: A Review. *J. Imaging Sci. Technol.* **2020**, *64*, 1–12. [CrossRef]
13. Qamar, S.; Jin, H.; Zheng, R.; Ahmad, P.; Usama, M. A variant form of 3D-UNet for infant brain segmentation. *Future Gener. Comput. Syst.* **2020**, *108*, 613–623. [CrossRef]
14. Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; Nsndi, K.A. Medical image segmentation using deep learning: A survey. *IET Image Process.* **2022**, *16*, 1243–1267. [CrossRef]
15. Wu, W.; Gao, L.; Duan, H.; Huang, G.; Ye, X.; Nie, S. Segmentation of pulmonary nodules in CT images based on 3D-UNET combined with three-dimensional conditional random field optimization. *Med. Phys.* **2020**, *47*, 4054–4063. [CrossRef] [PubMed]
16. Milletari, F.; Navab, N.; Ahmadi, S.A.; Net, V. Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), COUNTRY, Stanford, CA, USA, 15 June 2016. [CrossRef]
17. Guan, X.; Yang, G.; Ye, J.; Yang, W.; Xu, X.; Jiang, W.; Lai, X. 3D AGSE-VNet: An automatic brain tumor MRI data segmentation framework. *BMC Med. Imaging* **2022**, *22*, 6. [CrossRef] [PubMed]
18. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010. [CrossRef]

19. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 1–41. [CrossRef]
20. Chen, J.; Lu, Y.; Yu, Q.T. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**. [CrossRef]
21. Selvi, S.; Vishvaksenan, A.; Rajasekar, E. Cold metal transfer (CMT) technology-An overview. *Def. Technol.* **2018**, *14*, 28–44. [CrossRef]
22. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Nature Switzerland, Tel Aviv, Israel, 18 February 2023. [CrossRef]
23. Wenxuan, W.; Chen, C.; Meng, D.; Hong, Y.; Sen, Z.; Li, J. Transbts: Multimodal brain tumor segmentation using transformer. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 21 September 2021. [CrossRef]
24. Wu, Z.; Liu, Z.; Lin, J.; Lin, Y.; Han, S. Lite transformer with long-short range attention. *arXiv* **2004**. [CrossRef]
25. Yu, W.H.; Luo, M.; Zhou, P.; Si, C.Y.; Zhou, Y.C.; Wang, X.C.; Feng, J.S.; Yan, S.C. Metaformer is actually what you need for vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10819–10829. [CrossRef]
26. Hatamizadeh, A.; Tang, Y.C.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 574–584. [CrossRef]
27. Hatamizadeh, A.; Nath, V.; Tang, Y.C.; Yang, D.; Roth, H.R.; Xu, D.G. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*; Springer International Publishing: Cham, Switzerland, 2021; pp. 272–284. [CrossRef]
28. Xing, Z.H.; Yu, L.Q.; Wan, L.; Han, T.; Zhu, L. NestedFormer: Nested modality-aware transformer for brain tumor segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2022; pp. 140–150. [CrossRef]
29. Zhang, Y.; He, N.J.; Yang, J.W.; Li, Y.X.; Dong, W.; Huang, Y.W.; Zhang, Y.; He, Z.Q.; Zheng, Y.F. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2022; pp. 107–117. [CrossRef]
30. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 17–21 October 2016; pp. 424–432. [CrossRef]
31. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
32. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
33. Wang, L.B.; Li, R.; Zhang, C.; Fang, S.H.; Duan, C.X.; Meng, X.L.; Peter, M.A. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
34. Dolz, J.; Gopinath, K.; Yuan, J.; Lombaert, H.; Desrosiers, C.; Ayed, I.B. HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE Trans. Med. Imaging* **2018**, *38*, 1116–1126. [CrossRef] [PubMed]
35. Cardoso, M.J.; Li, W.; Brown, R.; Ma, N.; Kerfoot, E.; Wang, Y.; Murrey, B.; Myronenko, A.; Zhao, C.; Yang, D.; et al. Monai: An open-source framework for deep learning in healthcare. *arXiv* **2022**. [CrossRef]
36. Nam, H.; Ha, J.W.; Kim, J. Dual attention networks for multimodal reasoning and matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 299–307. [CrossRef]
37. Zhao, R.; Qian, B.; Zhang, X.; Li, Y.; Wei, R.; Liu, Y.; Pan, Y. Rethinking dice loss for medical image segmentation. In Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 17–20 November 2020. [CrossRef]
38. Zhang, Z.; Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 8792–8802. [CrossRef]
39. Loshchilov, I. Decoupled weight decay regularization. *arXiv* **2017**. [CrossRef]
40. Pham, Q.D.; Nguyen-Truong, H.; Phuong, N.N.; Nguyen, K.N.; Nguyen, C.D.; Bui, T.; Truong, S.Q. Segtransvae: Hybrid cnn-transformer with regularization for medical image segmentation. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022. [CrossRef]
41. 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. [CrossRef]
42. Lin, J.; Lu, C.; Chen, H.; Lin, H.; Zhao, B.; Shi, Z.; Qiu, B.; Pan, X.; Xu, Z.; Huang, B. CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *IEEE Trans. Med. Imaging* **2023**, *42*, 2451–2461. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.