ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240502

Modality-Aware and Shift Mixer for Multi-Modal Brain Tumor Segmentation

Zhongzhen Huang[a, b], Linda Wei[a], Shaoting Zhang[b, c] and Xiaofan Zhang [a, b,*]

[a]Shanghai Jiao Tong University
[b]Shanghai AI Laboratory
[c]SenseTime Research

Abstract. Combining images from multi-modalities is beneficial for exploring various information in computer vision, especially in the medical domain. As an essential part of clinical diagnosis, multi-modal brain tumor segmentation presents a set of distinct challenges for accurately delineating both the normal anatomy and the pathologic deviations caused by the tumor. In this paper, we aim to fuse information on different imaging modalities with the medical domain knowledge to segment tumors. We present MASM, a novel Modality Aware and Shift Mixer that integrates intra-modality and inter-modality dependencies of multi-modal images for effective and robust brain tumor segmentation. Specifically, we introduce a Modality-Aware (MA) module according to neuroimaging studies for modeling the specific modality pair relationships at low levels, and a Modality-Shift (MS) module with specific mosaic patterns is developed to explore the complex relationships that are not addressed by the MA module across modalities efficiently. Experimentally, we outperform previous state-of-the-art approaches on the public Brain Tumor Segmentation dataset. Further qualitative experiments demonstrate the effectiveness and robustness of MASM.

1 Introduction

Leveraging images from multi-modalities has shown promising potential in real-world scenarios due to the contribution of various information, especially in the medical domain, where multi-modal medical images are utilized to delineate anatomical structures and other abnormal entities. For instance, the Computed Tomography (CT) plain scan can be used to evaluate morphology and detect abnormalities, and the contrast-enhanced CT scan assesses the blood supply to potential tumors, aiding in distinguishing between benign and malignant lesions. Moreover, there are several Magnetic Resonance Imaging (MRI) sequences, such as T1-weighted (T1), T1-weighted with contrast-enhanced (T1-CE), T2-weighted (T2), and T2 Fluid Attenuation Inversion Recovery (T2-FLAIR) are combined to emphasize and distinguish different tissue properties and areas of tumors, as shown in Figure 1(a). As the most common cancer worldwide, elaborating on the characterization of brain tumors is vital for studying tumor progression and pre-surgical planning. However, different from other scenarios of multi-modal segmentation, multi-modal brain tumor segmentation presents a set of distinct challenges, which can be attributed to the complex nature of brain anatomy, the variability of tumor appearance, and the intricacies involved in integrating multi-modal imaging data.

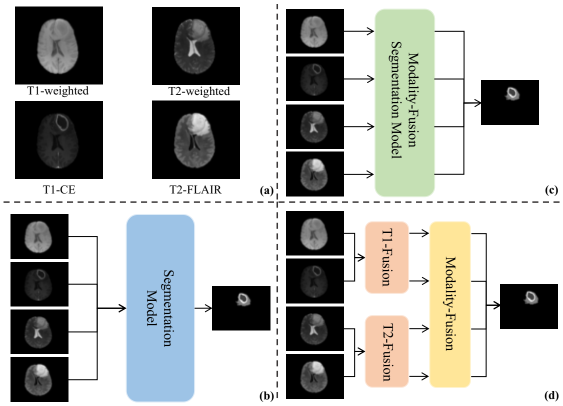

Figure 1. **(a)** Multi-modal brain tumor segmentation; **(b)** Early-Fusion strategy; **(c)** Later-Fusion strategy; **(d)** Our strategy

In recent years, promising advances in brain tumor segmentation have been primarily driven by the powerful capabilities of deep learning-based techniques. Convolution Neural Networks (CNNs) with the encoder-decoder architecture [15, 6, 12, 22, 14], which demonstrates state-of-the-art performances on various benchmarks, is an effective and reproducible solution for brain tumor delineation. Due to the capability of learning long-range dependencies, some transformer-based models [4, 10, 9] have been exploited for modeling the relation of model patches. Nevertheless, these paradigms overlook complementary information provided by different imaging modalities, which involve tumor morphology, metabolism, and blood flow. The use of multi-modal imaging data is crucial for further promoting the accuracy of delineating brain tumors.

To this end, previous methods adopted an early-fusion strategy where multi-modal images are merely concatenated at the input and processed jointly as a single stream of the network, as shown in Figure 1(b). Since brain tumors are highly heterogeneous, both in terms of their radiographic appearances, such early-fusion methods struggle to distinguish between the wide range of tumor tissue and normal variations in brain anatomy. Instead of concatenating all the modalities in the early stage, another volume of studies [17, 33, 11, 31] attempts to explore fusing features from multiple modalities in the later stage, as shown in Figure 1(c). They began to design specific fusion modules for multi-model information exchange and feature fusions. However, the exchange of information among modalities

[*] Corresponding Author. Email: xiaofan.zhang@sjtu.edu.cn

is conducted via conventional fusion techniques. The utilization of medical domain knowledge for integrating multi-modal imaging data is neglected, especially when the original T1, T2, T1-CE, and FLAIR images have different statistical properties due to significantly different image acquisition processes. Moreover, multi-scale representation interactions, which are vital for accurately delineating both brain anatomy and tumor regions, have also been disregarded. In this paper, we strive to utilize information on different imaging modalities with medical domain knowledge, as shown Figure 1(d).

Following these premises, we leverage the power of transformers and introduce a novel method dubbed as **M**odality **A**ware and **S**hift **M**ixer (**MASM**) as shown in Figure 2, which incorporates multi-scale feature modeling and multi-modal relationships mining for accurate segmentation. To model multi-modal features at low levels, we first propose a Modality-Aware module for more effective and reasonable information exchange across different modalities. Since different MRI sequences are often combined for diagnosis (e.g., T2 and FLAIR are together to observe water signals), the Modality-Aware module is carefully designed according to neuroimaging studies and attempts to establish the inter-modality dependencies between the specific pair of modalities, conforming to radiologists' process of analyzing combinations of two MRI sequences. For multi-modal features at high levels, a transformer-based Modality Shift module with specific mosaic patterns is introduced to explore and learn the relationship among modalities that are not addressed by the MA module (e.g., T1, T1-CE, and FLAIR). We endow transformers with the capability of modalities modeling without additional parameters and computational costs by shifting patches along the modality dimension for later self-attention. Each modality's patches are replaced with patches from other modalities according to the patterns. In our experiments, we measure the segmentation performance on the public Brain Tumor Segmentation Challenge (BraTS). MAMS shows robust segmentation performance across three tumor subcategories, outperforming the previous leading methods (generic and specialized). Our contributions can be summarized as follows:

- We propose a novel model that exploits the relationship and interaction among multi-modal imaging data with medical domain knowledge for accurately delineating brain tumors.
- The introduced Modality-Aware module enables reasonable information exchange on the multi-scale features according to neuroimaging studies, which are crucial for tumor segmentation. With the specific design, our introduced Modality-Shift explores the feature from multi-modalities without additional parameters or computational overhead.
- We validate the effectiveness of our method in the BraTS Challenge. MASM achieves state-of-the-art performance on the benchmark compared to the universal and specifically designed multi-modal methodologies.

2 Related Work

2.1 Medical Image Segmentation

Convolutional Neural Networks (CNNs) have demonstrated significant effectiveness in medical image segmentation tasks [27]. However, due to the local property of the convolutional kernels, the CNN-based segmentation models cannot learn long-range dependencies, which can severely impact the accurate segmentation of tumors that appear in various shapes and sizes. To cope with such an issue, another volume of transformer-based models has been exploited for powerful relation modeling. Chen *et al.* [4] proposed TransUNet, which introduced the self-attention mechanism to model the global context for high-level features. After Vision Transformer (ViT) [7] was shown to be a good visual feature extractor, there was a volume of new segmentation frameworks based on ViT. As a roadmap of utilizing a ViT as its encoder without relying on a CNN-based feature extractor, UNETR [10] has shown good performance in segmentation. Since multi-scale features play a pivotal role in medical image segmentation tasks such as tumor segmentation, a model is required to handle features across multiple scales effectively. To leverage the multi-scale features, SwinUNERT [9] was proposed to compute self-attention in an efficient schema. In medical fields, it is often necessary to use MRI scans to identify and locate brain tumors. However, a single MRI sequence, such as T1-weighted or T2-weighted images, may not provide sufficient information for accurate and robust segmentation results. Therefore, multi-modal segmentation methods should be employed, where those sequences offer complementary information about the tumor's location, size, and other characteristics, thereby enhancing the accuracy of the segmentation. However, the aforementioned methods fall short of effectively merging diverse imaging modalities. In this paper, we try reasonable and efficient modules to boost the performance in the multi-modal brain tumor segmentation task.

2.2 Multi-modal Segmentation

Recently, many methods [18, 18, 28] have been developed to tackle multi-modal image fusion in the natural image field. A common pipeline utilizes CNN-based feature extraction. The workflow involves extracting basic modality features via shared encoders and distinguishing modality-specific features via private encoders. However, these methods can hardly extract global information since CNN only extracts local information in a relatively small receptive field. Such limitations would be magnified when applied in the medical field, where 3D volume contains more contextual patches. One pointer work realized modality-specific and modality-shared feature extractions by a dual-branch Transformer-CNN feature extractor. Despite the remarkable performance, its dual-branch encoder can not be applied in brain tumor segmentation, where more modalities (the T1, T2, FLAIR, and T1-CE sequences) are utilized concurrently according to the real clinical scenario. Recently, our community has witnessed a wide adoption of deep-learning techniques to model the relationships among multi-modal images for medical image segmentation tasks. Lin *et al.* [17] proposed a clinical knowledge-driven model with a dual-branch hybrid encoder that splits the modalities into two groups based on the imaging principle as input. Xing *et al.* [31] performed nested multi-modal fusion for different modalities by establishing intra- and inter-modality coherence to build the long-range spatial dependencies across modalities. DBTrans [32] improves the segmentation process with dual-branch architectures for both the encoder and decoder, including a local branch and a global branch to capture both local and global information with linear computational complexity. Instead of applying one modality-fusion module for multi-scale features, we propose two novel modules (i.e., Modality-Aware and Modality-Shift) to exploit the relationships among multiple modalities across different scales.

3 Method

As illustrated in Figure 2, our MASM consists of a backbone based on U-Net, the Modality-Aware module, and the Modality-Shift module. We apply the Modality-Aware module and the Modality-Shift

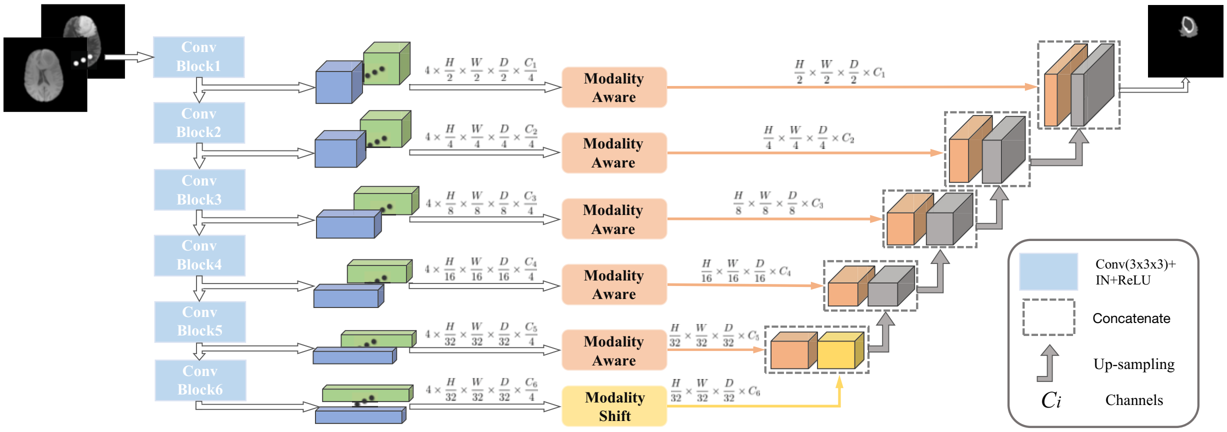
Conv Block1
Conv Block2
Conv Block3
Conv Block4
Conv Block5
Conv Block6
$4\times\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times\frac{C_1}{4}$
$4\times\frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}\times\frac{C_2}{4}$
$4\times\frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}\times\frac{C_3}{4}$
$4\times\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}\times\frac{C_4}{4}$
$4\times\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times\frac{C_5}{4}$
$4\times\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times\frac{C_6}{4}$
Modality Aware
Modality Aware
Modality Aware
Modality Aware
Modality Aware
Modality Shift
$\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times C_1$
$\frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}\times C_2$
$\frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}\times C_3$
$\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}\times C_4$
$\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times C_5$
$\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times C_6$
Conv(3x3x3)+ IN+ReLU
Concatenate
Up-sampling
Ci Channels

Figure 2. The overall architecture of *MASM* based on U-Net: 1) Modality-Aware module that handles the interaction between specific modalities, 2) Modality-Shift module that exploits high-level relationships among modalities.

module in the original skip connection for modeling the multi-modal relationships. Features from the first five layers are regarded as low-level features where the Modality-Aware is adopted, while the Modality-Shift module is only utilized in the last layer for modeling high-level features. Our experiments reveal the effectiveness of such a design. Given multi-modal images $M^i \in \mathbb{R}^{D\times H\times W}, i \in [1,4]$, our model aims to output the segmentation results $S \in \mathbb{R}^{D\times H\times W\times 3}$, where M^i represents modal T2, T1, T1-CE, and FLAIR respectively.

3.1 Backbone

Since U-Net [27] has been proven powerful in medical image segmentation, we adopt it as the backbone to extract multi-scale features from multi-modal magnetic resonance images. Instead of employing the modality-specific encoder to extract features for each modality, we leverage a single shared encoder to represent the image features. The input of each modal image individually goes through the shared module, and this design can effectively reduce the number of model parameters. It is worth noting that the parameter sharing in the encoder does not degrade the performance. In a similar way, we can formulate the extracted feature maps as:

$$F_j^i \in \mathbb{R}^{d_j\times h_j\times w_j\times \frac{C_j}{4}}$$

where $d_j, h_j, w_j = \frac{D}{2^j}, \frac{H}{2^j}, \frac{W}{2^j}, j \in [1,5]$. And the high-level embeddings from the last layer are obtained as $F_6^i \in \mathbb{R}^{\frac{D}{2^5}\times\frac{H}{2^5}\times\frac{W}{2^5}\times\frac{C_6}{4}}$.

3.2 Modality Aware Module

In clinical diagnosis, T1 images are usually combined with T1-CE images to get valuable information about the presence and nature of various abnormalities, and T2 images are often combined with FLAIR images for detecting different types of information about the tissues. Given features F_L^i in the layer L, we propose a Modality-Aware module to aggregate the features in a reasonable way, which models the relationships according to the neuroimaging studies to make the segmentation process conformance to the real scenario.

Moreover, the presence of redundant patches in each modality feature is a common occurrence for volumetric medical images, especially for multi-modal segmentation, where normal patches repeatedly appear in multiple modalities of images. Redundant patches in each modality feature may undermine the process of information exchange. Therefore, we selectively mask the uninformative tokens to obtain more discriminative features and reduce the computational scope. The Modality-Aware module produces a binary decision mask for each scale feature to decide which patches are redundant and can be pruned for each modality and substituted by alignment features from other modalities after feature mixing. As shown in the left part of Figure 3, we first flatten each feature F_L^i into a sequence $\hat{F}_L^i \in \mathbb{R}^{N_L\times\frac{C_L}{4}}$, and binary decision masks $\mathbf{D}_i \in \{0,1\}^{N_L}$ is maintained to indicate whether to mask each token or not, where $N_L = d_L \times h_L \times w_L$ is the number of the sequence embeddings. We input $\hat{F}_L^i$ to the mask prediction module and compute the decision mask $\mathbf{D}$ as follows:

$$\mathbf{y}_i^{\text{local}} = \text{MLP}(\hat{F}_L^i) \in \mathbb{R}^{N_L\times C'} \tag{1}$$

$$\mathbf{y}_i^{\text{global}} = \text{Agg}(\text{MLP}(\hat{F}_L^i)) \in \mathbb{R}^{C'}, \tag{2}$$

where C' is half the dimension of $\hat{F}_L^i$. $\mathbf{y}_i^{\text{local}}$ and $\mathbf{y}_i^{\text{global}}$ are the local and global features computed by MLP [8]. Agg is for aggregating the information from local features and can be implemented by an average pooling.

Intuitively, the local feature represents the information of each patch, and the global feature contains the context information. The global features will be expanded to the same length as local features and then concatenated with local features in the last dimension. Then, the concatenated vector is used to decide whether to mask the token:

$$\mathbf{y}_{ik} = \left[\mathbf{y}_{ik}^{\text{local}}, \mathbf{y}_{ik}^{\text{global}}\right], \quad 1 \le k \le N_L, \tag{3}$$

$$\boldsymbol{\pi}_i^{\text{gumbel}} = \text{Softmax}((\text{MLP}(\mathbf{y}_i) + G)/\tau) \in \mathbb{R}^{N_L\times 2} \tag{4}$$

$$\boldsymbol{\pi}_i^{\text{onehot}} = \text{onehot}\left(\text{argmax}\left(\boldsymbol{\pi}_{\text{gumbel}}\right)\right), \tag{5}$$

Here $G \in \mathbb{R}^{N_L\times 2}$ are i.i.d random samples drawn from the $Gumbel(0,1)$ distribution and τ is a learnable coefficient. Since it is non-differentiable to get a mask $\mathbf{D}$ sampling from $\boldsymbol{\pi}$ (i.e., one-hot), following [26, 29], we adopt the straight through trick [13] to sample from $\boldsymbol{\pi}$:

$$\mathbf{D}_i = \left(\boldsymbol{\pi}_i^{\text{onehot}}\right)^\top + \boldsymbol{\pi}_i^{\text{gumbel}} - \text{sg}\left(\boldsymbol{\pi}_i^{\text{gumbel}}\right), \in \{0,1\}^{N_L}, \tag{6}$$

where sg is the stop gradient operator and index 0 in $\mathbf{D}$ represents masking the corresponding patch. After attaining the masks, $\hat{F}_L^i$ can be pruned by the Hadamard product with $\mathbf{D}_i$:

$$\tilde{F}_L^i \leftarrow \hat{F}_L^i \odot \mathbf{D}_i, \tag{7}$$

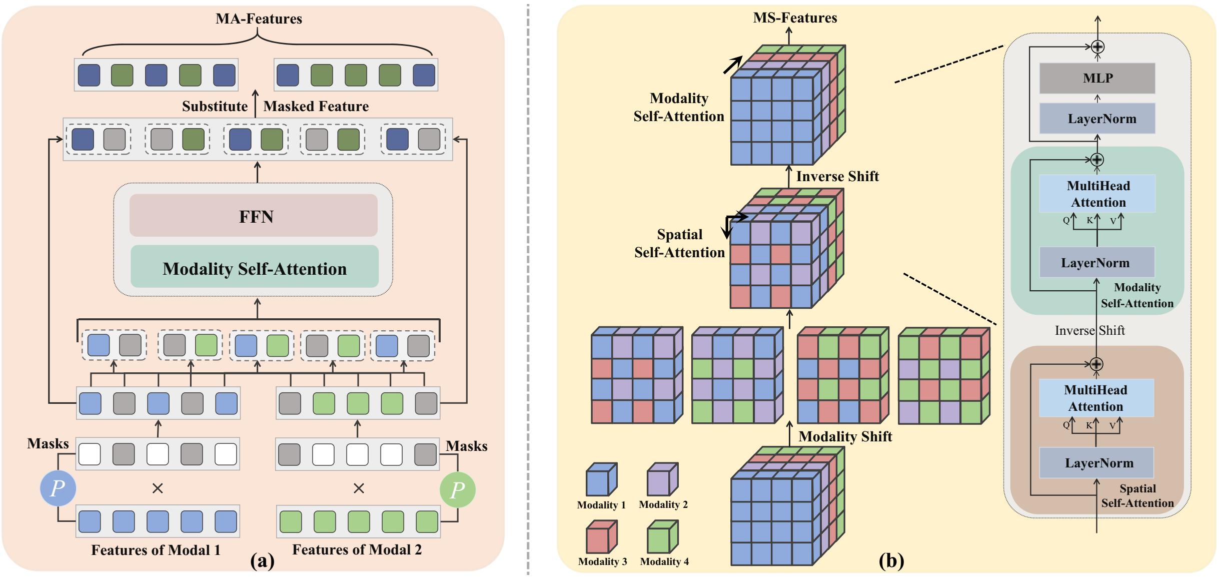

Figure 3. **(a)** The illustration of Modality-Aware; **(b)** The process of Modality-Shift. The Modality-Aware is performed on two modalities (i.e., T2 and FLAIR), and The Modality-Shift is applied to three modalities (i.e., T1, T1-CE, and FLAIR).

As shown in Figure 3, Modality Self-Attention is utilized to compute the correlation and interaction between the modalities $\tilde{F}_L^i$. We use $\tilde{F}_L^1$ and $\tilde{F}_L^4$ (i.e., T2 and FLAIR) as an example to illustrate the Modality Self-Attention. In this module, each image patch of $\tilde{F}_L^i$ (e.g. $\tilde{F}_{L1}^i$) is considered as one token, and tokens at the same location in different modalities form a paired modality sequence like $\tilde{\mathbf{z}}_1 = \{\tilde{F}_{L1}^1, \tilde{F}_{L1}^4\}$. Then the Modality Self-Attention is employed to the formed patches sequence $\tilde{\mathbf{z}} = [\tilde{F}_L^1, \tilde{F}_L^4]$. This operation can be performed efficiently by reshaping the input $\tilde{\mathbf{z}}$ with batch size B from $\mathbb{R}^{B\times 2\times N_L\times \frac{C_L}{4}}$ to $\mathbb{R}^{B\cdot N_L\times 2\times \frac{C_L}{4}}$ and leveraging the self-attention [30] scheme to compute the correlation for the above sequences:

$$Q = \tilde{\mathbf{z}}\,\mathrm{W}^Q, K = \tilde{\mathbf{z}}\,\mathrm{W}^K, V = \tilde{\mathbf{z}}\,\mathrm{W}^V \tag{8}$$

$$\text{Modality-SA}\left(\tilde{\mathbf{z}}\right) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{9}$$

$$H_L = \text{FFN}(\text{Modality-SA}\left(\tilde{\mathbf{z}}\right)) \tag{10}$$

$$\text{FFN}(x) = \max\left(0, x\,\mathrm{W}_0 + \mathrm{b}_0\right)\mathrm{W}_1 + \mathrm{b}_1 \tag{11}$$

where W^Q, W^K, $\mathrm{W}^V \in \mathbb{R}^{d\times d}$, $\mathrm{W}_0 \in \mathbb{R}^{d\times 4d}$ and $\mathrm{W}_1 \in \mathbb{R}^{4d\times d}$ are learnable parameters and b_0, b_1 are the bias terms. In our implementation, d is equal to $\frac{C_L}{4}$.

After getting the output H_L, we reshape it back to $\mathbb{R}^{B\cdot N_L\times 2\times \frac{C_L}{4}}$ and attain $[h_L^1, h_L^4]$. Then, these masked features in h_L^i are substituted with the corresponding token in h_L^j as shown in Figure 3 (a). For example, $\tilde{F}_{L3}^1$ is masked, we replace h_{L3}^1 with h_{L4}^1. The features are concatenated together to generate F_L' for the decoding stage.

3.3 *Modality Shift Module*

According to [23], later fusion of high-level features is better for the complex relationships between different modalities. Inspired by the notable performance of transformers in modeling relationships between different entities, we endow transformers with the capability of modality modeling without additional parameters and computational costs. To explore complex relationships among modalities in high-level, we introduce a Modality-Shift Module to fuse the highest level features F_6^i as shown in the right part of Figure 3.

Specific mosaic patterns are designed for patch shifting along the modality dimension. We can define a generic modality shift operation in transformers as follows:

$$\mathbf{Z}^1 = \left[\mathbf{z}_0^1, \mathbf{z}_1^1, \ldots, \mathbf{z}_N^1\right] \tag{12}$$

$$\mathbf{Z}^2 = \left[\mathbf{z}_0^2, \mathbf{z}_1^2, \ldots, \mathbf{z}_N^2\right] \tag{13}$$

$$\mathbf{A} = \left[\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_N\right] \tag{14}$$

$$\mathbf{Z} = \mathbb{I}_{\mathbf{A=1}} \odot \mathbf{Z}^1 + \mathbb{I}_{\mathbf{A=2}} \odot \mathbf{Z}^2 \tag{15}$$

where $\mathbf{Z}^1, \mathbf{Z}^2$ represent the patch features for modality 1 and modality 2, respectively. N is the number of patches, and $\mathbf{A}$ represents the matrix of shifting with $\mathbf{a}_i \in \{1, 2, 3, 4\}$ indicating the source of the shifting patch i. $\mathbb{I}$ is an indicator asserting the subscript condition. $\mathbf{Z}$ is the output image patches after shift operation.

Using the proposed modality shift operation, we can achieve information exchange among modalities at a high level. Since one specific modality has interacted with the other in the Modality-Aware module, we apply shift operation in the rest of the modality features in our case. To reduce the mixing space, we adopt fixed shift patterns. Namely, there is an invariant A_i for each M_i. For example, M^1 is interacted with M^4, thus, M^1 would be fused with M^2 and M^3 by the shift operation as follows:

$$\tilde{F}_6^1 = \mathbb{I}_{\mathbf{A_1=1}} \odot \hat{F}_6^1 + \mathbb{I}_{\mathbf{A_1=2}} \odot \hat{F}_6^2 + \mathbb{I}_{\mathbf{A_1=3}} \odot \hat{F}_6^3 \tag{16}$$

where $\hat{F}_6^i \in \mathbb{R}^{N_6\times C_6}$ is the flatten feature.

Then, we can exploit the complex relationships among modalities via the attention mechanism. The spatial self-attention and modality self-attention are employed sequentially in this module. The spatial self-attention can be performed by reshaping the input $\tilde{F}_6^i$ with batch size B from $\mathbb{R}^{B\times 4\times N_6\times \frac{C_6}{4}}$ to $\mathbb{R}^{B\cdot 4\times N_6\times \frac{C_6}{4}}$. Different from the Modalitiy-Aware module, we adopt MultiHead Attention (MHA) and LayerNorm (LN). The process can be formulated as follows:

$$\hat{X} = \mathrm{MHA}\left(\mathrm{LN}(X)\right) + X \tag{17}$$

$$\mathrm{MHA}(x) = \left[\mathrm{Att}_1(x), \ldots, \mathrm{Att}_n(x)\right] \mathrm{W}^{\mathrm{O}} \tag{18}$$

$$\mathrm{Att}_i(x) = \mathrm{Softmax}\left(\frac{x\,\mathrm{W}_i^{\mathrm{Q}}\left(x\,\mathrm{W}_i^{\mathrm{K}}\right)^T}{\sqrt{d_n}}\right) x\,\mathrm{W}_i^{\mathrm{V}} \tag{19}$$

where X denotes the input features. $\mathrm{W}_i^{\mathrm{Q}}, \mathrm{W}_i^{\mathrm{K}}, \mathrm{W}_i^{\mathrm{V}} \in \mathbb{R}^{d \times d_n}$ and $\mathrm{W}^{\mathrm{O}} \in \mathbb{R}^{d \times d}$ are learnable parameters, $d_n = d/n$, $[\cdot, \cdot]$ stands for concatenation operation. After self-attention, patches from different modalities are shifted back to their original locations as follows:

$$\tilde{S}_6^1 = \mathbb{I}_{\mathbf{A}_1=\mathbf{1}} \odot S_6^1 + \mathbb{I}_{\mathbf{A}_2=\mathbf{1}} \odot S_6^2 + \mathbb{I}_{\mathbf{A}_3=\mathbf{1}} \odot S_6^3 \tag{20}$$

where S_6^i is the output of the spatial self-attention and $\tilde{S}_6 \in \mathbb{R}^{B \times 4 \times N_6 \times \frac{C_6}{4}}$ is the visual feature after shifting back. Then, the modality self-attention is utilized to augment the feature fusion among modalities further. Similar to the operation in the Modality-Aware module, $\tilde{S}_6$ is reshaped to $\mathbb{R}^{B \cdot N_6 \times 4 \times \frac{C_6}{4}}$ and Eq 17 are applied. The output features are concatenated together along channels to generate F_6'.

3.4 Decoder

In the decoding stage, we first fold F_j' back to a $4D$ feature map $\mathbb{R}^{d_j \times w_j \times h_j \times C_j}$. Subsequently, with a 3D convolution and 2× upsampling operation, the resolution of the feature maps is increased by a factor of 2, and the outputs are concatenated with the outputs of the previous stage, a full resolution feature map is obtained and then converted to the final segmentation outputs by a sigmoid activation function. The soft Dice loss function [21] is adopted as follows:

$$\mathcal{L}(G, P) = 1 - \frac{2}{J} \sum_{j=1}^{J} \frac{\sum_{i=1}^{I} G_{i,j} P_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} P_{i,j}^2} \tag{21}$$

where I and J denote the number of voxels and classes, respectively. For class j at voxel i, $P_{i,j}$ denote the prediction of our model, and $G_{i,j}$ is the ground truth.

4 Experiments

In this section, we measure the performance of MASM on brain tumor segmentation and compare it to existing models (4.3). We ablate the proposed modules to show their importance (4.4). Finally, we visualize some cases to more intuitive demonstration.

4.1 Dataset

The BraTS dataset [20, 1, 2] is a public brain tumor segmentation dataset, including 1251 and 219 cases in the training and validation set, respectively. Each case contains four MRI modalities: a) T1-weighted, b) T1 contrasted-enhanced, c) T2-weighted, and d) T2 Fluid-attenuated Inversion Recovery (T2-FLAIR), which are rigidly aligned and resampled to the same resolution. The data were collected from multiple centers with different MRI scanners, and the labels in the training set were annotated by experts [2, 3]. The task of the dataset is to segment regions of brain tumors (i.e., whole tumor (WT), tumor core (TC), and enhancing tumor (ET)). Since the segmentation labels of the validation set are not publicly available, we adopt the training set for all the experiments.

4.2 Implementation Details

Our framework is implemented using PyTorch on an NVIDIA GTX 3090 GPU. We adopt U-Net with six layers as the backbone of our architecture. The channels C_j of each layer are $\{96, 128, 192, 256, 384, 512\}$. We also employ sinusoidal positional encodings [16] to represent the position information. AdamW [19] is utilized as the optimizer in all our experiments. The initial learning rate is 1×10^{-4}, and we decay it following the learning rate scheduling strategy of [9]. To ensure consistency with the experiment settings of previous work [31, 9, 33], each volume is cropped into patches with a size of $128 \times 128 \times 128$ and normalized to have zero mean and unit standard deviation according to non-zero voxels. Random mirroring, shift, and scale are applied for data augmentation. To gauge the performance, we employ the Dice score and 95% Hausdorff Distance (HD95) as evaluation metrics.

4.3 Comparison with SOTAs

We conduct experiments on a widely used split [24] where the 1251 MRI scans are split into 834, 208, and 209 for training, validation, and testing, respectively. To demonstrate the effectiveness, we first compare the performances of our model with a wide range of state-the-art models on BraTS21, including universal models (UNETR [10], SegTransVAE [25] and SwinUNETR [9]) and models designed for multi-modal imaging (MMEF-nnUNet [11], CKD-TransBTS [17] and NestedFormer [31]).

For a fair comparison, we adopt the results from the original papers. As illustrated in Table 1, MASM, with moderate model size and slight computation, can outperform all the state-of-the-art methods across all metrics and achieve the best segmentation performance. It can be observed that models with specific designs for multi-modal imaging attain a notable improvement compared to the universal models. The results indicate that exploring multi-modal features and dependencies is conducive to the tumor segmentation of MRI scans. Instead of considering single-modality spatial coherence and cross-modality coherence at high levels (i.e., NestedFormer), MASM introduces a more reasonable architecture for multi-scale features that is conformable to the property of each modality. Although with more parameters, it shows MASM not only improves the Dice score and HD95 score but also lowers the computations significantly with around 1/4 of the FLOPs. Such improvements demonstrate that our model can effectively learn multi-modal features and accurately identify the relationship between modalities. To further evaluate our method, we compare our method in the cross-validation split following [9] with several methods. The quantitative results are presented in Table 2, where our model outperforms the previous methods across all five folds. We conducted a significant test with SwinUNETR, which showed remarkable results. The P-values for ET, WT, and TC are **0.0001**, **0.0005** and **0.02**, respectively. The results demonstrate the effectiveness of our model even with only a modest improvement in Dice scores.

To provide a comprehensive validation of the effectiveness, we conduct experiments on BraTS23 and compare our method against the most current methods in multi-modal fusion for image segmentation in Table 3. MAMS outperforms the baselines by at least absolute 1.5% in all subclass segmentation. Our comparison includes CDDFuse, a popular multi-modal fusion method. By applying CDDFuse to fuse pairs of modalities (T1 with T1CE, and T2 with FLAIR) in a manner analogous to our Modality-Aware module, we highlight its limitations. The absence of cross-modality in-

Table 1. Quantitative comparison on BraTS 2021 dataset with respect to Dice score and 95% Hausdorff Distance. ET, WT and TC denote Enhancing Tumor, Whole Tumor and Tumor Core respectively.

Methods	Param (M)	FLOPs (G)	Dice↑				HD95↓			
			ET	WT	TC	Avg	ET	WT	TC	Avg
UNETR [10]	71.31	1159.0	0.852	0.922	0.866	0.880	12.26	7.78	7.73	9.26
SegTransVAE [25]	44.72	400.7	0.862	0.925	0.899	0.895	10.59	7.71	5.88	8.06
SwinUNETR [9]	62.19	774.8	0.871	0.925	0.899	0.897	11.06	7.62	6.86	8.51
MMEF-nnUNet [11]	76.85	208.1	0.872	0.928	0.900	0.900	9.68	8.29	**5.10**	8.29
CKD-TransBTS [17]	-	-	0.885	0.933	0.901	0.906	5.93	6.20	6.54	6.22
NestedFormer [31]	**10.57**	206.9	0.882	0.932	0.909	0.908	7.14	7.88	5.43	6.81
MASM (Ours)	24.89	**160.1**	**0.888**	**0.934**	**0.912**	**0.912**	**5.72**	**5.94**	5.40	**5.65**

Table 2. Mean Dice score for Enhancing Tumor, Whole Tumor, and Tumor Core in terms of five-fold cross-validation benchmarks. † denotes our implementation and * means we cite results from the original paper

	MASM			NestedFormer†			SwinUNETR*			nnU-Net*		
Dice	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
Fold 0	**0.891**	**0.933**	**0.915**	0.867	0.927	0.904	0.876	0.929	0.914	0.866	0.921	0.902
Fold 1	0.901	0.936	**0.919**	0.899	0.934	0.915	**0.908**	**0.938**	**0.919**	0.899	0.933	**0.919**
Fold 2	**0.891**	**0.933**	0.918	0.890	0.931	0.915	**0.891**	0.931	**0.919**	0.886	0.929	0.914
Fold 3	**0.890**	0.933	0.918	0.889	0.930	0.916	**0.890**	**0.937**	**0.920**	0.886	0.927	0.914
Fold 4	**0.891**	**0.935**	**0.919**	0.885	0.933	0.917	**0.891**	0.934	0.917	0.880	0.929	0.917
Avg.	**0.892**	**0.934**	**0.917**	0.886	0.931	0.913	0.891	0.933	**0.917**	0.883	0.927	0.913

formation exchange between T1 and T2 modalities in CDDFuse resulted in suboptimal performance, underscoring the critical importance of the Modality-Shift module within MAMS. As shown in the last four rows, MAMS surpasses the previous latest methods and demonstrates its superiority through rigorous statistical testing.

Table 3. Quantitative comparison on BraTS 2023 dataset with respect to Dice score and statistical tests (P-value).

Models	Dice↑				P-value
	ET	WT	TC	Avg.	
nnU-Net	0.873	0.920	0.904	0.899	4.61e-3
UNETR	0.870	0.919	0.901	0.896	1.47e-3
SwinUNETR	0.871	0.921	0.902	0.898	6.30e-3
CDDFuse [34]	0.875	0.932	0.911	0.906	1.05e-2
DBTrans [32]	0.877	0.933	0.913	0.907	2.45e-2
Q-CSL [5]	0.873	0.929	0.910	0.904	8.80e-3
MASM (Ours)	**0.886**	**0.936**	**0.918**	**0.913**	-

4.4 *Ablation Study*

To fully analyze our proposed modules, we conduct ablation studies on the five-fold cross-validation.

Table 4. Ablation study for proposed modules.

Module		Dice↑			Complexity↓	
Aware	Shift	ET	WT	TC	Param(M)	FLOPs(G)
		0.868	0.922	0.903	16.18	148.75
✓		0.879	0.929	0.911	24.36	160.12
	✓	0.876	0.927	0.908	16.72	149.14
✓	✓	**0.892**	**0.934**	**0.917**	24.89	160.14

4.4.1 *Effect of proposed modules*

To evaluate the effectiveness of our method, we conduct ablation studies for critical components (i.e., Modality-Aware and Modality-Shift). Table 4 summarizes the average results on the five-fold cross-validation benchmarks for the variants. We first remove the Modality-Aware and Modality-Shift modules in our MASM as the baseline in our experiments. Then, we apply Modality-Aware in all layers and Modality-Shift in the last three layers, respectively. One thing worth noticing is that the baseline differs from nnU-Net [12], where the multi-modal images are separated and fed to one single encoder sharing parameters. As can be seen, Modality-Shift boosts performance with a margin (e.g., 0.868 → 0.876) in Dice score for Enhanced tumor, and Modality-Aware brings a more considerable improvement (e.g., 0.868 → 0.879). The performance gain of the Modality-Aware module and the Modality-Shift module demonstrates that using our proposed modules helps construct the relationship information and enhance the dependency information among modalities. The reason why the combination of the Modality-Aware and Modality-Shift modules leads to significant improvement is that the design of Modality-Aware is aligned with the experience of radiologists in routine diagnosis and the highly non-linear relationships of the high-level features can be modeled by both spatial and channel-wise attention in the Modality-Shift module. Moreover, it is observed that the shift operation does not cause much increment of model parameters and computation. To validate the effectiveness of the two modules intuitively, we visualize results in Figure 5.

4.4.2 *Effect of different backbones*

In previous work, different backbones were typically applied to medical image segmentation tasks, such as UNETR and SwinUNETR. To further evaluate the impact of our proposed modules, we apply Modality-Aware and Modality-Shift on different backbones. We compare variations of different backbones as shown in Table 5. The modules are integrated into the specific layer of the framework where skip-connection is employed. It can be seen that replacing backbones does not boost performance. This could be partially attributed to the complex relationship modeling capability of the transformer-based backbone. It might degrade the interaction among multi-modal features. Moreover, comparing with the first three rows, we confirm that the proposed module is beneficial to medical multi-modal image segmentation.

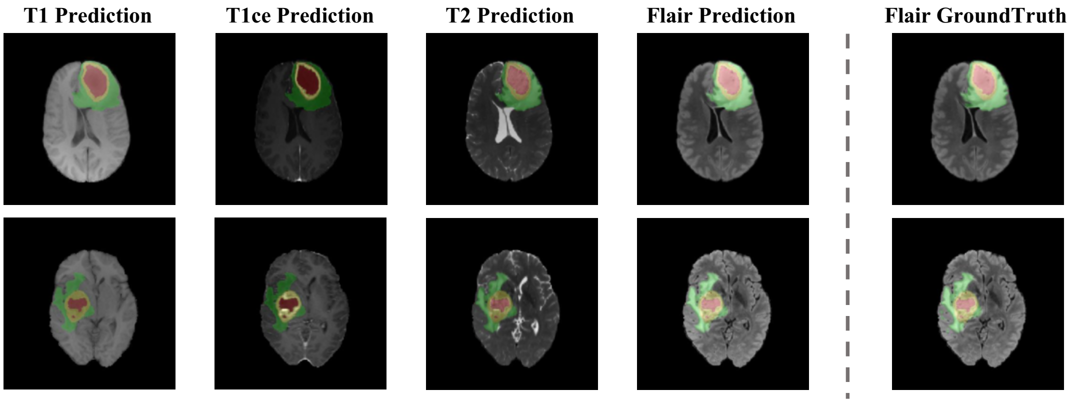

Figure 4. The visual comparison results on BraTs 2021. Segmentation examples of the predicted labels (ET, WT, TC) are overlaid on T1, T1ce, T2, and FLAIR MRI axial slices in each row. The right column is the Ground Truth.

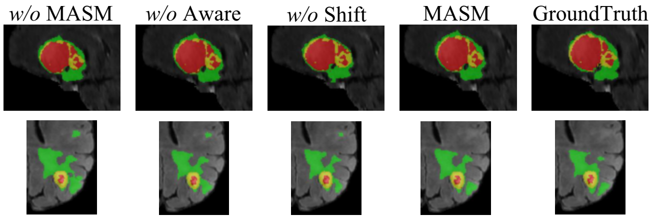

Figure 5. Visualizations of the results of designed modules.

Table 5. Ablation study for different backbones. The results are implemented by ourselves.

Backbones	Dice↑			
	ET	WT	TC	Avg
UNETR	0.867	0.921	0.892	0.893
SwinUNETR	0.861	0.927	0.898	0.895
UNet	0.866	0.924	0.906	0.898
MASM-UNETR	0.881	0.926	0.905	0.904
MASM-SwinUNETR	0.883	0.929	0.908	0.906
MASM-UNet	**0.892**	**0.934**	**0.917**	**0.914**

4.4.3 Effect of different ratios of two modules

In the encoder part, the input image first passes through a series of convolutional layers to capture fine-grained details and edge information, which can be regarded as low-level features. As the layer goes deeper, the receptive field becomes larger. These features represent higher-level abstractions. We design the Modality-Aware and the Modality-Shift modules for low-level and high-level features, respectively. To investigate the impact of the different ratios of two modules, we apply the different ratios of two modules to 6 block features, the results are shown in Table 6.

Table 6. Analysis for different ratios of two modules.

MA:MS	Dice↑			
	ET	WT	TC	Avg
3:3	0.870	0.929	0.905	0.901
4:2	0.889	0.932	0.911	0.910
5:1	**0.892**	**0.934**	**0.917**	**0.914**
6:0	0.888	0.930	0.911	0.910

4.5 Qualitative Analysis

To better understand the effectiveness of our model, we also visualize several segmentation results in Figure 4. Intuitively, the results predicted by MASM are accurate and robust, which shows better alignment with ground truth. As the figure shows, owing to the employment of the proposed Modality-Shift and Modality-Aware, our model is able to effectively fuse multi-modal MRIs and accurately segment brain tumors and peritumoral edema, even for small regions. Moreover, we include additional visual comparison results with models in Table 2 in Figure 6 (i.e., NestedFormer, SwinUNETR and nnUNet). Most methods suffer from segmentation target incompleteness-related failures and misclassification of background regions as tumors (false positives). MASM produces sharper boundaries and generates results that are more consistent with the ground truth in comparison with other models.

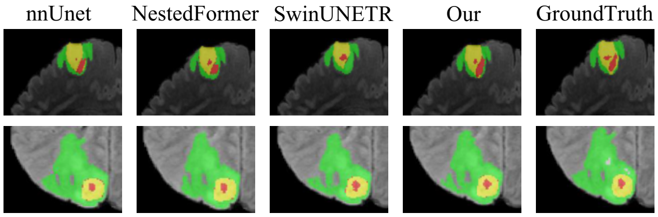

Figure 6. Visualizations of comparisons with other methods.

5 Conclusion

In this work, we present a simple yet effective approach MASM for segmenting multi-modal volumetric medical images. The key insight is to construct and identify the relationship across modalities accurately. Our proposed model uses a CNN-based U-shaped network as the encoder and decoder. Furthermore, MASM incorporates the Modality-Aware and Modality-Shift modules for learning intra- and inter-modality dependencies and is able to capture representations at multiple scales efficiently and effectively. Experimental results on the BraTS 2021 dataset validate the effectiveness of our approach. Ablation studies also prove the potential of the proposed parts. Overall, we hope this architecture can shed novel insights into learning from multi-modal medical images. More applications of MASM in medical image segmentation will be considered in future work.

References

[1] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[3] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

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

[5] J. Chen, G. Huang, X. Yuan, G. Zhong, Z. Zheng, C.-M. Pun, J. Zhu, and Z. Huang. Quaternion cross-modality spatial learning for multimodal medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[6] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao. S3d-unet: separable 3d u-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 358–368. Springer, 2019.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer, 2022.

[10] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[11] L. Huang, T. Denoeux, P. Vera, and S. Ruan. Evidence fusion with contextual discounting for multi-modality medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 401–411. Springer, 2022.

[12] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein. nnu-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*, pages 118–132. Springer, 2021.

[13] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[14] Z. Jiang, C. Ding, M. Liu, and D. Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5*, pages 231–241. Springer, 2020.

[15] X. Kong, G. Sun, Q. Wu, J. Liu, and F. Lin. Hybrid pyramid u-net model for brain tumor segmentation. In *Intelligent Information Processing IX: 10th IFIP TC 12 International Conference, IIP 2018, Nanning, China, October 19-22, 2018, Proceedings 10*, pages 346–355. Springer, 2018.

[16] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh. Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv:2105.09511*, 2021.

[17] J. Lin, J. Lin, C. Lu, H. Chen, H. Lin, B. Zhao, Z. Shi, B. Qiu, X. Pan, Z. Xu, et al. Ckd-transbts: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *arXiv preprint arXiv:2207.07370*, 2022.

[18] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022.

[19] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[20] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[21] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[22] A. Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 311–320. Springer, 2019.

[23] D. Nie, L. Wang, Y. Gao, and D. Shen. Fully convolutional networks for multi-modality isointense infant brain image segmentation. In *2016 IEEE 13Th international symposium on biomedical imaging (ISBI)*, pages 1342–1345. IEEE, 2016.

[24] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi. A robust volumetric transformer for accurate 3d tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 162–172. Springer, 2022.

[25] Q.-D. Pham, H. Nguyen-Truong, N. N. Phuong, K. N. Nguyen, C. D. Nguyen, T. Bui, and S. Q. Truong. Segtransvae: Hybrid cnn-transformer with regularization for medical image segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[26] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

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

[28] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

[29] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Z. Xing, L. Yu, L. Wan, T. Han, and L. Zhu. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 140–150. Springer, 2022.

[32] X. Zeng, P. Zeng, C. Tang, P. Wang, B. Yan, and Y. Wang. Dbtrans: A dual-branch vision transformer for multi-modal brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 502–512. Springer, 2023.

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

[34] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.