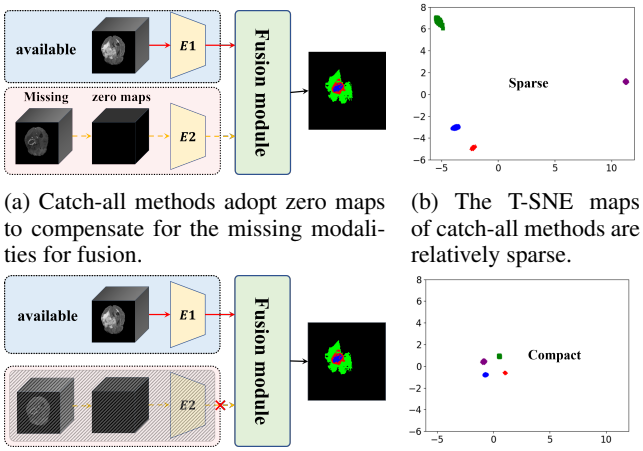
TMFormer: Token Merging Transformer for Brain Tumor Segmentation with Missing Modalities

Zheyu Zhang[1,*], Gang Yang[1,*], Yueyi Zhang[1,2,†], Huanjing Yue[4], Aiping Liu[1], Yunwei Ou[3,2,†], Jian Gong[3,2], Xiaoyan Sun[1,2]

[1]University of Science and Technology of China, Hefei 230026, China
[2]Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230088, China
[3]Beijing Tiantan Hospital, Capital Medical University, Beijing 100050, China
[4]Tianjin University, Tianjin 300072, China
zhyuey@ustc.edu.cn, ouyunwei@sina.com

Abstract

Numerous techniques excel in brain tumor segmentation using multi-modal magnetic resonance imaging (MRI) sequences, delivering exceptional results. However, the prevalent absence of modalities in clinical scenarios hampers performance. Current approaches frequently resort to zero maps as substitutes for missing modalities, inadvertently introducing feature bias and redundant computations. To address these issues, we present the **T**oken **M**erging trans**Former** (TMFormer) for robust brain tumor segmentation with missing modalities. TMFormer tackles these challenges by extracting and merging accessible modalities into more **compact** token sequences. The architecture comprises two core components: the Uni-modal Token Merging Block (UMB) and the Multi-modal Token Merging Block (MMB). The UMB enhances individual modality representation by **adaptively** consolidating spatially redundant tokens within and outside tumor-related regions, thereby refining token sequences for augmented representational capacity. Meanwhile, the MMB mitigates multi-modal feature fusion bias, exclusively leveraging tokens from present modalities and merging them into a unified multi-modal representation to accommodate varying modality combinations. Extensive experimental results on the BraTS 2018 and 2020 datasets demonstrate the superiority and efficacy of TMFormer compared to state-of-the-art methods when dealing with missing modalities.



(a) Catch-all methods adopt zero maps to compensate for the missing modalities for fusion.

(b) The T-SNE maps of catch-all methods are relatively sparse.

(c) Our proposed TMFormer only utilizes available modalities.

(d) The T-SNE map of proposed TMFormer.

Figure 1: Comparison between catch-all methods and our proposed TMFormer. Our proposed TMFormer reduces not only redundant information but also feature bias for cases of missing modalities. In (b) and (d), each point denotes a sample from the BraTS 2020. The points of the same color between (b) and (d) belong to the same case of missing modalities, while different colors mean different cases.

Introduction

Given the emergence of malignant brain tumors as a severe health threat, timely diagnosis is imperative for minimizing their impact. Brain tumor segmentation plays a pivotal role by identifying and delineating tumor boundaries in cerebral medical images (Havaei et al. 2017; Jia et al. 2020; She et al. 2023). Magnetic Resonance Imaging (MRI) sequences, including T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and Fluid Attenuation Inversion Recovery (FLAIR) modalities, are extensively employed for brain tumor segmentation. Several multi-modal techniques leverage these four MRI modalities to enhance brain tumor segmentation by integrating complementary information. However, in clinical settings, missing modalities are common due to image corruption, varying acquisition protocols, and contrast allergies (Liu et al. 2021a; Tran et al. 2017). Such absence of modalities significantly impairs the segmentation performance of multi-modal methods.

Various strategies have emerged to address diverse scenarios of missing modalities. One approach entails training dedicated networks for every potential combination of available modalities (Wang et al. 2021b; Zhang et al. 2021), yet this leads to extensive training costs and deployment space requirements. Some researchers seek to synthesize absent modalities to create complete multi-modal sets (Wang et al. 2018; Shen et al. 2020), but the segmentation accuracy

[*]Equal contribution.
[†]Corresponding authors.

is bound by the quality of the generated modalities, which can introduce unexpected noise and artifacts. Predominantly, catch-all methods prevail, utilizing a single model to handle all modality combinations (Havaei et al. 2016; Chen et al. 2019; Zhou et al. 2021). Nonetheless, these approaches use zero maps to compensate for missing modalities, introducing dramatic variations, called 'feature bias', in multi-modal feature fusion and superfluous computations in feature extraction and fusion for the absent modalities, as illustrated in Fig. 1. Hence, mitigating the impact of zero maps is crucial for enhancing brain tumor segmentation outcomes.

To address these challenges in this task, we introduce an innovative approach termed the **T**oken **M**erge Trans**former** (**TMFormer**), designed to tackle the diverse combinations of modalities essential for brain tumor segmentation. TMFormer interprets these varied amalgamations of modalities as sequences of tokens with varying and adaptable lengths. This architecture incorporates two pivotal components: the Uni-modal Token Merging Block (UMB), which focuses on information extraction from individual modalities, and the Multi-modal Token Merging Block (MMB), responsible for the fusion of multiple modalities. The UMB applies an adaptive token merging strategy that intelligently compresses spatially redundant tokens in regions associated with tumors. It retains more representative tokens in tumor-related regions and merges tokens more extensively in tumor-unrelated regions. This streamlined token sequence is subsequently refined to enhance global representation. Meanwhile, the MMB exclusively fuses and augments tokens from available modalities. The tokens from modalities that exert a relatively minor impact on segmentation into those with higher contributions to the segmentation process. The resultant multi-modal features are projected into a unified representative space, effectively mitigating feature biases that may arise when dealing with different scenarios involving missing modalities. This innovative TMFormer approach demonstrates its prowess in efficiently handling the intricate challenges of brain tumor segmentation by expertly managing diverse modalities and their nuanced interactions. In summary, the contributions of our work are as follows:

- We introduce a Token Merging Transformer aiming to alleviate feature bias in scenarios involving missing modalities and reduce redundant computations.
- We propose the UMB to reduce the spatially redundant information and augment the global representation of available modalities.
- We propose the MMB to merge multi-modal tokens based on the respective contributions of modalities to segmentation and project the fused tokens into a unified representative space.
- We conduct extensive experiments on the BraTS 2018 and 2020 datasets and demonstrate the superiority and effectiveness of our TMFormer for brain tumor segmentation with missing modalities.

Related Work

Multi-modal brain tumor segmentation with missing modalities. Several methods have been developed to address brain tumor segmentation with missing modalities, which can be divided into three categories: 1) *'dedicated'* methods that train a dedicated segmentation model for each possible combination of available modalities, 2) *'generative'* methods that synthesize missing modalities and training a segmentation model with complete modalities, and 3) *'catch-all'* methods that train a single model for various combinations of modalities.

For the dedicated methods, KDNet (Hu et al. 2020) distills knowledge from the multi-modal network to the dedicated network. Wang et al. (2021b) adopt a co-training strategy between the multi-modal network and the dedicated network, aligning the feature distribution in latent space. Since there are fifteen possible combinations of modalities, these methods suffer from high training costs.

For the generative methods, Hu et al. (2020) adopt a local-adaptive convolutional network to fuse the available modalities for generating the absent modalities. Shen et al. (2020) disentangle modality sequences into the content code and the style code, and the missing modalities are generated based on the content code. M3AE (Liu et al. 2023) adopts multi-modal masked auto-encoder and model inversion to build substitutes of multi-modal sequences, which performs a segmentation process on these substitutes. Nevertheless, as the generated modalities potentially have noises and artifacts, it is challenging to acquire an accurate segmentation result based on the synthesized modalities.

For the catch-all methods, HeMiS (Havaei et al. 2016) fuses the multi-modal features using their mean and variance, which obtains the segmentation based on the fusion. RFNet (Ding, Yu, and Yang 2021) conducts multi-modal fusion according to the tumor regions, and UNet-MFI (Zhao, Yang, and Sun 2022) builds the graph for fusing multi-modal features. Zhang et al. (2022) propose the hybrid CNN-Transformer architecture termed the mmFormer that utilizes multi-head self-attention to fuse multi-modal features only in the smallest scale. However, due to their convolutional operation defined by fixed-size kernels, these methods must use zero maps to compensate for the missing modalities during subsequent processing. This inadvertently introduces the feature bias. Besides, the feature extraction and fusion for zero maps lead to unnecessary computations. In contrast, we treat varying modality combinations as variable-length token sequences, which avoid the involvement of zero maps.

Efficient vision transformers. Since transformers have quadratic complexity, many works aim to improve their efficiency from different aspects. Some focus on the attention mechanisms (Huang et al. 2019; Liu et al. 2021b; Dong et al. 2022). Swin Transformer (Liu et al. 2021b) adopts self-attention in shifted local windows. Huang et al. (2019) and Dong et al. (2022) propose cross-shaped windows for computing attention. Besides, PVT (Wang et al. 2021a) employs the pyramid structure within the downsampled key and value tokens. Reducing the number of tokens processed in the network is an alternative way to improve efficiency. Several works attempt to prune less informative tokens based on the predicted importance (Rao et al. 2021) or token similarity (Liang et al. 2022; Fayyaz et al. 2022;

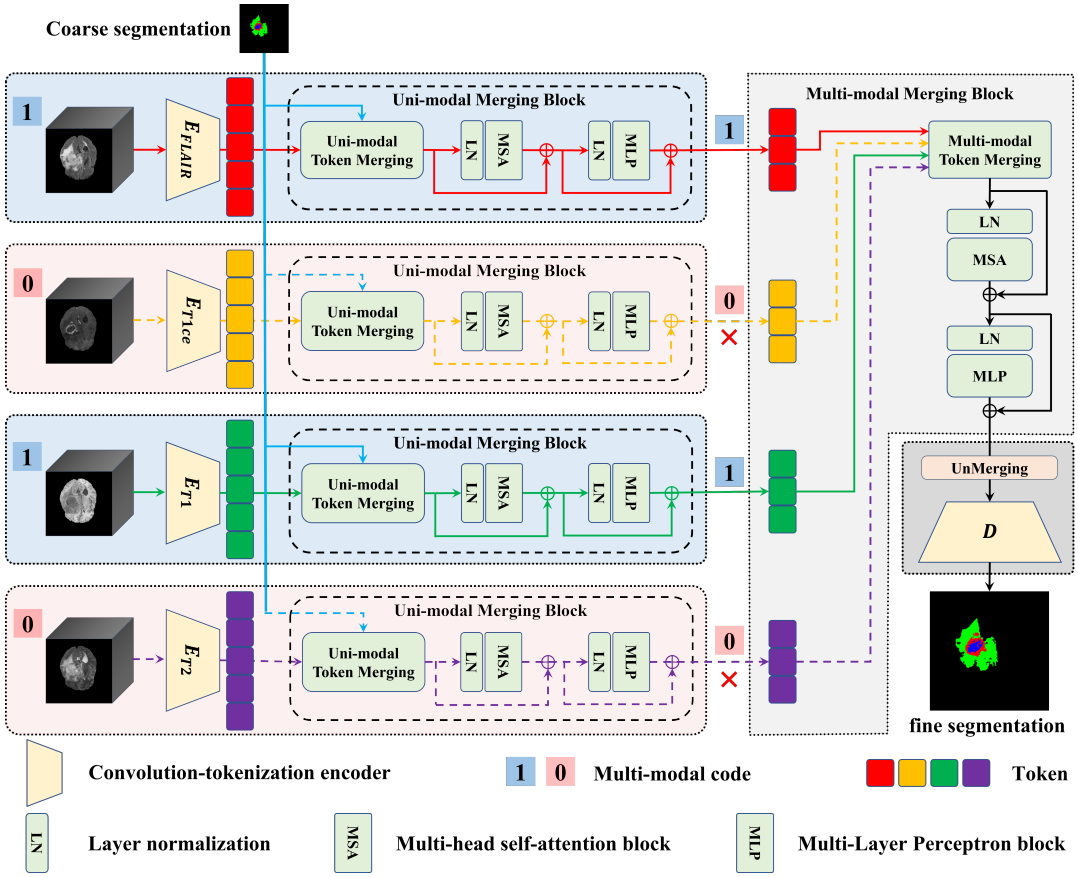


Figure 2: Overview of the proposed TMFormer, which is composed of four convolution-tokenization encoders, four UMBs, an MMB, an unmerging block, and a decoder.

Kong et al. 2023). Bolya et al. (2022) propose to merge adjacent similar tokens to accelerate the inference of ViT. Token Learner (Liang et al. 2022) fuse inattentive tokens that contribute less to the class token. The mentioned methods have achieved promising performance for image classification. Recently, Lu, de Geus, and Dubbelman (2023) propose to share values of tokens belonging to the same class for semantic segmentation. However, to the best of our knowledge, reducing redundant tokens remains unexplored within brain tumor segmentation, particularly in scenarios involving missing modalities. It is non-trivial to propose a token-reducing strategy as 3D MRI has redundant information in intra-modal and inter-modal spatial dimensions.

Method

In this section, we first briefly explain the motivation. Then we illustrate the overall architecture of our TMFormer in Fig. 2 and its components. Finally, we describe its corresponding optimization loss.

Motivation

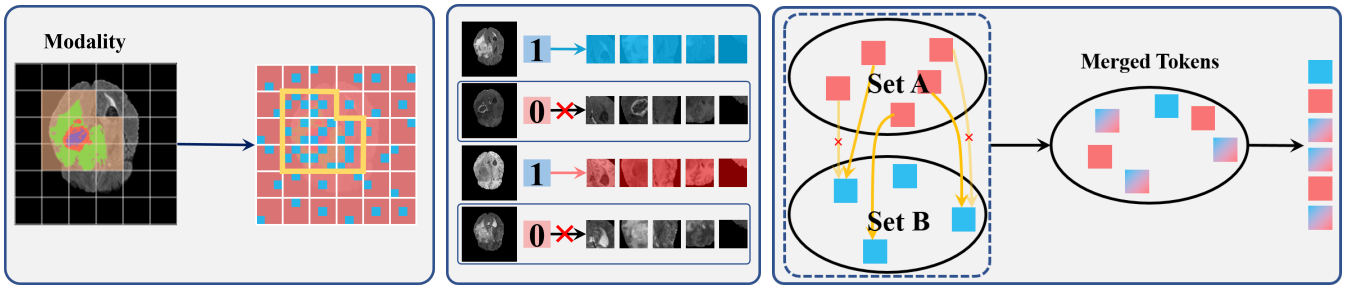
The prevalent catch-all methods utilize zero maps as substitutes for absent modalities, which inevitably introduce feature bias and the wastage of computational resources. Drawing inspiration from Transformer's proficiency in handling variable-length token sequences, we propose to employ the Transformer architecture to deal with diverse scenarios of missing modalities. The different combinations of available modalities are regarded as variable-length token sequences while the substitutes are no longer utilized.

Given the considerable length and potential redundancy of 3D medical image token sequences, we undertake the merging of modality tokens along intra-modal and inter-modal spatial dimensions. Furthermore, to alleviate feature bias across distinct scenarios involving absent modalities, we introduce the modeling of global dependencies within the compact token sequences and project them into a unified multi-modal representative space.

The Overall Architecture

A full-modal complete image set consists of four different modalities, i.e., FLAIR, T1ce, T1, and T2 modalities. In the task of missing modalities, a multi-modal data x is given with the dimension $M \times D \times H \times W$. D, H, and W are the depth, height, and width of the image, respectively, and M is the number of available modalities. Its multi-modal code is represented by $h = \{h_1, h_2, h_3, h_4\}$, where $h_i \in \{0, 1\}$ indicates whether the corresponding modality is available.

We illustrate the overall architecture in Fig. 2. The TMFormer employs a multi-scale network structure, facilitat-



(a) The partition for getting A_{uni} and B_{uni} in the UMB. (b) The partition for getting A_{mul} and B_{mul} in the MMB. (c) The token merging process after dividing A and B in UMB and MMB.

Figure 3: The token merging process of UMB and MMB. In (a) and (b), sample red tokens to set A and blue tokens to set B. In (c), merge uni-modal tokens from A_{uni} into B_{uni} in the UMB, while merge multi-modal tokens from A_{mul} into B_{mul} in the MMB. The most similar tokens are merged as mixed tokens, while the dissimilar tokens are preserved as red and blue tokens.

ing the integration of information across distinct hierarchical levels. At each scale, we first utilize encoders to project the data x into token sequences. Then, we adopt the UMB to decrease the spatial redundancy of token sequences and model the intra-modal global relationships for each available modality. After extracting features, we use MMB to decrease the redundancy among modalities and modal the inter-modal global relationship. Subsequently, the fused multi-modal token sequences are unmerged and re-arranged to their initial shape. Finally, the fused features are sent to convolutional decoders to yield the final segmentation result. Notably, as depicted in Fig. 2, the encoder branches marked as missing modalities by $h_i = 0$ are not involved in computations, which is different from the other methods dedicated to missing modalities (Yang et al. 2022; Zhang et al. 2022).

Convolution-tokenization Encoder We adopt individual encoders to extract local features for available modalities. Similar to the encoder part design of U-Net (Ronneberger, Fischer, and Brox 2015), we stack four convolutional blocks to extract multi-scale features, and each convolutional block consists of three convolutional layers. At each scale, the local features of each modality are projected into the token sequence $x_{tok} \in \mathbf{R}^{(\frac{D}{2^{l-1}} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}) \times C}$ through a patch embedding layer, where l indicates the scale, $\frac{D}{2^{l-1}} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}$ denotes the length of token sequence, and C is the channel dimension. We set the patch size to $1 \times 1 \times 1$ since the pixel-level information is essential to segmentation.

Uni-modal Merging Block (UMB) Due to the inherently local nature of convolutional networks, we aim to improve the uni-modal feature representation by capturing long-range dependencies. For each modality, the input for the UMB is $x_{tok} \in \mathbf{R}^{(\frac{D}{2^{l-1}} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}) \times C}$. At the $l = 0$ scale, the length of the sequence is $D \times H \times W$, which will lead to high complexity of computations. We firstly merge the token sequence with the UMB, which transforms the sequence length from $\frac{D}{2^{l-1}} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}$ to N_{l-1}, i.e.,

$$x_{uni} = UTM(x_{tok}|Seg_{coarse}), \tag{1}$$

where UTM denotes the uni-modal token merging process, and Seg_{coarse} is the coarse segmentation map predicted from the previous scale. Since the segmentation only focuses on the tumor region, we use Seg_{coarse} to identify the tumor-related region and the tumor-unrelated region to merge spatial tokens adaptively. Based on Seg_{coarse}, the uni-modal token merging process is as follows: 1) Partitioning the x_{tok} into parts that each part belongs to a cubic window in 3D dimensions; 2) Sampling tokens from parts that more tokens are sampled for the tumor-related part while fewer tokens are sampled for the tumor-unrelated part; 3) Putting the sampled tokens into B_{uni} and the rest tokens are in A_{uni}; 4) Calculating the similarity scores between A_{uni} and B_{uni}; 5) Selecting the most similar token in B_{uni} for each token in A_{uni}, which builds the similar token pairs and preserves its similar score; 6) Merging the top N_{l-1}^{uni} most similar token pairs that are sorted based on the similar score; 7) Appending the dissimilar tokens of A_{uni} to B_{uni}, and the length of dissimilar tokens is $N_{l-1} - N_{l-1}^{uni}$.

Significantly, a subset of tumor-unrelated tokens is allocated to B_{uni} as these tokens estimated by coarse segmentation maps, may include false negatives, which should be preserved. The merging process is shown in Fig. 3a and Fig. 3c. After this process, we then establish a global feature relationship for the merged token sequence $x_{uni} \in \mathbf{R}^{N_{l-1} \times C}$ via multi-head self-attention (MSA) and multi-layer perceptrons (MLP) to fully mine the feature information within the modality. The corresponding outputs are:

$$x_{MSA} = MSA(LN(x_{uni})) + x_{uni}, \tag{2}$$

$$x_{MLP} = MLP(LN(x_{MSA})) + x_{MSA}, \tag{3}$$

where $LN(\cdot)$ denotes the layer normalization, and the x_{MLP} is the final output of the UMB.

Multi-modal Merging Block (MMB) After the feature extraction by UMB, we obtain the token sequences $\mathcal{X}_{tok} \in \mathbf{R}^{M \times N_{l-1} \times C}$ for M available modalities. To enhance the multi-modal representation capacity, we propose the MMB to model the long-range dependencies among different modalities. In the MMB, we first merge the inter-modal redundant information, i.e.,

$$x_{mul} = MTM(\mathcal{X}_{tok}|h), \tag{4}$$

where $MTM(\cdot)$ denotes the multi-modal token merging process and h is the modality code.

Different from the UTM, the partition of the MTM is based on the observation that the FLAIR and T1ce modalities have a more pronounced impact on tumor segmentation (Ding, Yu, and Yang 2021; Yang et al. 2022). Consequently, we construct the multi-modal token sequences $\mathcal{X}_{tok}$ in the order of $[x_{tok}^{FLAIR}, x_{tok}^{T1ce}, x_{tok}^{T1}, x_{tok}^{T2}]$, maintaining sequential continuity even with the absence of modalities. For example, if the T1ce modality is missing, the multi-modal token sequences $\mathcal{X}_{tok}$ are organized as $[x_{tok}^{FLAIR}, x_{tok}^{T1}, x_{tok}^{T2}]$ that do not add zero maps. We partition the multi-modal token sequences $\mathcal{X}_{tok}$ into two parts, set A_{mul} and set B_{mul}, which are illustrated as follows:

$$\begin{cases} \mathrm{A}_{mul} = \mathcal{X}_{tok}[0:1], & \mathrm{B}_{mul} = \mathcal{X}_{tok}[2:M), & if M \geq 3, \\ \mathrm{A}_{mul} = \mathcal{X}_{tok}[0], & \mathrm{B}_{mul} = \mathcal{X}_{tok}[1], & if M = 2, \\ \mathrm{A}_{mul} = \mathcal{X}_{tok}, & \mathrm{B}_{mul} = \mathrm{None}, & if M = 1, \end{cases} \tag{5}$$

To reduce the inter-modal redundancy, we fuse the tokens of A_{mul} into the tokens of B_{mul} that are highly similar, concurrently appending dissimilar tokens of A_{mul} to B_{mul}. In summary, the MTM process is as follows: 1) Constructing the multi-modal token sequences $\mathcal{X}_{tok}$; 2) Partitioning the $\mathcal{X}_{tok}$ into A_{mul} and B_{mul} based on Eq. 5; 3) Calculating the similarity scores between A_{mul} and B_{mul}; 4) Selecting the most similar token in B_{mul} for each token in A_{mul}; 5) Merging the top N_{l-1}^{mul} most similar token pairs; 6) Appending the dissimilar tokens of A_{mul} to B_{mul}. The merging process is shown in Fig. 3b and Fig. 3c. If only one modality is available, we send it to the following layers without merging.

We then adopt MSA to facilitate information exchange between modalities. Finally, we utilize MLP to project the fused feature into a unified representation space for the following segmentation, which improves the robustness of segmentation in case of different scenarios of missing modalities. Similarly,the process is presented as follows:

$$x_{MSA} = MSA(LN(x_{mul})) + x_{mul}, \tag{6}$$

$$x_{MLP} = MLP(LN(x_{MSA})) + x_{MSA}. \tag{7}$$

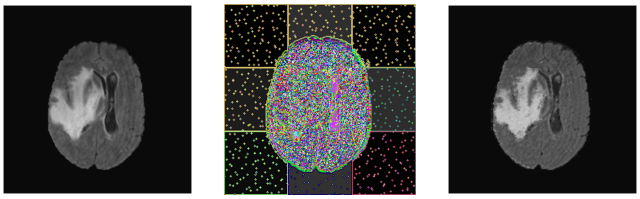
Unmerging Block and Decoder We employ the unmerging block to restore the initial length of the token sequence and re-arrange the token sequence into the feature maps of $\frac{D}{2^{l-1}} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}$. We use one multi-scale decoder to gradually restore spatial resolution, smoothly transitioning from the high-level latent space to the original mask space.

In the unmerging block, the tokens that are merged out share the identical value as the merged token, effectively reinstating the length of the enhanced token sequences to their original state before merging. In the decoder, the decoder blocks employ the feature maps for generating coarse-to-fine segmentation maps within our four-level structure. The coarse segmentation map plays a vital role in guiding the process of UTM within our UMB.

Optimization Loss

Following previous works (Dou et al. 2017; Ding, Yu, and Yang 2021; Zhao, Yang, and Sun 2022), we employ the weighted cross-entropy loss L_{CE} and Dice loss L_{Dice} to optimize our TMFormer at each scale, which is defined as:

$$L = L_{CE}(y, \hat{y}) + L_{Dice}(y, \hat{y}), \tag{8}$$



(a) Input (b) Similarity map (c) Merged visualization

Figure 4: Visualization of token merging. (a) depicts the initial input that has $D \times H \times W$ tokens. (b) illustrates token similarities, where tokens of the same color indicate their merging. (c) shows the visualization of merged tokens, whose length is near $\frac{1}{64}$ of the initial tokens.

where y and $\hat{y}$ are the segmentation prediction and the ground truth, respectively.

Experiments and Results

Datasets & Evaluation Metrics. The proposed method is evaluated on two publicly available multi-modal brain tumor segmentation (BraTS) datasets: BraTS 2018 (Bakas et al. 2018) and 2020 (Andres et al. 2020). The BraTS 2018 and 2020 datasets include 285 and 369 cases with ground truth publicly available, respectively. For the BraTS 2018 dataset, we split it into $199 : 29 : 57$ for training, validation, and testing, respectively. Additionally, we use a three-fold validation with the same split as (Dorent et al. 2019). For the BraTS 2020 dataset, following (Ding, Yu, and Yang 2021), we randomly split it into $219 : 50 : 100$.

The BraTS datasets consist of four different modalities: FLAIR, T1ce, T1, and T2 modalities. Each of them captures different properties of brain tumor subregions: GD-enhancing tumor (ET), peritumoral edema (ED), and the necrotic and non-enhancing tumor core (NCR/NET). Different subregions of brain tumors are combined into three nested subregions: the whole tumor (WT), the tumor core (TC), and the enhancing tumor (ET). Following (Ding, Yu, and Yang 2021; Zhao, Yang, and Sun 2022), we adopt Dice Similarity Coefficient (DSC) to evaluate the segmentation performance.

Implementation Details. We implement our TMFormer in the PyTorch framework (1.13) and train all parameters on four NVIDIA GeForce RTX 3090 GPUs for 500 epochs. Our method is optimized by the ADAM optimizer with a batch size of 4. The initial learning rate is 2×10^{-4} with the weight decay of 1×10^{-5}.

In the training stage, we randomly crop each volume to a fixed size of $80 \times 80 \times 80$ voxels that are further augmented with random flip, random rotation, and random intensity shift. In the inference stage, we crop the volumes from $240 \times 240 \times 155$ to $128 \times 128 \times 128$ with an overlap rate of 0.5.

Motivation Verification. As illustrated in motivation, we first regard the different combinations of available modalities as variable-length token sequences, avoiding unnecessary calculations on missing modalities. As is shown in

Methods	HeMiS	U-HVED	RFNet	UNet-MFI	mmFormer	M3AE	1 Mod.	2 Mods.	3 Mods.	4 Mods.
Param. (M)	0.57	1.25	8.98	34.12	36.65	40.42	8.93	8.93	8.93	8.93
GFLOPs	2.27	4.58	102.28	499.52	30.23	36.14	57.24	72.62	88.00	103.37

Table 1: The parameters and GFLOPs of compared methods on the BraTS 2020 dataset. '1/2/3/4 Mods.' means that we send 1/2/3/4 available modalities into our model, respectively.

M																	AVG
	F	●	○	○	○	●	●	●	○	○	○	●	●	●	○	●	
	T1ce	○	●	○	○	●	○	○	●	●	○	●	●	○	●	●	
	T1	○	○	●	○	○	●	○	●	○	●	●	○	●	●	●	
	T2	○	○	○	●	○	○	●	○	●	●	○	●	●	●	●	
WT	①	71.60	67.71	68.96	68.19	69.17	68.67	69.83	69.01	69.78	69.40	70.21	71.28	70.73	71.58	72.06	69.88
	②	69.85	46.82	46.77	54.03	61.45	58.25	64.50	62.91	65.76	64.29	66.99	69.70	68.38	70.35	71.41	62.76
	③	86.42	77.34	76.46	86.21	89.55	89.30	89.35	81.00	87.45	**87.95**	**90.39**	90.20	90.42	88.59	90.77	86.76
	④	82.27	73.18	72.10	82.45	83.64	84.34	84.85	77.30	83.44	83.52	85.45	85.70	85.85	84.20	84.93	82.21
	⑤	82.40	74.25	74.37	83.07	84.54	84.61	85.82	77.98	84.05	84.00	85.34	86.11	86.22	84.64	86.38	82.92
	⑥	86.53	73.85	76.71	86.09	89.48	89.38	89.25	78.11	87.37	87.20	89.99	90.18	90.42	**88.61**	90.56	86.25
	Ours	**87.45**	**78.53**	**78.94**	**86.46**	**89.67**	**89.64**	**89.98**	**81.97**	**88.01**	87.75	90.23	**90.53**	**90.51**	88.54	**90.83**	**87.27**
TC	①	53.43	51.41	51.56	51.11	51.70	51.08	51.85	51.88	52.35	51.51	52.95	53.76	52.97	54.38	55.03	52.46
	②	34.62	35.51	27.30	37.67	42.15	38.26	43.41	44.93	47.53	44.97	49.13	51.30	49.40	52.72	54.17	43.53
	③	65.04	82.37	64.31	68.47	**84.69**	71.45	72.62	**83.15**	84.06	72.11	**84.71**	**84.70**	74.28	**84.11**	84.74	77.39
	④	63.94	77.63	59.38	68.05	79.92	68.23	70.72	77.61	80.09	70.21	80.03	80.94	71.40	80.75	81.28	74.01
	⑤	66.19	77.96	61.17	69.18	80.36	69.58	71.55	79.93	80.79	70.90	80.18	81.31	72.02	81.12	81.22	74.90
	⑥	68.04	81.39	66.00	70.27	82.01	73.82	**74.95**	82.39	83.01	72.54	82.44	83.06	75.09	84.06	84.40	77.56
	Ours	**70.19**	**82.59**	**67.12**	**71.84**	84.62	**74.65**	74.76	83.13	**84.24**	**73.33**	84.69	84.64	**75.17**	84.00	**84.88**	**78.66**
ET	①	**43.77**	42.41	**41.59**	41.45	41.83	40.29	41.19	42.08	42.39	41.00	43.67	44.16	42.95	45.27	46.33	42.69
	②	12.88	24.94	7.27	24.26	30.02	21.95	29.40	33.64	36.18	32.12	39.39	40.91	38.09	43.18	45.33	30.64
	③	40.47	74.27	37.51	43.59	**76.45**	43.81	46.99	75.22	73.94	46.37	77.01	76.38	48.95	76.38	76.64	60.93
	④	39.70	69.42	29.38	46.00	70.13	40.06	48.69	69.25	72.32	45.71	71.28	70.88	46.55	72.00	71.41	57.52
	⑤	40.47	68.91	33.97	45.61	69.81	43.63	48.09	71.10	70.72	45.92	70.08	71.60	48.38	70.65	71.36	58.02
	⑥	40.49	72.43	39.93	45.97	74.66	43.20	47.30	75.42	76.81	46.63	75.94	77.08	48.19	77.40	78.00	61.30
	Ours	42.28	**76.21**	38.21	**46.94**	76.37	**48.20**	**51.67**	**78.68**	**78.25**	**48.81**	**78.64**	**78.45**	**51.23**	**78.51**	**78.98**	**63.43**

Table 2: Performance comparison (DSC%) with SOTA methods, including ① HeMiS (Havaei et al. 2016), ② U-HVED (Dorent et al. 2019), ③ RFNet (Ding, Yu, and Yang 2021), ④ UNet-MFI (Zhao, Yang, and Sun 2022), ⑤ mmFormer (Zhang et al. 2022) and ⑥ M3AE (Liu et al. 2023) on the BraTS 2020 dataset. Available and missing modalities are denoted by ● and ○, respectively.

Tab. 1, *our model presents variable GFLOPs with different combinations of modalities*. Conversely, the other methods exhibit fixed GFLOPs that consume redundant computations on zero maps for the missing modalities.

Secondly, we visualize an example of merging uni-modal tokens in Fig. 4. We obtain the similarity map for merging in the feature space and employ it on the input image for better understanding. The merged image demonstrates balanced preservation of edge details without excessive smoothing. Thus, the UMB preserves the essential edges for brain tumor segmentation while decreasing the redundant information.

Finally, as shown in Fig. 1, we obtain fused multi-modal features, which further go through the high-dimensional embedding pooling followed by the dimensionality reduction for 2D visualization via t-SNE on the BraTS 2020 dataset. We use 100 samples to simulate 4 cases of missing modalities, including 1/2/3/4 available modalities. After processing by our MMB, those samples tend to cluster more densely in Fig. 1d, in contrast to their sparser distribution of fused multi-modal features with zero maps in Fig. 1b. This observation demonstrates the capability of our MMB to unify multi-modal features, thus resulting in a consistent representation across varying modality combinations.

Comparison with the state-of-the-art (SOTA) methods on missing modalities. To demonstrate the superiority of our method, we compare our TMFormer with six state-of-the-art methods on different cases with missing MRI modalities.The involved methods contain HeMiS (Havaei et al. 2016), U-HVED (Dorent et al. 2019), RFNet (Ding, Yu, and Yang 2021), UNet-MFI (Zhao, Yang, and Sun 2022), mmFormer (Zhang et al. 2022), M3AE (Liu et al. 2023). For a fair comparison, all methods are trained under their recommended hyper-parameters within the same dataset split.

As shown in Tab. 2, our method achieves preferable results for most combinations of missing modalities. We achieve improvements of 0.5%, 1.1%, and 2.5% over the second-ranked method on the average DSC for WT, TC, and ET. In Fig. 5, we provide the segmentation visualizations that our method yields more accurate segmentation results in different combinations of modalities.

Ablation Study. We evaluate the proposed components on the BraTS 2020 dataset, employing the average DSC to measure the performance of the WT, TC, and ET. Ablative experiments are partitioned into three parts, i.e., the UMB, the MMB, and stages, as depicted in Tab. 3. We assess each part

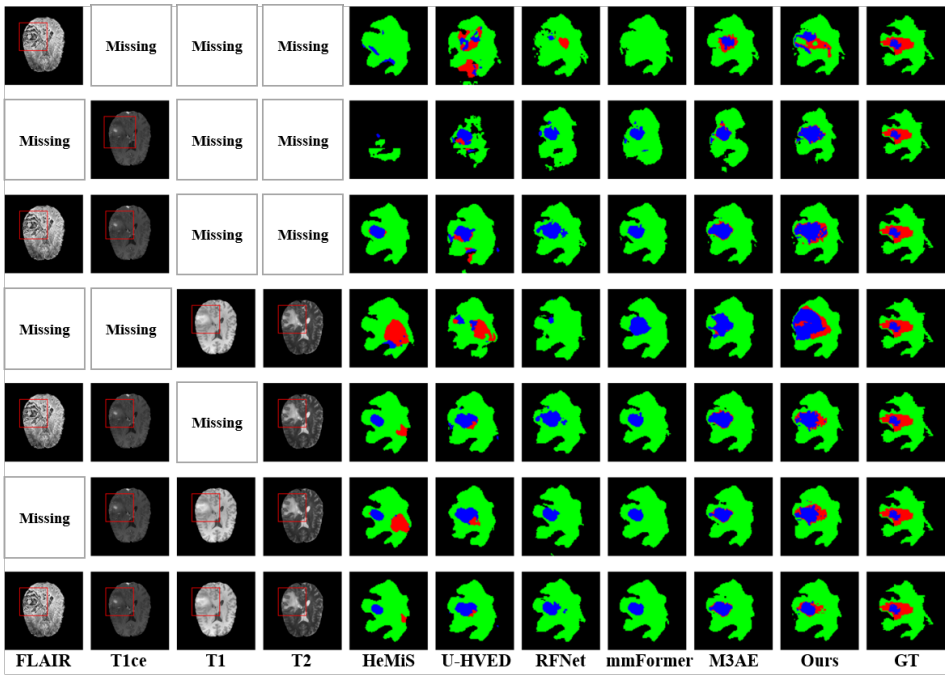


Figure 5: Segmentation results of different methods with various available modalities on the BraTS 2020 dataset.

Configuration		UMB						MMB			#Stage			
		Pooling	S_{conv}	SA	SWC	SRW	SRW+SAR	SA	DO	CO	1	2	3	4
DSC	WT	82.92	82.21	85.11	86.00	87.23	**87.27**	87.04	86.98	**87.27**	86.81	87.12	87.21	**87.27**
	TC	74.90	74.01	72.15	74.38	**78.97**	78.66	77.91	78.23	**78.66**	77.81	78.04	78.04	**78.66**
	ET	53.68	53.73	48.00	49.43	62.20	**63.43**	62.89	62.45	**63.43**	60.06	62.53	62.67	**63.43**

Table 3: Ablation results of proposed components on the BraTS 2020 dataset.

while keeping others constant.

To evaluate the UMB, we use 'Pooling' or 'S_{conv}' for merging uni-modal tokens within cubic windows. 'Pooling' computes the average mean, while 'S_{conv}' employs the convolution with a stride equal to the kernel size. In contrast, 'SA', 'SWC', 'SRW', and 'SRW+SAR' perform global token merging, albeit with token sampling within cubic windows. 'SA' means alternate token sampling. 'SWC' stands for central token sampling, and 'SRW' signifies random token sampling. Both 'SA' and 'SWC' involve fixed sampled tokens. 'SAR' denotes adaptive token sampling, and 'SRW+SAR' is our chosen approach in the UMB. From Tab. 3, we find that global merging outperforms local merging. The random sampling with the guidance of a coarse segmentation map also boosts the performance from 62.20 to 63.43 on ET.

To assess our MMB, we replace the 'CO' with both 'SA' and 'DO'. In MMB, 'SA' involves alternately sampling tokens from multi-modal sequences, without regard to the modality of the tokens. 'DO' constructs multi-modal sequences $\mathcal{X}_{tok}$ in a different order, i.e., $[x_{tok}^{T1}, x_{tok}^{T2}, x_{tok}^{FLAIR}, x_{tok}^{T1ce}]$. 'CO' represents constructing tokens in our proposed order, i.e., $[x_{tok}^{FLAIR}, x_{tok}^{T1ce}, x_{tok}^{T1}, x_{tok}^{T2}]$. The results show that our MMB improves the multi-modal feature fusion.

We verify the efficacy of TMFormer's multi-scale design by progressively integrating our UMB and MMB into each scale. The outcomes demonstrate that incorporating our proposed blocks within a multi-scale framework yields improvements in performance. More results are in Appendix.

Conclusion

In this paper, we introduce a novel **T**oken **M**erging trans**Former** (**TMFormer**) to tackle the challenge of missing modalities in brain tumor segmentation. This addresses the issues caused by using zero maps as substitutes, which lead to feature bias and redundant computations. TMFormer treats modalities' diverse combinations as variable-length token sequences, considering only available modalities. Our TMFormer comprises two pivotal modules: the UMB for uni-modal feature extraction and the MMB for multi-modal feature fusion. The UMB initially reduces spatially redundant tokens guided by a coarse segmentation map and models global dependencies for each available modality. The MMB merges uni-modal tokens based on modalities' contribution order to segmentation. The fused multi-modal token sequence is then projected into a unified representation to alleviate feature bias in different combinations of modalities. These proposed components collectively demonstrate the potential to mitigate feature bias and avoid unnecessary computations for missing modalities. Extensive experiments show the proficiency of our method.

Acknowledgments

This work was in part supported by the National Natural Science Foundation of China under grants 62032006 and 62021001.

References

Andres, E. A.; Fidon, L.; Vakalopoulou, M.; Lerousseau, M.; Carré, A.; Sun, R.; Klausner, G.; Ammari, S.; Benzazon, N.; Reuzé, S.; et al. 2020. Dosimetry-driven quality measure of brain pseudo computed tomography generated from deep learning for MRI-only radiation therapy treatment planning. *International Journal of Radiation Oncology* Biology* Physics*, 108(3): 813–823.

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2018. 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*.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Chen, C.; Dou, Q.; Jin, Y.; Chen, H.; Qin, J.; and Heng, P.-A. 2019. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, 447–456. Springer.

Ding, Y.; Yu, X.; and Yang, Y. 2021. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3975–3984.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12124–12134.

Dorent, R.; Joutard, S.; Modat, M.; Ourselin, S.; and Vercauteren, T. 2019. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 74–82. Springer.

Dou, Q.; Yu, L.; Chen, H.; Jin, Y.; Yang, X.; Qin, J.; and Heng, P.-A. 2017. 3D deeply supervised network for automated segmentation of volumetric medical images. *Medical image analysis*, 41: 40–54.

Fayyaz, M.; Koohpayegani, S. A.; Jafari, F. R.; Sengupta, S.; Joze, H. R. V.; Sommerlade, E.; Pirsiavash, H.; and Gall, J. 2022. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 396–414. Springer.

Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Pal, C.; Jodoin, P.-M.; and Larochelle, H. 2017. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35: 18–31.

Havaei, M.; Guizard, N.; Chapados, N.; and Bengio, Y. 2016. Hemis: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 469–477. Springer.

Hu, M.; Maillard, M.; Zhang, Y.; Ciceri, T.; La Barbera, G.; Bloch, I.; and Gori, P. 2020. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 772–781. Springer.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 603–612.

Jia, H.; Xia, Y.; Cai, W.; and Huang, H. 2020. Learning high-resolution and efficient non-local features for brain glioma segmentation in MR images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, 480–490. Springer.

Kong, Z.; Ma, H.; Yuan, G.; Sun, M.; Xie, Y.; Dong, P.; Meng, X.; Shen, X.; Tang, H.; Qin, M.; et al. 2023. Peeling the onion: Hierarchical reduction of data redundancy for efficient vision transformer training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8360–8368.

Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2022. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*.

Liu, H.; Wei, D.; Lu, D.; Sun, J.; Wang, L.; and Zheng, Y. 2023. M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1657–1665.

Liu, Y.; Fan, L.; Zhang, C.; Zhou, T.; Xiao, Z.; Geng, L.; and Shen, D. 2021a. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. *Medical Image Analysis*, 69: 101953.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, C.; de Geus, D.; and Dubbelman, G. 2023. Content-aware Token Sharing for Efficient Semantic Segmentation with Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23631–23640.

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

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

She, D.; Zhang, Y.; Zhang, Z.; Li, H.; Yan, Z.; and Sun, X. 2023. EoFormer: Edge-Oriented Transformer for Brain Tumor Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 333–343. Springer.

Shen, L.; Zhu, W.; Wang, X.; Xing, L.; Pauly, J. M.; Turkbey, B.; Harmon, S. A.; Sanford, T. H.; Mehralivand, S.; Choyke, P. L.; et al. 2020. Multi-domain image completion for random missing input data. *IEEE transactions on medical imaging*, 40(4): 1113–1122.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1405–1414.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021a. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wang, Y.; Zhang, Y.; Liu, Y.; Lin, Z.; Tian, J.; Zhong, C.; Shi, Z.; Fan, J.; and He, Z. 2021b. Acn: Adversarial co-training network for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, 410–420. Springer.

Wang, Y.; Zhou, L.; Yu, B.; Wang, L.; Zu, C.; Lalush, D. S.; Lin, W.; Wu, X.; Zhou, J.; and Shen, D. 2018. 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. *IEEE Transactions on Medical Imaging*, 38(6): 1328–1339.

Yang, Q.; Guo, X.; Chen, Z.; Woo, P. Y.; and Yuan, Y. 2022. D 2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 41(10): 2953–2964.

Zhang, Y.; He, N.; Yang, J.; Li, Y.; Wei, D.; Huang, Y.; Zhang, Y.; He, Z.; and Zheng, Y. 2022. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 107–117. Springer.

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

Zhao, Z.; Yang, H.; and Sun, J. 2022. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 183–192. Springer.

Zhou, T.; Canu, S.; Vera, P.; and Ruan, S. 2021. Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. *IEEE Transactions on Image Processing*, 30: 4263–4274.