M^{2}FTrans: Modality-Masked Fusion Transformer for Incomplete Multi-Modality Brain Tumor Segmentation

Junjie Shi, Li Yu, *Senior Member, IEEE*, Qimin Cheng, Xin Yang, *Member, IEEE*, Kwang-Ting Cheng, *Fellow, IEEE*, and Zengqiang Yan, *Member, IEEE*

***Abstract*—Brain tumor segmentation is a fundamental task and existing approaches usually rely on multi-modality magnetic resonance imaging (MRI) images for accurate segmentation. However, the common problem of missing/incomplete modalities in clinical practice would severely degrade their segmentation performance, and existing fusion strategies for incomplete multi-modality brain tumor segmentation are far from ideal. In this work, we propose a novel framework named M^{2}FTrans to explore and fuse cross-modality features through modality-masked fusion transformers under various incomplete multi-modality settings. Considering vanilla self-attention is sensitive to missing tokens/inputs, both learnable fusion tokens and masked self-attention are introduced to stably build long-range dependency across modalities while being more flexible to learn from incomplete modalities. In addition, to avoid being biased toward certain dominant modalities, modality-specific features are further re-weighted through spatial weight attention and channel-wise fusion transformers for feature redundancy reduction and modality re-balancing. In this way, the fusion strategy in M^{2}FTrans is more robust to missing modalities. Experimental results on the widely-used BraTS2018, BraTS2020, and BraTS2021 datasets demonstrate the effectiveness of M^{2}FTrans, outperforming the state-of-the-art approaches with large margins under various incomplete modalities for brain tumor segmentation.**

***Index Terms*—Incomplete multi-modality segmentation, transformer, fusion token, masked self-attention.**

Manuscript received 20 June 2023; revised 27 September 2023; accepted 17 October 2023. Date of publication 20 October 2023; date of current version 5 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62202179, 62271220, and 42271352, in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585, and in part by the National Natural Science Foundation of China/Research Grants Council Joint Research Scheme under Grant N_HKUST627/20. *(Corresponding author: Zengqiang Yan.)*

Junjie Shi, Li Yu, Qimin Cheng, Xin Yang, and Zengqiang Yan are with the Hubei Key Laboratory of Smart Internet Technology, Hubei 430074, China, and also with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: shijunjie@hust.edu.cn; hustlyu@hust.edu.cn; chengqm@hust.edu.cn; xinyang2014@hust.edu.cn; z_yan@hust.edu.cn).

Kwang-Ting Cheng is with the School of Engineering, Hong Kong University of Science and Technology, Kowloon 999077, Hong Kong (e-mail: timcheng@ust.hk).

Code is available at https://github.com/Jun-Jie-Shi/M2FTrans.

Digital Object Identifier 10.1109/JBHI.2023.3326151

I. Introduction

ACCURATE brain tumor segmentation is crucial for quantitative assessment of tumor progression and surgery treatment planning. Magnetic resonance imaging (MRI) provides various tissue contrast views and spatial resolutions for brain examination, making it possible to quantitatively categorize brain tumor regions into heterogeneous subregions by comparing MRI modalities with different contrast levels (i.e., T1, T1c, T2, and Flair) [1], [2], [3], [4], [5]. Thus, multi-modality MRI imaging becomes a standard routine as different modalities complement each other in understanding brain structure and physiopathology. However, in clinical practice, MRI sequences may be incomplete due to image degradation, patient motion-related artifacts, incorrect acquisition settings, short scan times, etc. Insufficient brain information will be the bottleneck of brain tumor segmentation. How to reconstruct complete brain information given incomplete modalities is of great clinical value, which is formulated as incomplete multi-modality brain tumor segmentation.

Existing approaches mainly differ in fusion strategies as illustrated in Fig. 1. One typical way is to calculate the mean and variance of each accessible modality and fuse the corresponding features with equal importance, which may fail to effectively aggregate features with missing modalities. Another solution is to directly fuse modality features through convolution followed by a modality re-weighting mechanism based on the relative weights across modalities or between tumor regions and different modalities, but such a convolution-based fusion strategy may be insufficient to incorporate global information. It is due to the inductive bias of locality and weight sharing of convolutional operations. Comparatively, transformer, as a sequence-to-sequence framework, builds pair-wise dependency for each pair of tokens/patches. By dividing an input image into patches/tokens, the features of each patch can be re-weighted and refined based on the interactions with all other patches through transformers. In this way, global information and long-range dependency are well captured. Therefore, introducing a transformer [6], [7] module to complement global information is a straightforward solution, but it struggles to minimize the impact of missing values brought by incomplete modalities in attention calculation. How to construct a more compatible transformer-style fusion strategy under the existence of incomplete/missing modalities is demanding yet under-explored.

Inspired by the class tokens in vision transformers (ViT) [7] and masked attention [8], we propose M^{2}FTrans to exploit transformers for feature fusion under incomplete multi-modal settings. Specifically, M^{2}FTrans first consists of four encoders

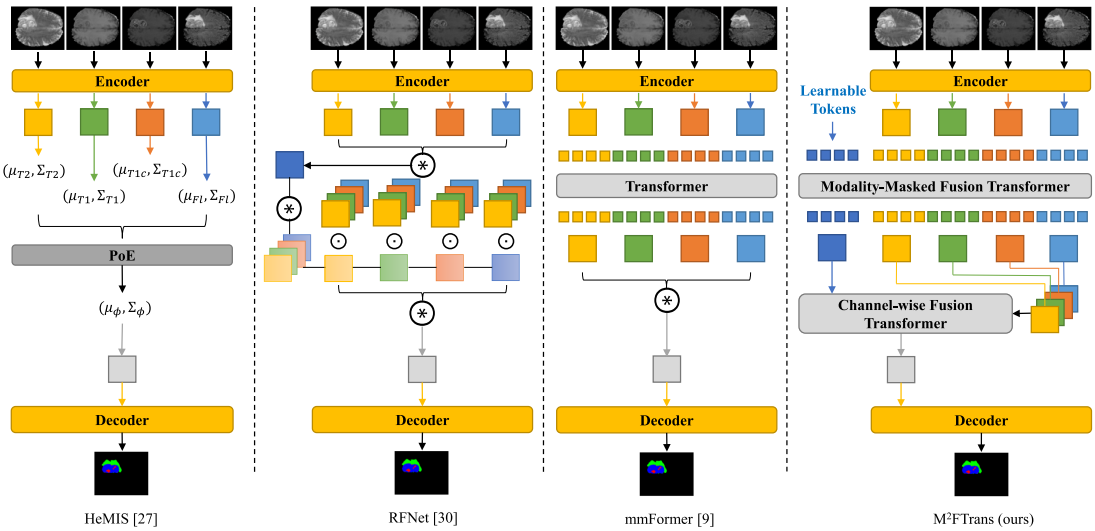


Fig. 1. Representative fusion strategies for incomplete multi-modality brain tumor segmentation. Given modality-specific features from different modalities, HeMIS [27] shares statistical indicators but is less effective, RFNet [30] adopts convolution but suffers from insufficient global information, mmFormer [9] introduces vanilla self-attention to enrich global information but struggles to well deal with missing modalities, and our M^{2}FTrans introduces learnable fusion tokens and calculate masked self-attention for better cross-modality interaction and is more robust to missing modalities. Here, PoE is short for Products of Experts developed in HeMIS-like approaches, corresponding to the Abstraction Layer in the original HeMIS.

and one shared decoder to extract modality-specific features from the four modalities separately. To minimize the negative influence of missing modalities, instead of direct inter-modality feature fusion through cross-attention like mmFormer [9], we introduce learnable fusion tokens to a modality-masked fusion transformer (MMFT) to adaptively incorporate accessible cross-modality features. As modalities contribute differently to tumor regions, both spatial weight attention (SWA) and channel-wise fusion transformers (CFT) are jointly adopted to re-weight modalities during fusion from different perspectives. In this way, fusion features are less likely to be dominated by certain modalities, being more robust to various incomplete multi-modality scenarios especially when dominant modalities are missing. Experimental results on the widely-used BraTS2018 and BraTS2020 datasets demonstrate the superiority of M^{2}FTrans against the state-of-the-art approaches for brain tumor segmentation under various settings (i.e., complete and incomplete modalities). The main contributions are summarized as follows:

- A transformer-style fusion strategy to deal with incomplete multi-modality scenarios for brain tumor segmentation with richer global information.
- Modality-masked fusion transformers with learnable fusion tokens for effective feature fusion while minimizing the negative influence of missing modalities.
- Spatial weight attention and channel-wise fusion transformers to adaptively re-weight modalities to avoid dominant modalities given the incomplete/missing modality problem.

The rest of this article is organized as follows. Section II reviews related works on brain tumor segmentation with incomplete/missing modalities and Section III describes the proposed M^{2}FTrans in detail. We present a thorough evaluation against the state-of-the-art methods in Section IV and ablation studies in Section V. Section VI concludes this article.

II. Related Work

A. Multi-Modality Brain Tumor Segmentation

Brain tumor segmentation is a fundamental task in medical image analysis, and related works can be roughly categorized as follows:

1) Complete Multi-Modality: Brain tumor segmentation with complete/full modalities follows the same pipeline of medical image segmentation, where both convolutional neural networks (CNN) [1], [2], [4], [10], [11], [12], [13] and transformers [3], [14], [15], [16], [17] have been extensively studied. More recently, SF-Net [37] introduced pixel-level image fusion as an auxiliary task to regularize feature learning while both deep semantics and edge information are jointly fused in [38]. SegTransVAE [39] exploited transformers with a variational autoencoder (VAE) branch to reconstruct the input images jointly with segmentation. TbraTS [40] quantified the voxel-wise uncertainty for brain tumor segmentation by introducing the confidence level for image segmentation to disease diagnosis. UMMNet [41] developed an uncertainty-aware multi-dimensional mutual learning framework to learn different dimensional networks simultaneously, providing useful soft labels as supervision to the others for improving model generalizability. The main challenge is how to balance modalities rather than being biased toward certain modalities. As these frameworks are trained by full modalities, they would encounter severe performance degradation when dealing with incomplete/missing modalities, making them less attractive in clinical practice.

2) Incomplete Multi-Modality: Solutions to missing/incomplete modalities include knowledge distillation (KD) [18], [19], [20], [21], [22], [23], generative adversarial networks (GAN) [24], [25], [26], and shared representation learning [9], [27], [28], [29], [30]. In terms of KD-based approaches, ACN [21] and SMU-Net [23] are the most representative ones, where ACN trained a teacher-student framework for each missing situation and SMU-Net proposed a style matching

mechanism to reconstruct missing information from the full-modality network. Both of the two approaches are of high complexity, as they trained a separate model for each multi-modality setting (i.e., 15 models in total) to deal with missing modalities. Despite the efficiency, KD-based approaches may encounter instabilities in training during knowledge transfer. RA-HVED [24] is the most recent GAN-based approach using a region-of-interest attentive discriminator to learn segmentation-relevant shared latent representations. One limitation of GAN-based approaches lies in unstable training, often resulting in less competitive performance.

Given the limitations of knowledge distillation and generative adversarial learning in feature fusion especially when dominant modalities are missing, shared representation learning, to complement missing information by sharing a spatial mapping across modalities, becomes mainstream in incomplete multi-modality brain tumor segmentation. Specifically, HeMIS [27] computed variance statistics (i.e., mean and variance) to construct a uniform representation for segmentation, and U-HVED [28] employed a multi-modal variational auto-encoder (MVAE) [31] to embed all observed modalities. RobustMseg [29] performed fusion by feature disentanglement and modality re-weighting via a gating strategy, and RFNet [30] performed region-aware fusion using attention gating modules by exploring each modality's contribution to different tumor regions. Unfortunately, all these approaches fail to build global/remote dependencies across modalities. In addition, MAML [42] used multiple U-Net architectures with the same structure to extract the features of different modalities, and in the final stage fused the extracted modality-specific features by convolutional weighting. It further introduced a modality-aware mutual learning strategy to make its architecture robust to incomplete/missing modality scenarios. U-Net-MFI [43] introduced graph convolution networks for incomplete brain tumor segmentation. It treated modalities as graph nodes and indicated the presence or absence of each modality through the introduced multi-modal code, which in turn guided the model to adaptively fuse complementary modalities' features in different incomplete/missing modality scenarios. M^{3}AE [45] designed a two-stage architecture including self-supervised pre-training and self-distillation to reduce the parameters of convolutional neural networks for brain tumor segmentation against missing modalities. mmFormer [9] is the first work to introduce transformers to exploit both intra- and inter-modality dependence for feature fusion. However, adopting vanilla mutual attention computation in mmFormer would encounter difficulties with missing modality features.

B. Masked Attention in Transformers

Despite the remarkable success of pair-wise self-attention in transformers, its computation-extensive nature, and alternative solutions using masked attention have been proposed to make self-attention more focused on specific regions/features. Mask2Former [8] effectively constrains cross-attention to localized regions, thereby enabling the Transformer to efficiently process high-dimensional output. Zorro [32] approach has demonstrated impressive performance in audio classification tasks by employing a meticulously designed mask transformer. They strategically divide the information flow into uni-modal and multi-modal streams, subsequently generating corresponding outputs through comparative analysis and supervised guidance.

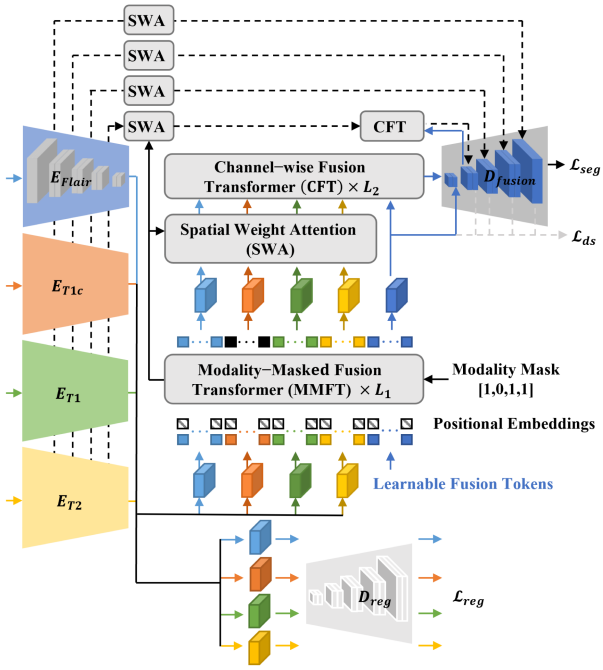


Fig. 2. Overview of M^{2}FTrans with incomplete modalities for brain tumor segmentation. The encoders E_{T1c}, E_{T1}, and E_{T2} share the same architectures with E_{Flair}.

The idea of these methods motivates us to design a missing-modality-based mask transformer to achieve robust multi-modal feature fusion for brain tumor segmentation tasks with incomplete modalities.

III. Methodology

The complete framework of M^{2}FTrans is presented in Fig. 2. The key idea is to re-weight independent modalities from both the spatial and channel dimensions, realized by spatial weight attention (SWA) and channel-wise fusion transformers (CFT), and introduce learnable fusion tokens for cross-modality interaction realized by modality-masked fusion transformers (MMFT). In the following, we detail each component of M^{2}FTrans.

A. Modality-Specific Feature Extraction

Let M={Flair, T1c, T1, T2} denote the complete set of modalities. Given any 3D input modality path x_m of each modality $m \in M$ and the ground-true annotation y, an independent modality-specific encoder E_m following 3D U-Net is trained to extract modality-specific features $F_m \in \mathbb{R}^{c\times h\times w\times d}$.

In typical multi-modality frameworks, features of different modalities are directly fused and fed to a decoder for segmentation. The decoder tends to select the most discriminative modality as the main modality for brain tumor segmentation, which would suffer from severe performance degradation especially when the main modality is missing. To avoid modality bias and balance different modalities, a shared decoder D_{reg} is introduced for regularization as illustrated in Fig. 2. In this way, modality-specific features are projected to a shared latent space, and each modality is trained separately for segmentation so as to reduce the negative influence of missing modalities. The four

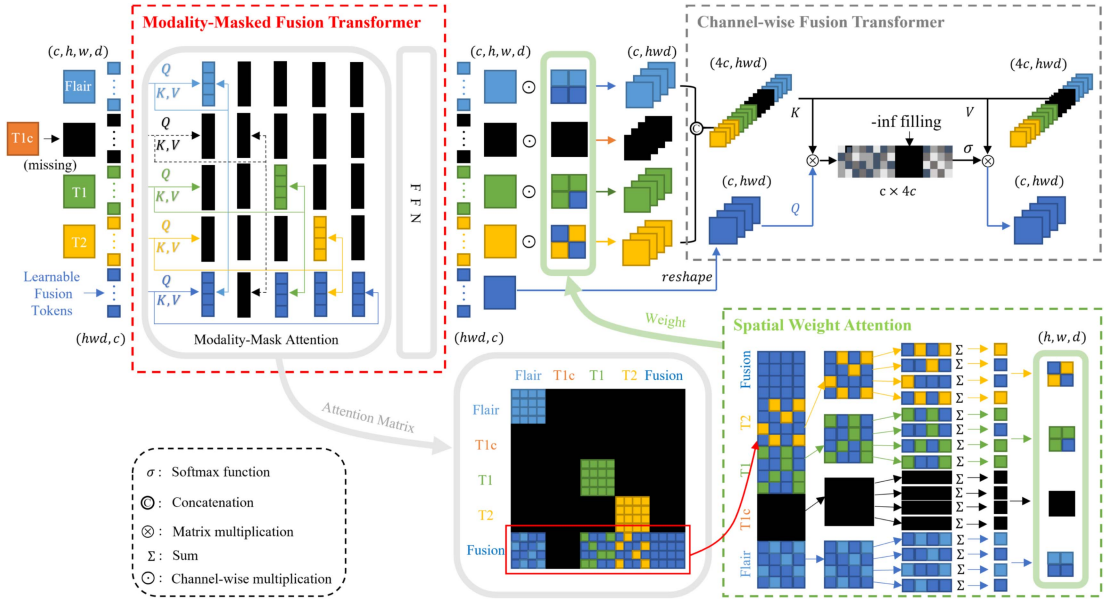


Fig. 3. Attention mechanisms in M^{2}FTrans for modality fusion and re-weighting, including modality-masked fusion transformers for cross-modality feature fusion of accessible modalities, spatial weight attention for re-weighting modality-specific features, and channel-wise fusion transformers for cross-modality feature fusion along the channel dimension among accessible modalities.

encoders and the shared decoder are trained by

$$\mathcal{L}_{reg} = \sum_{m \in M} \mathcal{L}_{Dice}(D_{reg}(E_m(x_m)), y) + \mathcal{L}_{WCE}(D_{reg}(E_m(x_m)), y), \quad (1)$$

where $\mathcal{L}_{Dice}$ and $\mathcal{L}_{WCE}$ denote the Dice loss and the weighted cross-entropy loss respectively.

B. Modality-Masked Fusion Transformer

Despite the effectiveness of transformers in capturing long-range dependency for inter-modality feature exploration, directly concatenating and projecting the features of different modalities into queries (Q), keys (K), and values (V) for self-attention calculation is quite sensitive to missing modalities [7], [9]. Inspired by the learnable class tokens in ViT [7] to generate output embedding vectors by interacting with input tokens for classification, we introduce learnable fusion tokens for inter-modality feature fusion and propose a modality-masked fusion transformer (MMFT) to deal with missing modalities as shown in Fig. 3.

Given the modality-specific features $F_m \in \mathbb{R}^{c \times h \times w \times d}$ of each modality $m \in M$ produced by the corresponding encoder E_m, we reshape it into $F_m \in \mathbb{R}^{N \times c}$ where $N = h \times w \times d$, and introduce learnable fusion tokens $F_{fusion} \in \mathbb{R}^{N \times c}$. Then, F_m and F_{fusion} are concatenated into $F_{multi} \in \mathbb{R}^{5N \times c}$. After combined with learnable positional embeddings $PE \in \mathbb{R}^{5N \times c}$, the feature embeddings Z_0, written as

$$Z_0 = F_{multi} + PE, \quad (2)$$

would be fed to MMFT.

The core component of MMFT is modality-masked attention to make modality-specific tokens focus only on interacting within each modality and the fusion tokens interact with all accessible modality-specific tokens through self-attention for inter-modality feature fusion. Specifically, given Z_0, it is projected to Q, K, and V following vanilla self-attention. To deal with missing modalities, we build a binary attention mask $\mathcal{M} \in \{0,1\}^{5N \times 5N}$, corresponding to the pair-wise dependencies across tokens in Z_0. Given any position (i, j) in $\mathcal{M}$, $\mathcal{M}_{i,j}$ is to determine whether filter out the relationship between $q_i \in Q$ and $k_j \in K$ is determined according to:

- If q_i and k_j belong to the same existing modality, $\mathcal{M}_{i,j} = 1$ is set to refine the modality-specific features.
- If q_i belongs to the fusion tokens F_{fusion} and k_j comes from an existing/accessible modality, $\mathcal{M}_{i,j} = 1$ is set to fuse modality-specific features by cross-attention.
- If either q_i or k_j comes from missing modalities, $\mathcal{M}_{i,j} = 0$ is set to avoid building impossible inter-modality dependency.

According to $\mathcal{M}_{i,j}$, modality-masked attention (MA) is written as

$$\mathrm{MA}(Q, K, i, j) = \frac{\mathcal{M}_{i,j} \exp\left(\frac{q_i (k_j)^T}{\sqrt{c/H}}\right)}{\sum_{j', \mathcal{M}_{i,j'}=1} \exp\left(\frac{q_i (k_{j'})^T}{\sqrt{c/H}}\right)}, \quad (3)$$

where H is the number of heads in multi-head modality-masked attention (MHMA) implemented by concatenation just like vanilla self-attention in ViT [7]. Through a feed-forward network (FFN), the original feature embeddings Z_0 after one MMFT layer are written as

$$\begin{aligned} Z_0' &\leftarrow \mathrm{MA}(Q, K) \cdot V, \\ Z_1 &\leftarrow \mathrm{MHMA}(Z_0') + Z_0, \\ Z_1 &\leftarrow \mathrm{FFN}(\mathrm{LN}(Z_1)) + Z_1, \end{aligned} \quad (4)$$

where LN is layer normalization. Through L_1 MMFT layers, the learnable fusion tokens $F_{fusion} \in \mathbb{R}^{N\times c}$ are updated and reshaped to $\hat{F}_{fusion} \in \mathbb{R}^{c\times h\times w\times d}$ possessing cross-modality information, and the modality-specific feature tokens $F_m \in \mathbb{R}^{N\times c}$ are updated and reshaped to $\hat{F}_m \in \mathbb{R}^{c\times h\times w\times d}$ containing more global features within each modality.

C. Spatial Weight Attention

In modality-masked attention, tokens of different modalities are treated with equal importance, where each row corresponding to each fusion token of the attention matrix (i.e., the last rows of the modality-masked attention matrix as illustrated in Fig. 3) model the relationship/similarity between the fusion token and tokens from existing modalities. However, modalities, as well as tokens belonging to the same modality, can contribute differently to feature fusion. Including all modality-specific tokens equally to update the fusion tokens can be counter-productive. Therefore, it is necessary to re-weight tokens from accessible modalities for better feature fusion. As discussed in [33], given a self-attention matrix, the column vectors can somewhat reflect the importance of individual tokens. Inspired by this, we propose spatial weight attention for weighted cross-attention between fusion tokens and modality-specific tokens and gradually exploit important tokens in each modality along the spatial dimension as illustrated in Fig. 3.

Specifically, given L modality-masked fusion layers as described above, the modality-masked attention matrix of the first layer is used for token importance modeling. It is based on the observation that self-attention matrices are more likely to be uniform in deep layers [34]. Given the computed modality-masked attention matrix MA_1, the weight of each token j is calculated by summing over the column vectors, namely

$$Col(j) = \sum_{H} \sum_{i=4N+1}^{5N} \mathrm{MA}_1(i,j), j \in [1, 4N], \tag{5}$$

where H is the number of modality-masked attention heads. After calculating $Col \in \mathbb{R}^{1\times 4N}$ for all modalities, token weights of each modality $m \in M$ are obtained via slicing, i.e., $Col_m = \mathrm{Split}(Col) \in \mathbb{R}^{1\times N}$. Then, Col_m is reshaped to the same size as $\hat{F}_m$ to indicate spatial importance $I_m \in \mathbb{R}^{1\times h\times w\times d}$, and modality-specific features $\hat{F}_m$ are re-weighted and updated to $\widetilde{F}_m = \hat{F}_m \odot I_m$ where $\odot$ is element-wise multiplication. Noting that, for missing modalities, I_m is set to zero and spatial weight attention is applied to the skip connections of all stages by upsampling.

D. Channel-Wise Fusion Transformer

Though spatial redundancy among tokens is alleviated through spatial weight attention, there may exist redundancy along the channel dimension, i.e., across feature maps of each token. To address this, channel-wise fusion transformers (CFT) are constructed between fusion tokens and modality-specific tokens as described in Fig. 3. Specifically, given the enhanced modality-specific features $\widetilde{F}_m \in \mathbb{R}^{c\times h\times w\times d}$ of each modality $m \in M$ after both MMFT and SWA, they are first reshaped along the spatial dimension to $\widetilde{Z}_0^m \in \mathbb{R}^{c\times N}$ and concatenated as $\widetilde{Z}_0 = \mathrm{Concat}(\widetilde{Z}_0^m, m \in M) \in \mathbb{R}^{4c\times N}$. Then, $\widetilde{Z}_0$ is projected into K and V for self-attention calculation. Similarly, the fusion features $\hat{F}_{fusion} \in \mathbb{R}^{c\times h\times w\times d}$ after MMFT are reshaped into $\hat{Z}_0 \in \mathbb{R}^{c\times N}$ and projected into Q. As attention is channel-wisely computed and the original spatial information will be lost if performing a fully connected layer along the spatial dimension for projection, we adopt convolutional locally-connected projection grouped by each modality in the channel dimension. The projection process is formulated as

$$\begin{aligned} Q &= \Psi_q\left(\hat{Z}_0\right), \\ K &= \mathrm{Concat}\left(\Psi_k^m\left(\widetilde{Z}_0^m\right), m \in M\right), \\ V &= \mathrm{Concat}\left(\Psi_v^m\left(\widetilde{Z}_0^m\right), m \in M\right), \end{aligned} \tag{6}$$

where $\Psi_{\alpha\in\{q,k,v\}}$ is expressed as:

$$\Psi_{\alpha\in\{q,k,v\}}(\cdot) = \phi_\alpha\left(\psi_\alpha\left(\varphi_\alpha(\cdot)\right)\right), \tag{7}$$

where ϕ_α and φ_α correspond to point-wise convolution with a convolution kernel size of 1 to map each token to a corresponding feature preview, and ψ_α corresponds to depth-wise convolution with a convolution kernel size of 3 to map each token to a local interval.

Similar to modality-masked attention, we introduce a binary attention mask $\mathcal{G} \in \{0,1\}^{c\times 4c}$ along the channel dimension, when performing channel-wise attention calculation. Given any position (i,j), $\mathcal{G}_{i,j} \in \mathcal{G}$ corresponds to the relationship between q_i and k_j in Q and K respectively. If k_j corresponds to a missing modality, then $\mathcal{G}_{i,j}$ is set to 0. Otherwise, $\mathcal{G}_{i,j}$ is set to 1. Based on this, channel-wise masked attention (CMA) is calculated by

$$\mathrm{CMA}(Q,K,i,j) = \frac{\mathcal{G}_{i,j}\exp\left(\frac{q_i(k_j)^T}{\sqrt{N}}\right)}{\sum_{j',\mathcal{G}_{i,j'}=1}\exp\left(\frac{q_i(k_{j'})^T}{\sqrt{N}}\right)}, \tag{8}$$

where $N = h \times w \times d$. Following CMA, Ψ_o with the same structure as $\Psi_{\alpha\in\{q,k,v\}}$ is adopted for projection, namely

$$\hat{Z}_0 \leftarrow \Psi_o\left(\mathrm{CMA}(Q,K)\cdot V\right) + \hat{Z}_0. \tag{9}$$

In terms of feed-forward, we add a convolutional kernel of group convolution of size 3 between two fully connected layers in the channel dimension to obtain richer local feature information and update the original fusion features $\hat{Z}_0$ into $\widetilde{Z}_1$ through one CFT layer, namely

$$\widetilde{Z}_1 \leftarrow \widetilde{\mathrm{FFN}}\left(\hat{Z}_0\right) + \hat{Z}_0, \tag{10}$$

where $\widetilde{\mathrm{FFN}}$ is the feed-forward network. After L_2 CFT layers, the fusion features $\hat{F}_{fusion} \in \mathbb{R}^{c\times h\times w\times d}$ are further enhanced into $\widetilde{F}_{fusion} \in \mathbb{R}^{c\times h\times w\times d}$ based on the spatially and channel-wisely re-weighted modality-specific features.

E. Decoding With Deep Supervision

For decoding and segmentation, a five-stage decoder D_{fusion} with a similar structure as D_{reg} is adopted, consisting of multiple convolutional blocks stacked together. The main difference between D_{fusion} and D_{reg} lies in the skip connections where features from the encoders would go through spatial weight attention and channel-wise masked attention before concatenation

in D_{fusion} as shown in Fig. 3. Specifically, given the modality-specific features from encoders, modality-masked fusion transformers (MMFT), spatial weight attention, and channel-wise fusion transformers are sequentially applied to learn high-quality fusion features, namely $\hat{F}_{fusion}$ and $\widetilde{F}_{fusion}$. Then, the fused features are concatenated, upsampled, and fed to channel-wise fusion transformers to interact with the re-weighted skip connection features from encoders according to spatial weight attention. Starting from the bottom two layers, the fused features would be upsampled and concatenated with the re-weighted skip connection features progressively till the final output layer. D_{fusion} is trained by a segmentation loss,

$$\mathcal{L}_{seg} = \mathcal{L}_{Dice}\left(D_{fusion}(\text{Concat}(E_m(x_m))), y\right) + \mathcal{L}_{WCE}\left(D_{fusion}(\text{Concat}(E_m(x_m))), y\right). \quad (11)$$

To stabilize the training process and regularize the fusion features, deep supervision is adopted and trained by

$$\mathcal{L}_{ds} = \sum_{l=1}^{5} \mathcal{L}_{Dice}(Up_{2^{l-1}}(\hat{F}^l_{fusion}), y) + \mathcal{L}_{WCE}(Up_{2^{l-1}}(\hat{F}^l_{fusion}), y), \quad (12)$$

where $Up_{2^{l-1}}$ represents $2^{l-1}\times$ upsampling and $\hat{F}^l_{fusion}$ denotes the fusion features from the l-th stage of D_{fusion}.

F. Overall Loss

The overall loss includes the segmentation regularization loss $\mathcal{L}_{reg}$ for modality-specific feature extraction, the deep supervision loss $\mathcal{L}_{ds}$, and the final segmentation loss $\mathcal{L}_{seg}$, written as

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{reg} + \mathcal{L}_{ds}. \quad (13)$$

IV. Evaluation

A. Experimental Setup

1) Datasets: Two datasets from the Multimodal Brain Tumor Segmentation Challenge (BRATS) [35] are adopted for evaluation, namely BRATS2018 and BRATS2020, both of which contain data from four MRI modalities including Flair, T1c, T1, and T2. Following [28], [29], [30], we excised the black background regions outside the brain and normalized each MRI modality to zero mean and unit variance. For a fair comparison on the BRATS2018 dataset consisting of 285 training samples, we used the same data split lists in [28], [30] and split the data into 199, 29, and 57 subjects for training, validation, and test respectively. In terms of the BRATS2020 dataset containing 369 training samples, we split the data into the same 219, 50, and 100 subjects for training, validation, and test respectively by strictly following [30]. For both datasets, the Dice similarity coefficient (DSC) and the Hausdorff distance (HD) are utilized for evaluation.

2) Implementation Details: The framework was implemented in Pytorch and trained using an AdamW [36] optimizer with an initial learning rate of 2e-4, a weight decay of 1e-4, and a batch size of 2 on two 24 G NVIDIA Geforce RTX 3090 GPUs for 1000 epochs. Specifically, we adopted a warm-up learning rate adjustment strategy and a poly decay strategy with $p = 0.9$ during training. Following [28], [29], [30], modality masks were introduced to discard modalities and simulate various missing-modality cases. For training, each volume was randomly cropped to $80\times80\times80$ pixels and augmented by random rotation, intensity change, and mirror flip.

For the positional embedding in M^{2}FTrans, we follow the use of learnable positional embeddings in mmFormer's Inter-modal Transformers, which are also added for the learnable fusion tokens just like a larger-size class token in ViT.

B. Comparison With SOTA Approaches

Based on the availability of source codes and data splits, five state-of-the-art approaches, adopting the same publicly-available data splits for incomplete multi-modality brain tumor segmentation, are included for comparison, including CNN-based (i.e., HeMIS [27], U-HVED [28], RobustMSeg [29], and RFNet [30]) and transformer-based (i.e., mmFormer [9]).

1) Quantitative Comparison: Quantitative comparison results under all fifteen multi-modality combinations on the BraTS2018 dataset are summarized in Table I. Among the comparison approaches, RFNet achieves the best overall segmentation performance for WT, TC, and ET, outperforming mmFormer under 40 out of 45 modality settings and achieving an average increase of 0.66%, 1.48%, and 0.85% in DSC respectively. Comparatively, M^{2}FTrans consistently outperforms both RFNet and mmFormer under all 45 modality settings for WT, TC, and ET respectively, leading to an average increase of 0.94%, 1.38%, and 4.58% in DSC compared to RFNet and an average increase of 1.60%, 2.86%, and 5.43% in DSC compared to mmFormer.

Similar observations are found on the BraTS2020 dataset as summarized in Table II. Among the comparison approaches, RFNet achieves better segmentation performance across all three tumor types compared to mmFormer, leading to an average increase of 0.83%, 0.50%, and 0.48% in DSC respectively. M^{2}FTrans consistently outperforms other approaches under all 45 modality settings (i.e., 15 for each tumor type) with an average increase of 1.02%, 1.18%, and 4.98% in DSC compared to RFNet and 1.85%, 1.68%, and 5.46% against mmFormer for WT, TC, and ET respectively.

One interesting observation across both datasets is that the more difficult the tumor type is to segment (*e.g.*, WT < TC<ET), the more performance gains M^{2}FTrans achieves. It further validates the effectiveness of M^{2}FTrans in exploiting richer cross-modality information which is more beneficial for challenging cases. In addition, as summarized in Tables I and II, M^{2}FTrans achieves statistically significant (i.e. all p-value scores being lower than 0.05) performance improvements against other approaches on both datasets, validating its stability under various modality settings.

In addition to the Dice coefficient, Hausdorff distance is included for shape evaluation on both datasets as summarized in Table III. Compared to other approaches, M^{2}FTrans achieves the best performance on shape preservation, which is crucial in clinical diagnosis. Statistical analysis indicates that the performance improvements of M^{2}FTrans are statistically significant (i.e., all p-value scores being lower than 0.05).

2) Qualitative Comparison: Qualitative results of mmFormer, RFNet, and M^{2}FTrans under different modality combinations are illustrated in Fig. 4. Given only one modality, though all three approaches encounter performance degradation, M^{2}FTrans achieves better segmentation performance, especially for ET. It validates the robustness of M^{2}FTrans in exploring

TABLE I
QUANTITATIVE COMPARISON RESULTS MEASURED IN DSC (%) ON BRATS2018

Type	Flair	○	○	○	●	○	○	●	○	●	●	●	●	●	○	●	Avg.	p-value
	T1	○	○	●	○	○	●	●	●	○	○	●	●	○	●	●		
	T1c	○	●	○	○	●	●	○	○	○	●	●	○	●	●	●		
	T2	●	○	○	○	●	○	●	●	●	○	○	●	●	●	●		
WT	HeMIS	78.31	55.82	53.23	75.16	80.62	63.28	81.14	80.30	83.13	82.24	83.77	84.82	85.19	81.83	85.85	76.98	<0.001
	U-HVED	80.11	61.41	57.03	77.30	82.92	66.82	82.59	82.06	85.42	84.07	85.64	86.19	87.37	83.36	87.68	79.33	<0.001
	RobustMSeg	83.43	70.66	67.91	80.20	85.56	74.55	85.75	85.21	87.76	86.48	88.73	88.26	88.33	85.96	88.65	83.07	<0.001
	RFNet	84.68	76.33	76.16	85.69	86.69	79.54	88.05	86.60	88.20	88.35	88.76	89.01	89.19	87.17	89.46	85.59	<0.001
	mmFormer	84.28	75.24	73.36	85.01	86.10	78.60	87.39	86.00	88.00	88.11	88.51	88.49	89.01	86.61	89.19	84.93	<0.001
	M²FTrans	**86.92**	**77.78**	**77.21**	**87.15**	**88.07**	**81.06**	**88.37**	**87.45**	**89.24**	**88.85**	**88.95**	**89.39**	**89.78**	**88.00**	**89.73**	**86.53**	-
TC	HeMIS	56.68	62.49	33.18	46.25	74.95	66.28	51.90	58.73	59.21	73.60	75.31	60.34	77.11	75.93	77.45	63.29	<0.001
	U-HVED	58.83	67.79	41.68	43.66	76.26	70.77	51.88	60.89	60.89	75.23	76.30	62.26	77.95	76.99	78.37	65.32	<0.001
	RobustMSeg	65.86	77.43	55.61	55.73	83.91	80.72	68.37	70.45	70.51	81.11	82.26	72.39	82.70	84.02	83.18	74.28	<0.001
	RFNet	69.69	81.88	65.92	68.14	84.02	82.36	73.92	72.54	73.06	82.87	83.89	74.68	83.69	84.77	84.48	77.73	0.016
	mmFormer	67.97	79.01	62.06	64.80	82.26	81.37	72.72	71.38	71.93	82.04	83.42	74.09	83.23	83.43	84.00	76.25	<0.001
	M²FTrans	**72.37**	**82.60**	**66.24**	**69.89**	**85.23**	**83.45**	**74.08**	**74.45**	**75.40**	**84.78**	**85.26**	**76.48**	**85.29**	**85.46**	**85.67**	**79.11**	-
ET	HeMIS	30.06	57.08	6.60	20.63	63.96	59.17	14.83	29.88	31.48	65.62	68.16	29.74	64.66	64.82	66.69	44.89	<0.001
	U-HVED	30.85	59.49	13.18	13.40	64.66	64.18	18.98	32.98	32.73	64.29	66.56	31.84	66.60	67.21	68.46	46.36	<0.001
	RobustMSeg	37.13	63.99	26.30	28.92	66.93	67.24	36.24	40.54	40.26	66.92	67.90	42.38	65.70	68.87	69.36	52.58	0.001
	RFNet	38.11	74.47	36.26	36.98	76.72	73.49	39.84	42.09	42.85	77.65	78.26	44.52	74.55	76.84	76.65	59.29	0.006
	mmFormer	37.19	75.37	32.45	31.59	74.47	76.30	38.76	40.26	41.09	76.83	79.53	43.01	77.17	74.88	77.69	58.44	0.002
	M²FTrans	**46.41**	**78.92**	**37.24**	**37.98**	**80.93**	**80.77**	**43.48**	**47.23**	**49.12**	**82.05**	**82.19**	**49.79**	**80.56**	**80.82**	**80.61**	**63.87**	-

The best and second-best results under each modality setting are marked in bold and underlined.

TABLE II
QUANTITATIVE COMPARISON RESULTS MEASURED IN DSC (%) ON THE BRATS2020 DATASET

Type	Flair	○	○	○	●	○	○	●	○	●	●	●	●	●	○	●	Avg.	p-value
	T1	○	○	●	○	○	●	●	●	○	○	●	●	○	●	●		
	T1c	○	●	○	○	●	●	○	○	○	●	●	○	●	●	●		
	T2	●	○	○	○	●	○	●	●	●	○	○	●	●	●	●		
WT	HeMIS	76.07	58.26	51.23	79.52	80.69	64.69	83.74	79.46	84.63	83.56	85.55	85.97	87.26	82.35	88.00	78.07	<0.001
	U-HVED	80.02	62.31	55.13	79.88	82.30	64.74	83.05	81.59	86.63	84.73	85.80	87.16	87.78	82.77	88.06	79.46	<0.001
	RobustMSeg	83.00	71.61	67.73	82.42	86.10	76.31	87.19	85.64	88.34	87.75	88.69	89.02	89.28	86.49	89.57	83.94	<0.001
	RFNet	86.55	76.74	76.82	86.97	88.16	80.50	89.37	87.97	89.63	89.43	90.24	90.42	90.38	88.54	90.89	86.84	<0.001
	mmFormer	85.37	74.86	74.91	86.27	87.35	79.61	88.91	87.19	89.04	89.03	89.61	89.96	89.82	87.85	90.31	86.01	<0.001
	M²FTrans	**87.20**	**78.80**	**79.15**	**88.70**	**88.67**	**82.40**	**90.30**	**88.34**	**90.56**	**90.38**	**91.00**	**90.91**	**91.16**	**89.01**	**91.36**	**87.86**	-
TC	HeMIS	56.71	66.35	34.81	53.31	76.34	70.49	60.29	59.60	63.82	73.87	75.63	65.10	77.79	77.41	78.34	65.98	<0.001
	U-HVED	62.35	69.70	43.57	51.92	78.68	73.50	58.17	65.10	65.31	76.05	77.93	66.89	80.04	79.68	80.49	68.62	<0.001
	RobustMSeg	63.87	77.95	53.29	57.28	83.55	81.51	67.01	69.65	69.35	81.93	82.47	70.64	83.17	84.38	83.39	73.96	<0.001
	RFNet	69.85	81.72	64.78	68.82	**84.75**	82.42	73.38	72.03	73.70	85.46	84.62	74.15	**85.47**	84.10	85.09	78.02	0.018
	mmFormer	70.21	80.74	64.24	67.80	84.30	82.00	71.83	72.61	72.82	84.44	84.59	73.90	84.63	84.28	84.49	77.52	<0.001
	M²FTrans	**72.31**	**81.85**	**66.75**	**72.20**	84.62	**83.70**	**74.44**	**73.56**	**75.42**	**85.54**	**85.82**	**76.14**	85.27	**84.90**	**85.43**	**79.20**	-
ET	HeMIS	30.52	61.70	12.47	26.25	68.60	65.40	29.57	32.66	35.87	67.56	68.16	36.73	68.21	70.01	69.39	49.54	<0.001
	U-HVED	37.30	65.77	19.95	19.32	69.56	67.70	28.84	38.79	38.18	68.03	70.21	39.07	70.94	70.11	72.41	51.74	<0.001
	RobustMSeg	40.14	74.16	25.42	32.67	74.80	73.98	38.61	42.21	42.41	75.62	76.99	45.10	73.39	74.68	74.42	57.64	<0.001
	RFNet	47.75	75.65	34.85	40.40	76.02	78.18	44.62	47.82	47.96	75.17	78.42	48.86	76.38	78.27	76.50	61.79	<0.001
	mmFormer	46.12	76.45	34.78	38.39	75.29	77.12	41.17	48.07	47.98	77.22	77.13	48.77	76.36	76.69	78.11	61.31	<0.001
	M²FTrans	**51.50**	**82.57**	**40.87**	**43.39**	**82.35**	**83.81**	**47.04**	**49.90**	**53.87**	**83.07**	**84.12**	**53.33**	**81.23**	**82.36**	**82.17**	**66.77**	-

The best and second-best results under each modality setting are marked in bold and underlined.

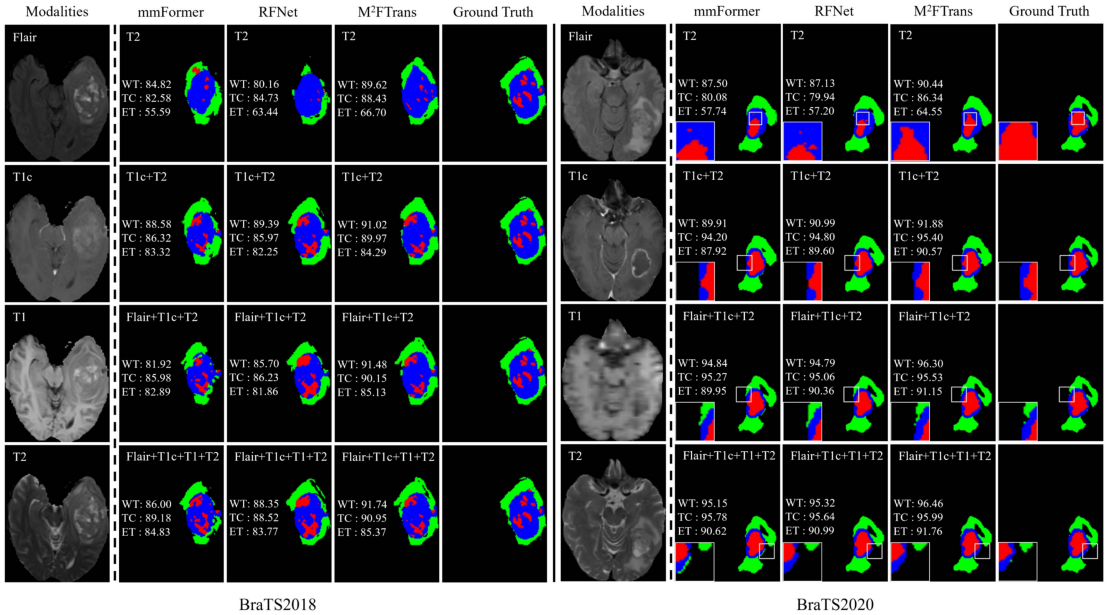


Fig. 4. Qualitative results of mmFormer [9], RFNet [30], and the proposed M²FTrans on BraTS2018 and BraTS2020. Detailed regions are zoomed in for better visualization and comparison. In addition, the DSC scores of the three tumor types in each image are provided.

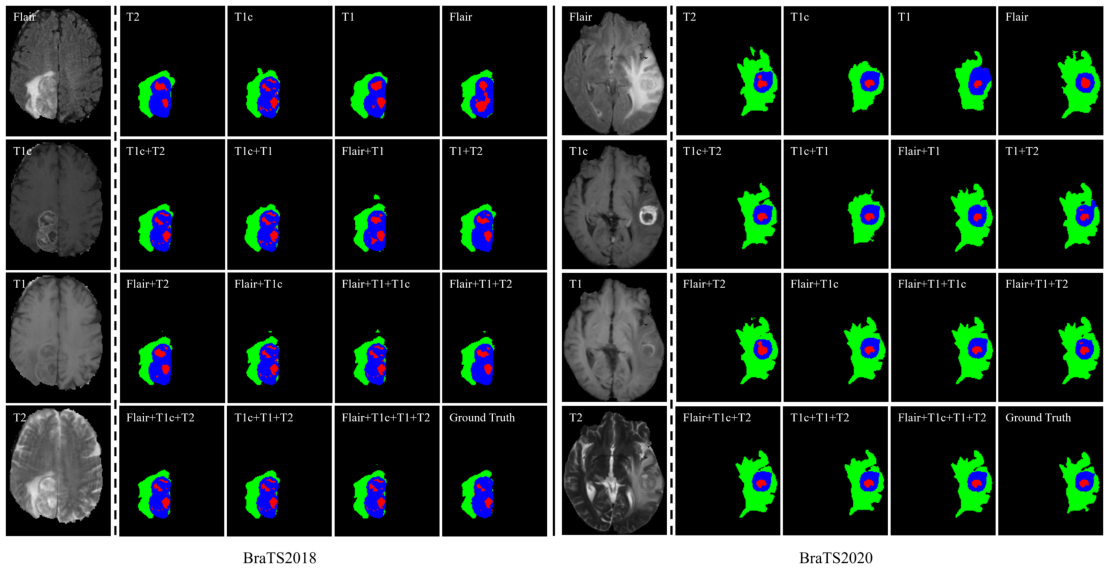


Fig. 5. Qualitative segmentation results of M^{2}FTrans on BraTS2018 and BraTS2020 under all fifteen multi-modality settings.

TABLE III
QUANTITATIVE COMPARISON RESULTS MEASURED IN AVERAGE HAUSDORFF DISTANCE (AVG. HD) ON BRATS2018 AND BRATS2020

Type	Method	BraTS2018		BraTS2020	
		Avg.	p-value	Avg.	p-value
WT	HeMIS	26.72	<0.001	27.32	<0.001
	U-HVED	25.10	<0.001	28.00	<0.001
	RobustMSeg	11.37	<0.001	13.05	<0.001
	RFNet	7.24	<0.001	8.42	<0.001
	mmFormer	7.30	0.002	7.71	<0.001
	M^{2}FTrans	**6.38**	-	**5.68**	-
TC	HeMIS	27.99	<0.001	25.27	<0.001
	U-HVED	25.18	<0.001	23.77	<0.001
	RobustMSeg	11.74	<0.001	12.70	<0.001
	RFNet	8.29	0.004	8.72	<0.001
	mmFormer	8.42	<0.001	8.09	<0.001
	M^{2}FTrans	**6.60**	-	**6.49**	-
ET	HeMIS	15.48	<0.001	16.70	<0.001
	U-HVED	13.48	<0.001	14.86	<0.001
	RobustMSeg	8.28	0.012	9.04	<0.001
	RFNet	7.13	0.032	6.66	0.006
	mmFormer	7.31	0.003	6.11	0.014
	M^{2}FTrans	**5.95**	-	**5.02**	-

The best results are marked in bold.

intra-modality information for segmentation. With the introduction of more modalities, all approaches would benefit from additional information, among which M^{2}FTrans achieves the best performance with better shape preservation, validating its effectiveness.

We further visualize the segmentation results of M^{2}FTrans under all fifteen modality settings as shown in Fig. 5. Given only one modality, T1c is the most informative modality for the segmentation of ET, while Flair seems more suitable for WT and TC. Under two-modality combinations, all segmentation results have been effectively improved, of which Flair and T1c obtain slightly better results. With the introduction of more modalities, segmentation results become more similar and stable under various modality combinations. The above results validate the effectiveness of M^{2}FTrans on cross-modality feature fusion.

V. ABLATION STUDY

A. On Components of M^{2}FTrans

For a more comprehensive evaluation, we conduct component -wise ablation studies as summarized in Table IV. Given only the encoders and decoder without additional regularization, the segmentation performance across three tumor types is far from satisfactory, worse than both RFNet and mmFormer according to the quantitative results in Table II. Through a shared decoder for regularization, intra-modality features are learned independently, which somewhat makes the decoder less biased to certain dominant modalities and improves the overall segmentation performance. With the introduction of MMFT for cross-modality feature fusion, significant performance gains are achieved especially for ET, validating its effectiveness. Coupling either SWA or CFT with MMFT for modality re-weighting is helpful, as it would better balance feature fusion in dealing with missing/incomplete modalities. After all, jointly utilizing all components achieves the best segmentation performance, leading to an average increase of 2.49%, 3.53%, and 6.99% in DSC for WT, TC, and ET respectively compared to the baseline. According to the component-wise complexity as summarized in Table IV, the backbone networks contribute most model parameters and GFLOPs. Of different components, MMFT owns the most model parameters and the highest GFLOPs. Therefore, adopting a lightweight backbone is expected to significantly improve the model efficiency of M^{2}FTrans, which is not the main focus of this work and will be explored in future work.

In M^{2}FTrans, spatial weight attention is applied to all skip connections for re-weighting, while channel-wise fusion transformers are just introduced to the bottom two layers. To figure out how CFT works on different levels of skip connections, we conduct additional ablation studies as summarized in Table V. Compared to the baseline, only applying SWA is slightly helpful, leading to an average increase of 0.35% in Avg. DSC, while jointly introducing one CFT to the bottom layer achieves much better results, leading to an average increase of 0.82% in Avg. DSC. When applying CFT to the bottom two layers, the overall

TABLE IV
COMPONENT-WISE ABLATION STUDY OF M²FTRANS ON BRATS2020

Components				Avg. DSC (%)				Complexity	
Reg	MMFT	CFT	SWA	WT	TC	ET	Avg.	Params	GFLOPs
×	×	×	×	85.37	75.67	59.78	73.61	29.00	197.818
✓	×	×	×	86.16	77.68	61.85	75.23	+0.00	+0.000
✓	✓	×	×	87.15	78.50	64.24	76.63	+10.44	+7.851
✓	✓	✓	×	87.38	78.63	65.53	77.18	+13.87	+11.750
✓	✓	×	✓	87.45	78.95	64.55	76.98	+10.44	+7.856
✓	✓	✓	✓	**87.86**	**79.20**	**66.77**	**77.94**	+13.87	+11.754

Model complexity during inference is evaluated by Params (*i.e.*, the number of parameters measured in millions) and GFLOPS (*i.e.*, the number of floating-point operations per second measured in billions). For comparison, the increase in model complexity of each component combination against the baseline is reported separately.

TABLE V
ABLATION STUDY ON WHERE TO USE CFT IN M²FTRANS ON BRATS2020

Stage	SWA	Avg. DSC (%)				Complexity	
		WT	TC	ET	Avg.	Params	GFLOPs
0	×	87.15	78.50	64.24	76.63	39.44	205.670
0	✓	87.45	78.95	64.55	76.98	+0.00	+0.005
1	✓	87.62	78.88	65.86	77.45	+2.63	+1.151
2	✓	**87.86**	**79.20**	**66.77**	**77.94**	+3.43	+3.903
3	✓	87.60	78.99	65.39	77.33	+3.69	+11.241
4	✓	87.39	78.54	66.11	77.35	+3.79	+33.253
5	✓	87.32	78.72	66.14	77.39	+3.83	+106.621

Stage indicates how many stages/skip connections are combined with CFT (from bottom to top skip connections of the encoders in fig. 2). model complexity during inference is evaluated by params (*i.e.*, the number of parameters measured in millions) and GFLOPS (*i.e.*, the number of floating-point operations per second measured in billions). For comparison, the increase in model complexity of each setting against the baseline is reported separately.
The best segmentation results under each evaluation metric are marked in bold.

TABLE VI
ABLATION STUDY ON THE NUMBER OF MMFT LAYERS (I.E., L_1) IN M²FTRANS ON BRATS2020

L_1	Avg. DSC (%)				Complexity	
	WT	TC	ET	Avg.	Params	GFLOPs
0	86.16	77.68	61.85	75.23	29.00	197.818
1	87.69	79.11	65.35	77.38	+9.14	+5.851
2	87.77	**79.20**	66.34	77.77	+11.50	+8.803
3	**87.86**	**79.20**	**66.77**	**77.94**	+13.87	+11.754

Model complexity during inference is evaluated by Params (*i.e.*, the number of parameters measured in millions) and GFLOPS (*i.e.*, the number of floating-point operations per second measured in billions). For comparison, the increase in model complexity of each setting against the baseline is reported separately
The best segmentation results under each evaluation metric are marked in bold.

TABLE VII
ABLATION STUDY ON THE NUMBER OF CFT LAYERS (I.E., L_2) IN M²FTRANS ON BRATS2020

L_2	Avg. DSC (%)				Complexity	
	WT	TC	ET	Avg.	Params	GFLOPs
0	87.45	78.95	64.55	76.98	39.44	205.674
1	87.58	79.05	65.93	77.52	-0.01	+0.076
2	**87.86**	79.20	**66.77**	**77.94**	+3.43	+3.899
3	87.68	79.20	66.10	77.66	+6.86	+7.721
4	87.64	**79.24**	65.65	77.51	+10.30	+11.544

Model complexity during inference is evaluated by Params (*i.e.*, the number of parameters measured in millions) and GFLOPS (*i.e.*, the number of floating-point operations per second measured in billions). For comparison, the increase in model complexity of each setting against the baseline is reported separately
The best segmentation results under each evaluation metric are marked in bold.

segmentation performance is further improved. However, progressively adding CFT to more skip connection layers would degrade the segmentation performance. It is because the features from the upper layers are less discriminative, making it difficult to re-weight feature maps along the channel dimension which in turn affects the fusion features. In terms of model complexity, SWA would barely bring additional computation burdens. When introducing CFT to more stages, model parameters would increase relatively slower while the increase of GFLOPs is much faster as summarized in Table V. This is because computation on higher-resolution feature maps is far more complicated and will increase in a non-linear manner when introducing CFT to more stages.

B. On Hyper-Parameters of M² FTrans

To further evaluate the performance of modality-masked fusion transformers (MMFT), we conduct additional ablation studies by using different numbers of MMFT for comparison on the BraTS2020 dataset as summarized in Table VI. Introducing a one-layer MMFT effectively outperforms the baseline with a large margin, leading to an average increase of 2.15% in Avg. DSC. One interesting observation is that introducing more MMFT layers would not necessarily bring significant performance gains. It is because training transformers is data-intensive and stacking more MMFT layers may not be helpful as deep layers can be uniform as discussed in [34]. In terms of model complexity, it is a natural observation that using more MMFT layers would introduce additional model parameters and GFLOPs as summarized in Table VI. It should be noted that compared to the baseline networks, model complexity brought by MMFT layers is relatively limited.

Another important hyper-parameter is the number of channel-wise fusion transformers, L_2. To figure out how it matters, we conduct ablation studies on L_2 as summarized in Table VII. As discussed in Section III.D, the main motivation of CFT is to reduce additional redundancy along the channel dimension by re-weighting feature maps. Therefore, gradually introducing more CFT layers is helpful to learn more compact modality-specific features for fusion. However, using too many CFT layers is harmful. It is because adding more CFT layers will continuously re-weight feature maps along the channel dimension and is more likely to lose spatial information.

Similar to MMFT layers, adopting more CFT layers would gradually increase both model complexity and GFLOPs as summarized in Table VII. One interesting observation is that adopting a one-layer CFT module is even more lightweight than the baseline. It is because we followed the use of skip connections in mmFormer for the stages without CFT. The decrease in Params indicates that using a one-layer CFT module at the bottom stage needs fewer parameters than mmFormer's convolutional skip connection.

TABLE VIII
QUANTITATIVE COMPARISON WITH DEDICATED APPROACHES ON BRATS2018 MEASURED IN AVG. DSC (%)

Type	Flair	○	○	○	●	○	○	●	○	●	●	●	●	●	○	●	Avg.
	T1	○	○	●	○	○	●	●	●	○	○	●	●	○	●	●	
	T1c	○	●	○	○	●	●	○	○	○	●	●	○	●	●	●	
	T2	●	○	○	○	●	○	○	●	●	○	○	●	●	●	●	
WT	ACN	85.4	79.8	**78.7**	87.3	84.9	79.6	86.0	84.4	86.9	87.8	88.4	87.4	87.2	86.6	89.1	85.3
	SMU-Net	**85.7**	**80.3**	78.6	87.5	86.1	80.3	87.3	85.6	87.9	88.4	88.2	88.3	88.2	86.5	88.9	85.9
	M²FTrans	83.5	78.9	78.0	**87.7**	**86.3**	**82.5**	**88.9**	**86.0**	**88.7**	**89.4**	**89.5**	**89.2**	**89.6**	**87.1**	**89.6**	**86.3**
TC	ACN	66.8	83.3	70.9	66.4	83.2	83.9	70.4	72.8	70.7	82.9	83.3	67.7	82.9	83.2	84.8	76.8
	SMU-Net	67.2	84.1	69.5	**71.8**	85.0	84.4	71.2	73.5	71.2	84.1	84.2	67.9	82.5	84.4	87.3	77.9
	M²FTrans	**70.8**	**87.7**	**71.0**	70.8	**88.2**	**88.4**	**75.4**	**74.6**	**73.7**	**88.3**	**88.5**	**75.8**	**88.2**	**88.6**	**88.4**	**81.3**
ET	ACN	41.7	78.0	41.8	42.2	74.9	75.3	42.5	46.5	44.3	77.5	75.1	42.8	73.8	75.9	78.2	60.7
	SMU-Net	43.1	78.3	**42.8**	**46.1**	75.7	75.1	**44.0**	47.7	46.0	77.3	76.2	43.1	75.4	76.2	79.3	61.8
	M²FTrans	**47.5**	**79.0**	39.7	32.8	**79.4**	**79.8**	41.0	**48.6**	**48.4**	**79.2**	**79.5**	**49.1**	**79.4**	**79.6**	**79.4**	**62.8**

M 2 ftrans adopts an input size of 128×128×128 pixels while ACN [21] and SMU-Net [23] use the same input size of 160×192×128 pixels. The best results under each modality setting are marked in bold.

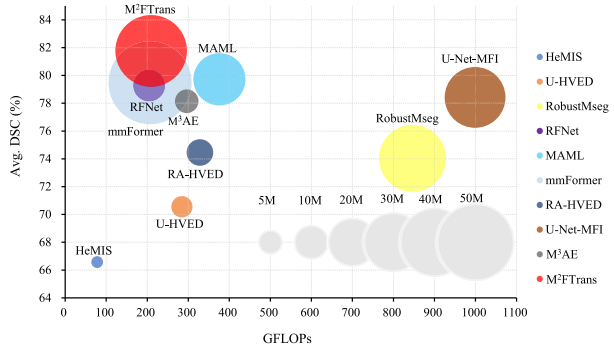
C. Vs. Dedicated Approaches

In Section IV, we mainly focus on the state-of-the-art single-model-based approaches for comparison, especially against the most-advanced approach RFNet [30]. As discussed in Section II, the most representative KD-based approaches, ACN [21] and SMU-Net [23], trained a separate model for each modality setting (i.e., 15 models in total) and achieved good segmentation performance. For a more comprehensive evaluation, we further compare M^{2}FTrans against ACN and SMU-Net. It should be noted that ACN and SMU-Net adopt a much larger input size of $160 \times 192 \times 128$ pixels while M^{2}FTrans uses a smaller input size of $80 \times 80 \times 80$ pixels in Section IV (the same as RFNet). For a fair comparison, we increase the input size of M^{2}FTrans to $128 \times 128 \times 128$ pixels (limited by GPU memories) and conduct a comparison against ACN and SMU-Net. As both data splits and evaluation methods of ACN and SMU-Net are different from those in Section IV.A, we strictly follow the data splits and source code of SMU-Net to re-run M^{2}FTrans for comparison as summarized in Table VIII. It should be noted that the quantitative results of both ACN and SMU-Net are from [23].

Despite using a smaller input size (i.e. $128 \times 128 \times 128$ vs. $160 \times 192 \times 128$ pixels), M^{2}FTrans stably outperforms both ACN and SMU-Net. Specifically, M^{2}FTrans achieves better segmentation performance under 39 out of 45 modality settings compared to ACN and under 38 out of 45 modality settings against SMU-Net, leading to an average increase of 1.0%, 4.5%, and 2.1% for ACN and 0.4%, 3.4%, and 1.0% for SMU-Net in the segmentation of WT, TC, and ET respectively. It should be noted that we can expect more performance improvements given an even larger input size for M^{2}FTrans. The above comparison results further validate the effectiveness of M^{2}FTrans in exploiting richer cross-modality features and being robust to various incomplete multi-modality scenarios.

D. On Model Efficiency

Model efficiency of different approaches is visualized in Fig. 6. Compared to the best transformer-based method mmFormer, M^{2}FTrans achieves better performance with fewer model parameters and similar GFLOPs. Compared to the best CNN-based approach MAML, M^{2}FTrans is of more model parameters but lower GFLOPs, due to the inherent property of transformers involving more model parameters. It should be noted that model efficiency is not the main focus of M^{2}FTrans and can be further improved through lightweight designs.



Method	HeMIS	U-HVED	RobustMSeg	RFNet	MAML	mmFormer	RA-HVED	U-Net-MFI	M³AE	M²FTrans
Params	1.17	3.79	37.58	8.40	22.71	57.61	5.89	30.91	4.70	42.87
GFLOPs	77.88	284.37	846.54	204.57	375.92	206.83	328.43	999.04	296.11	209.57

Fig. 6. Comparison of different approaches on model efficiency. The size of each circle indicates the number of model parameters (i.e., Params) measured in millions. GFLOPs is the abbreviation of floating-point operations per second measured in billions.

E. Cross-Training Evaluation

For a more comprehensive evaluation on model generalizability, we have conducted cross-training validation by directly applying models trained on BraTS2018 to BraTS2020. As BraTS2020 is a superset of BraTS2018, all BraTS2018 samples in BraTS2020 are excluded for generalizability evaluation and performance comparison as summarized in Table IX. Compared to all comparison approaches, M^{2}FTrans achieves the best overall performance across all tumor types. More importantly, compared to previous comparisons conducted on each dataset separately, the performance gap between M^{2}FTrans and other approaches becomes more significant on model generalizability, demonstrating the robustness of M^{2}FTrans on unseen data.

F. Vs. More SOTA Approaches on BraTS2021

For a more comprehensive evaluation, we have conducted additional experiments on the latest BraTS2021 dataset and introduced more SOTA approaches for comparison, including MAML [42], RA-HVED [44], U-Net-MFI [43], and M^{3}AE [45] as summarized in Table X. Among comparison methods, RFNet achieves the best performance on the segmentation of WT while MAML outperforms others on the segmentation of TC and ET. In terms of shape preservation measured in HD as summarized in Table X, M^{3}AE outperforms other comparison approaches.

TABLE IX
CROSS VALIDATION OF DIFFERENT APPROACHES WHEN APPLYING THE MODELS TRAINED ON BRATS2018 DIRECTLY TO BRATS2020

Type	Method	DSC (%)	p-value	HD (mm)	p-value
WT	HeMIS	81.89	<0.001	23.55	<0.001
	U-HVED	83.12	<0.001	20.01	<0.001
	RobustMSeg	86.43	<0.001	10.04	<0.001
	RFNet	88.63	<0.001	5.90	<0.001
	mmFormer	87.93	<0.001	5.33	<0.001
	M²FTrans	**90.00**	-	**4.31**	-
TC	HeMIS	71.03	<0.001	25.58	<0.001
	U-HVED	73.38	<0.001	19.47	<0.001
	RobustMSeg	78.12	<0.001	9.87	<0.001
	RFNet	83.05	<0.001	5.73	<0.001
	mmFormer	82.64	<0.001	6.30	<0.001
	M²FTrans	**84.82**	-	**4.67**	-
ET	HeMIS	59.67	<0.001	16.08	<0.001
	U-HVED	62.15	<0.001	10.53	<0.001
	RobustMSeg	66.54	<0.001	6.89	<0.001
	RFNet	72.13	<0.001	5.10	0.018
	mmFormer	71.15	<0.001	5.22	0.002
	M²FTrans	**73.61**	-	**4.28**	-

The best segmentation results under each evaluation metric are marked in bold.

TABLE X
QUANTITATIVE COMPARISON RESULTS MEASURED IN DSC (%) AND HD (MM) ON BRATS2021

Type	Method	DSC (%)	p-value	HD (mm)	p-value
WT	HeMIS	78.75	<0.001	24.62	<0.001
	U-HVED	80.75	<0.001	24.54	<0.001
	RobustMSeg	83.64	<0.001	22.17	<0.001
	RFNet	87.16	<0.001	9.64	<0.001
	MAML	86.51	<0.001	11.63	<0.001
	mmFormer	87.13	<0.001	7.51	<0.001
	RA-HVED	84.33	<0.001	9.06	<0.001
	U-Net-MFI	86.01	<0.001	8.86	<0.001
	M³AE	86.40	<0.001	7.14	<0.001
	M²FTrans	**88.33**	-	**6.37**	-
TC	HeMIS	66.05	<0.001	23.62	<0.001
	U-HVED	70.50	<0.001	24.62	<0.001
	RobustMSeg	74.25	<0.001	12.54	<0.001
	RFNet	80.56	<0.001	7.50	<0.001
	MAML	81.49	<0.001	7.29	<0.001
	mmFormer	80.90	<0.001	6.84	<0.001
	RA-HVED	74.69	<0.001	9.30	<0.001
	U-Net-MFI	79.18	<0.001	8.13	<0.001
	M³AE	79.75	<0.001	6.62	<0.001
	M²FTrans	**83.20**	-	**5.38**	-
ET	HeMIS	54.93	<0.001	17.61	<0.001
	U-HVED	60.41	<0.001	18.03	<0.001
	RobustMSeg	64.23	<0.001	10.07	<0.001
	RFNet	70.13	<0.001	6.18	<0.001
	MAML	71.16	<0.001	6.27	<0.001
	mmFormer	70.40	<0.001	5.78	<0.001
	RA-HVED	64.34	<0.001	7.56	<0.001
	U-Net-MFI	70.07	<0.001	6.41	<0.001
	M³AE	68.27	<0.001	5.71	<0.001
	M²FTrans	**73.76**	-	**4.88**	-

The best and second-best results are marked in bold and underlined.

Comparatively, M²FTrans consistently achieves superior performance under all modality settings for all tumor types, outperforming both RFNet and MAML. In addition, statistical analysis in Table X validates the effectiveness of M²FTrans on performance improvement.

G. Limitations and Future Work

Based on the quantitative results on BraTS2018, BraTS2020, and BraTS2021, we find that performance improvements of M²FTrans are more significant given fewer modalities. On the one hand, it validates the effectiveness of M²FTrans on rebalancing modalities, especially for non-dominant modalities. On the other hand, it indicates that M²FTrans may affect the feature extraction of dominant modalities and in turn affect the performance given more modalities. In addition, as discussed in Section V.F, model efficiency is not a major concern of M²FTrans, which shall be carefully studied in our future work.

According to the design of M²FTrans, we believe it is widely extendible to a more general feature fusion task in both complete and incomplete multi-modality scenarios. For instance, M²FTrans can be applied to text-image learning tasks by replacing one encoder for text inputs. One main challenge is how to balance the feature/token dimensions across different types of modalities, which will be explored based on M²FTrans in our future work.

VI. CONCLUSION

In this article, we present a novel framework named M²FTrans for incomplete multi-modality brain tumor segmentation. For cross-modality feature fusion, a modality-masked fusion transformer is designed to explore long-range dependency across modalities while minimizing the negative influence of missing modalities via masked self-attention. To reduce redundancy in modality-specific features, in addition to the regularization from a shared decoder, both spatial weight attention and channel-wise fusion transformer are proposed to re-weight each modality and its intra-modality features/tokens. In this way, the weights of modalities are more balanced, making M²FTrans more robust to missing/incomplete modalities. Extensive experiments on widely-used datasets demonstrate the superiority of M²FTrans against the state-of-the-art approaches under various incomplete multi-modality settings.

ACKNOWLEDGMENT

The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

REFERENCES

[1] M. Havaei et al., "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.
[2] F. Isensee, P. F. Jaeger, S. AA. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.
[3] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. MICCAI*, 2021, pp. 109–119.
[4] D. Zhang et al., "Exploring task structure for brain tumor segmentation from multi-modality MR images," *IEEE Trans. Image Process.*, vol. 29, pp. 9032–9043, 2020.
[5] M. J. Graves and D. G. Mitchell, "Body MRI artifacts in clinical practice: A physicist's and radiologist's perspective," *J. Magn. Reson. Imag.*, vol. 38, no. 2, pp. 269–287, 2013.
[6] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[8] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1290–1299.
[9] Y. Zhang et al., "mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 107–117.
[10] S. Chen, C. Ding, and M. Liu, "Dual-force convolutional neural networks for accurate brain tumor segmentation," *Pattern Recognit.*, vol. 88, no. 90–100, 2019.

[11] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. 19th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 424–432.

[12] L. Fidon et al., "Scalable multimodal convolutional networks for brain tumour segmentation," in *Proc. 20th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2017, pp. 285–293.

[13] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu, "TBraTS: Trusted brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 503–513.

[14] A. Hatamizadeh et al., "UNETR: Transformers for 3d medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 574–584.

[15] H. -Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnFormer: Volumetric medical image segmentation via a 3D transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 4036–4045, 2023.

[16] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3 D tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 162–172.

[17] Z. Xing, L. Yu, L. Wan, T. Han, and L. Zhu, "NestedFormer: Nested modality-aware transformer for brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 140–150.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[19] S. Vadacchino, R. Mehta, N. M. Sepahvand, B. Nichyporuk, J. J. Clark, and T. Arbel, "Had-Net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images," in *Proc. Med. Imag. Deep Learn.*, 2021, pp. 787–801.

[20] Q. Wang, L. Zhan, P. Thompson, and J. Zhou, "Multimodal learning with incomplete modalities by knowledge distillation," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 1828–1838.

[21] Y. Wang et al., "ACN: Adversarial co-training network for brain tumor segmentation with missing modalities," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 410–420.

[22] Q. Yang, X. Guo, Z. Chen, P. Y. M. Woo, and Y. Yuan, "D^2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2953–2964, Oct. 2022.

[23] R. Azad, N. Khosravi, and D. Merhof, "SMU-Net: Style matching U-net for brain tumor segmentation with missing modalities," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 48–62.

[24] S. Jeong, H. Cho, J. Kwon, and H. Park, "Region-of-interest attentive heteromodal variational encoder-decoder for segmentation with missing modalities," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 132–148.

[25] A. Sharma and G. Hamarneh, "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1170–1183, Apr. 2020.

[26] B. Yu, L. Zhou, L. Wang, J. Fripp, and P. Bourgeat, "3D cGAN based cross-modality MR image synthesis for brain tumor segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, 2018, pp. 626–630.

[27] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Heteromodal image segmentation," in *Proc. 19th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 469–477.

[28] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren, "Heteromodal variational encoder-decoder for joint modality completion and segmentation," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 74–82.

[29] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P. -A. Heng, "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 447–456.

[30] Y. Ding, X. Yu, and Y. Yang, "RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3975–3984.

[31] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, 2019, pp. 2915–2921.

[32] A. Recasens et al., "Zorro: The masked multimodal transformer," 2023, *arXiv:2301.09595*.

[33] Y. Liu, H. Wang, Z. Chen, K. Huangliang, and H. Zhang, "TransUNet : Redesigning the skip connection to enhance features in medical image segmentation," *Knowl. Based Syst.*, vol. 256, 2022, Art. no. 109859.

[34] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 32–42.

[35] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.

[37] Y. Liu, F. Mu, Y. Shi, and X. Chen, "SF-Net: A multi-task model for brain tumor segmentation in multimodal MRI via image fusion," *IEEE Signal Process. Lett.*, vol. 29, pp. 1799–1803, 2022.

[38] Z. Zhu et al., "Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI," *Inf. Fusion*, vol. 91, pp. 376–387, 2023.

[39] Q. D. Pham et al., "Segtransvae: Hybrid CNN-transformer with regularization for medical image segmentation," in *Proc. IEEE 19th Int. Symp. Biomed. Imag.*, 2022, pp. 1–5.

[40] K. Zou et al., "Tbrats: Trusted brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 503–513.

[41] J. Zhao et al., "Uncertainty-Aware Multi-Dimensional Mutual Learning for Brain and Brain Tumor Segmentation," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 9, pp. 4362–4372, Sep. 2023.

[42] Y. Zhang et al., "Modality-aware mutual learning for multi-modal medical image segmentation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 589–599.

[43] Z. Zhao, H. Yang, and J. Sun, "Modality-adaptive feature interaction for brain tumor segmentation with missing modalities," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2022, pp. 183–192.

[44] S. Jeong, H. Cho, J. Kwon, and H. Park, "Region-of-interest attentive heteromodal variational encoder-decoder for segmentation with missing modalities," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 3707–3723.

[45] H. Liu et al., "M3AE: Multimodal representation learning for brain tumor segmentation with missing modalities," in *Proc. AAAI*, 2023, pp. 1657–1665.