Bridging the Gap in Missing Modalities: Leveraging Knowledge Distillation and Style Matching for Brain Tumor Segmentation

Shenghao Zhu[1⋆], Yifei Chen[2⋆], Weihong Chen[1], Yuanhan Wang[1], Chang Liu[1], Shuo Jiang[1], Feiwei Qin[1⋆⋆], and Changmiao Wang[3⋆⋆]

[1] Hangzhou Dianzi University, Hangzhou, China
[2] Tsinghua University, Beijing, China
[3] Shenzhen Research Institute of Big Data, Shenzhen, China
qinfeiwei@hdu.edu.cn,cmwangalbert@gmail.com

Abstract. Accurate and reliable brain tumor segmentation, particularly when dealing with missing modalities, remains a critical challenge in medical image analysis. Previous studies have not fully resolved the challenges of tumor boundary segmentation insensitivity and feature transfer in the absence of key imaging modalities. In this study, we introduce MST-KDNet, aimed at addressing these critical issues. Our model features Multi-Scale Transformer Knowledge Distillation to effectively capture attention weights at various resolutions, Dual-Mode Logit Distillation to improve the transfer of knowledge, and a Global Style Matching Module that integrates feature matching with adversarial learning. Comprehensive experiments conducted on the BraTS and FeTS 2024 datasets demonstrate that MST-KDNet surpasses current leading methods in both Dice and HD95 scores, particularly in conditions with substantial modality loss. Our approach shows exceptional robustness and generalization potential, making it a promising candidate for real-world clinical applications. Our source code is available at https://github.com/Quanato607/MST-KDNet.

Keywords: Missing Modalities · Knowledge Distillation · Style Matching · Neuroglioma · Brain Tumor Segmentation · Multi-modality MRI.

1 Introduction

Brain tumor segmentation is a critical task in medical neuroimaging, playing a vital role in diagnosis, treatment planning, and prognosis assessment. Among brain tumors, gliomas stand out as particularly aggressive, exhibiting complex biological behaviors and significant heterogeneity, which complicate clinical treatment [19]. Magnetic resonance imaging (MRI), utilizing multiple modalities, is the preferred method for visualizing and segmenting brain tumors due to its ability

⋆ Co-first author.
⋆⋆ Corresponding author.

to capture diverse and complementary information [27,20,3]. Each MRI modality contributes unique insights: T1 and T2 modalities are effective in identifying angioedema in subacute strokes; T1Gd highlights vascular structures and the blood-brain barrier; and FLAIR provides a broad overview of stroke lesion characteristics. Together, these modalities complement one another, offering detailed information about tumor size, location, and morphology [26].

Missing modalities, caused by scan corruption, imaging artifacts, and varying machine settings, are a common challenge in clinical settings [22,11,6]. Furthermore, in clinical practice, the number of available modalities and their correct labeling for algorithmic use is uncertain, limiting segmentation accuracy [15]. Several methods have been proposed to handle missing modalities, such as Ding *et al.*'s [5] region-aware fusion module and Zhao *et al.*'s [25] modality-adaptive feature interaction. Zhang *et al.* [23] and Dai *et al.* [4] introduced techniques to reconstruct missing modalities by aggregating features from available ones. Whereas, the absence of critical modalities often leads to performance degradation [17]. To tackle this, Wang *et al.* [17] proposed a cross-modal knowledge distillation framework to identify key modalities. In contrast, Liu *et al.* [10] developed M3AE using self-distillation to handle missing-modality. Other approaches, such as Huo *et al.*'s [8] bi-directional distillation and Xing *et al.*'s [21] contrastive distillation, have proved the efficacy of knowledge distillation in improving model performance despite missing modalities. However, past methods remain inadequate in handling modality inconsistency and correlation due to limited knowledge adaptability and the inability to align features between modalities.

To address the difficulty of understanding the inter-modal correlation and inconsistency in the model, this paper proposes **MST-KDNet**, which significantly improves cross-modal semantic extraction, maintaining superior segmentation accuracy and robust generalization capabilities across varying modality combinations. The main contributions of this work are summarized as follows: 1) We propose the Multi-Scale Transformer Knowledge Distillation (MS-TKD), which enhances segmentation performance by extracting attention weights and features across different resolutions; 2) We present the Dual-Modal Logit Distillation (DMLD), which leverages Logit Alignment and Normalized Kullback-Leibler Distillation to enhance knowledge adaptation and ensure robust learning even in the absence of certain modalities; 3) We develop the Global Style Matching Module (GSME) to combine feature matching with adversarial learning to strengthen the model's performance across missing modality scenarios.

2 Proposed Method

2.1 Baseline Network

The baseline network structure of the model is based on 3D convolution and multi-scale Transformer architecture [7]. First, the input $x \in \mathbb{R}^{H\times W\times D\times C}$ undergoes one layer of 3D convolution to extract the initial features and then reduces the feature map size layer-by-layer by three layers of 3D convolutional

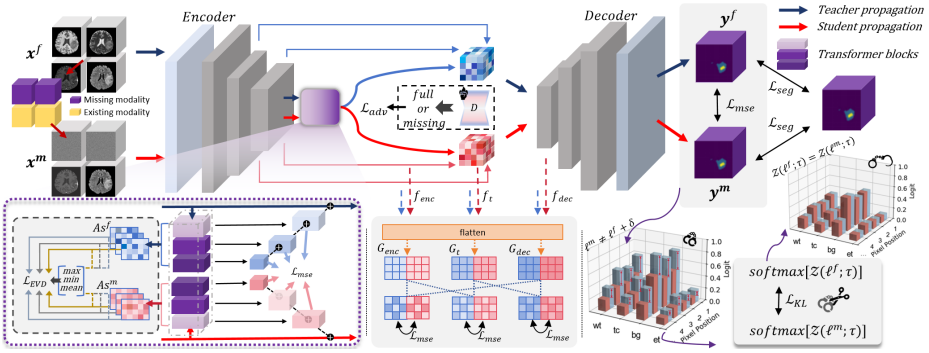

Fig. 1. The overall framework of MST-KDNet. The teacher propagation processes all available modalities, while the student propagation accommodates incomplete inputs.

downsampling with a step size of 2, to extract the rich spatial and semantic information in the external encoder stage. After downsampling, the feature map size is $x \in \mathbb{R}^{(H/8)\times(W/8)\times(D/8)\times 32C}$ and is converted to a 1D sequence. The volume is divided into non-overlapping (P, P, P) chunks to obtain $x_v \in \mathbb{R}^{(N\times(P^3\cdot 32C))}$, with a sequence length of $N = \frac{(H/8)\cdot(W/8)\cdot(D/8)}{P^3}$. These chunks are projected through a linear layer into the K dimensional embedding space and enter the Transformer processing. The Transformer architecture contains multiple Transformer blocks using the Multihead Self-Attention (MSA) mechanism. Each MSA sublayer has n parallel self-attention heads, and the attention weight A is computed by querying (q) and key (k) similarity with the following formula:

$$A = \text{softmax}\left(\frac{q\,k^{\top}}{\sqrt{K_h}}\right); \text{SA}(z) = A\ v; \text{MSA}(\mathbf{z}) = [\text{SA}_1(\mathbf{z}); ...; \text{SA}_n(\mathbf{z})]\,W, \quad (1)$$

where $K_h = \frac{K}{n}$ serves as a scaling factor to keep the number of parameters consistent across key dimensions. Here, v denotes the value mapping in the sequence z, and $W \in \mathbb{R}^{(n\cdot K_h)\times K}$ denotes the trainable parameter weights of the multi-head self-attention sublayer. At different resolution stages, multiple z_i representations are extracted, sized as $\frac{H\times W\times D}{P^3}\times K$ and reshaped as $\frac{H}{P}\times\frac{W}{P}\times\frac{D}{P}\times K$ tensor after $3\times 3\times 3$ convolution and normalization. Thereafter, the feature maps are combined with the feature maps of the corresponding Transformer blocks through jump connections. The inverse convolution extends the size. Finally, the external decoder extracts the deep semantic information through 3D convolution and integrates the multi-scale jump-connected features, and up-sampling gradually restores the spatial resolution.

2.2 Multi-scale Transformer Knowledge Distillation

As illustrated in Fig. 1, we extract the attention weights A from each resolution layer as a sequence for the Extreme Value Distillation (EVD) process. We calculate the maximum, minimum, and mean values of these attention weights at

each pixel position along the C dimension:

$$A_{\max} = \max(As, \text{axis} = 1); A_{\min} = \min(As, \text{axis} = 1); A_{\text{mean}} = \frac{1}{n}\sum_{i=1}^{n} A_i. \quad (2)$$

These values are then used as weights to generate three sequences by multiplying them with the corresponding attention weights:

$$EV_1 = A_{\max} \cdot A; EV_2 = A_{\min} \cdot A; EV_3 = A_{\text{mean}} \cdot A. \quad (3)$$

For the complete modality teacher model and the student model with missing modalities, we apply knowledge distillation using mean square error (MSE) loss. Additionally, MSE is applied to each reshaped tensor, ensuring consistency between the models. The multiscale Transformer-based knowledge distillation loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{MS-TKD}} &= \alpha\mathcal{L}_{\text{EVD}}(EV^f, EV^m) + \beta\mathcal{L}_{\text{MST}} \\ &= \alpha\frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{n}\left(EV_{i,j}^f - EV_{i,j}^m\right)^2 + \beta\frac{1}{N}\sum_{i=1}^{3}\sum_{j=1}^{N}(z_{i,j}^f - z_{i,j}^m)^2. \end{aligned} \quad (4)$$

2.3 Dual-mode Logit Distillation

Logit Discrepancy Distillation. To enhance the student's ability to mimic the teacher model's style, we apply MSE loss to align features from the complete modality with those of the missing modality. The specific formula for this calculation is as follows:

$$\mathcal{L}_{\text{mse}}(l^f, l^m) = \frac{1}{n}\sum_{j=1}^{n}\left(l_j^f - l_j^m\right)^2, \quad (5)$$

where l^f and l^m represent the logit outputs from the teacher and student models, respectively, with N indicating the dimension of the logit vector.

Logit Standardization KL Distillation. Traditional knowledge distillation applies a global temperature factor to align the logit ranges of both student and teacher networks. This rigid coupling limits the student's ability to adapt, especially when there is a significant disparity in model capacity, reducing the student's learning potential. To address this limitation, we introduce logit normalization into the distillation process [14]. Specifically, the logits (l) are first normalized using the $\mathcal{Z}$-score normalization function before being passed through the softmax function:

$$\mathcal{Z}(l;\tau) = \frac{1-\mu}{(\sigma + 10^{-7})\tau}; q(l) = \text{softmax}[\mathcal{Z}(l;\tau)] \quad \text{where} \quad l \in \{l^f, l^m\}, \quad (6)$$

where μ represents the mean, σ denotes the standard deviation, and τ is the temperature coefficient. The normalized logits are then passed into the KL divergence loss function, defined as:

$$\mathcal{L}_{\text{KL}}(l^f || l^m) = \sum_{k=1}^{K} q(l^f)\log\left(\frac{q(l^f)}{q(l^m)}\right). \quad (7)$$

Logit Standardization KL Distillation removes the need for a globally shared temperature, allowing for flexible adjustments, and preserves the core distribution relationship between the teacher's and student's logits without rigidly matching the teacher's output magnitude. Thus, the Dual-Mode Logit Distillation loss function combines Logit Discrepancy loss and Logit Standardization KL loss, expressed as:

$$\mathcal{L}_{\text{logit}} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{KD}}\tau^2\mathcal{L}_{\text{KL}}. \tag{8}$$

2.4 Global Style Matching Module

The structural and stylistic variations inherent to different MRI modalities often pose challenges for decoding networks, particularly when some modalities are missing. To address this issue, our GSME integrates Mean Square Error MSE loss with adversarial learning to mitigate these challenges. In GSME, we first take the max-pooled feature output from the penultimate convolutional layer, f_{enc}, concatenate it with the output of the transformer block, f_{t}, and feed the combined features into the decoder. The decoder processes these features and outputs f_{dec}, while the fused features, $f_{\text{enc\&t}}$, are simultaneously passed into a feature discriminator D to compute the adversarial loss:

$$\mathcal{L}_{\text{adv}} = \log(1 - D(f^f_{\text{enc\&t}})) + \log(D(f^m_{\text{enc\&t}})), \tag{9}$$

where $f_{\text{enc\&t}}$ represents the input to the decoder. The decoder's first convolutional output, f_{dec}, is further processed by reshaping the spatial dimensions H, W, and D of $f_{\text{enc}}, f_{\text{t}}$, and f_{dec} into two-dimensional tensors $G_{\text{enc}}, G_{\text{t}}, G_{\text{dec}}$. These tensors are then subjected to a feature fusion operation:

$$M_1 = G_{\text{enc}}G^T_{\text{dec}}; M_2 = G_{\text{enc}}G^T_{\text{t}}; M_3 = G_{\text{dec}}G^T_{\text{t}}. \tag{10}$$

The fusion results in the feature sequence $M \in \{M_1, M_2, M_3\}$, which is used to compute the GSME loss, incorporating both adversarial and MSE losses:

$$\mathcal{L}_{\text{GSME}} = \epsilon\mathcal{L}_{\text{adv}} + \frac{\theta}{4n^2}\sum_{i=1}^{3}\sum_{j}^{n}\left(M^f_{i,j} - M^m_{i,j}\right)^2. \tag{11}$$

2.5 Total Loss

The total loss function utilized during training integrates four key components, each contributing to different aspects of the model's optimization:

$$\mathcal{L}_{\text{joint}} = \lambda_1\mathcal{L}_{\text{MS-TKD}} + \lambda_2\mathcal{L}_{\text{logit}} + \lambda_3\mathcal{L}_{\text{GSME}} + \lambda_4\mathcal{L}_{\text{Dice}}, \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are weighting coefficients that balance the contributions of each individual loss term to the overall objective.

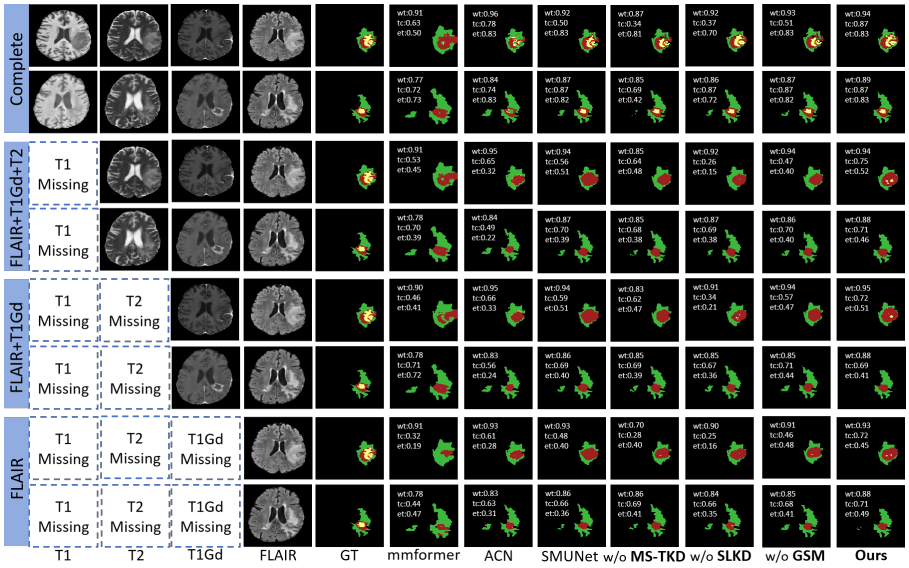

Fig. 2. Comparison of segmentation results under four missing-modality scenarios. Color legend: WT = red + yellow + green, TC = red + yellow, ET = red.

3 Experiments

3.1 Dataset and Implementation

BraTS 2024 dataset [16] and FeTS 2024 dataset [12]. Both datasets offer 3D multimodal MR brain images accompanied by corresponding accurate labels. The BraTS 2024 comprises 1,350 cases, and the FeTS 2024 comprises 1,250 cases, with each case including MR images in four modalities: T1, T2, T1Gd, and FLAIR. The images are classified into three distinct regions: Enhancing Tumor (ET) region, Core Tumor (CT) region, and Whole Tumor (WT) region. We randomly divided the dataset into two parts for our analysis, using 80% for training and the remaining 20% for testing. We applied data augmentation techniques, such as random flips, rotations, and cropping and each MRI was resized to dimensions of $160 \times 192 \times 128$ to standardize the data.

Implementation details. All experiments were conducted on a Tesla V100 GPU using PyTorch 2.4.1 as the foundational framework. The batch size for training was set to 1, and the model was trained for 250 epochs. Model parameters were optimized using the Adam optimization algorithm, with an initial learning rate of 0.0001. The hyperparameters β_1 and β_2 were configured to 0.9 and 0.99, respectively, to ensure stable and efficient convergence.

3.2 Experiment Results

Comparative Experiments. As shown in Tables 1 and 2, we performed multimodal ablation experiments on the BraTS 2024 dataset. Input modes were ran-

domly excluded to simulate various scenarios, resulting in unimodal, bimodal, trimodal, and quaternary input, for a total of 16 combinations. Compared with several state-of-the-art models, MST-KDNet achieved optimal or suboptimal results in key tumor regions (WT, TC, and ET), indicating superior voxel overlap and spatial distance performance. Notably, MST-KDNet maintains robust performance even under extreme conditions, such as when only one or two modalities were available, showing strong robustness. MST-KDNet also performs impressively on the FeTS 2024 dataset, as shown in Table 3. MST-KDNet consistently outperforms the other methods in most cases of missing modal combinations.

Table 1. Comparison of Dice score for various state-of-the-art models (BraTS 2024). Red represents the optimal value, and Blue represents the suboptimal value.

Type	FLAIR / T1 / T1Gd / T2	○ ○ ○ ●	○ ○ ● ○	○ ● ○ ○	● ○ ○ ○	○ ○ ● ●	○ ● ○ ●	● ○ ○ ●	○ ● ● ○	● ○ ● ○	● ● ○ ○	● ● ○ ●	● ● ● ○	○ ○ ● ●	● ● ○ ●	● ● ● ●	Avg.
WT	RA-HVED [9]	75.4	51.3	9.5	71.4	77.5	53.4	72.9	76.1	80.1	72.9	72.9	80.6	80.4	77.7	80.1	68.8
	RMBTS [2]	70.1	51.2	51.8	65.0	75.3	60.6	76.4	75.0	77.3	76.0	79.3	79.7	80.3	76.1	80.9	71.7
	mmformer [24]	72.6	55.5	61.3	72.7	74.3	65.4	79.2	75.1	79.6	78.3	80.0	80.7	81.0	75.6	81.3	74.2
	M2FTrans [13]	72.5	58.8	62.0	73.0	73.9	64.2	77.4	73.6	78.9	77.0	77.3	78.5	79.5	74.2	78.8	73.3
	ACN [18]	69.6	58.7	60.1	80.7	71.8	63.6	82.1	72.2	82.3	81.3	82.0	82.8	82.0	72.5	82.5	75.0
	SMUNet [1]	75.0	67.9	69.6	84.2	76.7	70.6	84.6	77.1	85.2	85.2	85.4	85.6	86.0	77.2	86.0	79.7
	MST-KDNet	77.2	72.9	73.5	84.7	79.8	75.1	85.7	79.3	85.8	86.4	86.5	86.1	86.9	80.0	86.8	81.8
TC	RA-HVED [9]	26.5	54.2	9.4	41.1	61.3	54.8	41.9	29.2	40.5	61.9	62.5	43.2	64.0	61.9	65.0	47.8
	RMBTS [2]	10.9	36.5	12.6	11.2	40.4	37.6	16.8	15.2	14.5	38.9	40.1	17.4	40.4	40.9	40.6	27.6
	mmformer [24]	47.2	52.3	44.4	33.1	62.6	60.6	49.6	51.1	49.6	60.6	64.3	52.6	65.5	65.3	67.0	55.1
	M2FTrans [13]	46.6	53.3	43.3	33.8	60.0	57.7	46.7	48.5	48.3	57.8	60.0	49.6	61.5	60.8	62.0	52.7
	ACN [18]	21.2	54.2	19.5	22.5	58.8	57.9	26.1	23.2	26.7	60.0	63.8	28.3	62.6	62.7	64.1	43.4
	SMUNet [1]	29.3	64.1	28.2	28.8	67.3	67.1	32.6	31.5	32.5	66.9	70.4	33.7	69.4	69.1	69.8	50.7
	MST-KDNet	47.3	68.3	44.5	33.9	70.3	71.3	50.1	41.5	50.2	72.0	74.1	53.6	72.5	72.6	73.1	59.5
ET	RA-HVED [9]	35.8	37.8	9.24	39.8	42.3	36.6	42.6	43.8	44.4	44.1	43.9	48.4	46.8	40.7	45.9	40.1
	RMBTS [2]	7.9	37.8	10.0	8.2	41.9	40.1	13.1	11.8	10.8	40.6	43.5	14.0	42.3	44.1	55.2	28.1
	mmformer [24]	44.9	50.5	42.3	31.4	61.3	59.0	45.3	49.4	46.6	59.3	63.0	49.6	63.6	64.2	65.7	53.1
	M2FTrans [13]	47.1	54.2	44.6	34.0	62.6	60.0	47.5	49.4	49.3	60.2	62.7	50.4	64.5	63.4	65.0	54.3
	ACN [18]	18.0	55.2	16.9	19.6	59.8	59.6	22.2	19.2	22.4	60.8	65.1	23.9	64.0	64.3	65.9	42.5
	SMUNet [1]	25.5	64.8	25.0	25.1	67.9	68.1	28.6	27.6	28.6	67.9	70.6	29.7	69.8	70.1	70.8	49.3
	MST-KDNet	48.3	68.6	32.0	40.6	70.0	72.3	48.5	50.1	51.1	72.4	74.9	52.5	72.8	73.1	73.9	59.8

Ablation Study. As shown in Table 4, on the BraTS 2024 dataset, excluding MS-TKD decreased Dice scores by 2.0% for WT, 5.1% for TC, and 5.6% for ET, while increasing HD95, highlighting the importance of multiscale attentional alignment. Removing GSME reduced Dice scores by 3.5% for WT, 3.4% for TC, and 6.4% for ET, emphasizing the role of global style and texture compensation. The absence of SLKD caused Dice score drops of 1.8% for WT, 3.4% for TC, and 4.6% for ET, indicating the importance of flexible teacher-student distribution matching. A similar pattern was observed in the FeTS 2024 dataset. These results indicate that MS-TKD, GSME, and SLKD have their respective and complementary roles, which significantly improve the model's segmentation accuracy and stability under missing mode conditions.

Table 2. Comparison of HD95 score for various state-of-the-art models (BraTS 2024). Red represents the optimal value, and Blue represents the suboptimal value.

Type	FLAIR / T1 / T1Gd / T2															Avg.
WT	RA-HVED [9]	22.1	40.2	57.7	23.8	19.8	34.8	20.9	17.4	16.9	21.2	20.5	15.0	16.3	18.6	15.9	24.1
	RMBTS [2]	39.1	63.6	57.7	59.4	36.1	50.1	41.7	33.1	37.4	47.8	34.8	33.2	35.3	34.1	34.0	42.5
	mmformer [24]	19.5	52.0	40.7	18.2	18.8	34.5	13.9	16.8	13.1	15.5	13.4	12.9	12.2	16.8	11.8	20.7
	M2FTrans [13]	43.8	51.8	47.0	47.3	42.4	44.5	43.0	42.6	42.1	41.9	41.3	41.3	40.7	40.8	40.5	43.4
	ACN [18]	11.6	28.4	29.6	11.8	13.5	20.4	11.4	15.6	10.3	13.2	11.7	10.2	11.5	15.1	10.3	15.0
	SMUNet [1]	9.1	13.3	5.9	12.2	5.9	7.6	11.2	5.4	7.7	5.1	5.2	5.3	4.9	4.8	8.0	7.4
	MST-KDNet	8.1	11.1	11.0	6.7	5.3	9.2	6.1	5.2	4.6	6.2	5.1	5.0	4.7	4.7	5.3	6.6
TC	RA-HVED [9]	25.3	30.4	57.1	22.5	15.8	26.8	20.9	23.1	19.7	15.9	14.4	21.6	13.3	16.2	12.5	22.4
	RMBTS [2]	24.8	23.1	47.1	24.1	19.8	25.8	23.7	21.9	19.1	18.5	16.3	20.0	15.6	14.0	13.7	21.8
	mmformer [24]	27.7	62.1	39.1	24.3	25.6	38.7	19.7	24.1	19.3	20.5	17.3	18.7	15.4	22.1	14.7	26.0
	M2FTrans [13]	79.4	79.2	82.6	82.4	76.3	76.3	79.7	79.2	79.5	78.5	77.5	78.3	77.0	77.0	76.3	78.6
	ACN [18]	15.7	9.2	19.3	18.2	6.4	8.5	17.3	17.0	15.7	6.6	6.2	17.6	5.8	6.2	5.8	11.7
	SMUNet [1]	14.0	6.3	14.0	13.4	4.4	5.0	12.2	12.1	12.0	4.8	4.3	11.9	4.2	4.5	4.6	8.5
	MST-KDNet	12.0	4.9	11.2	12.1	3.7	4.3	10.5	10.8	11.0	3.6	3.4	10.0	3.7	3.3	4.0	7.2
ET	RA-HVED [9]	12.9	25.0	47.0	15.2	14.9	23.7	13.2	10.9	10.8	14.0	14.2	11.0	12.8	15.4	12.2	16.9
	RMBTS [2]	23.8	21.9	44.8	23.7	19.2	24.2	22.4	21.9	19.5	17.2	15.1	19.5	15.2	13.5	13.3	21.0
	mmformer [24]	26.4	59.8	37.6	23.2	24.0	36.7	18.6	22.2	18.4	18.3	16.4	17.7	14.5	20.4	14.0	24.5
	M2FTrans [13]	23.4	31.5	21.5	24.1	16.1	16.2	16.2	19.4	20.9	16.8	13.3	18.5	15.3	14.2	13.9	18.8
	ACN [18]	14.7	8.0	19.3	18.1	6.1	7.6	16.6	16.4	14.9	5.9	5.3	17.2	5.2	5.3	5.2	11.1
	SMUNet [1]	13.5	5.4	14.0	13.0	3.9	4.3	11.8	11.5	12.0	4.1	3.7	11.3	3.7	4.0	4.0	8.0
	MST-KDNet	11.7	4.5	10.5	11.9	3.3	3.8	9.8	10.3	10.6	3.2	3.0	9.8	3.3	2.9	3.0	6.8

Table 3. Comparison of average Dice and HD95 scores for various state-of-the-art models (FeTS 2024). Due to space limits, full 16 results are in our GitHub.

Method	Average Dice Score (%)			Average HD95 Score (mm)		
	WT	TC	ET	WT	TC	ET
RA-HVED [9]	69.7	60.0	50.9	22.0	20.6	19.8
RMBTS [2]	75.2	60.4	65.6	8.6	25.2	19.1
mmformer [24]	68.9	54.6	48.6	26.7	27.5	34.0
M2FTrans [13]	82.0	74.3	63.0	26.5	14.8	20.8
ACN [18]	84.9	78.8	67.3	8.5	8.4	16.5
SMUNet [1]	87.5	82.9	72.1	6.4	6.3	5.5
MST-KDNet	88.4	84.3	73.4	5.9	5.7	5.4

4 Conclusion

In this study, we propose MST-KDNet, a novel framework for incomplete multi-modality brain tumor segmentation. MST-KDNet effectively captures cross-modality correlations and significantly enhances tumor region representations for robust segmentation, even with significant missing modalities. The framework employs global and local feature refinement mechanisms to align available modalities, effectively compensating for the missing ones and improving feature distribution. Extensive experiments on the BraTS and FeTS 2024 benchmarks

Table 4. Comparison of average Dice and HD95 scores for ablation studys.

Method	Average Dice Score (%)			Average HD95 Score (mm)		
	WT	TC	ET	WT	TC	ET
BraTS 2024 [16]						
w/o MS-TKD	79.8	54.4	54.2	7.5	8.3	7.8
w/o GSME	78.3	55.1	53.4	9.6	9.7	9.5
w/o SLKD	80.0	56.1	55.2	8.1	8.7	8.0
MST-KDNet	81.8	59.5	59.8	6.6	7.2	6.8
FeTS 2024 [12]						
w/o MS-TKD	87.0	81.8	72.6	7.3	6.8	5.5
w/o GSME	86.1	82.9	72.6	7.3	6.6	5.9
w/o SLKD	87.5	82.1	72.9	6.5	6.6	5.8
MST-KDNet	88.2	84.3	73.4	5.9	5.7	5.4

demonstrate MST-KDNet's superiority and robustness, consistently outperforming state-of-the-art methods, especially in incomplete modality settings.

Acknowledgements This work was supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK259909299001-006), Anhui Provincial Joint Construction Key Laboratory of Intelligent Education Equipment and Technology (No. IEET202401), the Guangxi Key R&D Project (No. AB24010167), the Project (No. 20232ABC03A25), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617), Shenzhen Medical Research Fund (No. C2401036), and Hospital University United Fund of The Second Affiliated Hospital, School of Medicine, The Chinese University of Hong Kong, Shenzhen (No. HUUF-MS-202303).

Disclosure of Interests The authors declare no competing interests.

References

1. Azad, R., Khosravi, N., Merhof, D.: Smu-net: Style matching u-net for brain tumor segmentation with missing modalities. In: International Conference on Medical Imaging with Deep Learning. pp. 48–62. PMLR (2022)
2. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 447–456. Springer (2019)
3. Chen, Y., Zhu, S., Fang, Z., Liu, C., Zou, B., Qiu, L., Wang, Y., Chang, S., Jia, F., Qin, F., Fan, J., Peng, Y., Wang, C.: Toward robust early detection of alzheimer's disease via an integrated multimodal learning approach. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025)
4. Dai, Q., Wei, D., Liu, H., Sun, J., Wang, L., Zheng, Y.: Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1445–1453 (2024)
5. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3975–3984 (2021)
6. Ge, R., Yu, X., Chen, Y., Zhou, G., Jia, F., Zhu, S., Jia, J., Zhang, C., Sun, Y., Zeng, D., et al.: Tc-kanrecon: High-quality and accelerated mri reconstruction via adaptive kan mechanisms and intelligent feature scaling. arXiv preprint arXiv:2408.05705 (2024)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
8. Huo, F., Xu, W., Guo, J., Wang, H., Guo, S.: C2kd: Bridging the modality gap for cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16006–16015 (2024)
9. Jeong, S., Cho, H., Kwon, J., Park, H.: Region-of-interest attentive heteromodal variational encoder-decoder for segmentation with missing modalities. In: Proceedings of the Asian Conference on Computer Vision. pp. 3707–3723 (2022)
10. Liu, H., Wei, D., Lu, D., Sun, J., Wang, L., Zheng, Y.: M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1657–1665 (2023)
11. Liu, M., Jiao, Y., Lu, J., Chen, H.: Anomaly detection for medical images using teacher-student model with skip connections and multi-scale anomaly consistency. IEEE Transactions on Instrumentation and Measurement (2024)
12. Pati, S., Baid, U., Zenk, M., Edwards, B., Sheller, M., Reina, G.A., Foley, P., Gruzdev, A., Martin, J., Albarqouni, S., et al.: The federated tumor segmentation (fets) challenge. arXiv preprint arXiv:2105.05874 (2021)
13. Shi, J., Yu, L., Cheng, Q., Yang, X., Cheng, K.T., Yan, Z.: M2ftrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. IEEE Journal of Biomedical and Health Informatics (2023)
14. Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15731–15740 (2024)

15. Varsavsky, T., Eaton-Rosen, Z., Sudre, C.H., Nachev, P., Cardoso, M.J.: Pimms: permutation invariant multi-modal segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 201–209. Springer (2018)
16. de Verdier, M.C., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N.H., Foltyn-Dumitru, M., Zhang, J., Alafif, M., Baig, S., et al.: The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. arXiv preprint arXiv:2405.18368 (2024)
17. Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., Carneiro, G.: Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 216–226. Springer (2023)
18. Wang, Y., Zhang, Y., Liu, Y., Lin, Z., Tian, J., Zhong, C., Shi, Z., Fan, J., He, Z.: Acn: adversarial co-training network for brain tumor segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 410–420. Springer (2021)
19. Weller, M., Wick, W., Aldape, K., Brada, M., Berger, M., Pfister, S.M., Nishikawa, R., Rosenthal, M., Wen, P.Y., Stupp, R., et al.: Glioma. Nature Reviews Disease Primers **1**(1), 1–18 (2015)
20. Wu, C., Chen, Y., Du, Y., Zong, J., Dong, J., Liu, M., Peng, Y., Fan, J., Qin, F., Wang, C.: Towards practical alzheimer's disease diagnosis: A lightweight and interpretable spiking neural model. arXiv preprint arXiv:2506.09695 (2025)
21. Xing, X., Zhu, M., Chen, Z., Yuan, Y.: Comprehensive learning and adaptive teaching: Distilling multi-modal knowledge for pathological glioma grading. Medical Image Analysis **91**, 102990 (2024)
22. Zhang, C., Chen, Y., Fan, Z., Huang, Y., Weng, W., Ge, R., Zeng, D., Wang, C.: Tc-diffrecon: Texture coordination mri reconstruction method based on diffusion model and modified mf-unet method. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2024)
23. Zhang, J., Zhang, S., Shen, X., Lukasiewicz, T., Xu, Z.: Multi-condos: Multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 76–95 (2023)
24. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 107–117. Springer (2022)
25. Zhao, Z., Yang, H., Sun, J.: Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 183–192. Springer (2022)
26. Zhu, S., Chen, Y., Jiang, S., Chen, W., Liu, C., Wang, Y., Chen, X., Ke, Y., Qin, F., Wang, C., Zhu, Z.: Xlstm-hved: Cross-modal brain tumor segmentation and mri reconstruction method using vision xlstm and heteromodal variational encoder-decoder. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2025)
27. Zhu, Z., Wang, Z., Qi, G., Mazur, N., Yang, P., Liu, Y.: Brain tumor segmentation in mri with multi-modality spatial information enhancement and boundary shape correction. Pattern Recognition **153**, 110553 (2024)