Computers in Biology and Medicine 163 (2023) 107234

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Joint learning-based feature reconstruction and enhanced network for incomplete multi-modal brain tumor segmentation

Yueqin Diao, Fan Li *, Zhiyuan Li

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
Yunnan Key Laboratory of Artificial Intelligence, Kunming 650500, China

ARTICLE INFO

Keywords:
Incomplete multimodal learning
Brain tumor segmentation
Joint learning
Feature reconstruction
Feature enhancement

ABSTRACT

Multimodal Magnetic Resonance Imaging (MRI) can provide valuable complementary information and substantially enhance the performance of brain tumor segmentation. However, it is common for certain modalities to be absent or missing during clinical diagnosis, which can significantly impair segmentation techniques that rely on complete modalities. Current advanced methods attempt to address this challenge by developing shared feature representations via modal fusion to handle different missing modality situations. Considering the importance of missing modality information in multimodal segmentation, this paper utilize a feature reconstruction method to recover the missing information, and proposes a joint learning-based feature reconstruction and enhancement method for incomplete modality brain tumor segmentation. The method leverages an information learning mechanism to transfer information from the complete modality to a single modality, enabling it to obtain complete brain tumor information, even without the support of other modalities. Additionally, the method incorporates a module for reconstructing missing modality features, which recovers fused features of the absent modality through utilizing the abundant potential information obtained from the available modalities. Furthermore, the feature enhancement mechanism improves shared feature representation by utilizing the information obtained from the missing modalities that have been reconstructed. These processes enable the method to obtain more comprehensive information regarding brain tumors in various missing modality circumstances, thereby enhancing the model's robustness. The performance of the proposed model was evaluated on BraTS datasets and compared with other deep learning algorithms using Dice similarity scores. On the BraTS2018 dataset, the proposed algorithm achieved a Dice similarity score of 86.28%, 77.02%, and 59.64% for whole tumors, tumor cores, and enhanced tumors, respectively. These results demonstrate the superiority of our framework over state-of-the-art methods in missing modalities situations.

1. Introduction

Brain tumors are serious diseases that could pose a significant risk to human health. Accurate segmentation of brain tumors has positive implications for clinical assessment and treatment. Magnetic resonance imaging (MRI) is a widely used technique for clinical brain tumor detection due to its ability to clearly show different areas of soft tissue lesions with minimal invasiveness. Common MRI used for brain tumor imaging include Fluid Attenuated Inversion Recovery (FLAIR), T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), and T2-weighted (T2). Each modality sequence provides complementary information for analyzing different regions of the tumor. Fig. 1 shows how different modality sequences highlight different regions of the brain tumor. FLAIR highlights the whole tumor, while T1ce highlights the tumor core. Therefore, leveraging the complementarity of multimodal images enhances the reliability of information acquired using a single modality and improves precision in clinical diagnosis and segmentation. Traditionally, clinicians manually delineate lesion regions based on clinical experience, which is not only time-consuming but also prone to errors. Therefore, there is an urgent need to promote the development of automated brain tumor image segmentation techniques to improve both the accuracy of clinical diagnosis and its efficiency.

Several brain tumor segmentation methods have been proposed in recent years [1–8], yielding encouraging results in reducing labor costs and improving efficiency. Examples of several convolutional neural networks (CNNs) including Unet [2], Nested Unet [3], and Attention Unet [4] have achieved satisfactory results in medical image segmentation by utilizing convolution and pooling operations. These CNNs make full use of the inherent properties of convolution, using skip-join and continuous upsampling techniques to integrate low-level,

* Corresponding author at: Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China.
E-mail addresses: diaoyueqin@stu.kust.edu.cn (Y. Diao), lifan198686@163.com (F. Li), lizy@stu.kust.edu.cn (Z. Li).

https://doi.org/10.1016/j.compbiomed.2023.107234
Received 6 April 2023; Received in revised form 12 June 2023; Accepted 1 July 2023
Available online 4 July 2023

fine-grained features in the encoder with high-level, coarse-grained multi-layered semantic features in the decoder, thereby enhancing the overall accuracy of segmentation. Although CNNs have remarkable feature extraction capabilities, their limited receptive fields in convolutional operations hinder the explicit establishment of long-range dependencies in the global feature space, thus limiting the network's ability to capture anatomical features with global contextual semantic information. Local and global features play a crucial role in intensive prediction tasks, particularly in 3D medical image segmentation. Transformers employ self-attention mechanisms to model global information, which have been successful in natural language processing [9] and computer vision [10]. As a result, several recent studies have applied Transformers to the medical image domain [11–16] to explore their potential in improving segmentation performance. For instance, TransBTS [7] explored how to effectively combine 3D-CNN and Transformer for MRI brain tumor segmentation. Nestedformer [16] introduced a Transformer encoder embedded with modality awareness and a CNN decoder to segment tumor images.

While the methods discussed employ rich complementary information from different modalities for complete brain tumor segmentation, real clinical practice can introduce incomplete or missing data due to the complexity of the diagnostic environment. This incompleteness can negatively impact the segmentation performance. Thus, there is a need to improve model robustness to accommodate missing modalities. Currently, methods to tackle missing modalities in brain tumor images can be grouped into three categories: (a) training methods based on different cases of missing modalities, in which models are trained for each missing modality case, resulting in good segmentation but are time-consuming and resource-intensive [17–20]; (b) methods based on missing modality generation, in which missing modalities are synthesized and using the obtained complete modalities for segmentation. However, this method involves an additional generative network training, with the quality of the results directly impacting the segmentation quality [21–25]; (c) modality-based fusion approaches, which involve fusing available modalities into a latent space to learn a shared feature representation that is used for brain tumor segmentation [26–34]. Among these approaches, the third approach is more efficient since it only requires learning an end-to-end network, avoiding the effect of the quality of the generated modality on the segmentation accuracy. However, due to the lack of effective supervision, the method does not always result in a complete shared feature representation, leading to less accurate segmentation results. In this paper, we propose a feature reconstruction and enhancement method based on joint learning to address the challenges of incomplete information in brain tumor images. The method guides a single modality to learn the rich information from a complete modality through joint learning, alleviating the issue of incomplete information. Additionally, we reconstruct missing modality fusion features using the shared feature representation of brain tumors, further enhancing feature discrimination capability.

The method proposed in this paper consists of three key stages: Uni-Modal information Learning (UML), Missing Modality Feature Reconstruction (MMFR), and Shared Feature Representation Enhancement (SFRE). In the UML phase, we address the problem of dependency among modalities in feature extraction stage by using complete modality information to guide each modality. This approach helps to alleviate the issue of relying on other modalities due to the limitations of information carried by a single modality. In the MMFR stage, we fuse and project available modality information into the shared feature representation space to reconstruct the missing modality features, thus recovering the information of the missing modality. Furthermore, to enhance the discriminative features of brain tumors, the reconstructed features are fully fused with the shared features in the SFRE stage, leading to a more comprehensive representation of brain tumor information. Moreover, a shared decoder is employed to align the potential features from two pathways, which further enhances the shared feature representation of brain tumors. This approach effectively addresses the issue of incomplete information caused by missing modalities, thereby improving the robustness of the model in clinical practice. The main contributions of this paper can be summarized as follows:

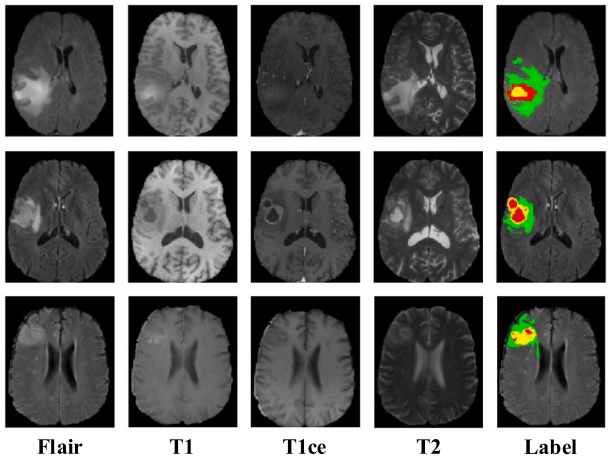

Fig. 1. Four modalities and corresponding labels for three cases. Red areas on the label denote the necrotic and non-enhancing tumor core (NCR/NET); yellow areas denote the GD-enhancing tumor (ET); green areas denote the peritumoral edema (ED).

1. The proposed method uses a joint learning strategy in which the complete modality supervises the missing modality path and guides each modality to extract complete brain tumor information. This approach mitigates the need to rely on other modalities to compensate for incomplete information, thus improving segmentation accuracy.

2. To address the issue of missing modality information, we implemented a mechanism that interacts with available modal brain tumor features to improve the features reconstruction of missing modalities, facilitating information recovery.

3. We proposed a feature enhancement method to address incomplete modality brain tumor segmentation. This method leverages the reconstructed missing modality information to improve the shared feature representation of brain tumors, resulting in more comprehensive brain tumor features.

2. Related work

2.1. Incomplete modality brain tumor segmentation

In recent years, multimodal learning has garnered significant interest and has been applied to various computer vision tasks [35–37]. Complementary information between multiple modalities is essential for improving the performance of various computer vision tasks. An essential aspect of multimodal learning is exploring effective methods for multimodal fusion. However, in most real-world scenarios, certain modalities may not be available or missing, which has led to the development of incomplete modality learning research. Our work aims to address the challenging problem of brain tumor segmentation with missing modalities. This problem is more practical but significantly more difficult than previous brain tumor segmentation tasks.

To address the issue of missing modalities, previous work proposed the use of fusion mechanisms to aggregate available modality features. For example, U-HeMIS [27] proposed a method that computes the mean and variance of available modality features, while Shen Y [38] used multiple encoders to extract features for different modalities, and then

fused the feature maps using skip connections. RFNet [29] used a tumor region aware module to adaptively aggregate multimodal features from different regions to establish the relationship between modality and tumor. However, these methods utilize only the fused features of available modalities for segmentation, ignoring the rich information contained in the missing modalities. In contrast, CoCa-GAN [22] and MGM-GAN [25]generate missing modalities using generative adversarial networks (GAN). Lee [39] and Shen L [40] recover missing modality images through domain translation while at the same time learning the shared features across modalities for multimodal segmentation. However, these approaches require complex computations and an additional generative network to be trained, with the generated quality affecting segmentation performance. More recently, knowledge distillation-based networks have shown promising performance in dealing with missing modality problems. For example, ACN [17] proposed an adversarial knowledge distillation network to align the potential representations of complete and missing modalities. However, this approach does not consider reconstructing texture or style information in missing modalities. To address this limitation, SMU-Net [18] decomposes the representation space into content and style representations, and uses a matching module to reconstruct missing information. However, each of these methods requires a model to be trained for each missing modality situation, which is a process that is both resource-intensive and time-consuming. This study proposes a novel incomplete modality brain tumor segmentation method that leverages missing modality feature reconstruction for recovering missing information. Our approach utilizes available modality fusion information to reconstruct missing modality features, rendering the recovery of missing information feasible. Unlike prior methods, our approach is both efficient and less resource-intensive because it does not necessitate training specific models for individual cases.

2.2. Shared feature representation learning

In the field of multimodal learning, shared feature representation has been extensively studied [41–44]. The goal is to learn invariant representations between modalities by jointly projecting multiple modality information into a shared subspace. Hazarika et al. [41] proposed a shared subspace to learn potential commonalities between modalities and reduce the impact of modal gaps. Liu et al. [44] introduced a discrete shared space to capture fine-grained representations and improve the accuracy of cross-modal retrieval. Prior research has demonstrated the efficacy of modality-invariant features in bridging inter-modal gaps. Consequently, learning shared feature representations across multiple modalities can ameliorate the impact of missing modalities.

The HVED [28] method uses independent encoders to extract first-order moment features from each modality and models a Gaussian distribution over the encoded features to form a shared representation space. Although this space provides a shared representation for all modalities, it is limited in its ability to model mode-dependent features, and may not perform well when multiple modes are missing. RobustSeg [34] decomposes features into mode-invariant and mode-specific features and uses mode-invariance to generate segmentation results based on a gating strategy. Yang et al. [31] proposes a modality disentanglement and a tumor-region disentanglement to capture the correlation between modality and tumor region and extract sufficient information for segmentation. Zhang et al. [32] applied Transformers to the incomplete modality brain tumor segmentation task, constructing intra-modal and inter-modal encoders to establish and align global correlations between different modalities, and extracting modality-invariant representations. Our proposed method differs from existing approaches,as it utilizes complete modality latent features for supervising the learning of shared feature representations. In addition, it employs an information interaction mechanism to facilitate the transfer and fusion of information between available modalities. Moreover, the shared feature representation of brain tumors is improved by leveraging reconstructed information from missing modalities.

3. Proposed method

3.1. Overview

To overcome the challenges posed by missing brain tumor modalities in clinical practice, this paper proposes a joint learning-based feature reconstruction and enhancement method, which consists of three phases as illustrated in Fig. 2: unimodal information learning, missing modality feature reconstruction, and brain tumor shared feature representation enhancement. In UML stage, intact modality is used as supervision to guide each modality in learning the brain tumor information present in the intact modality. In the MMFR stage, we use the shared feature representation to reconstruct the features of the missing modality and recover its information. The SFRE stage utilizes the reconstructed missing modality features to enhance the shared feature representation of brain tumors to obtain more comprehensive tumor information. Furthermore, to ensure that the shared feature representation aligns with the full modality, both the missing modality and complete modality paths use a shared convolutional decoder. This improves the feature representation capability and ensures consistency in the shared feature representation.

3.2. Unimodal information learning

To achieve accurate segmentation results in multimodal brain tumor segmentation, it is essential to leverage the complementary information from different modalities. However, the performance of segmentation models may suffer when multiple modalities are damaged or missing, which presents a challenge to accurate segmentation. To tackle this issue, we propose a joint learning approach with two independent learning paths: the first learning path uses all available modalities as input, while the second learning path uses the modalities that are intact. The goal is to transfer the rich feature information from the complete modality path to the missing modality path, while also encouraging the missing modality path to reconstruct the missing information. Different from previous distillation methods, the proposed approach does not require training a separate model for each missing modality situation, and uses complete modality information to guide each modality to alleviate the inter-modality dependency issue in feature extraction stage. The complete modality path and missing modality path use the same encoding and decoding structure. In the encoding stage, we utilize convolutional operations to extract shallow details and local information, as well as self-attention to establish long-range dependencies, which improves the ability to model global context while retaining low-level details. We define the full modality as $M = \{Flair, T1ce, T1, T2\}$, and define the input data for each modality as $\boldsymbol{X}_m = \{X_i, i \in M\}$. The input data for the full modality path is defined as $\boldsymbol{X}_f = \{\forall \boldsymbol{X}_i, i \in M\}$, and in our setting, $\boldsymbol{X}_i \in \mathbb{R}^{1\times H\times W\times D}$ is a 3D modality image. In the unimodal information learning stage, our method uses the full modality information to supervise the feature extraction of each modality, guiding each modality to learn the tumor information from the complete modality, achieving the transfer of rich semantics from the full modality to the unimodal modality.

To extract local detailed features within each modality, both paths utilize an encoder composed of multiple convolutional blocks in the feature extraction stage. We define the modality-specific encoder and the full modality encoder as E_m and E_f, respectively. Therefore, the feature maps at different levels for both the single modality and the full modality can be represented as:

$$\boldsymbol{F}_m^l = E_m\left(\boldsymbol{X}_m; \theta_m^{conv}\right) \tag{1}$$

$$\boldsymbol{F}_f^l = E_f\left(\boldsymbol{X}_f; \theta_f^{conv}\right) \tag{2}$$

Where $\boldsymbol{F}_f^l \in \mathbb{R}^{C\times H\times W\times D}$ and $\boldsymbol{F}_m^l \in \mathbb{R}^{C\times H\times W\times D}$ are the feature maps at different levels of the complete modality and each single modality, l

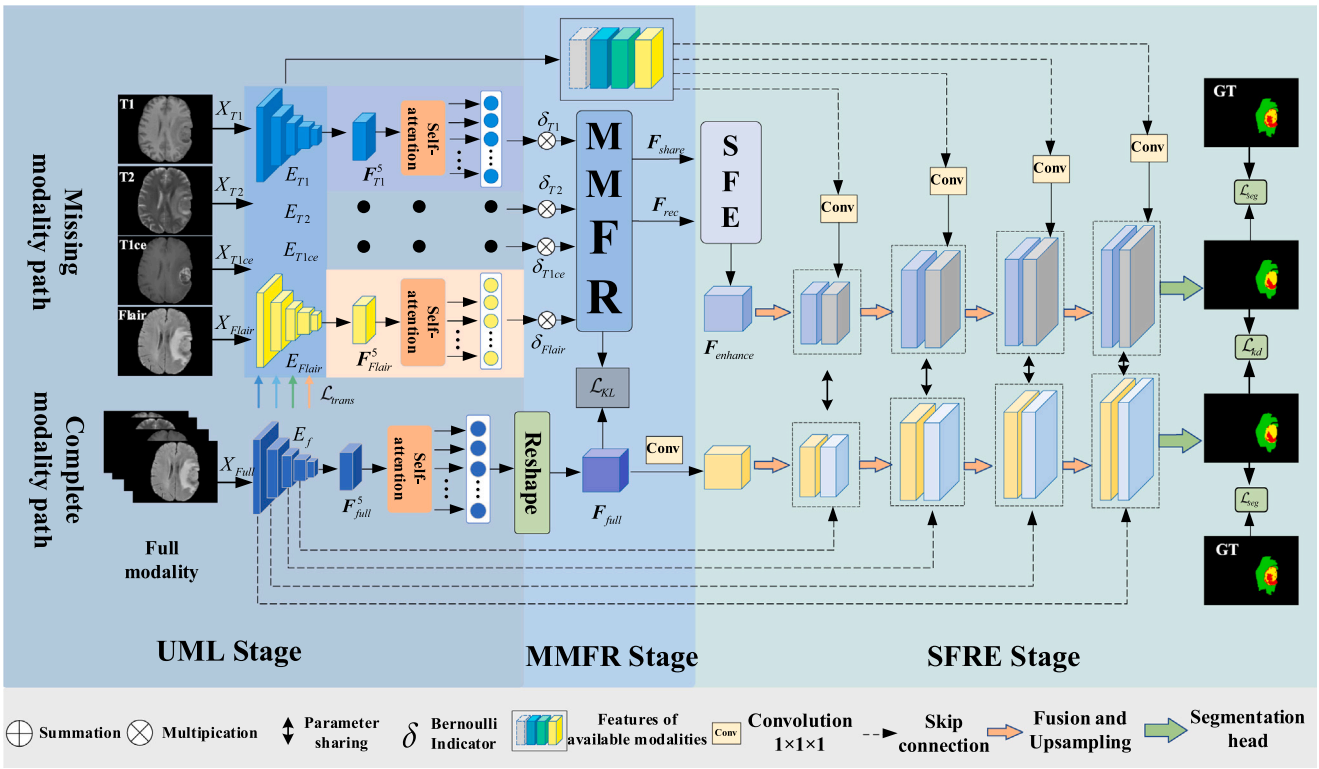

Fig. 2. The overview of our proposed network architecture, consisting of a complete modal path and a missing modal path. E_f and E_m $(m \in Flair, T1ce, T1, T2)$ extract full modality image features and m modality image features respectively. The two paths utilize a shared decoder.

denotes the lth feature extraction block and there are five in total, θ is the parameter of the encoder, H and W are the height and width of the input image, and D is the number of slices. Specifically, our encoder is comprised of five feature extraction blocks, each consisting of cascaded group normalization, LeakyRelu, and convolutional layers with a kernel size of $3 \times 3 \times 3$. Between two connected blocks, the feature maps are downsampled using convolutional layers with a stride of 2. The number of channels in each stage is 8, 16, 32, 64, and 128, respectively.

To address the issue of incomplete information from single modality, we utilize the complete modality to supervise the feature extraction of each modality, enabling the transfer of rich multimodal information to the single modality. Our method guides the feature extractor of each modality to learn the complete tumor features of the multimodalities by using L1 parameterization and KL divergence. This process is defined as:

$$\mathcal{L}_{trans} = \frac{1}{4}\sum_{i=4}^{l}\left(\left\|F_f^i - F_m^i\right\|_1 + \mathcal{L}_{kl}\left(F_f^i, F_m^i\right)\right) \tag{3}$$

Where $i = 4$ represents the fourth feature extraction block, and F^i represents the output of the ith feature extraction block. Due to the inherent limitations of convolutional neural networks, convolutional operations cannot effectively capture global contextual information. Therefore, we use self-attention operations to capture global contextual information. The self-attention module can be formulated as follows:

$$\boldsymbol{q}_m = \boldsymbol{F}_m^5 w_{mq}, \boldsymbol{k}_m = \boldsymbol{F}_m^5 w_{mk}, \boldsymbol{v}_m = \boldsymbol{F}_m^5 w_{mv} \tag{4}$$

$$\boldsymbol{F}_m^{global} = \boldsymbol{F}_m^5 + Softmax\left(\frac{\boldsymbol{q}_m \boldsymbol{k}_m^T}{\sqrt{d_{mk}}}\right)\boldsymbol{v}_m \tag{5}$$

Where $\boldsymbol{F}_m^5$ denotes the output of the modality encoder at the 5-th feature extraction block, and $\boldsymbol{F}_m^{global}$ denotes the output with global relevance established through self-attention. w_{mi} $(i = [q, k, v])$ denotes the parameter matrix of a linear projection, and $\boldsymbol{q}_m, \boldsymbol{k}_m, \boldsymbol{v}_m$ respectively represent query, key, and value.

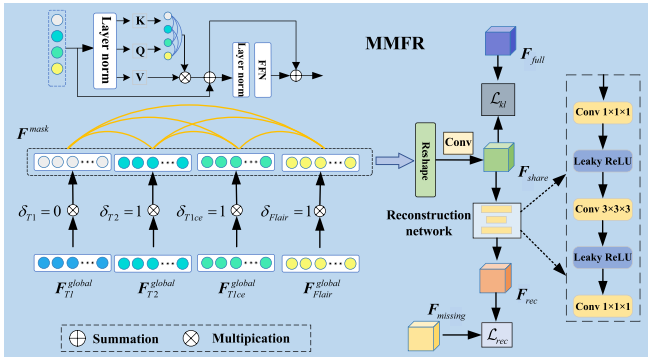

Fig. 3. Missing modality feature reconstruction module.

3.3. Missing modality feature reconstruction

Previous methods for handling missing modalities often involve reconstructing complete modalities and then fusing the information of the complete modalities for segmentation. As a result, the final segmentation performance is significantly influenced on the quality of the reconstructed modalities. As shown in Fig. 3, unlike previous modality reconstruction methods, the current study uses feature reconstruction to reconstruct the fused features of the missing modality by utilizing the shared feature representation of the potential space to enhance the shared feature representation. This approach not only circumvents the impact of reconstructed modality quality but also eliminates the need to train a modal generation network. As there is a strong correlation between different tumor modalities, this study proposes an interactive fusion method that leverages the available correlation information among modal tumor features to enhance shared feature representation.

In the interactive fusion of available modality information, we denote the current modality as T_c and the other modalities except

the current modality as $T_o \in \{Flair, T1ce, T1, T2\}$. We also denote the information flowing to the current modalities as $T_c \leftarrow T_o$. In this process, we use the principle of cross-attention to achieve the information interaction between different modalities. The output after randomly applying a mask to $\boldsymbol{F}_{T_o}^{global}$ is denoted as $\boldsymbol{F}_{T_o}^{mask}$.

$$\boldsymbol{F}^{mask} = \left[\boldsymbol{F}_{Flair}^{mask}, \boldsymbol{F}_{T1ce}^{mask}, \boldsymbol{F}_{T1}^{mask}, \boldsymbol{F}_{T2}^{mask}\right] \tag{6}$$

Where $\boldsymbol{F}_{T_o}^{mask} = \delta_{T_o} \boldsymbol{F}_{T_o}^{global}$, $[\cdot,\cdot]$ denote concatenation operation. During the training process, we randomly set δ_m to 0 to simulate the missing modality. The process of in multimodal information interaction can be represented as:

$$\boldsymbol{H}_{T_c \leftarrow T_o}^{i} = \mathrm{CrossAtt}\left(\boldsymbol{Q}_{T_c}^{i}, \boldsymbol{K}_{T_o}^{i}, \boldsymbol{V}_{T_o}^{i}\right) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{T_c}^{i}(\boldsymbol{K}_{T_o}^{i})^{T}}{\sqrt{d_k}}\right)\boldsymbol{V}_{T_o}^{i} \tag{7}$$

Where $\boldsymbol{H}_{T_c \leftarrow T_o}^{i}$ indicates the result of the ith attention head in $T_c \leftarrow T_o$. $\boldsymbol{Q}_{T_c}^{i} = LN\left(\boldsymbol{F}_{T_c}^{global}\right)\boldsymbol{W}_{T_c,\boldsymbol{Q}}^{i}$, $\boldsymbol{K}_{T_o}^{i} = LN\left(\boldsymbol{F}_{T_o}^{global}\right)\boldsymbol{W}_{T_o,K}^{i}$, $\boldsymbol{V}_{T_o}^{i} = LN\left(\boldsymbol{F}_{T_o}^{global}\right)\boldsymbol{W}_{T_o,Q}^{i}$, $\boldsymbol{W}_{m,l}^{i} \in \mathbb{R}^{C' \times d}(m = T_c, T_o; l = \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ is the parameter matrix of the linear projection, and LN is layer normalization. After obtaining features of the ith attention head $\boldsymbol{H}_{T_c \leftarrow T_o}^{i}$, the complete expression of $T_c \leftarrow T_o$ is:

$$\boldsymbol{F}_{T_c \leftarrow T_o} = \left[\boldsymbol{H}_{T_c \leftarrow T_o}^{1}, \boldsymbol{H}_{T_c \leftarrow T_o}^{2}, \ldots, \boldsymbol{H}_{T_c \leftarrow T_o}^{N_h}\right]\boldsymbol{W}_{T_c \leftarrow T_o} \tag{8}$$

Where $\boldsymbol{W}_{T_c \leftarrow T_o}$ is the linear projection matrix. In this paper, the number of heads in the attention mechanism is N_h, which set to 8 [32]. The features after interaction are then fused with the available modal features:

$$\boldsymbol{F}_{fuse} = \boldsymbol{F}^{mask} + \left[\boldsymbol{F}_{Flair \leftarrow T_o}, \boldsymbol{F}_{T1ce \leftarrow T_o}, \boldsymbol{F}_{T1 \leftarrow T_o}, \boldsymbol{F}_{T2 \leftarrow T_o}\right] \tag{9}$$

The fused features $\boldsymbol{F}_{fuse}$ are obtained by passing the input through a Feed Forward Network (FFN) consisting of a linear layer, activation function GELU, and dropout. Then, the global shared feature representation $\boldsymbol{F}_{share} \in \mathbb{R}^{C' \times H \times W \times D}$ is obtained by reshaping the output.

To obtain a better shared feature representation of brain tumors, we force the missing modality path to learn the rich feature representation of the complete modality path, making their potential feature representations as close as possible. To achieve this, we use the KL divergence as a loss function between them:

$$\mathcal{L}_{kl} = D_{kl}\left(\boldsymbol{F}_{share}, \boldsymbol{F}_{full}\right) \tag{10}$$

Aligning the shared feature representation with the full modal potential feature distribution ensures that the missing modality path learn rich feature information from the complete modality, leading to enhanced reconstruction of the missing modality information.

Since the obtained shared feature representation of the brain tumor is obtained under the supervision and guidance of the full modality, it enables recovery of the missing modality information. To reconstruct the fused features of the missing modality, we use a reconstruction network composed of three convolutional layers. To ensure accurate missing modality features, we use a pre-trained network to integrate missing modality features to obtain $\boldsymbol{F}_{missing}$ and utilize it for supervision. The pre-trained network has the same structure as the missing modality path and aims to guide the network to recover the missing information using the shared feature representation. To ensures high quality features can be reconstructed, we used $\mathcal{L}_{rec}$ to supervise the feature reconstruction:

$$\mathcal{L}_{rec} = \left\|\boldsymbol{F}_{missing} - \boldsymbol{F}_{rec}\right\|_1 \tag{11}$$

3.4. Shared feature representation enhancement

We fuse the reconstructed missing modality features with our shared features to obtain a more comprehensive representation of the brain tumor and further enhance the shared feature representation to its fullest potential. As illustrated in Fig. 4, we use cross-attention to compute the correlation between the two sets of features and capture their tumor-related features more accurately. Specifically, the global shared feature representation $\boldsymbol{F}_{share}$ is first projected to a key ($\boldsymbol{k}$) and value ($\boldsymbol{v}$) vector using linear projection, while the reconstructed missing modality fusion feature is projected to a query vector ($\boldsymbol{q}$). To strengthen the correlation between the two sets of features, we calculate the similarity between $\boldsymbol{q}$ and $\boldsymbol{k}$ using cross-attention, and then use similarity-weighted multiplication by $\boldsymbol{v}$. We fuse it with the missing modality features in an additive manner to fully exploit the complementary information of the missing modality. We then apply self-attention to establish the global inter-pixel correlation, and finally add it to the global shared feature representation to obtain a more complete shared feature representation of the brain tumor. This process can be expressed as:

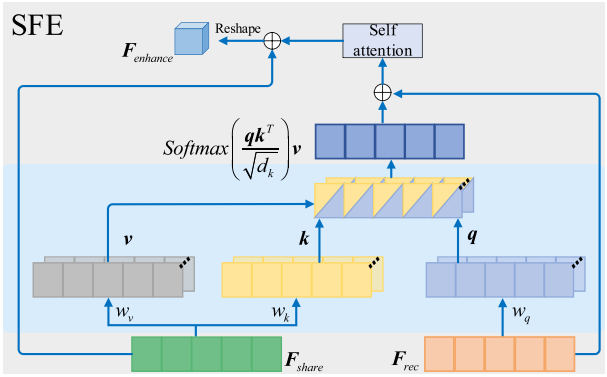

Fig. 4. Shared feature enhancement module.

$$\boldsymbol{q} = \boldsymbol{F}_{rec} w_q, \boldsymbol{k} = \boldsymbol{F}_{share} w_k, \boldsymbol{v} = \boldsymbol{F}_{share} w_v \tag{12}$$

$$\boldsymbol{F}_{enhance} = \boldsymbol{F}_{share} + self attention\left(\boldsymbol{F}_{rec} + Softmax\left(\frac{\boldsymbol{q}\boldsymbol{k}^T}{\sqrt{d_k}}\right)\boldsymbol{v}\right) \tag{13}$$

To further enhance the shared feature representation, we employ a shared decoder for both the missing and complete modality paths, which enforces the shared feature representation to align with the potential features of the full modality. The enhanced shared feature representation is then progressively upsampled to recover the original resolution size and generate the segmentation results. In addition, to obtain spatially detailed features rich in available modalities, we fuse the shallow feature maps obtained from each level of the encoder for each modality, and fuse the encoder and decoder features using a skip connection for finer segmentation. Finally, we transfer useful knowledge from the complete modality path to the missing modality path by using the segmentation results generated by the former as soft labels to supervise the segmentation results of the latter. The knowledge distillation loss is defined as follows:

$$\mathcal{L}_{kd} = \mathcal{L}_{WCE}(\boldsymbol{P}_{missing}, \boldsymbol{P}_{full}) + \mathcal{L}_{Dice}(\boldsymbol{P}_{missing}, \boldsymbol{P}_{full}) \tag{14}$$

$\boldsymbol{P}_{full}$ and $\boldsymbol{P}_{missing}$ represent the predictions of the complete modality path and missing modality path, respectively. We utilize the combination of $\mathcal{L}_{WCE}$ and $\mathcal{L}_{Dice}$ to quantify the similarity between the two segmentation outcomes.

3.5. Loss function

As depicted in Fig. 2, the decoder is employed to make the ultimate segmentation prediction, and the weighted cross-entropy loss and Dice loss are utilized to optimize the network's segmentation performance:

$$\mathcal{L}_{seg} = \sum_{path} \mathcal{L}_{WCE}(\boldsymbol{P}_{path}, \boldsymbol{Y}) + \mathcal{L}_{Dice}(\boldsymbol{P}_{path}, \boldsymbol{Y}) \tag{15}$$

Table 1
Results of the proposed method and state-of-the-art unified models on BraTS 2018 dataset [29,37]. Dice similarity scores (%) is employed for evaluation with every combination settings of modalities. • and ∘ denote available and missing modalities, respectively.

M																	Avg
	Flair	∘	∘	∘	•	∘	∘	•	∘	•	•	•	•	•	∘	•	
	T1ce	∘	∘	•	∘	∘	•	•	•	∘	∘	•	•	∘	•	•	
	T1	∘	•	∘	∘	•	•	∘	∘	∘	•	•	∘	•	•	•	
	T2	•	∘	∘	∘	•	∘	∘	•	•	∘	∘	•	•	•	•	
ET	U-HeMIS [27]	25.63	10.16	62.02	11.78	32.39	66.22	66.10	67.83	30.22	10.71	68.54	68.72	31.07	69.92	70.27	46.10
	U-HVED [28]	22.82	8.60	57.64	23.8	24.29	61.11	68.36	67.83	32.31	27.96	68.60	68.93	32.34	67.75	69.03	46.76
	RobustSeg [34]	28.97	17.29	67.07	25.69	32.01	69.06	70.30	69.71	33.84	32.13	70.78	70.88	70.78	70.10	71.13	51.02
	D^2-Net [31]	16.00	8.10	66.30	8.10	16.50	70.70	64.80	68.70	17.40	9.50	65.70	66.40	19.40	68.30	68.40	42.30
	mmformer [32]	**43.05**	32.35	72.60	**39.33**	**44.99**	74.04	75.05	74.51	**47.52**	42.96	75.47	**75.67**	**47.70**	74.75	**77.61**	**59.85**
	RFNet [29]	40.71	34.43	69.43	33.92	41.42	70.73	72.84	73.01	43.77	39.68	73.17	73.13	73.17	72.56	73.21	57.12
	Ours	40.22	**36.94**	**74.79**	36.96	44.97	**76.75**	**75.63**	**75.48**	42.54	**43.47**	**75.66**	74.80	45.08	**75.88**	75.39	59.64
TC	U-HeMIS [27]	57.20	37.39	65.29	26.06	60.92	72.46	71.49	76.64	57.68	41.12	76.01	77.53	60.32	78.96	79.48	62.57
	U-HVED [28]	54.67	33.90	59.59	57.90	56.26	67.55	75.07	73.92	62.70	61.14	77.05	76.75	63.14	75.28	77.71	64.84
	RobustSeg [34]	57.49	47.9	76.83	53.57	62.19	78.72	80.62	80.20	61.16	60.68	81.06	80.72	81.06	80.33	80.86	69.78
	D^2-Net [31]	56.70	16.80	65.10	47.30	63.20	78.20	80.80	80.30	62.60	61.60	80.90	80.70	63.70	79.00	80.10	66.50
	mmformer [32]	64.20	56.55	75.41	61.21	69.42	78.59	77.88	78.61	69.75	65.91	79.80	79.55	71.52	80.39	85.78	72.97
	RFNet [29]	67.62	64.42	**80.99**	64.89	70.89	82.22	82.94	**83.27**	70.82	71.59	83.77	**83.54**	73.09	**83.97**	**84.02**	76.53
	Ours	**68.49**	**66.03**	80.91	**67.04**	**71.56**	**82.66**	**83.31**	83.26	**71.92**	**72.12**	**83.82**	83.34	**73.47**	83.61	83.69	**77.02**
WT	U-HeMIS [27]	80.96	57.62	61.53	52.48	82.41	68.47	68.99	82.48	82.95	64.62	72.31	83.85	83.43	83.94	84.74	74.05
	U-HVED [28]	79.83	49.51	53.62	84.39	81.56	64.22	85.93	81.32	87.58	85.71	86.72	88.09	88.07	82.32	88.46	79.16
	RobustSeg [34]	82.24	70.11	74.93	85.69	84.78	77.18	88.51	85.19	88.28	88.24	88.73	89.27	88.73	86.01	89.45	84.39
	D^2-Net [31]	76.30	15.50	42.80	84.20	80.10	62.10	87.50	84.10	87.90	87.30	87.70	88.80	88.40	80.90	88.80	76.20
	mmformer [32]	81.15	67.52	72.22	86.10	82.20	74.42	87.30	82.99	87.59	87.06	87.33	88.14	87.75	82.71	89.64	82.94
	RFNet [29]	84.30	74.68	74.93	86.46	86.15	78.59	88.78	86.39	89.12	88.78	89.71	90.06	89.71	86.78	90.26	85.67
	Ours	**85.38**	**77.24**	**74.79**	**87.83**	**86.55**	**79.38**	**89.48**	**86.86**	**89.84**	**89.37**	**89.60**	**90.38**	**90.04**	**87.11**	**90.36**	**86.28**

$\boldsymbol{Y}$ represents the ground truth. $\boldsymbol{P}_{path} \in \{\boldsymbol{P}_{full}, \boldsymbol{P}_{missing}\}$, $\mathcal{L}_{WCE}$ and $\mathcal{L}_{Dice}$defined as:

$$\mathcal{L}_{Dice}(\boldsymbol{P},\boldsymbol{Y}) = 1 - \sum_{k\in\Omega}\left(\frac{2\sum_{j=1}^{N} \boldsymbol{y}_j^k \boldsymbol{p}_j^k}{\sum_{j=1}^{N} \boldsymbol{y}_j^k \boldsymbol{y}_j^k + \sum_{j=1}^{N} \boldsymbol{p}_j^k \boldsymbol{p}_j^k + \varepsilon}\right) \tag{16}$$

$$\mathcal{L}_{\mathrm{WCE}}(\boldsymbol{P},\boldsymbol{Y}) = -\sum_{k\in\Omega}\left(\sum_{\mathrm{j}=1}^{\mathrm{N}} \omega^{\mathrm{k}} \cdot \mathbf{y}_{\mathrm{j}}^{\mathrm{k}} \log \mathbf{q}_{\mathrm{j}}^{\mathrm{k}}\right) \tag{17}$$

Where $\Omega = \{BG(background), NCR/NET, ED, ET\}$, $\boldsymbol{y}_j^k$, $\boldsymbol{q}_j^k$ and $\boldsymbol{p}_j^k$ represent the ground truth, probability prediction and one-hot output of voxel j on class k, respectively. ω^k is the weight of region k, $\varepsilon = 1{\times}10^{-5}$, $N = H \times W \times D$.

Therefore, the overall objective function of our framework can be expressed as:

$$\mathcal{L}_{overall} = \mathcal{L}_{seg} + \mathcal{L}_{kd} + \mathcal{L}_{trans} + \mathcal{L}_{kl} + \mathcal{L}_{rec} \tag{18}$$

Our framework incorporates standard Dice loss and multi-class cross-entropy loss to accurately perform segmentation tasks.

4. Experiments

4.1. Datasets and evaluation metrics

To evaluate the effectiveness of our proposed method, we conducted experiments on two widely used brain tumor datasets, BraTS2018 and BraTS2020 [45]. The BraTS2018 dataset consists of scans from 210 patients with high-grade glioma and 75 patients with low-grade glioma, each containing 4 modalities: Flair, T1ce, T1, and T2. The dataset is labeled in 4 categories, namely tumor enhancing area, edema area, necrotic area, and healthy tissue. The performance of our method was evaluated by the segmentation results of whole tumor (WT), tumor core (TC), and enhanced tumor (ET). WT includes tumor enhancement zone, edema zone, and necrotic zone, TC includes tumor enhancement zone and necrotic zone, and ET includes tumor enhancement zone. Consistent with standard practice in the literature [29,32,34], we randomly divided our dataset into three parts, including a training set of 190 cases and a validation set of 95 cases. We performed experiments in a three-fold cross-validation manner. In addition, the provided datasets have undergone various preprocessing operations, such as co-registered, skull-stripped, and re-sampled to 1 mm^3 resolution.

In order to assess the efficacy of our proposed method, we employed the Dice similarity score as a quantitative measure of segmentation accuracy. The Dice similarity score quantifies the level of overlap between the predicted segmentation results and the true labels, where 0 indicates no overlap and 1 indicates perfect overlap. Higher Dice similarity scores indicate better segmentation performance. The Dice similarity score is defined as follows:

$$Dice_{\bar{k}}(\hat{\boldsymbol{y}},\boldsymbol{y}) = \frac{2\cdot\left\|\hat{\boldsymbol{y}}_{\bar{k}} \cap \boldsymbol{y}_{\bar{k}}\right\|_1}{\left\|\hat{\boldsymbol{y}}_{\bar{k}}\right\|_1 + \left\|\boldsymbol{y}_{\bar{k}}\right\|_1} \tag{19}$$

Where $\bar{k}$ denotes different tumor classes, including whole tumor, core tumor, and enhanced tumor. The Dice similarity score of tumor class $\bar{k}$ is denoted as $Dice_{\bar{k}}$. A higher Dice similarity score indicates better segmentation accuracy, as it indicates that the predicted result is closer to the ground truth.

4.2. Implement details

In accordance with the approach described in [28], we applied a normalization operation to each modality in the dataset to ensure a consistent distribution of grayscale values in the MR images. All experiments were conducted on two NVIDIA GTX3090 GPUs (24G) under the Pytorch 1.12.1 framework. During the training phase, we randomly cropped the input images to 128 × 128 × 128 and utilized random rotation, intensity shift, and mirror flip for data augmentation. Additionally, we randomly set certain modality to a zero matrix to simulate the scenario of missing modality [20,28,31,34]. The network was trained for 1500 epochs with a batch size of 2, and was optimized using the Adam optimizer [46] with exponential decay rates β_1 and β_2 of 0.9 and 0.999, respectively. The learning weight decay was set to $1e^{-4}$. Furthermore, the learning rate of each epoch can be denoted as $lr = base\_lr \times \left(1 - \frac{epoch}{end\_epoch}\right)^p$, using "poly" to dynamically adjust the learning rate, where $p = 0.9$, $base\_lr = 1e^{-4}$ as the baseline learning rate, and the total epoch number $end\_epoch = 1500$.

4.3. Comparison with state-of-the-arts

To assess the effectiveness of our proposed method, we conducted extensive comparative experiments between our method and existing state-of-the-art methods on the BraTS2018 and BraTS2020 datasets.

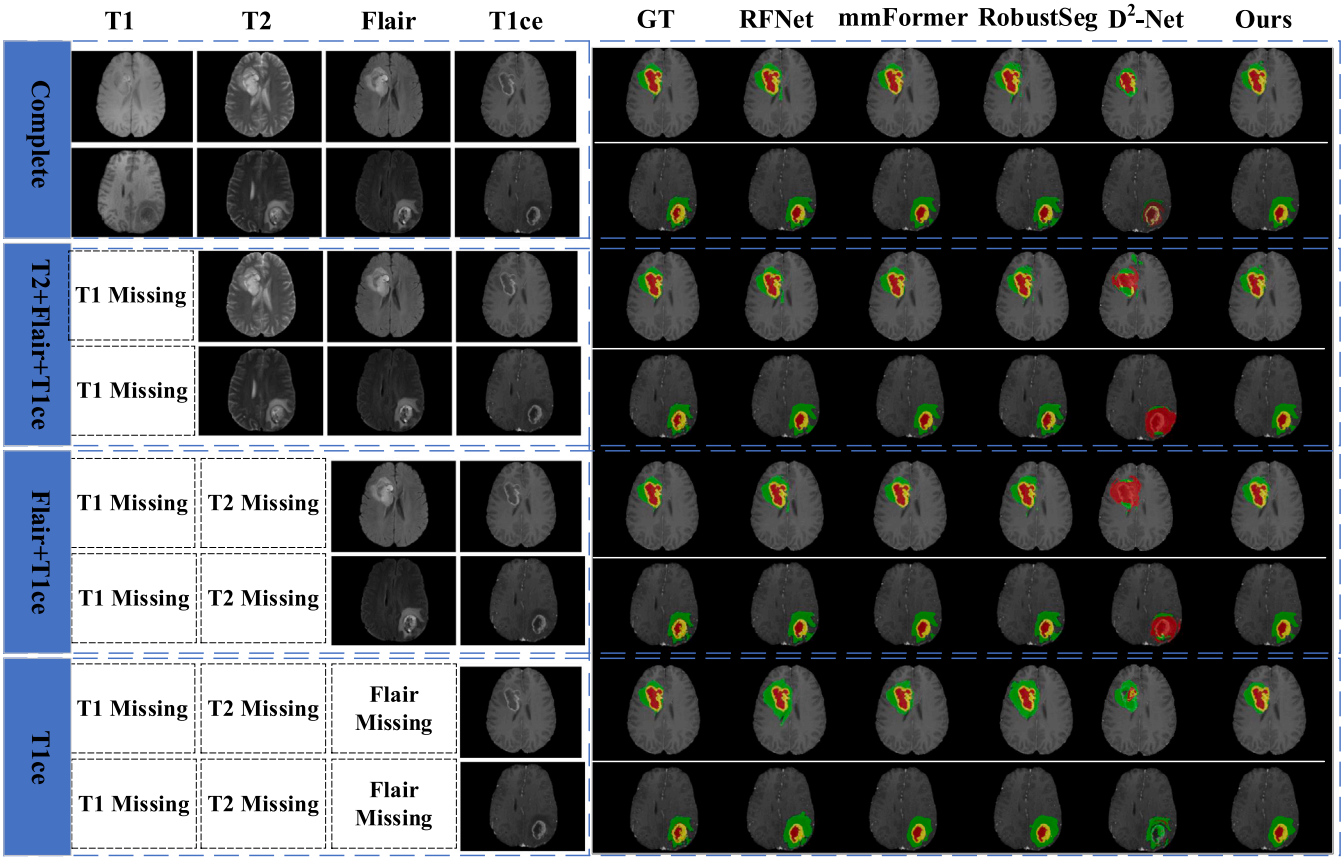

Fig. 5. Comparison of segmentation results on four cases of missing modalities: complete modalities; FLAIR, T1ce, T2; FLAIR, T1ce; T1ce. From the left to right are four MRI modalities: T1, T2, FLAIR and T1ce; The fifth column presents the Ground Truth of two patients, the sixth to ninth columns perform the results of state-of-the-art approaches, and the right column shows our segmentation results;.

To ensure a fair comparison, we adopted the experimental setup used in [29,32,34] and employed a three-fold cross-validation strategy to evaluate the performance of our method on both datasets.

4.3.1. Comparison on BraTS2018

We conducted extensive experiments on the BraTS2018 and BraTS2020 datasets to evaluate the performance of our proposed method. We compared our results with those of D^2-Net [31], mmFormer [32], RFNet [29] and other state-of-the-art methods using Dice Score as the evaluation metric. As shown in Table 1, our method significantly outperformed other methods in terms of Dice scores 86.28%, 77.02%, and 59.64% on WT, TC, and ET, respectively, for each of the 15 cases of missing modalities. Our network architecture proved to be effective, as demonstrated by these results, which outperforms current state-of-the-art methods, such as RFNet and mmFormer, in most modality missing cases. Compared with RFNet, we obtained an average Dice score improvement of 0.61% and 0.49% for the whole tumor and tumor core regions, respectively, and a remarkable 2.52% improvement for the most challenging enhanced tumor region. This result indicates that our method has a greater improvement for the enhanced tumor region, which is the most challenging region among these three types of tumor regions. Our proposed method outperformed other methods in all fifteen cases for the segmentation of WT brain tumor regions, indicating that using the self-attention mechanism to establish inter-modal correlations and reconstruct missing modality features can well establish correlations under global semantics and facilitate the modeling of global contexts. Furthermore, under the four unimodal modalities, our proposed method outperformed other methods in most cases, indicating that multimodal information can be effectively transferred to a single modality in our unimodal information learning phase, thus compensating for the problem of incomplete unimodal information and exhibiting robust performance under unimodal modality. We also analyzed the different effects of various missing modality cases and found that the whole tumor performance decreases significantly under the missing Flair modality, which affects the whole tumor region by about 3%. In the absence of T1, T2 or one of the Flair modalities, having the T1ce modality has good segmentation performance for the enhanced tumor and tumor core regions, because the T1ce modality is the main modality that shows the enhanced tumor and tumor core. In the absence of T1 or T2, the segmentation performance of all regions was slightly decreased.

4.3.2. Comparison on BraTS2020

In addition to conducting experiments on BraTS2018, we also performed experiments on BraTS2020 using three-fold cross-validation to obtain average results. We then compared these results with advanced methods, and for the sake of fairness, we directly cited the results of RFNet [29]. Table 2 displays the results, which demonstrate that our method produces remarkable improvements in segmentation performance. Specifically, our method achieves optimal Dice scores of 63.56%, 79.44%, and 87.03% for ET, TC, and WT, respectively. In comparison to RFNet, our method outperforms it by 2.09%, 1.21%, and 0.05% for ET, TC, and WT, respectively. These results validate the superiority of our method.

4.3.3. Visual effect comparison of segmentation results

To demonstrate the effectiveness of our method, we provide quantitative comparisons with other state-of-the-art methods on the BraTS2018 dataset in Fig. 5, presenting results for the Complete modality, Flair+T1ce+t2, Flair+T1ce, and T1ce cases. The results reveal that

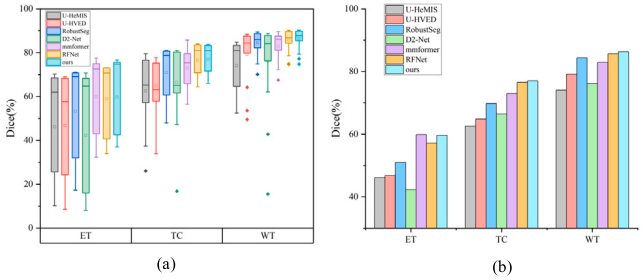

Fig. 6. Boxplots and histogram of the Dice score of the comparison experiment results. (a) Shows the results for 15 cases of missing modalities. (b) shows the average results of state-of-the-art approaches and ours.

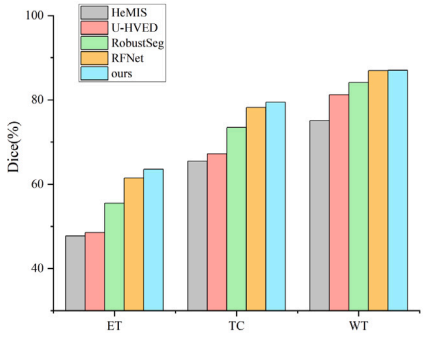

Fig. 7. Histogram of the Dice score of the comparison experiment results on BraTS2020.

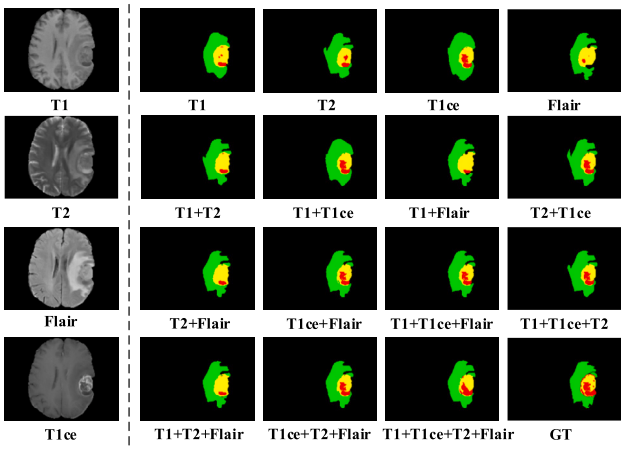

Fig. 8. Visualization of the predicted segmentation maps. Left: four image modalities. Right: segmentation maps predicted by our approaches from all fifteen combinations of image modalities and the corresponding ground truth.

our method achieves accurate segmentation results in most cases. Fig. 6 presents the comparison of our method with other methods on the BraTS2018 dataset using box plots of Dice coefficients obtained from the segmentation results of 15 missing modality cases and histograms generated from average results. The figure reveals the distribution characteristics of segmentation results of different methods in various cases. The results on the BraTS2018 test set are generally maintained at a balanced level and outperforms other methods in most cases. Fig. 7 illustrates the comparison of our method with other methods on the BraTS2020 dataset, demonstrating that our method achieves outstanding results on this dataset.

To visualize the segmentation results for each missing case, we provide visualizations for 15 missing cases in Fig. 8. The visualizations

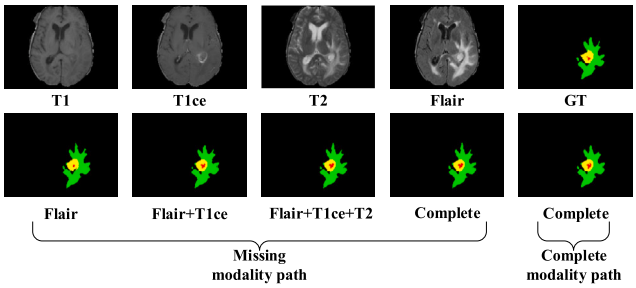

Fig. 9. Visualizations of the predicted segmentation maps of two paths.

Table 2
Comparisons with the state-of-the-art on BRATS2020.

Method	Dice score (%)		
	ET	TC	WT
U-HeMIS [27]	47.73	65.45	75.10
U-HVED [28]	48.55	67.19	81.24
RobustSeg [34]	55.49	73.45	84.17
RFNet [29]	61.47	78.23	86.98
Ours	**63.56**	**79.44**	**87.03**

Table 3
Ablation study on components of our method.

Models	Dice score (%)			
	ET	TC	WT	Mean
Baseline(w/o self-attention)	53.98	71.88	82.55	69.47
Baseline	55.46	73.21	83.55	70.74
+UML(w/o L1)	57.30	75.76	83.89	72.32
+UML(w/o KL)	58.66	75.54	84.32	72.84
+UML	59.16	75.93	84.89	73.33
+UML+MMFR(w/o interactive fusion)	59.66	75.64	85.19	73.50
+UML+MMFR	60.54	76.27	85.27	74.03
+UML+MMFR+SFE(Ours)	**61.18**	**76.59**	**85.60**	**74.46**

highlight that our method delivers superior segmentation results in various missing cases, particularly in unimodal cases, and can predict more accurate segmentation results. Furthermore, we investigate the effect of joint learning on missing modal paths by visualizing the results of complete modality paths and missing modality paths in Fig. 9. The figure reveals that our method can achieve comparable segmentation results for different cases of missing modality paths, indicating that joint learning can enhance feature extraction of missing modality paths and improve segmentation results.

4.4. Ablation study

In this study, we introduce a method composed of three essential components: UML, MMFR, and SFE. The effectiveness of each module was validated through ablation experiments, using a three-fold cross-validation on BraTS2018 dataset. During the experiments, the three main components were removed, and the baseline model was used to aggregate the encoded features of available modalities using only $1 \times 1 \times 1$ convolutional layers. To verify the effectiveness of the self-attention mechanism, we remove self-attention from Baseline and find that the performance degrades in all regions. Subsequently, the three modules were incrementally added, i.e., Baseline+UML, Baseline+UML +MMFR, and Baseline+UML+MMFR+SFE. As shown in Table 3, the results demonstrate the effectiveness of each module and demonstrate the effectiveness of the proposed method in improving segmentation performance.

Effectiveness of UML: In this paper, we propose a method that supervises each modality with the full modality to learn complete brain tumor information. This approach addresses the incomplete information of single modality and reduces the dependence on additional

Table 4
The comparison results between Baseline and Baseline+UML in all four single modality cases.

Dice score (%)		Baseline	Baseline +UML
WT	Flair	85.77	**87.74**
	T1	70.16	**72.72**
	T1ce	68.09	**72.18**
	T2	81.64	**83.11**
TC	Flair	61.03	**64.18**
	T1	56.69	**65.05**
	T1ce	68.09	**78.99**
	T2	62.85	**63.96**
ET	Flair	33.11	**35.84**
	T1	26.83	**33.97**
	T1ce	70.83	**76.29**
	T2	35.43	**37.12**

modalities to improve brain tumor segmentation results. To validate the effectiveness of our approach, we compared the performance of Baseline and Baseline+UML. As shown in Table 3, the Dice scores of ET, TC, and WT reached 59.16%, 75.93%, and 84.89%, respectively, which were improved by 3.70%, 2.72%, and 1.34% compared to Baseline. We also evaluated the segmentation performance under single modality and compared it with the baseline. As shown in Table 4, our proposed approach, Baseline+UML, outperforms the baseline in all four single modal cases, highlighting the effectiveness of UML in improving the segmentation performance. To assess the benefits and necessity of the two loss items in Eq. (3), we removed the KL divergence and L1 loss separately. As shown in Table 3, the Dice scores of ET, TC, and WT all decreased. The KL measures the difference between the probability distributions of two features, while the L1 is used to measure the difference between two features. Therefore, the joint use of both measures can achieve optimal results.

Effectiveness of MMFR: MMFR module achieves the fusion of available modality brain tumor information and reconstructs missing modality features by incorporating information from available modalities. The introduction of MMFR resulted in improved Dice scores of ET, TC, and WT by 1.38%, 0.34%, and 0.38%, respectively, compared to Baseline+UML, as shown in Table 3. This improvement validates the effectiveness of the MMFR module. We also removed the interactive fusion and instead employed $1 \times 1 \times 1$ convolutional layers to collect and process encoded features from available modalities. The Dice scores of ET, TC, and WT were decreased by 0.50%, 0.63%, and 0.80%, respectively, which were validates the effectiveness of the interactive fusion.

Effectiveness of SFE: SFE enhances the shared feature representation of brain tumors by utilizing the complementary nature of reconstructed missing modality features. This results in more complete information about brain tumors and improves the segmentation results. The performance of SFE was evaluated by comparing the Dice scores of ET, TC, and WT obtained with Baseline+UML+MMFR and Baseline+UML+MMFR+SFE, as shown in Table 3. The results demonstrate that SFE can further improve the segmentation results, with an increase of 0.64%, 0.32%, and 0.33% for ET, TC, and WT, respectively, compared to Baseline+UML+MMFR. These findings validate the effectiveness of the SFE module in improving the performance of the proposed method.

In order to demonstrate the effectiveness of each module on segmentation results, we provide visualizations of the four different modal missing cases using various methods. As shown in Fig. 10, compared to the baseline, the segmentation results of other methods have been improved, and our method is particularly successful in approximating the ground truth. In addition, Fig. 11 displays box and histogram plots of Dice coefficients obtained using the segmentation results of 95 validation set samples, revealing the distribution characteristics of the segmentation results for the four methods.

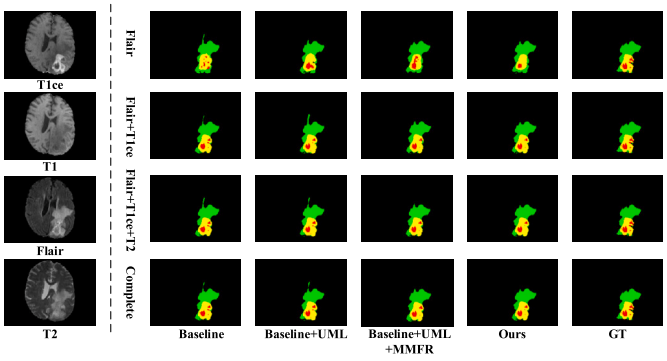

Fig. 10. Visual comparison of the effectiveness of different components.

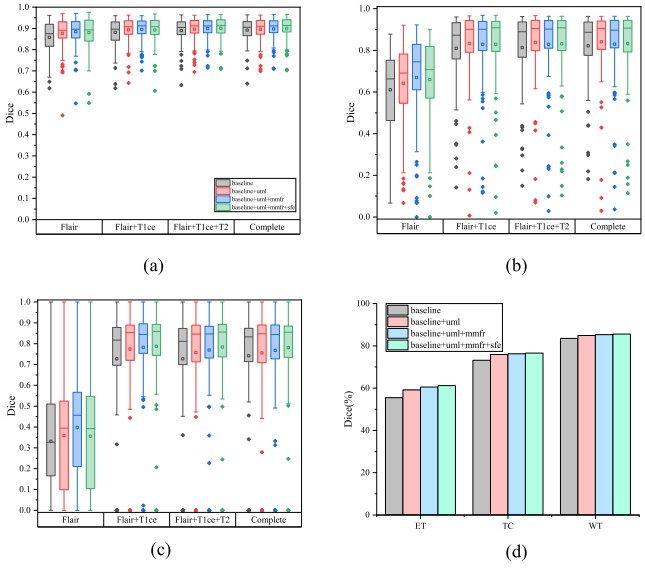

Fig. 11. Boxplots and histogram of the Dice score of the ablation experiment results. From (a) to (c) are the Dice of WT, TC, and ET, respectively. (d) is the column plot contrast result of different components.

5. Conclusion

This paper proposes a novel method for feature reconstruction and enhancement based on joint learning. Our approach employs multimodal guided unimodal learning to transfer multimodal information to unimodal and alleviate the limitations of unimodal carried information. The proposed method considers the complementary nature of missing modality information, and utilizes interaction mechanisms to transfer and fuse information between available modalities for better reconstruction of missing modality features, and consequently, the recovery of missing information. Moreover, a feature enhancement mechanism is used to improve the shared feature representation using the recovered information, resulting in more complete brain tumor information and improved segmentation performance of the network in brain tumor segmentation. The effectiveness of the proposed method is demonstrated through extensive experiments on the BraTS2018 and BraTS2020 datasets, which show that our model achieves excellent performance and robustness in handling incomplete modality brain tumor segmentation.

CRediT authorship contribution statement

Yueqin Diao: Conception and design, Data acquisition, Analysis and interpretation, Drafting and critically revising the manuscript.

Fan Li: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. **Zhiyuan Li:** Visualization, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62161015), and Science and Technology Department of Yunnan Province (Grant No. 202101AT070136).

References

[1] X. Guo, C. Yang, P.L. Lam, P.Y. Woo, Y. Yuan, Domain knowledge based brain tumor segmentation and overall survival prediction, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer, 2020, pp. 285–295, http://dx.doi.org/10.1007/978-3-030-46643-5_28.

[2] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[3] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11, http://dx.doi.org/10.1007/978-3-030-00889-5_1.

[4] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, 2018, http://dx.doi.org/10.48550/arXiv.1804.03999, arXiv preprint.

[5] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer, 2022, pp. 272–284, http://dx.doi.org/10.1007/978-3-031-08999-2_22.

[6] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, UNETR: Transformers for 3D medical image segmentation, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2022, pp. 574–584, http://dx.doi.org/10.1109/WACV51458.2022.00181.

[7] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer, 2021, pp. 109–119, http://dx.doi.org/10.1007/978-3-030-87193-2_11.

[8] T. Zhou, S. Ruan, P. Vera, S. Canu, A tri-attention fusion guided multi-modal segmentation network, Pattern Recognit. 124 (2022) 108417, http://dx.doi.org/10.1016/j.patcog.2021.108417.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) http://dx.doi.org/10.1007/978-3-031-08999-2_22.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, 2021, http://dx.doi.org/10.48550/arXiv.2010.11929.

[11] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, http://dx.doi.org/10.48550/arXiv.2102.04306, arXiv preprint.

[12] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer, 2021, pp. 14–24, http://dx.doi.org/10.1007/978-3-030-87193-2_2.

[13] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer, 2021, pp. 36–46, http://dx.doi.org/10.1007/978-3-030-87193-2_4.

[14] X. Huang, Z. Deng, D. Li, X. Yuan, Y. Fu, Missformer: An effective transformer for 2D medical image segmentation, IEEE Trans. Med. Imaging (2022) 1, http://dx.doi.org/10.1109/TMI.2022.3230943.

[15] B. Chen, Y. Liu, Z. Zhang, G. Lu, A.W.K. Kong, TransAttUnet: Multi-level attention-guided U-net with transformer for medical image segmentation, 2021, http://dx.doi.org/10.48550/arXiv.2107.05274, CoRR.

[16] Z. Xing, L. Yu, L. Wan, T. Han, L. Zhu, Nestedformer: Nested modality-aware transformer for brain tumor segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022, Springer, 2022, pp. 140–150, http://dx.doi.org/10.1007/978-3-031-16443-9_14.

[17] Y. Wang, Y. Zhang, Y. Liu, Z. Lin, J. Tian, C. Zhong, Z. Shi, J. Fan, Z. He, ACN: adversarial co-training network for brain tumor segmentation with missing modalities, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Springer, 2021, pp. 410–420, http://dx.doi.org/10.1007/978-3-030-87234-2_39.

[18] R. Azad, N. Khosravi, D. Merhof, SMU-net: Style matching U-net for brain tumor segmentation with missing modalities, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 48–62, http://dx.doi.org/10.48550/arXiv.2204.02961.

[19] M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, P. Gori, Knowledge distillation from multi-modal to mono-modal segmentation networks, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020, Springer, 2020, pp. 772–781, http://dx.doi.org/10.1007/978-3-030-59710-8_75.

[20] Q. Wang, L. Zhan, P. Thompson, J. Zhou, Multimodal learning with incomplete modalities by knowledge distillation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1828–1838, http://dx.doi.org/10.1145/3394486.3403234.

[21] P. Huang, D. Li, Z. Jiao, D. Wei, B. Cao, Z. Mo, Q. Wang, H. Zhang, D. Shen, Common feature learning for brain tumor MRI synthesis by context-aware generative adversarial network, Med. Image Anal. 79 (2022) 102472, http://dx.doi.org/10.1016/j.media.2022.102472.

[22] Y. Wang, L. Zhou, B. Yu, L. Wang, C. Zu, D.S. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen, 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis, IEEE Trans. Med. Imaging 38 (6) (2018) 1328–1339, http://dx.doi.org/10.1109/TMI.2018.2884053.

[23] T. Zhou, H. Fu, G. Chen, J. Shen, L. Shao, Hi-net: hybrid-fusion network for multi-modal MR image synthesis, IEEE Trans. Med. Imaging 39 (9) (2020) 2772–2781, http://dx.doi.org/10.1007/978-3-031-08999-2_22.

[24] Y. Li, K.K. Singh, U. Ojha, Y.J. Lee, Mixnmatch: Multifactor disentanglement and encoding for conditional image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8039–8048, http://dx.doi.org/10.48550/arXiv.1911.11758.

[25] B. Zhan, D. Li, X. Wu, J. Zhou, Y. Wang, Multi-modal MRI image synthesis via GAN with multi-scale gate mergence, IEEE J. Biomed. Health Inf. 26 (1) (2021) 17–26, http://dx.doi.org/10.1109/JBHI.2021.3088866.

[26] A. Jog, A. Carass, S. Roy, D.L. Pham, J.L. Prince, Random forest regression for magnetic resonance image synthesis, Med. Image Anal. 35 (2017) 475–488, http://dx.doi.org/10.1016/j.media.2016.08.009.

[27] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, Hemis: Hetero-modal image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016, Springer, 2016, pp. 469–477, http://dx.doi.org/10.1007/978-3-319-46723-8_54.

[28] R. Dorent, S. Joutard, M. Modat, S. Ourselin, T. Vercauteren, Hetero-modal variational encoder-decoder for joint modality completion and segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019, Springer, 2019, pp. 74–82, http://dx.doi.org/10.1007/978-3-030-32245-8_9.

[29] Y. Ding, X. Yu, Y. Yang, Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3975–3984, http://dx.doi.org/10.1109/ICCV48922.2021.00394.

[30] T. Zhou, S. Canu, P. Vera, S. Ruan, Latent correlation representation learning for brain tumor segmentation with missing MRI modalities, IEEE Trans. Image Process. 30 (2021) 4263–4274, http://dx.doi.org/10.1109/TIP.2021.3070752.

[31] Q. Yang, X. Guo, Z. Chen, P.Y. Woo, Y. Yuan, D 2-net: Dual disentanglement network for brain tumor segmentation with missing modalities, IEEE Trans. Med. Imaging 41 (10) (2022) 2953–2964, http://dx.doi.org/10.1109/TMI.2022.3175478.

[32] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, Y. Zheng, Mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022, Springer, 2022, pp. 107–117, http://dx.doi.org/10.1007/978-3-031-16443-9_11.

[33] Z. Zhao, H. Yang, J. Sun, Modality-adaptive feature interaction for brain tumor segmentation with missing modalities, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022, Springer, 2022, pp. 183–192, http://dx.doi.org/10.1007/978-3-031-16443-9_18.

[34] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, P.-A. Heng, Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019, Springer, 2019, pp. 447–456, http://dx.doi.org/10.1007/978-3-030-32248-9_50.

[35] M. Ma, J. Ren, L. Zhao, D. Testuggine, X. Peng, Are multimodal transformers robust to missing modality? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18177–18186, http://dx.doi.org/10.1109/CVPR52688.2022.01764.

[36] L. Tang, J. Yuan, J. Ma, Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, Inf. Fusion 82 (2022) 28–42, http://dx.doi.org/10.1016/j.inffus.2021.12.004.

[37] L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, Piafusion: A progressive infrared and visible image fusion network based on illumination aware, Inf. Fusion 83 (2022) 79–92, http://dx.doi.org/10.1016/j.inffus.2022.03.007.

[38] Y. Shen, M. Gao, Brain tumor segmentation on MRI with missing modalities, in: Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26, Springer, 2019, pp. 417–428, http://dx.doi.org/10.1007/978-3-030-20351-1_32.

[39] D. Lee, W.-J. Moon, J.C. Ye, Which contrast does matter? towards a deep understanding of mr contrast using collaborative gan, 2019, http://dx.doi.org/10.48550/arXiv.1905.04105, arXiv preprint.

[40] L. Shen, W. Zhu, X. Wang, L. Xing, J.M. Pauly, B. Turkbey, S.A. Harmon, T.H. Sanford, S. Mehralivand, P.L. Choyke, et al., Multi-domain image completion for random missing input data, IEEE Trans. Med. Imaging 40 (4) (2020) 1113–1122, http://dx.doi.org/10.1109/TMI.2020.3046444.

[41] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia(MM'20), 2020, pp. 1122–1131, http://dx.doi.org/10.1145/3394171.3413678.

[42] R. Guerrero, H.X. Pham, V. Pavlovic, Cross-modal retrieval and synthesis (x-mrs): Closing the modality gap in shared subspace learning, in: Proceedings of the 29th ACM International Conference on Multimedia(MM'21), 2021, pp. 3192–3201, http://dx.doi.org/10.1145/3474085.3475465.

[43] D. Jia, A. Hermans, B. Leibe, 2D vs. 3D lidar-based person detection on mobile robots, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2022, pp. 3604–3611, http://dx.doi.org/10.1109/IROS47612.2022.9981519.

[44] A.H. Liu, S. Jin, C.-I.J. Lai, A. Rouditchenko, A. Oliva, J. Glass, Cross-modal discrete representation learning, 2021, http://dx.doi.org/10.48550/arXiv.2106.05438, arXiv preprint.

[45] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024, http://dx.doi.org/10.1109/TMI.2014.2377694.

[46] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, http://dx.doi.org/10.1007/978-3-031-08999-2_22, arXiv preprint.