Big Data and Information Analytics

https://www.aimspress.com/journal/BDIA

BDIA, 8: 109–128.
DOI:10.3934/bdia.2024006
Received: 21 September 2024
Revised: 23 November 2024
Accepted: 11 December 2024
Published: 20 December 2024

Research article

Frequency filtering prompt tuning for medical image semantic segmentation with missing modalities

Yaru Cheng and Yuanjie Zheng[*]

School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

* **Correspondence:** Email: yjzheng@sdnu.edu.cn.

Abstract: Most multimodal brain tumor segmentation methods assume the availability of all modalities. However, models trained on complete modality data often experience a significant performance drop when certain modalities are missing, posing a major challenge for real-world applications. In this study, we address this issue by maximizing the use of information from the remaining modalities to reduce inter-modal dependency, allowing the encoder to extract robust features from the available data for accurate tumor segmentation. To this end, we propose a novel framework, the discriminative prompt optimization network (DPONet), that incorporates frequency filtering prompts and spatial perturbation prompts to enhance image representation space during feature extraction and fusion. To handle various missing modality scenarios, we also introduce a probability-based missing data simulation method. We evaluate DPONet on two public brain tumor segmentation datasets, BraTS2018 and BraTS2020. Experimental results demonstrate that DPONet outperforms state-of-the-art methods in terms of Dice score, HD95, and sensitivity, proving its effectiveness under both complete and incomplete modality conditions.

Keywords: prompt-based learning; brain tumor segmentation; multi-modal; deep learning; missing modalities

1. Introduction

Brain tumors are abnormal cell growths located in or near brain tissue that damage the nervous system, causing symptoms such as headaches, dizziness, dementia, seizures, and other neurological signs [1]. Magnetic resonance imaging (MRI)—including T1-weighted (T1), post-contrast T1-weighted (T1CE), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) sequences—is a prevalent diagnostic tool for brain tumors due to its sensitivity to soft tissue and high image contrast, as shown in Figure 1. Physicians utilize MRI for lesion diagnosis, but accuracy can be hindered by factors such as fatigue and emotional state. Automated methods have garnered extensive attention in the medical field due to their capability to objectively and accurately analyze imaging information.

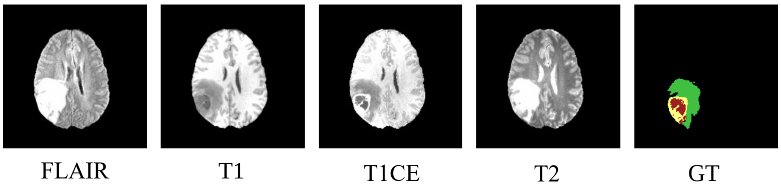


Figure 1. The samples of four MRI modalities and ground-truth of brain tumors image. FLAIR, T1, T1CE, T2 represent the four input samples respectively, and GT represents the ground truth.

Most multimodal approaches assume complete data availability; however, in reality, missing modalities are common. As illustrated in Figure 2, various missing scenarios can occur during both training and inference stages. The absence of certain MRI sequences may fail to capture tumor characteristics, thereby limiting a comprehensive understanding of the tumor [2]. Therefore, it is crucial for multimodal learning methods to maintain robustness in the presence of missing modalities during inference.

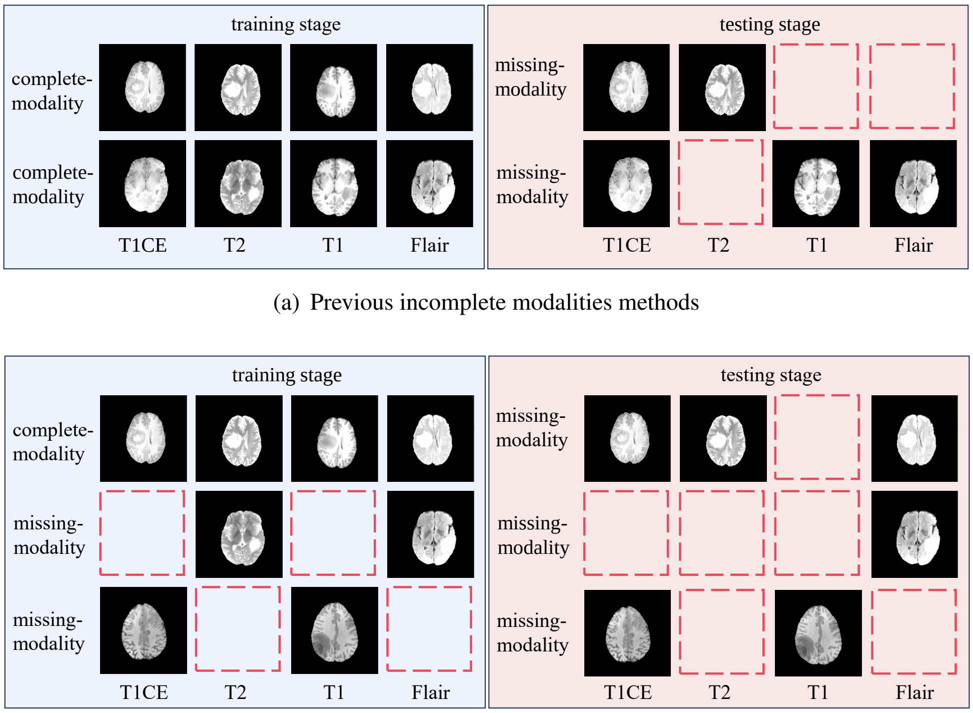
Currently, a prevalent approach to tackle segmentation arising from missing modality is knowledge distillation [3, 4], where information is transferred from a teacher-student network to recover missing data, but this can be computationally intensive. Another method is image synthesis [5], leveraging generative models to reconstruct the missing data. However, synthetic images may introduce noise to the task. Additionally, mapping available modalities into a common latent subspace aims to compensate for or recover the missing information [6–8]. However, existing approaches often require training multiple sets of parameters to address various missing modality scenarios, thereby escalating the model's complexity and computational overhead.

With the expansion of data scale and enhancement of computational resources, researchers favor general neural networks for diverse tasks, minimizing the need for task-specific model design and training. Recently, transformer [9] has shown great potential in natural language processing, visual recognition, intensive prediction. However, its complex architecture and high computational demands limit comprehensive fine-tuning for downstream tasks, especially accurate segmentation, potentially leading to overfitting and reduced generalization ability.

Inspired by recent advancements in prompt learning [10–12] and efficient fine-tuning techniques [13–15], we introduce a novel brain tumor segmentation framework, called DPONet. This framework employs an encoder-decoder structure for the segmentation network, enhancing performance in both incomplete and complete modality scenarios. Specifically, we leverage image frequency information as frequency filtering prompt (FFP) to facilitate the pre-trained model in extracting discriminative features. Furthermore, by learning a series of spatial perturbation prompt (SPP), we map these discriminative features into a common latent space, mitigating the challenges of modality fusion in the decoder. Finally, we validate the robustness of our approach on two commonly used public datasets. To sum up, our main contributions are as follows:

- We propose a new framework for incomplete-modal image segmentation that effectively handles common cases of missing modalities. This approach requires only 7% of the trainable parameters to adjust the pre-trained model, thereby avoiding the heavy fine-tuning typically necessary for transformers.

- We introduce a frequency filtering prompt to extract spatial frequency components from images. This method addresses the model's oversight of target domain features and enhances its adaptation to brain tumor datasets.
- We propose a spatial perturbation prompt that incorporates learnable parameters into a spatial modulation model. This aims to achieve consistent multimodal feature embeddings even in the presence of missing partial modalities.



(a) Previous incomplete modalities methods

(b) Our

Figure 2. We compared our method with others in terms of incomplete modality scenarios encountered during training and testing. While other methods utilize a complete dataset for training and a dataset with missing modalities for testing, our method employs datasets with missing modalities for both training and testing.

2. Related works

2.1. Incomplete multi-modal

Incomplete multimodal learning refers to scenarios in multimodal learning tasks where partial modality information is missing or incomplete. This issue becomes particularly prominent in brain tumor segmentation tasks, where medical imaging data is typically composed of multiple MRI sequences. The absence of one modality results in the challenge of incomplete modality information learning. Many studies [16–18] are devoted to solving this problem, demonstrating impressive performance in various incomplete multimodal learning tasks. Zhou et al. [16] showed that there exists a certain correlation within the latent representations of modalities, which can be utilized to describe missing modalities by calculating the correlation between modalities in a latent space. Ting et al. [17] combines available modality information to estimate the latent features of missing modalities. Liu et al. [18] explicitly considers the relationship between modalities and regions, giving different attention to different modalities for each region. However, these models require full fine-tuning of the pre-trained model, which increases the computational burden and leads to a decrease in generalization ability.

2.2. Fourier transform

The task of most neural networks is to learn the optimal points in functions. Fourier Transform establishes the transformation relationship between the function in the spatial domain and the frequency domain, so that we can analyze a function by the frequency component to approximate the objective function more effectively [19]. The frequency of an image represents the intensity of gray change in the image. Fourier transform analyzes the features by analyzing the coefficients of each frequency component [20]. The performance of computer vision models is significantly affected by the Fourier statistical properties of the training data and show a certain sensitivity to the Fourier basis direction, and their robustness can be improved by learning this sensitivity [21]. For example, Fang et al. [22] and Xu et al. [23] argued that different parts of the same organ in MRI images exhibit regularity and that high-frequency structural information can more effectively capture these similarities and regularities.

2.3. Prompt learning

Prompt learning is an effective transfer learning approach in natural language processing [10,24,25], which fine-tunes pre-trained models on source tasks by embedding contextual prompts. Recently, prompts have also been employed in computer vision tasks [26–28] and multimodal learning tasks [11,29,30], introducing self-adaptation in the input space to optimize the target task. For instance, Jia et al. [26] proposed the Pyramid Vision Transformer model (PVT), achieving downstream performance comparable to full fine-tuning by adding a small number of learnable prompt embeddings on the patch embedding. Different from the PVT model, Bahng et al. [27] further proposed a method to learn a single disturbance to adjust the pixel space and affect the model output. These studies suggest that continuously adjusting and optimizing prompts can enhance the adaptability of model. Lee et al. [29] treats different scenarios of missing modalities as different types of inputs and employs learnable prompts to guide the predictions of model under various missing conditions. Qiu et al. [30] utilizes an intermediate classifier to generate a prompt for each missing scenario based on intermediate features for segmentation prediction. The difference is that our work does not require learning a set of prompts for each missing scenario but aims to learn generic visual prompts and generalize them to modulate feature space in missing scenes.

3. Materials and method

In this paper, we focus on brain tumor segmentation under common missing modality scenarios. We simulate real-world data incompleteness by assuming absences of one or multiple modalities

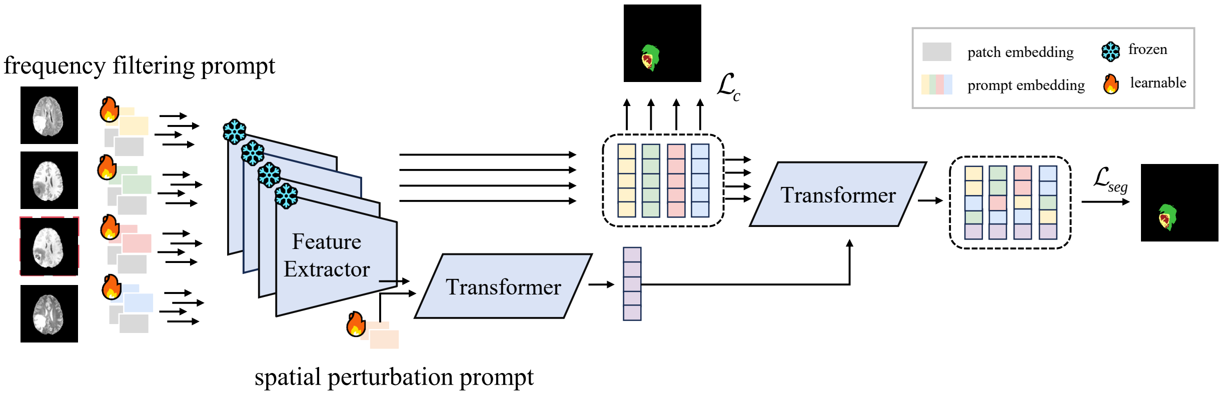


Figure 3. The proposed DPONet framework. It takes MRI images as input. Each image is combined with frequency filtering prompts and fed into a pre-trained transformer network to extract discriminative features. Subsequently, the intermediate features extracted by four encoders are integrated with spatial perturbation prompts to learn consistent features within a shared latent space. Finally, the fused discriminant features and consistent features are input into the decoder to get the segmentation map.

(Figure 2). Additionally, due to the difficulty of fully training a pre-trained transformer with limited computational resources, we design a discriminative prompt optimization network that avoids fine-tuning the entire pre-trained model. In this section, we will elaborate on the framework and its components.

3.1. Preliminary and notation

The pyramid vision transformer (PVT) [31] introduces a progressive shrinking strategy within the transformer block to control the scale of feature maps for dense prediction tasks. We chose the backbone is initialized with the weights pre-trained on ImageNet. PVT comprises four stages, each consisting of a patch embedding layer and l transformer encoder layers, which generate feature maps of different scales. Given an input image $X \in \mathbb{R}^{H\times W\times C}$, the patch embedding layer divides the sample X into $\frac{HW}{p_i}$ non-overlapping patches, where p_i represents the patch size of the i-th layer. As the stage progresses, the patch size decreases accordingly. The flattened patches are then fed into a linear projection to obtain embedded patches. The embedded patches, along with positional embedding information, are subsequently input into the transformer encoder to produce a feature map x of size $\frac{H}{p_i} \times \frac{W}{p_i} \times C$.This process can be described as follows:

$$x^l = MLP(LN(SRA(x^{l-1}))), \tag{3.1}$$

where $\mathbf{x}^{l-1}$ represents the feature map output from the previous layer, $SRA(\cdot)$ denotes the spatial reduction attention proposed in PVT, and $LN(\cdot)$ and $MLP(\cdot)$ refer to normalization and multi-layer perceptron operations, respectively. SRA is like multi-head attention. The formula is as follows:

$$SQA = Attention(QW^Q, SRA(K)W^K, SR(V)W^V), \tag{3.2}$$

where W^Q, W^K, and W^V are the parameters of the linear projections. $SRA(\cdot)$ is used to reduce the spatial dimension. This can be expressed as:

$$SRA(x) = LN(Reshape(x_i, r_i)W^S), \tag{3.3}$$

The r_i represents the feature map reduction rate for stage i.

The $Reshape(\cdot)$ operation reshapes the input $x \in \mathbb{R}^{h_i \times w_i \times c_i}$ to $\frac{h_i w_i}{r_i^2} \times (r_i^2 c_i))$. The W^S is a linear projection that reduces the dimensionality of the input. The attention calculation is as follows:

$$Attention(q, k, v) = Softmax(\frac{qk^T}{\sqrt{d}})v, \tag{3.4}$$

where q, k and v are the query, key, and value transform matrices, and d is the dimension.

3.2. Problem definition

We consider a multimodal dataset consisting of $N(N = 4)$ modalities, M = FLAIR, T1CE, T1 and T2. The dataset is denoted as $\mathcal{D} = \mathcal{D}_1 4, \mathcal{D}_1 3, \ldots, \mathcal{D}_i, \ldots, \mathcal{D}_0$, where $\mathcal{D}_1 4$ represents the complete set of modalities, and other sets represent missing modalities subsets, such as $\mathcal{D}_0 = X_0^F, X_0^{T1c}, X_0^{T1}, X_1^{T2}$ indicating only T2 mode is available. X_k^m represents the input sample, where m denotes the modality type, and k represents the modal state. For the model, it is unaware of which specific modality is missing, therefore, we introduce placeholder values (set to 0) to assign to the missing modality data $X_0^F, X_0^{T1c}, X_0^{T1}, X_0^{T2}$ to ensure the format of the multimodal input.

3.3. Overall framework

We propose a novel discriminative prompt optimization network, as shown in Figure 3, which aims to provide natural insertion points for intermediate features of the network while preserving the integrity of the pre-trained model and enabling fine-tuning for downstream tasks. We adapt a pre-trained transformer as feature extractor and keep it frozen during training. Multimodal images $\mathrm{D} = \{\mathrm{X}_m^k\}^{k=[0,1]}$ are fed into four extractors, and task-relevant information is aggregated through discriminative prompts to fully exploiting the discriminative features. Next, a spatial perturbation prompt module is introduced, which hierarchically fuses the discriminative features of available modalities and maps them to a shared feature representation space to learn cross-modal shared information. Furthermore, the fused features are mapped back to the original input size through up-sampling in the decoder, and the resulting segmentation masks are obtained from these feature maps. Notably, during training, the trainable parameters are confined to the prompt components and the decoder.

3.4. Frequency filtering prompt

The frequency filtering prompt method, as illustrated in Figure 4, utilizes Fourier transform to extract frequency features and jointly modulates the intermediate features with image embeddings. The frequency processing method decomposes images into different frequency components, which are distributed across different spatial locations of the image, encouraging the model to focus on critical information of the image [21]. The core idea is to remodulate the intermediate features using

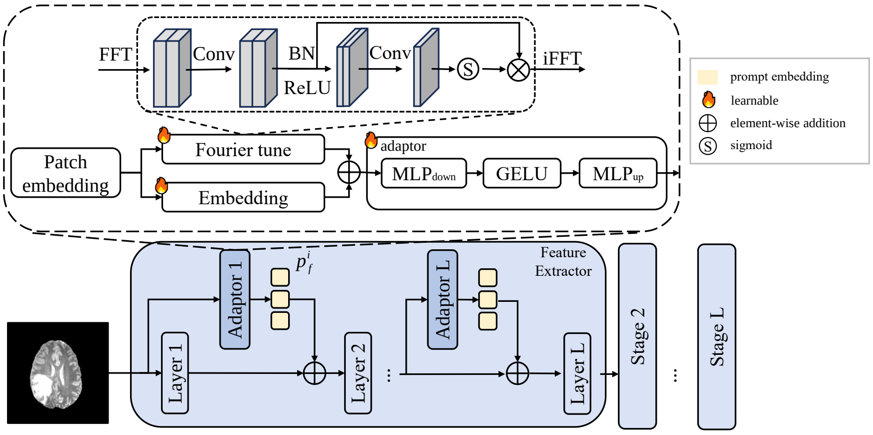


Figure 4. The architecture of the proposed frequency filtering prompt (FFP). The image is mapped into patch embeddings through a linear layer. The frequency filtering prompt method splits these embeddings into two branches for processing. One branch undergoes frequency filtering operations to obtain high-frequency features, while the other branch remains unprocessed. The combination of these two branches will generate prompts through an adaptor. The frequency filtering prompt and the image embeddings go through transformer blocks to extract discriminative features.

frequency domain information, shifting the distribution from the pre-trained dataset to the target dataset. Furthermore, since there may be commonalities between features of different modalities, even if image data from a particular modality is missing, the remaining modalities still contain corresponding frequency information, which enhances the robustness of the information to a certain extent. Taking a single branch as an example, for a given image, we apply the fast Fourier transform (FFT) along the spatial dimension to obtain frequency components corresponding to different spatial locations. FFT is applied to each channel to convert the spatial domain representation into a frequency representation in the frequency domain, and filtering operations are performed in the frequency domain. Then, an attention mask is learned in the frequency domain to analyze the dominant frequency components in the feature map. Finally, the feature representation is transformed back to the spatial domain using inverse FFT (iFFT). The transformation from the spatial domain to the frequency domain is expressed as follows:

$$\mathcal{F}(x)(\mu, \upsilon) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h, w)\, e^{-i2\pi\left(\frac{h\mu}{H} + \frac{w\upsilon}{W}\right)}, \tag{3.5}$$

After obtaining the frequency representation, different frequency components are modulated by filtering through the attention mechanism. Specifically, the attention mechanism compresses information across channels through convolution and a sigmoid function. The expression of the frequency filtering mechanism is as follows:

$$\mathcal{F}'(x) = F_x \otimes \sigma(conv([AvgPool(F_x), Maxpool(F_x)])), \tag{3.6}$$

where, σdenotes the Sigmoid function, $AvgPool(\cdot)$ and $MaxPool(\cdot)$ represent the average pooling and max pooling operations respectively.

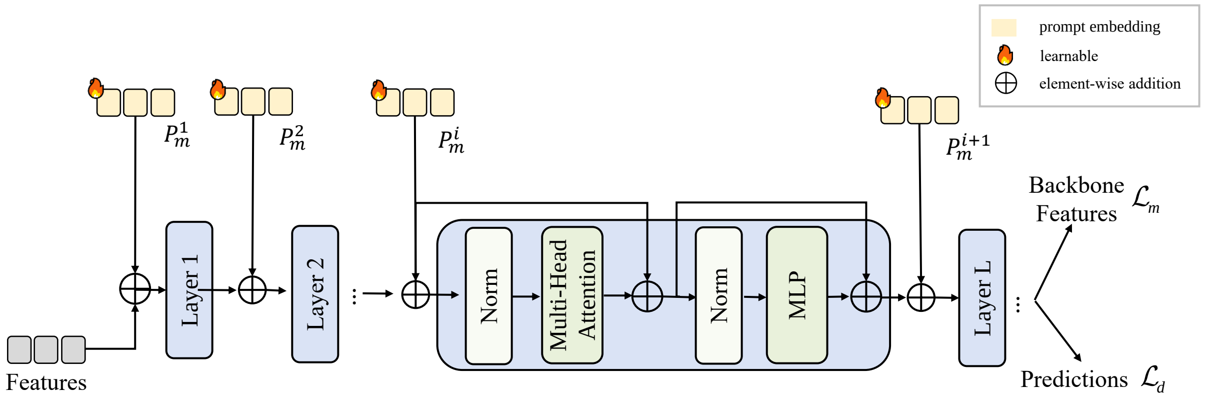


Figure 5. The architecture of the proposed spatial perturbation prompt (SPP). Intermediate features and prompt embeddings are combined with input into the transformer block and utilizing consistency loss to facilitate the learning of prompts.

Finally, the inverse FFT is used to transform back to the spatial domain features:

$$x'(h,w) = \frac{1}{H \cdot W}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\mathcal{F}'(x)\, e^{i2\pi\left(\frac{h\mu}{H}+\frac{wv}{W}\right)}, \tag{3.7}$$

Inspired by AdaptFormer [32], we employ a frequency enhancement adaptor, a bottleneck structure that limits the number of parameters. It takes the combination of filtered frequency features and image features as input and generates relevant frequency prompts through a down-projection layer, a lightweight multi-layer perceptron, and an up-projection layer. Formally, this process can be expressed as:

$$p_f^i = MLP_{up}(GELU(MLP_{down}^i(x' + x))), \tag{3.8}$$

Thirdly, the generated prompts are appended to the transformer layers to facilitate the model in learning more representative and discriminative image features.

3.5. Spatial perturbation prompt

To enable the model to handle missing modalities, we employ null values for filling, however, such null values are likely to disturb the feature space and result in failure of modal feature fusion. Therefore, we propose learnable spatial perturbation prompts, as show in Figure 5, aiming to learn a task-specific visual prompt (P) within a latent space that encourages the sharing of cross-modal information. Prompts interact dynamically with input features, facilitating adaptive modal fusion rather than simply injecting fixed information using learning prompts.

First, the extracted discriminative features are concatenated through element-wise addition $f_c^i = [f_{\mathrm{f}}^i, f_{t1c}^i, f_{t1}^i, f_{t2}^i]$ and then passed through a 3 × 3 convolutional layer followed by a Sigmoid activation function to generate prompt weights $\omega_{\mathrm{i}} \in [0,1]$. These weights describe the importance of each spatial data point in the input. Inspired by EVP [27], we add random visual embeddings of the same size as the transformer tokens, train only these random embeddings in the training phase, and the trained visual prompts as the guidance for the model, denoted as $F^i = (F_{token}^i, p_m^i)$. The process can be described as:

$$\omega_i = \sigma(conv([f_{\mathrm{f}}^i, f_{t1c}^i, f_{t1}^i, f_{t2}^i])), \tag{3.9}$$

$$p_m^i = conv(\sum_{c=1}^{N} \omega_i p_c^i), \tag{3.10}$$

$$F^i = transformer(f_c^i + p_m^i), \tag{3.11}$$

where, σ is the Sigmoid function. Finally, the cross-modal information features (F) are fed into Transformer encoder block to establish cross-modal long-range dependencies.

We introduce a consistency loss to optimize the prompts to capture task-shared knowledge and transform it into representations that are beneficial for the task. Specifically, we map the feature maps obtained from the transformer encoder stages to the same size as the input image and use mean squared error ensuring that the model learns coherent and consistent information at each stage. Note that, since shallower layers may lack sufficient semantic information, we apply the consistency loss only in the last two stages of the transformer encoder.

$$\mathcal{L}_m = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}(\hat{f}_i - f_i^m)^2, \tag{3.12}$$

where, N is the number of samples, M is the number of decoder layers, and the rescaled features of images in transformer layer m, and their average is denoted as $\hat{f}_i = \frac{1}{m}\sum_{k=1}^{m} f_i^k$.

In addition, we mapped the feature map into a segmentation map, and calculated Dice loss from the ground truth to prompt the model capture consistent feature representations.

$$\mathcal{L}_d = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} Dice(y_i - f({x_i}^m)), \tag{3.13}$$

where, y_i denotes the ground-truth labels of the image x_i, and $f({x_i}^m)$ denotes the prediction corresponding to the m-th layer features of the image.

The feature consistency loss and prediction consistency loss are combined to supervise prompt generation.

$$\mathcal{L}_c = \gamma\mathcal{L}_m + (1-\gamma)\mathcal{L}_d, \tag{3.14}$$

where, γ is the weight parameter used to balance the two losses. We experiment with different values of γ and found that $\gamma = 0.3$ gives the best result.

3.6. Convolutional decoder

The convolutional decoder gradually restores the spatial resolution from the fused features to the original segmentation space. The convolutional decoder employs skip connections to merge features from different modalities at specific hierarchical levels into the encoder, to preserve more low-level details. Therefore, the overall processing steps are as follows:

$$D_i = conv(upsample(conv(f_c^i, D_{i-1}))), \tag{3.15}$$

where D_i is the feature map from the i-th layer of the convolutional decoder, and f_c^i is the combined feature from multiple encoder layers.

3.7. Loss function

We employ a hybrid loss to measure the difference between the predictions and the ground truth. Dice Loss is used to calculate the similarity between the predicted segmentation result and the true segmentation result. Cross-entropy loss measures the prediction performance by quantifying the difference between the predicted probability distribution and the true probability distribution. Gradients are calculated based on the feedback of the sum of the two losses to update the parameters. The definition is as follows:

$$\mathcal{L}_{Dice} = -\frac{2\sum_i^N y_i f(x_i)}{\sum_i^N y_i + \sum_i^N f(x_i)}, \tag{3.16}$$

$$\mathcal{L}_{CE} = -\sum\nolimits_i^N y_i \log p(f(x_i)), \tag{3.17}$$

where $f(x_i)$ and y_i represent the prediction and ground-truth labels, respectively. Besides, N is the number of pixels, $p(\cdot)$ is the SoftMax of the prediction. Last, our hybid loss function $\mathcal{L}_{seg}$ can be given by

$$\mathcal{L}_{seg} = \mathcal{L}_c + \mathcal{L}_{Dice} + \mathcal{L}_{CE}, \tag{3.18}$$

4. Experiments

4.1. Datasets

We use two public datasets from the Multimodal Brain Tumor Segmentation Challenge (BraTS) to demonstrate the effectiveness of the proposed method, BraTS 2018 and BraTS 2020 [33–35]. BraTS 2018 contains 285 cases of patients for training, while BraTS 2020 includes 369 cases for training and 125 for validation. In these datasets, each case comprises four MRI modalities: Flair, T1ce, T1, and T2. The volume of each modality is 240 × 240 × 155, aligned within the same spatial space. Medical experts provide manual pixel-level annotations of three mutually inclusive tumor regions in each image, namely, whole tumor (WT), tumor core (TC), and enhancing tumor (ET). WT encompasses all tumor tissues, while TC comprises ET, necrosis, and non-enhancing tumor core.

4.2. Data preprocessing

Data preprocessing is performed on the two datasets before training. For each dataset, we slice along the axial plane of the 3D medical images. To eliminate non-informative slices and irrelevant background regions, thereby saving training efficiency and time, we use central slices as the training data and reshape each 2D slice to 224 × 224. We design a simulation method for missing modalities. The MRI modalities are randomly removed from the input. The missing modality can be any one or multiple modalities, and the missing rate for each modality is random. The purpose of this is to simulate the scenario where missing modalities may occur in real-world situations.

4.3. Implementation details and evaluation metrics

In this study, our method is implemented in Pytorch utilizing a single NVIDIA Tesla V100 32 GB GPU. We adopt the U-Net architecture composed of transformer blocks as the benchmark, and the transformer is pre-trained on ImageNet-1K. We utilize the SGD optimizer with an initial learning rate of 0.01. After many experiments and parameter tuning, we set our model to train 100 epochs with an initial learning rate of $1e{-}2$ and a batch size of 12. For the segmentation task, we use the Dice coefficient (which computes the similarity of two sets), the Hausdorff distance (HD95, which measures the distance between two sets), and the sensitivity (the ratio of the number of positive samples correctly identified by the model to the number of all true positive samples) as performance metrics to evaluate various methods.

5. Results

We focus on exploring the robustness of discriminative optimization networks to general incompleteness in multimodal image without fine-tuning the entire pretraind model. In this chapter, we first introduce the excellent results obtained by our method. Subsequently, a series of ablation experiments on the proposed components. Considering that the BraTS 2020 dataset contains many patient cases and is representative, we experimented with it in the ablation study.

Table 1. Quantitative results of state-of-the-art unified models (Ding [36], Zhang [37], Ting [17], Qiu [30]), and our DPONet on the BraTS2020 dataset. ✓ indicates available modalities. Bold indicates optimal, underline indicates sub-optimal.

Modalities				Dice (%) ↑														
				Complete					Core					Enhancing				
F	T1	T1c	T2	D	Z	T	Q	Our	D	Z	T	Q	Our	D	Z	T	Q	Our
			✓	86.1	86.1	86.5	86.7	93.9	71.0	70.9	71.5	71.0	93.3	46.3	46.3	45.6	47.2	76.1
		✓		76.8	78.5	77.4	79.5	91.6	81.5	84.0	83.4	84.3	95.3	74.9	80.1	78.9	81.4	88.4
	✓			77.2	78.0	78.1	79.5	89.1	66.0	65.9	66.8	67.7	91.9	37.3	38.0	41.3	39.1	71.6
✓				87.3	87.4	89.1	86.9	95.2	69.2	68.8	69.3	69.9	93.5	38.2	42.4	43.6	42.8	74.6
		✓	✓	87.7	87.8	88.4	88.4	94.5	83.5	84.8	86.4	86.3	95.8	75.9	79.4	81.7	80.1	88.9
	✓	✓		81.1	81.8	81.2	83.1	92.1	83.4	83.6	85.2	85.8	95.4	78.0	80.1	79.2	81.7	88.3
✓	✓			89.7	89.8	89.9	89.8	95.5	73.1	73.8	73.9	74.4	94.3	41.0	45.9	48.2	46.8	77.3
	✓		✓	87.7	87.8	88.0	87.9	94.4	73.1	73.4	73.3	72.9	94.1	45.7	46.8	50.1	47.3	77.5
✓			✓	89.9	89.9	90.5	90.1	95.5	74.1	74.6	75.5	74.5	94.1	49.3	48.6	48.6	49.5	76.6
✓		✓		89.9	89.3	90.0	90.0	95.6	84.7	84.8	85.5	86.6	95.9	76.7	81.9	81.8	81.2	88.9
✓	✓	✓		90.7	90.1	90.7	90.6	95.6	85.1	85.2	86.5	86.7	95.8	76.8	82.1	81.8	81.8	88.8
✓	✓		✓	90.6	90.6	90.3	90.6	95.7	75.2	75.6	75.9	75.8	94.7	49.9	50.3	52.5	51.1	78.0
✓		✓	✓	90.7	90.4	90.6	90.8	95.8	85.0	85.3	86.4	86.4	96.0	77.1	78.7	81.0	80.0	88.9
	✓	✓	✓	88.3	88.2	88.7	88.9	94.6	83.5	84.2	86.5	86.5	95.8	77.0	79.3	78.5	82.1	88.9
✓	✓	✓	✓	91.1	90.6	90.6	91.0	95.9	85.2	84.6	87.4	86.4	95.9	78.0	79.9	81.6	81.0	88.9
Average				87.0	87.1	87.3	<u>87.6</u>	**94.3**	78.2	78.6	79.6	<u>79.7</u>	**94.8**	61.5	64.0	<u>64.9</u>	<u>64.9</u>	**82.8**

5.1. Comparison with other methods

As shown in Table 1, our method achieves remarkable performance in Dice score on both the modality-complete and modality-missing scenarios. For example, our proposed approach has significantly better mean Dice scores for whole tumors, tumor cores, and enhanced tumors than suboptimal approaches. From the experimental results in Table 2, we observed that the baseline model generally exhibited unsatisfactory performance on the T1 modality. However, our model achieved significant improvements in this aspect, effectively enhancing the performance under the T1 modality. In Figures 6 and 7, we present the visualization of segmentation results. Furthermore, Table 3 clearly exhibits that our method outperforms other approaches in terms of HD95 and sensitivity under complete modality testing, further validating the superior performance of our approach.

We further conducted experiments to analyze the robustness of our proposed method to varying missing modality rates between the training and testing phases. As shown in Figure 8(a), we trained the model with a 70% missing rate and randomly removed multiple modalities to simulate modality missing scenarios for testing. We found that, compared to the baseline, our DPONet method was robust to different missing rates during testing. Moreover, in Figure 8(b), where we used 10%, 70%, and 90% to represent the degree of missingness during training (through many experiments, we found that these missing rates are representative), we observed that when training with more complete modality data, the performance was significantly higher when testing with low missing rates. In this paper, the experiments based on the general reality that collecting complete modality data cannot be guaranteed. However, there are still some publicly available datasets with complete modalities. Therefore, we trained the models using complete data, as shown in Figure 8(c), where the baseline model could not handle data missing, our method consistently improved upon the baseline.

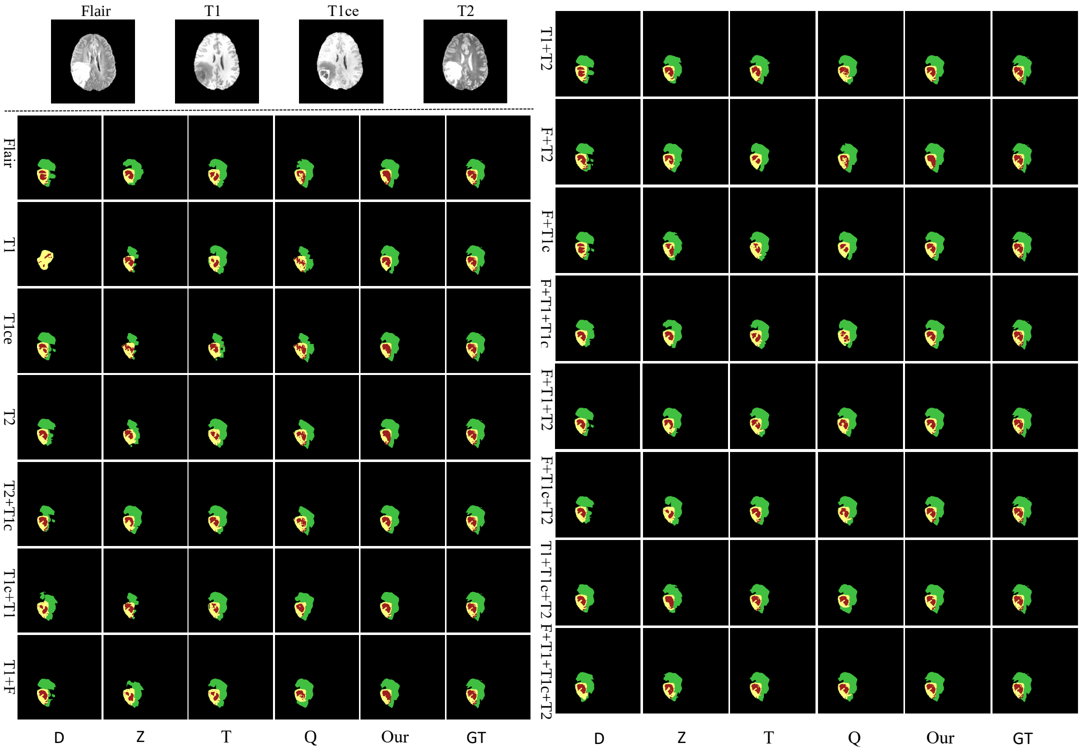


Figure 6. Visual comparison results of state-of-the-art unified models and our proposed DPONet on the BraTS2020 dataset.

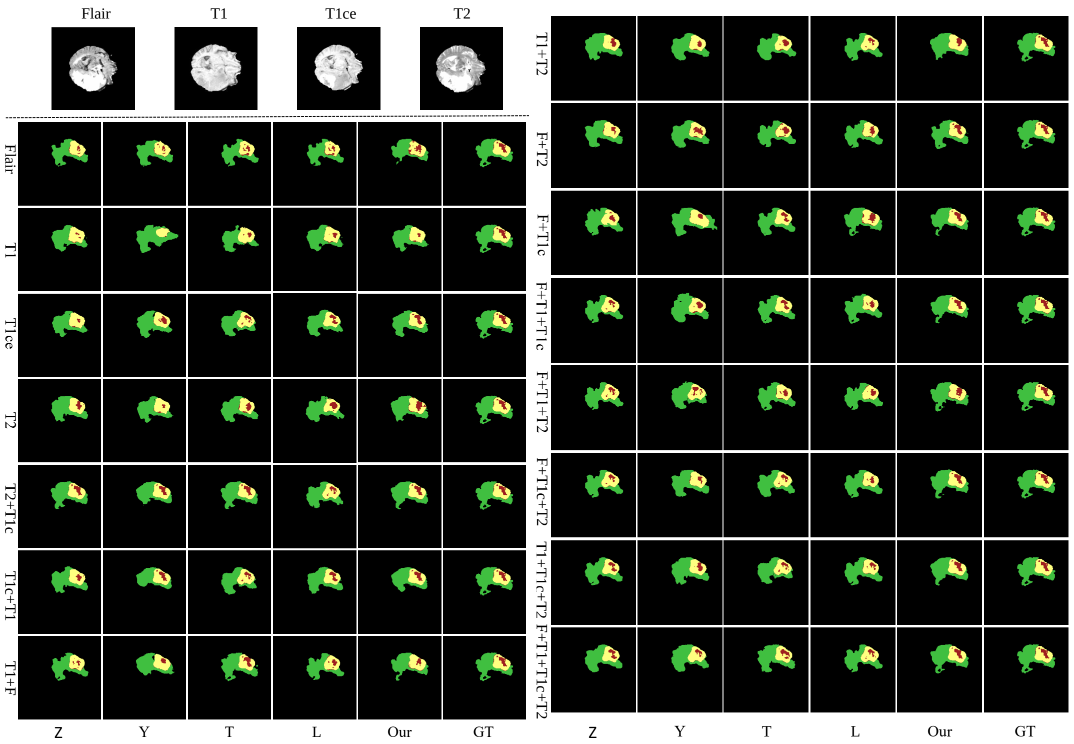
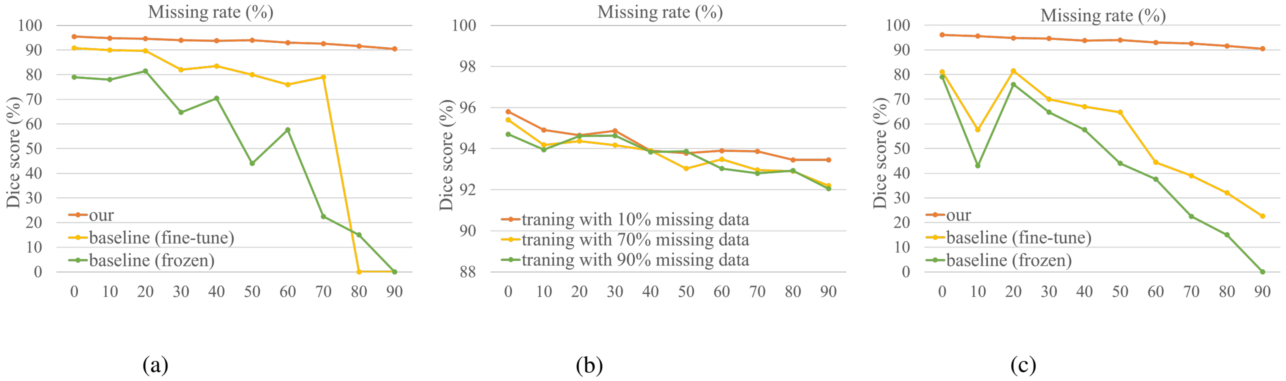


Figure 7. Visual comparison results of state-of-the-art models and our proposed DPONet on the BraTS2020 dataset.



(a)

(b)

(c)

Figure 8. Study on the robustness of DPONet to testing missing rates under different scenarios (where the absence of one, two, or three modalities is random, to account for the possible missing modalities during testing). (a) All models are trained under a 70% missing rate and evaluated under varying missing rates. (b) Training with different missing rates scenarios with 10%, 70%, and 90% missing rates (through many experiments, we found that these missing rates are representative), representing data with higher modality completeness, balanced data, and data with lower modality completeness, respectively. (c) All models are trained with modality-complete data.

Table 2. Quantitative results of state-of-the-art unified models (Zhang [37], Yang [38], Ting [17], Liu [18] and our DPONet on the BraTS2018 dataset. ✓ indicates available modalities. Bold indicates optimal, underline indicates sub-optimal.

Modalities				Dice (%) ↑														
				Complete					Core					Enhancing				
F	T1	T1c	T2	Z	Y	T	L	Our	Z	Y	T	L	Our	Z	Y	T	L	Our
			✓	81.2	76.3	86.6	84.8	94.3	64.2	56.7	68.8	69.4	94.4	43.1	16.0	41.4	47.6	76.2
		✓		72.2	42.8	77.8	75.8	92.6	75.4	65.1	81.5	82.9	95.4	72.6	66.3	75.7	73.7	89.2
	✓			67.5	15.5	78.7	74.4	90.9	56.6	16.8	65.6	66.1	93.2	32.5	8.1	44.5	37.1	74.7
✓				86.1	84.2	88.4	88.7	95.2	61.2	47.3	66.7	66.4	94.2	39.3	8.1	40.5	35.6	74.8
		✓	✓	83.0	84.1	88.2	86.3	95.0	78.6	80.3	84.8	84.2	96.1	74.5	68.7	77.7	75.3	90.0
	✓	✓		74.4	62.1	81.8	77.2	93.1	78.6	78.2	83.5	83.4	95.7	74.0	70.7	77.1	74.7	89.5
✓	✓			87.1	87.3	89.7	89.0	95.6	65.9	61.6	72.0	70.8	95.2	43.0	9.5	44.4	41.2	77.9
	✓		✓	82.2	84.2	88.4	88.7	94.9	61.2	47.3	66.7	66.4	95.1	45.0	16.5	47.7	48.7	77.7
✓			✓	87.6	87.9	90.3	89.9	95.9	69.8	62.6	71.8	70.9	95.1	47.5	17.4	48.3	45.4	78.1
✓		✓		87.1	87.5	89.5	89.7	95.6	77.9	80.8	84.8	84.4	96.1	75.1	64.8	76.8	75.0	90.0
✓	✓	✓		87.3	87.7	90.4	88.9	95.7	79.8	80.9	85.2	84.1	96.2	75.5	65.7	77.4	74.0	90.0
✓	✓		✓	87.8	88.4	89.7	89.9	96.0	71.5	63.7	74.1	72.7	95.5	47.7	19.4	50.0	44.8	78.7
✓		✓	✓	88.1	88.8	90.6	90.4	96.0	79.6	80.7	85.8	84.6	96.3	75.7	66.4	76.6	73.8	90.1
	✓	✓	✓	82.7	80.9	88.4	86.1	95.1	80.4	79.0	85.8	84.4	96.2	74.8	68.3	78.5	75.4	90.1
✓	✓	✓	✓	89.6	88.8	90.6	90.1	96.1	85.8	80.1	85.9	84.5	96.3	77.6	68.4	80.4	75.5	90.0
Average				82.9	76.4	87.3	86.0	**94.8**	72.4	65.4	77.5	77.0	**95.4**	59.9	42.3	62.5	59.9	**83.8**

Table 3. Quantitative results of the state-of-the-art unified models and our proposed DPONet on the BraTS2020 dataset. The models are evaluated using Dice, HD95, and sensitivity scores. Baseline (fine-tune) means that the pre-trained transformer feature extractor is fully fine-tuned on the target dataset. Baseline (frozen) indicates that the pre-trained transformer feature extractor is frozen.

Method	Dice ↑				HD95 ↓				Sensitivity ↑			
	WT	TC	ET	Avg	WT	TC	ET	Avg	WT	TC	ET	Avg
Ding et al.	86.13	71.93	58.98	72.35	-	-	-	-	-	-	-	-
Zhang et al.	87.08	78.69	64.08	76.62	2.90	6.21	44.64	17.92	99.60	99.81	99.82	99.74
Ting et al.	90.71	84.60	79.07	84.79	4.05	5.78	33.77	14.53	90.98	83.90	77.68	84.18
Qiu et al.	87.58	79.67	64.87	77.37	2.82	5.71	43.92	17.48	99.66	99.83	99.81	99.77
baseline(fine-tune)	77.63	78.94	70.85	93.56	2.61	2.09	2.39	2.36	86.28	86.50	82.74	85.17
baseline(frozen)	58.11	61.09	40.88	89.16	2.83	2.29	2.97	2.70	81.41	84.68	85.90	84.00
our	94.96	94.12	89.98	93.02	2.58	2.09	2.21	2.29	96.81	96.32	93.01	95.38

5.2. Ablation study

We explored the effects of frequency filtering prompts and spatial perturbation prompts, the results showing in Table 4, our method achieved a higher Dice score of 93.02. The term *baseline (fine-tune)* refers to a pre-trained transformer that is comprehensively fine-tuned on the BraTS dataset. The term *baseline (frozen)* refers to a baseline model where the pre-trained backbone parameters are frozen.

We introduced frequency filtering prompts into the baseline model, the model achieved comparable performance to fine-tuned model, demonstrating the efficiency of proposed component. Furthermore, as shown in Figure 9, during training with complete modalities, when a significant portion of modalities were absent during inference (i.e., retaining only one modality), the baseline model suffered a severe performance degradation. Excitingly, when prompts were introduced, the model was able to perform image segmentation normally even with a single modality input, indicating that the proposed visual prompts facilitated the encoder to learn discriminative features across modalities.

We introduced the spatial perturbation prompts module into the baseline, the overall robustness of the model was improved. As shown in Table 4, our method achieved a higher Dice score of 93.02, exceeding the baseline model by 17.21. Furthermore, the Dice score for the ET region saw a significant increase, indicating that the spatial perturbation prompt facilitated the fusion of inter-modal information and preserved more edge details and small-scale information. Figure 10 visualizes the segmentation results before and after using the spatial perturbation prompt, clearly demonstrating that more small-scale lesion areas are preserved.

Table 4. Ablation study of our proposed DPONet on the BraTS2020 dataset. The models are evaluated using Dice, HD95, and sensitivity scores. Baseline (fine-tune) means that the pre-trained transformer feature extractor is fully fine-tuned on the target dataset. Baseline (frozen) indicates that the pre-trained transformer feature extractor is frozen.

Method	Dice ↑				HD95 ↓				Sensitivity ↑			
	WT	TC	ET	Avg	WT	TC	ET	Avg	WT	TC	ET	Avg
baseline (fine-tune)	77.63	78.94	70.85	75.81	2.61	2.09	2.39	2.36	86.28	86.50	82.74	85.17
baseline (frozen)	58.11	61.09	40.88	53.36	2.83	2.29	2.97	2.70	81.41	84.68	85.90	84.00
baseline + FFP	93.65	92.40	85.08	90.38	2.45	2.04	2.16	2.22	96.54	96.11	91.26	94.64
baseline + SPP	94.56	94.40	87.37	92.11	2.47	2.05	2.22	2.25	96.59	96.07	90.53	94.40
baseline + FFP + SPP	94.96	94.12	89.98	93.02	2.58	2.09	2.21	2.29	96.81	96.32	93.01	95.38

Table 5. The number of model parameters (10^6) before and after adding the learnable prompt component.

Method	Param (M)	Tunable Param (M)
baseline (fine-tune)	194.82	194.82
baseline (frozen)	194.82	49.30
baseline + FFP	160.42	58.97
baseline + SPP	173.93	48.69
baseline + FFP + SPP	153.43	10.58

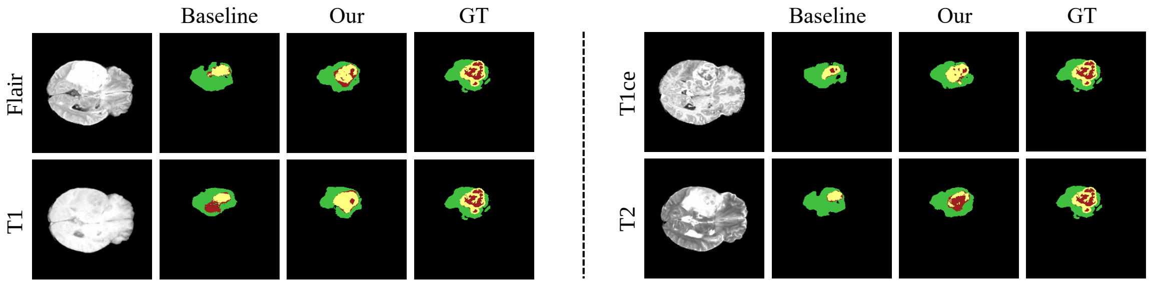


Figure 9. Qualitative results from state-of-the-art models and our DPONet, which was trained using the complete modal dataset of BraTS2020 and randomly missing three modalities with a 70% miss rate during the test phase.

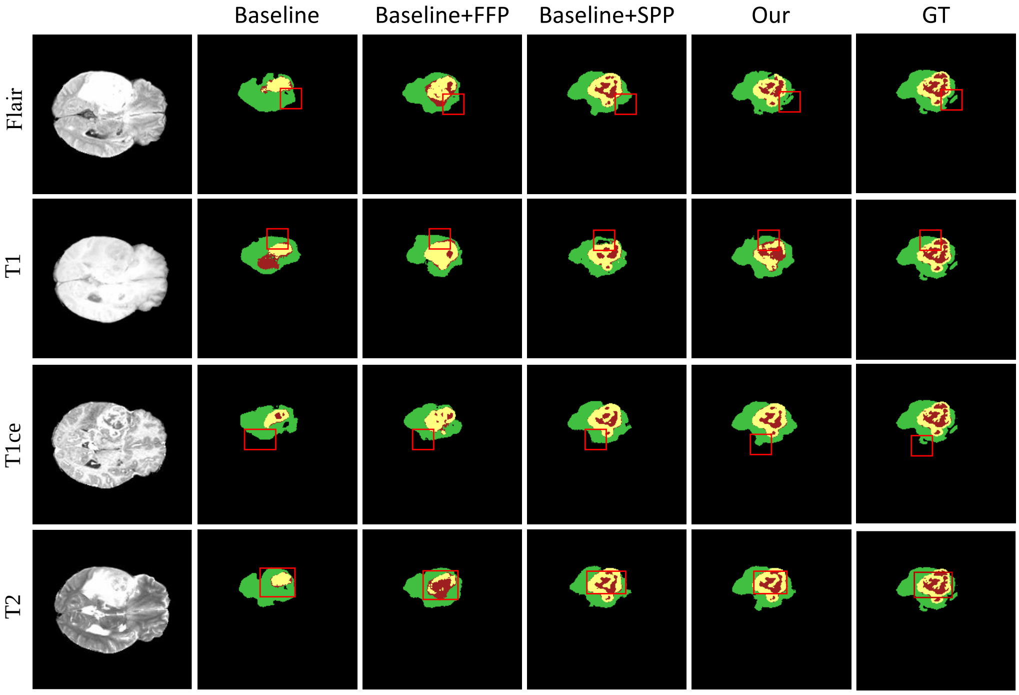


Figure 10. Qualitative results from DPONet, which was trained using the complete dataset of BraTS2020 and randomly missing three modalities with a 70% miss rate during the test phase. The red box indicates the progress of DPONet.

Additionally, in Table 5, we described the parameter information before and after adding the module. It indicates that our method only introduced approximately 7% of the total trainable parameters but achieved excellent segmentation performance. Once extended to large models with billions of parameters, our proposed method will be more favorable and suitable for multimodal downstream tasks with missing modalities, achieving a favorable trade-off between computational cost and performance.

6. Conclusions

In this paper, we introduce a parameter-efficient and discriminatively optimized segmentation network that exhibits robust adaptability to generalized missing modality inputs. Our model filters frequency features to generate discriminative visual cues and introduces learnable spatial perturbation prompts into shared feature representations, effectively addressing the challenge of incomplete multimodal brain tumor segmentation. Compared to fine-tuning the entire transformer model, our approach requires only 7% of the trainable parameters while demonstrating superior performance in handling real-world scenarios with missing modality data. Extensive experiments and ablation studies on the publicly available BraTS2018 and BraTS2020 datasets validate the effectiveness of our proposed method.

7. Limitations and future works

In this work, we investigate a parametrically efficient incomplete modal image segmentation method for brain tumors. Although our model successfully captures consistent features by mapping robust multimodal features to the same potential space, we must point out that our model cannot recover information about missing modalites from available multimodal inputs. Therefore, our next plan will study how to use the available multimodal image to estimate the missing modal information to obtain rich image information.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (No.U24A20231, No.62272283), New Twentieth Items of Universities in Jinan (No.2021GXRC049).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Soomro T, Zheng L, Afifi A, Ali A, Soomro S, Yin M, et al. (2022) Image segmentation for MR brain tumor detection using machine learning: A review. *IEEE Rev Biomed Eng* 16: 70–90. https://doi.org/10.1109/RBME.2022.3185292

2. Li S, Du C, Zhao Y, Huang Y, Zhao H, (2023) What makes for robust multi-modal models in the face of missing modalities?. preprint, arXiv:2310.06383. https://doi.org/10.48550/arXiv.2310.06383

3. Choi Y, Al-Masni M, Jung K, Yoo RE, Lee SY, Kim DH, (2023) A single stage knowledge distillation network for brain tumor segmentation on limited MR image modalities. *Comput Methods Programs Biomed* 240: 107644. https://doi.org/10.1016/j.cmpb.2023.107644

4. Wang S, Yan Z, Zhang D, Wei H, Li Z, Li R, (2023) Prototype knowledge distillation for medical segmentation with missing modality. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095014

5. Chen Q, Zhang J, Meng R, Zhou L, Li Z, Feng Q, et al. (2024) Modality-specific information disentanglement from multi-parametric MRI for breast tumor segmentation and computer-aided diagnosis. *IEEE Trans Med Imaging* 43: 1958–1971. https://doi.org/10.1109/TMI.2024.3352648

6. Chen C, Dou Q, Jin Y, Chen H, Qin J, Heng PA, (2019) Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019* 11766: 447–456. https://doi.org/10.1007/978-3-030-32248-9_50

7. Zhang C, Chu X, Ma L, Zhu Y, Wang Y, Wang J, et al. (2022) M3care: Learning with missing modalities in multimodal healthcare data. *Assoc Comput Mach* 2022: 2418–2428. https://doi.org/10.1145/3534678.353938

8. Wang H, Chen Y, Ma C, Avery J, Hull L, Carneiro G, (2023) Multi-modal learning with missing modality via shared-specific feature modelling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15878–15887.

9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) Attention is all you need. Advances in neural information processing systems. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000–6010.

10. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. (2020) Language models are few-shot learners. *Adv Neural Inf Process Syst* 33: 1877–1901.

11. Zhou K, Yang J, Loy CC, Liu Z, (2022) Learning to prompt for vision-language models. *Int J Comput Vision* 130: 2337–2348. https://doi.org/10.1007/s11263-022-01653-1

12. Zhou K, Yang J, Loy CC, Liu Z, (2022) Conditional prompt learning for vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16816–16825. https://doi.org/10.1109/CVPR52688.2022.01631

13. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y, (2022) Coca: Contrastive captioners are image-text foundation models. preprint, arXiv:2205.01917. https://doi.org/10.48550/arXiv.2205.01917

14. Wang Q, Mao Y, Wang J, Yu H, Nie S, Wang S, et al. (2023) Aprompt: Attention prompt tuning for efficient adaptation of pre-trained language models. *Assoc Comput Linguist* 2023: 9147–9160. https://doi.org/10.18653/v1/2023.emnlp-main.567

15. Dehghani M, Djolonga J, Mustafa B, Padlewski P, Heek J, Gilmer J, et al. (2023) Scaling vision transformers to 22 billion parameters. In: *Proceedings of Machine Learning Research* 202: 7480–7512.

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

17. Ting H, Liu M, (2023) Multimodal transformer of incomplete MRI data for brain tumor segmentation. *IEEE J Biomed Health Inf* 28: 89–99. https://doi.org/10.1109/JBHI.2023.3286689

18. Liu H, Wei D, Lu D, Sun J, Wang L, Zheng Y, (2023) M3AE: Multimodal representation learning for brain tumor segmentation with missing modalities. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37: 1657–1665. https://doi.org/10.1609/aaai.v37i2.25253

19. Tran A, Mathews A, Xie L, Ong CS, (2021) Factorized fourier neural operators. preprint, arXiv:2111.13802. https://doi.org/10.48550/arXiv.2111.13802

20. Chen Y, Ren Q, Yan J, (2022) Rethinking and improving robustness of convolutional neural networks: A shapley value-based approach in frequency domain. *Adv Neural Inf Process Syst* 35: 324–337. https://dl.acm.org/doi/10.5555/3600270.3600294

21. Krishnamachari K, Ng S, Foo C, (2023) Fourier sensitivity and regularization of computer vision models. preprint, arXiv:2301.13514. https://doi.org/10.48550/arXiv.2301.13514

22. Fang C, Zhang D, Wang L, Zhang Y, Cheng L, Han J, (2022) Cross-modality high-frequency transformer for mr image super-resolution. In: *Proceedings of the 30th ACM International Conference on Multimedia* 1584–1592. https://doi.org/10.1145/3503161.3547804

23. Xu Z, Gong H, Wan X, Li H, (2023) Asc: Appearance and structure consistency for unsupervised domain adaptation in fetal brain mri segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 14226: 325–335. https://doi.org/10.1007/978-3-031-43990-2_31 .

24. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. (2019) Parameter-efficient transfer learning for NLP. In: *International Conference on Machine Learning* 2790–2799.

25. Li X, Liang P, (2021) Prefix-tuning: Optimizing continuous prompts for generation. *Assoc Comput Linguist* 4582–4597. https://doi.org/10.18653/v1/2021.acl-long.353

26. Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, et al. (2022) Visual prompt tuning. In: *Proceedings of the 17th European Conference on Computer Vision, Springer* 13693: 709–727. https://doi.org/10.1007/978-3-031-19827-4_41

27. Bahng H, Jahanian A, Sankaranarayanan S, Isola P, et al. (2022) Exploring visual prompts for adapting large-scale models. preprint, arXiv:2203.17274. https://doi.org/10.48550/arXiv.2203.17274

28. Wang Z, Zhang Z, Lee CY, Zhang H, Sun R, Ren X, et al. (2022) Learning to prompt for continual learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 139–149.

29. Lee Y, Tsai Y, Chiu W, et al. (2023) Multimodal prompting with missing modalities for visual recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 14943–14952.

30. Qiu Y, Zhao Z, Yao H, Chen D, Wang Z, (2023) Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. *Assoc Comput Mach* 2023: 3228–3239. https://doi.org/10.1145/3581783.361171

31. Wang W, Xie E, Li X, Fan DP, Song K, Liang D, et al. (2021) Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* 568–578.

32. Chen S, Ge C, Tong Z, Wang J, Song Y, Wang J, et al. (2022) Adaptformer: Adapting vision transformers for scalable visual recognition. *Adv Neural Inf Process Syst* 35: 16664–16678.

33. Menze B, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. (2014) The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* 34: 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

34. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. (2017) Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data* 4: 1–13. https://doi.org/10.1038/sdata.2017.117

35. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. (2018) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. preprint, arXiv:1811.02629. https://doi.org/10.48550/arXiv.1811.02629

36. Ding Y, Yu X, Yang Y, (2021) RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: *IEEE/CVF International Conference on Computer Vision (ICCV)* 3975–3984. https://doi.org/10.1109/ICCV48922.2021.00394

37. Zhang Y, He N, Yang J, Li Y, Wei D, Huang Y, et al. (2022) Mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* 13435: 107–117. https://doi.org/10.1007/978-3-031-16443-9_1

38. Yang Q, Guo X, Chen Z, Woo PYM, Yuan Y, (2022) D2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Trans Med Imaging* 41: 2953–2964. https://doi.org/10.1109/TMI.2022.3175478