FreqDF: Frequency-guided Disentanglement Framework for Multi-modal Brain Tumor Segmentation

Ruibin Li[1], Meiyi Wei[1], Chudi Hu[1], Zhongling Wu[1] and Gang Chen[1*]
[1]School of Cyber Science and Engineering, Wuhan University, Wuhan, China
{liruibin, meiyiwei_mia, chudihu, 2024202210034, chenzuolin}@whu.edu.cn

***Abstract*—Accurate segmentation of brain tumors from multi-modal Magnetic Resonance Imaging (MRI) is critical for clinical diagnosis and treatment planning. Multi-modal learning aims to capture modality-shared global shapes while preserving modality-specific contrasts and textures. However, existing approaches often extract and integrate features indiscriminately, leading to feature redundancy and blurred boundaries. To address these issues, we introduce a novel Frequency-guided Disentanglement Framework (FreqDF) for multi-modal brain tumor segmentation. Our framework employs a dual-path encoding architecture to explicitly model modality-shared anatomical structures and modality-specific texture details in parallel. Based on these decoupled representations, we design a two-stage Enhancement-Fusion Module to collaboratively integrate features, ensuring low-frequency structural coherence while selectively refining high-frequency fine-grained edges. To guarantee effective disentanglement, we introduce a Frequency-Domain Constraint strategy. During training, we impose consistency on low-frequency components to extract unified anatomical shapes while maximizing the discrepancy between high-frequency components to preserve unique contrasts and textures. Extensive experiments and comprehensive ablation studies demonstrate that our method achieves superior segmentation accuracy. The code will be available soon.**

***Index Terms*—Medical Image Segmentation, Multi-modal Learning, Feature Disentanglement, Frequency Domain Analysis**

I. Introduction

Brain tumors are critical neurological conditions characterized by high morbidity and significant mortality. Treatment strategies depend heavily on the tumor's type, size, and precise location. Therefore, accurate diagnosis and segmentation are critical prerequisites for effective treatment planning. Magnetic Resonance Imaging (MRI) is widely used in clinical practice for the diagnosis of brain tumors. Typical MRI comprises multiple imaging modalities, including T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR) [1]. These modalities provide a comprehensive clinical basis as T1 offers structural reference, T1ce highlights tumor cores, while T2 and FLAIR accentuate edema and inflammation [2]. Consequently, precise automated segmentation that effectively exploits these complementary features has become a critical research focus for assisting diagnosis and treatment planning [2]–[4].

Over the past decade, encoder-decoder CNN architectures like 3D UNet [5], V-Net [6], and nnUNet [7] have established a strong foundation for volumetric medical segmentation. To further exploit complementary information from multi-modal data, feature disentanglement has been introduced to decompose data into independent latent factors. Seminal works like SDNet [8] pioneered the separation of anatomical spatial factors from appearance, while recent approaches [2], [9]–[11] further explored shared and specific representations to reduce redundancy. However, these methods often rely on implicit adversarial training or complex auxiliary tasks lacking of explicit guidance. Effectively fusing these decoupled features to reconstruct sharp boundaries remains a critical challenge, as naive aggregation frequently leads to information loss.

Due to the strong correlation between frequency and image semantics [12], [13], frequency-domain analysis is a compelling solution. Specifically, low-frequency components primarily encode global spatial shapes, such as the overall outline of brain structures, which tend to be consistent across modalities and correspond to modality-shared features. In contrast, high-frequency components capture fine-grained contrasts, textures, and edges that reflect unique signal intensities in specific sequences, thereby relating to modality-specific features [14]. This inherent spectral decomposition aligns with the objective of separating shared shapes and specific textural details, providing a prior for feature disentanglement.

In this paper, we propose a novel **Frequency-guided Disentanglement Framework (FreqDF)** based on this spectral prior. During inference, our framework employs a dual-path encoding architecture to decompose global shapes and texture details in parallel. To effectively merge these distinct information streams, we design a two-stage **Enhancement-Fusion Module (EFM)**. This module recalibrates the shared features to maintain structural shape integrity while selectively incorporating specific details to sharpen boundaries. During training, we introduce an explicit **Frequency-Domain Constraint (FDC)** strategy to guide the learning process. By transforming intermediate feature maps into the spectral domain, we impose consistency constraints on low-frequency components to encourage a unified shape representation, while simultaneously maximizing the discrepancy of high-frequency components to prevent information leakage and preserve unique textural contrasts. In summary, our key contributions are as follows:

- We propose a **Frequency-guided Disentanglement Framework (FreqDF)** that explicitly separates multi-

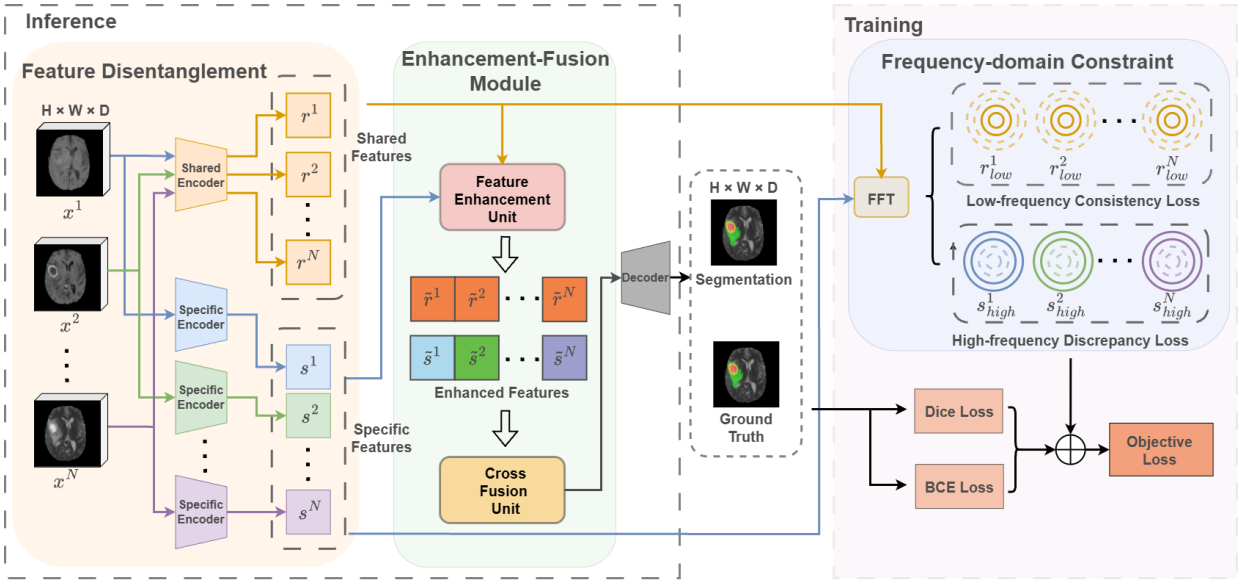


Fig. 1: Overview of FreqDF. The framework disentangles inputs into modality-shared and specific features, which are integrated via the Enhancement-Fusion Module (EFM) for final segmentation. A Frequency-Domain Constraint is employed during training to explicitly supervise this separation via low-frequency consistency and high-frequency discrepancy losses.

modal data into consistent anatomical shapes and unique textural details, supervised by a novel **Frequency-Domain Constraint (FDC)** strategy.

- We design a two-stage **Enhancement-Fusion Module (EFM)**, comprising a Feature Enhancement Unit (FEU) and a Cross-Fusion Unit (CFU), which dynamically recalibrates features and collaboratively merges complementary streams to ensure precise boundary delineation.
- Extensive experiments and ablation studies demonstrate that our framework achieves superior segmentation accuracy, validating the effectiveness of the proposed frequency-guided disentanglement and fusion strategies.

II. Related Works

A. Multimodal Medical Image Segmentation

Medical image segmentation has evolved from standard CNN-based architectures, such as U-Net [15], to Transformer-based hybrid models like TransUNet [16] and UNETR [17]. In multimodal tasks, the prevailing *input-level fusion* strategy implicitly assumes semantic alignment across sequences, often forcing the network to learn a mixed distribution that obscures unique modality features such as edema visible in T2 sequences. Although recent approaches utilize sophisticated attention or state-space mechanisms for dynamic fusion [18]–[21], most methods still lack an explicit mechanism to model *modality-shared* anatomy and *modality-specific* textures, limiting the exploitation of cross-modal complementarity.

B. Disentangled Representation Learning

Disentangled representation learning is widely used in computer vision to separate structural content and textural style, and has recently been extended to multimodal medical imaging to address feature redundancy. Early works focused on semi-supervised cardiac [22] or brain tumor segmentation [23], [24]. Recent studies [9], [25] have further refined these strategies using contrastive or style-mixup mechanisms. However, such methods often rely on complex adversarial training or auxiliary losses. Moreover, the definition of *shared* and *specific* factors in these frameworks remains implicit and learned in a black-box manner, lacking a direct and explicit guidance mechanism.

C. Frequency-Domain Learning

Frequency-Domain analysis provides a different perspective from spatial-domain convolutions, capturing global context and repeating patterns efficiently [26]. In medical imaging, spectral representations have been widely used for reconstruction and synthesis [27], [28] to preserve high-frequency details. However, the potential of frequency analysis for semantic segmentation, particularly for feature disentanglement in multimodal settings, remains under-explored. Recent studies in natural image analysis suggest that neural networks exhibit spectral bias [12], [14], with low-frequency components correlating with shape and structure, and high-frequency components corresponding to texture and noise.

III. Methods

A. Overall Architecture

In this section, we present the Frequency-guided Disentanglement Framework(FreqDF) as shown in Figure 1, comprising two key components: a disentanglement network that separates modality-shared and modality-specific features, and an Enhancement-Fusion Module (EFM) that integrates these representations. During training, we introduce a Frequency-Domain Constraint (FDC) to explicitly supervise disentanglement.

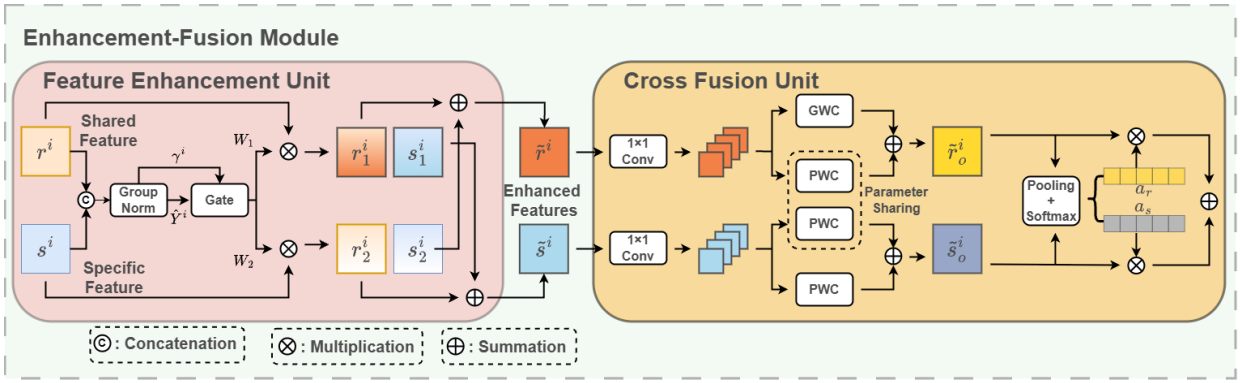


Fig. 2: Architecture of the Enhancement-Fusion Module (EFM). It comprises a Feature Enhancement Unit (FEU) for cross-branch feature recalibration via Group Norm-based gating, and a Cross-Fusion Unit (CFU) that employs differentiated convolutions and adaptive attention for final feature integration.

a) Disentanglement Strategy: Let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ denote the input set of N modalities. As illustrated in Figure 1, our framework utilizes a dual-path architecture to construct disentangled subspaces, consisting of a unified Modality-Shared Encoder (E^{sha}) and N independent Modality-Specific Encoders (E_i^{sp}). In the shared path, all inputs share the weights of E^{sha} to extract shared features $r^i \in \mathbb{R}^{C\times H^L\times W^L\times D^L}$. In parallel, each modality is processed by its specific encoder E_i^{sp} to generate specific features s^i of identical dimensions.

b) Integration Strategy: Effective integration of the disentangled shared (r^i) and specific (s^i) features is critical. Since naive concatenation or summation ignores the inherent semantic and spectral gaps between these representations, we propose a two-stage Enhancement-Fusion Module (EFM). Comprising a Feature Enhancement Unit (FEU) and a Cross-Fusion Unit (CFU), this module recalibrates and collaboratively fuses the disentangled features to ensure structural coherence and boundary precision.

B. Enhancement-Fusion Module

1) Feature Enhancement Unit: As shown in Figure 2, the FEU leverages Group Normalization (GN) statistics to suppress noise and selectively emphasize informative regions. Let $Y^i = \text{Concat}[s^i, r^i]$ denote the concatenated features. We apply GN to normalize Y^i:

$$\hat{Y}_c^i = \gamma \frac{Y_c^i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta, \tag{1}$$

where c denotes the c-th channel, μ and σ are the group mean and standard deviation, ϵ is a constant for numerical stability, and $\gamma, \beta \in \mathbb{R}^C$ are learnable affine parameters. Crucially, we interpret the scale parameter γ_c as an indicator of channel significance, where larger magnitudes correspond to richer spatial information rather than noise.

We derive a channel-wise weight vector w_c^i to modulate activation by normalizing γ and applying non-linear scaling to the features $\hat{Y}_c^i$:

$$w_c^i = \sigma\left(\hat{Y}_c^i \odot \frac{\gamma_c^i}{\sum_{j=1}^{C}\gamma_j^i + \epsilon}\right), \tag{2}$$

where C is the number of channels, $\sigma(\cdot)$ is the Sigmoid function, and $\odot$ denotes element-wise multiplication.

Subsequently, a gating mechanism generates complementary weight masks $W_1^i, W_2^i \in \mathbb{R}^{C\times H^L\times W^L\times D^L}$ based on a threshold $\tau = 0.5$:

$$W_1 = \begin{cases} 1 & \text{if } w > \tau \\ w & \text{otherwise,} \end{cases} \quad W_2 = \begin{cases} 0 & \text{if } w > \tau \\ w & \text{otherwise,} \end{cases} \tag{3}$$

These generated weights are employed to recalibrate the input features. To facilitate information exchange while preserving distinctiveness, we employ a cross-injection strategy. Let Ω_{sp} and Ω_{sha} denote the channel indices for specific and shared features within Y^i. We apply W_1^i to highlight dominant information and W_2^i to extract complementary context. The enhanced specific ($\tilde{s}^i$) and shared ($\tilde{r}^i$) features are derived as:

$$\begin{aligned} s_1^i &= \left(W_1^i \odot Y^i\right)_{\Omega_{sp}}, & r_1^i &= \left(W_1^i \odot Y^i\right)_{\Omega_{sha}}, \\ s_2^i &= \left(W_2^i \odot Y^i\right)_{\Omega_{sp}}, & r_2^i &= \left(W_2^i \odot Y^i\right)_{\Omega_{sha}}, \\ \tilde{s}^i &= s_1^i + r_2^i, & \tilde{r}^i &= r_1^i + s_2^i, \end{aligned} \tag{4}$$

where $(\cdot)_\Omega$ denotes channel slicing. These ensures the specific branch absorbs necessary structural context from the shared branch without being overwhelmed by it, and vice versa.

2) Cross-Fusion Unit: To seamlessly merge the enhanced features, we introduce the Collaborative Fusion Unit (CFU). First, the enhanced features $\tilde{s}^i$ and $\tilde{r}^i$ are squeezed into $\tilde{s}_z^i, \tilde{r}_z^i \in \mathbb{R}^{\frac{C}{2}\times H^L\times W^L\times D^L}$ via 1×1 convolutions.

Considering the distinct nature of the two branches, we apply differentiated convolution strategies: Group-wise Convolution (GWC) for the shared branch to aggregate correlated global structures, and Point-wise Convolution (PWC) for the specific branch to preserve independent edge details without blurring. A parameter-shared PWC (PWC$'$) is also applied to both streams to facilitate cross-stream information flow.

$$\tilde{r}_o^i = GWC(\tilde{r}_z^i) + PWC'(\tilde{r}_z^i),$$
$$\tilde{s}_o^i = PWC(\tilde{s}_z^i) + PWC'(\tilde{s}_z^i), \tag{5}$$

Finally, an adaptive selection mechanism fuses these streams. We first capture global context via Global Average Pooling to generate descriptors P_s^i and P_r^i. Channel-wise attention vectors $a_r^i, a_s^i \in \mathbb{R}^C$ are then computed via Softmax:

$$a_r^i = \frac{e^{P_r^i}}{e^{P_r^i} + e^{P_s^i}}, \quad a_s^i = \frac{e^{P_s^i}}{e^{P_r^i} + e^{P_s^i}} \tag{6}$$

where $a_r^i + a_s^i = 1$. The final fused feature F_o^i is obtained by dynamically weighting the branches:

$$F_o^i = \tilde{r}_o^i \odot a_r^i + \tilde{s}_o^i \odot a_s^i. \tag{7}$$

This allows the network to dynamically balance modality-specific details and modality-shared structures, achieving a robust and precise representation.

C. Frequency Constraint Loss

To explicitly decouple features, we introduce a frequency-domain supervision scheme. Let $F \in \mathbb{R}^{C \times D^L \times H^L \times W^L}$ denote an intermediate feature map. We transform F into the frequency domain and shift the zero-frequency component to the center via the 3D Fast Fourier Transform (FFT). To separate components, we define a binary mask $\mathcal{M}$ active only in the central region determined by a bandwidth ratio $\rho = 0.5$. For a frequency coordinate (u, v, w), $\mathcal{M}(u, v, w) = 1$ if the coordinate lies within the central ρ-proportion, and 0 otherwise. We apply this mask and perform Inverse FFT (IFFT) to generate shape-oriented low-frequency maps (F_{low}) and texture-oriented high-frequency maps (F_{high}):

$$\begin{aligned} \mathcal{Z} &= \text{FFT}(F), \\ F_{low} &= \text{IFFT}(\mathcal{Z} \odot \mathcal{M}), \\ F_{high} &= \text{IFFT}(\mathcal{Z} \odot (\mathbf{1} - \mathcal{M})), \end{aligned} \tag{8}$$

where $\mathcal{Z}$ is the complex spectrum and $\text{FFT}(\cdot)$ denotes the Fourier transformation operation, $\mathbf{1}$ denotes an all-ones matrix with the same size as $\mathcal{M}$ and $\text{IFFT}(\cdot)$ is the inverse Fourier transform. Through this process, we effectively separate the signal into shape-oriented low-frequency components and texture-oriented high-frequency components.

Low-frequency Consistency Loss: To ensure that shared features capture invariant anatomical structures, we impose a consistency constraint by minimizing the Mean Squared Error (MSE) between the spatially reconstructed low-frequency maps of different modalities:

$$\mathcal{L}_{low} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{1}{d} \left\| r_{low}^i - r_{low}^j \right\|_2^2 \right), \tag{9}$$

where $d = C \times D^L \times H^L \times W^L$ represents the feature dimensions, N is the number of modalities, r_{low}^i denotes the low-frequency feature map from the i-th shared encoder, and $\| \cdot \|_2$ denotes the L_2 norm.

High-frequency Discrepancy Loss: Conversely, to prevent information leakage and encourage modality-specific details (e.g., textures), we maximize the discrepancy between specific features. This is formulated as minimizing the negative Sigmoid of their pairwise MSE:

$$\mathcal{L}_{high} = - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sigma \left(\frac{1}{d} \left\| s_{high}^i - s_{high}^j \right\|_2^2 \right), \tag{10}$$

where s_{high}^i represents the high-frequency map from the i-th specific encoder. The Sigmoid function $\sigma(\cdot)$ acts as a bounded mapping to stabilize optimization as the distance increases.

Total Objective: The overall training objective integrates voxel-wise segmentation supervision with these frequency-guided constraints:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \alpha \cdot \mathcal{L}_{low} + \beta \cdot \mathcal{L}_{high}, \tag{11}$$

where $\mathcal{L}_{seg} = \mathcal{L}_{Dice} + \mathcal{L}_{BCE}$ combines Dice and Binary Cross-Entropy losses. The hyperparameters α and β balance the contribution of the consistency and discrepancy constraints.

IV. Experiments

A. Experimental Setup

1) Datasets and Evaluation Metrics: We evaluate our method on the BraTS 2018 and 2020 datasets [29], comprising multi-modal scans (T1, T1ce, T2, FLAIR; $240 \times 240 \times 155$). Both datasets are randomly split into training, validation, and testing sets (7:1:2) to segment the Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC). Segmentation performance is measured using the Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95). The ablation studies are mainly performed on the BraTS 2018.

2) Implementation Details: Our framework utilizes a ResNet50 backbone with Feature Pyramid Encoders, implemented in PyTorch 1.13.1 on Python 3.9.18. Models trained from scratch on two Tesla V100 GPUs for 40,000 iterations via Adam (batch size 2/GPU). We set the initial learning rate to 1×10^{-4} with a 2,000-iteration warm-up and cosine decay. Standard augmentations (random flipping, cropping, and intensity shifting) are applied to mitigate overfitting.

B. Comparison with Other Methods

We compare our approach against representative CNN baselines, including CNN-based models: 3D U-Net, Residual 3D U-Net [5], and a ResNet50-based backbone [30]. Additionally, we compare against CNN-Transformer hybrid architectures, including UNETR [17], SwinUNETR [31] and TransBTS [32].

Table I summarizes the quantitative results on the BraTS 2018 and 2020 datasets. Overall, our framework achieves the best performance with the highest Mean DSC (84.93% and 87.87%) and the lowest Mean HD95 (3.27mm and 3.55mm). Compared to the ResNet50 baseline, our method yields substantial improvements across metrics, notably a 2.45% DSC increase in the 2018 Tumor Core (TC) and significant HD95 reductions on both datasets. This performance gain validates our frequency-guided supervision: low-frequency constraints

TABLE I: Quantitative comparison on the BraTS 2018 and BraTS 2020 datasets. The best results are **bolded** and the second-best are underlined. ↑ indicates higher is better, and ↓ indicates lower is better.

Methods	BraTS 2018								BraTS 2020							
	Dice Score (%) ↑				HD95 (mm) ↓				Dice Score (%) ↑				HD95 (mm) ↓			
	ET	TC	WT	Mean	ET	TC	WT	Mean	ET	TC	WT	Mean	ET	TC	WT	Mean
3D U-Net	72.91	84.64	90.95	82.84	2.35	4.22	4.54	3.70	80.76	89.64	92.55	87.65	3.07	6.39	7.39	5.62
Residual 3D U-Net	71.42	76.01	91.12	79.51	4.24	8.13	6.36	6.24	79.78	88.21	92.24	86.74	5.17	7.61	8.53	7.10
ResNet50 (Baseline)	**74.97**	85.44	91.26	83.89	2.25	4.30	4.73	3.76	80.56	89.42	**92.74**	87.57	2.71	6.28	5.38	4.79
UNETR	69.69	80.68	89.55	79.97	3.52	8.19	9.96	7.22	79.01	86.55	91.29	85.62	3.86	7.56	11.19	7.54
SwinUNETR	71.34	75.04	90.94	79.11	2.47	6.37	7.19	5.34	80.71	88.38	91.24	86.77	4.90	8.01	10.56	7.82
TransBTS	68.35	70.55	87.90	75.60	5.18	10.77	9.49	8.48	79.79	85.91	91.30	85.67	5.28	12.16	10.73	9.39
Ours	74.86	**87.89**	**92.03**	**84.93**	**2.14**	**4.21**	**3.46**	**3.27**	**80.96**	**89.90**	92.73	**87.87**	**2.60**	**3.17**	**4.86**	**3.55**

preserve holistic structures (aiding WT), while high-frequency discrepancy sharpens boundaries and internal variations (improving TC and HD95). Although Enhancing Tumor (ET) performance remains comparable to the baseline, the superior aggregated metrics confirm the robustness of our collaborative fusion strategy.

As shown in Figure 3, our method achieves higher boundary precision, successfully delineating fine-grained details. The predicted tumor margins are sharper and align closely with the ground truth, even in complex transition regions.

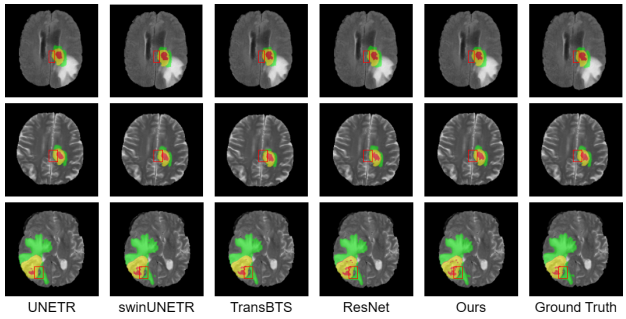


Fig. 3: Visual comparison of segmentation results

C. Ablation Study

We analyze the contribution of each component, the selection of FDC function, and the impact of hyperparameters.

TABLE II: Ablation study on the effectiveness of different components and the bandwidth ratio of FDC.

DF	FDC	EFM	DSC (%)			
			ET	TC	WT	Mean
-	-	-	**74.97**	85.44	91.26	83.89
✓	-	-	72.56	87.18	90.83	83.52
✓	✓	-	73.72	88.06	90.94	84.24
✓	-	✓	73.36	87.80	91.88	84.35
✓	✓	✓	74.86	**87.89**	**92.03**	**84.93**
✓	$\rho = 0.2$	✓	70.86	82.98	88.65	80.83
✓	$\rho = 0.8$	✓	71.26	83.33	89.42	81.34

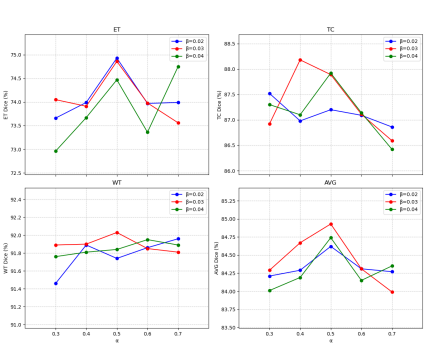
1) Effectiveness of Key Components: Table II summarizes the ablation of the Disentanglement Framework (*DF*), Frequency-Domain Constraints (*FDC*), and Enhancement-Fusion Module (*EFM*).

As shown, naively applying *DF* causes a slight performance drop due to feature ambiguity. However, adding *FDC* resolves this by explicitly separating modality-specific textures (benefiting ET/TC) from shared shapes, improving Mean DSC to 84.24%. Integrating *EFM* to collaboratively fuse these refined features yields the highest Mean DSC of **84.93%** with a notable **2.45%** TC gain over the baseline. Furthermore, varying the FDC bandwidth ratio (ρ) reveals that extreme values ($\rho = 0.2$ or 0.8) significantly degrade accuracy, confirming that a balanced frequency partition ($\rho = 0.5$) is crucial for optimal disentanglement.

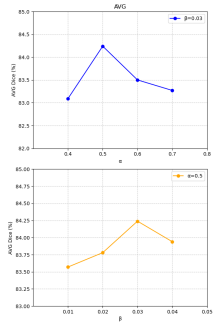
TABLE III: Ablation study on frequency loss functions.

Loss Function	DSC (%)			
	ET	TC	WT	Mean
Cosine Similarity	74.84	87.56	91.91	84.77
KL divergence	**74.90**	**88.14**	91.72	84.92
Mean Squared Error(Ours)	74.86	87.89	**92.03**	**84.93**

2) Selection of Frequency Loss Function: As shown in Table III, we evaluate three metrics for frequency constraints: Cosine Similarity (COS), KL Divergence (KL), and Mean Squared Error (MSE). While distance-based metrics (KL and MSE) consistently outperform the direction-based COS, MSE achieves the highest overall Mean DSC of **84.93%** by most effectively capturing spectral magnitude discrepancies.



(a) Grid search of parameters



(b) Sensitivity analysis of parameters

3) Impact of Hyperparameters: Grid search yields optimal hyperparameters of $\alpha = 0.5$ and $\beta = 0.03$ (Figure 4a). Further sensitivity analysis (Figure 4b) shows that performance degrades significantly if weights exceed these ranges, suggesting that excessive auxiliary loss weighting disrupts the gradient flow of the primary segmentation task.

V. Conclusion

In this paper, we propose a novel Frequency-guided Disentanglement Framework (FreqDF) for multi-modal brain tumor segmentation. By exploiting spectral analysis as a physical prior, our method explicitly decouples modality-shared anatomical structures from modality-specific edge details. To guarantee robust separation, we introduce the Frequency-Domain Constraint (FDC), which imposes consistency in low-frequency shapes while maximizing discrepancy in high-frequency textures. Finally, the Enhancement-Fusion Module (EFM) collaboratively integrates disentangled features, ensuring structural coherence and precise boundary delineation. Extensive experiments and comprehensive ablation studies demonstrate that our method achieves superior segmentation accuracy and effective disentanglement.

References

[1] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI brainlesion workshop*. Springer, 2018, pp. 311–320.

[2] T. Zhou, "Multi-modal brain tumor segmentation via disentangled representation learning and region-aware contrastive learning," *Pattern Recognition*, vol. 149, p. 110282, 2024.

[3] D. Li, B. Yang, W. Zhan, and X. He, "Multi-category graph reasoning for multi-modal brain tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 445–455.

[4] Z. Zhang, Y. Lu, F. Ma, Y. Zhang, H. Yue, and X. Sun, "Incomplete multi-modal brain tumor segmentation via learnable sorting state space model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 982–25 992.

[5] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[6] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[7] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnU-Net: Self-adapting framework for U-Net-based medical image segmentation (2018)," *arXiv preprint arXiv:1809.10486*, 1809.

[8] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical Image Analysis*, vol. 58, p. 101535, 2019.

[9] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4237–4248, 2020.

[10] Z. Che, Z. Zhang, Y. Wu, and M. Wang, "Disentangle and then fuse: A cross-modal network for synthesizing gadolinium-enhanced brain MR Images," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[11] T. Zhou, Z. Wang, S. Ruan, Y. Meng, J. Duan, and B. Lei, "DFuse-Net: Disentangled feature fusion with uncertainty-aware learning for reliable multi-modal brain tumor segmentation," *Medical Image Analysis*, p. 103923, 2025.

[12] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," *arXiv preprint arXiv:1901.06523*, 2019.

[13] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4085–4095.

[14] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 5301–5310.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[17] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[18] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnFormer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[19] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, p. 1034, 2024.

[20] M. Zhou and F. Khalvati, "ClinicalFMamba: Mamba-based multimodal medical image fusion for enhanced clinical diagnosis."

[21] Q. Chen, J. Li, and X. Fang, "Dual triple attention guided CNN-VMamba for medical image segmentation," *Multimedia Systems*, vol. 30, no. 5, p. 275, 2024.

[22] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 490–498.

[23] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P.-A. Heng, "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 447–456.

[24] X. Liu, P. Sanchez, S. Thermos, A. Q. O'Neil, and S. A. Tsaftaris, "Learning disentangled representations in the imaging domain," *Medical Image Analysis*, vol. 80, p. 102516, 2022.

[25] Z. Cai, J. Xin, C. You, P. Shi, S. Dong, N. C. Dvornek, N. Zheng, and J. S. Duncan, "Style mixup enhanced disentanglement learning for unsupervised domain adaptation in medical image segmentation," *Medical Image Analysis*, vol. 101, p. 103440, 2025.

[26] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in neural information processing systems*, vol. 34, pp. 980–993, 2021.

[27] Y. Chen, X. Zhang, L. Peng, Y. He, F. Sun, and H. Sun, "Medical image segmentation network based on multi-scale frequency domain filter," *Neural Networks*, vol. 175, p. 106280, 2024.

[28] X. Zhang, X. Wang, and T. Niu, "Ct image segmentation using frequency domain feature-assisted selective long memory state space model," *Sensing and Imaging*, vol. 26, no. 1, p. 74, 2025.

[29] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 272–284.

[32] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 109–119.