DSS-NET: DOWNSTREAM TASK-DRIVEN MULTI-MODAL MEDICAL IMAGE FUSION VIA DECOUPLED SPATIAL-SPECTRAL INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient multi-modal medical image fusion is vital for precision diagnostics. However, existing methods face a dilemma: models focusing solely on image fusion often prioritize visual aesthetics while neglecting semantic details required for downstream task. Meanwhile, unified fusion-downstream networks often suffer from feature entanglement and progressive semantic attenuation due to insufficient cross-task interaction. To address this, we propose DSS-Net, a novel framework that harmonizes high-fidelity fusion with downstream task performance. Central to our architecture are two specialized modules designed to reconcile representational conflicts: the Spatial Complementary Interaction (SCI) and the Global Spectral Integration (GSI). Specifically, SCI enables inter-modal complementary modulation to extract spatial visual features, while GSI executes spectral-domain aggregation to derive modality-invariant semantic representations. These decoupled features are then processed by dual-path decoders to achieve high-fidelity fusion and superior downstream utility. By decoupling representations into vision-driven spatial enhancement and semantic-driven spectral alignment, DSS-Net preserves intricate anatomical details while boosting downstream effectiveness. Extensive experiments demonstrate state-of-the-art performance in fusion quality and downstream task accuracy.

1 INTRODUCTION

Multimodal Medical Image Fusion (MMIF) aims to integrate complementary information from diverse modalities into a cohesive image, supporting precise diagnosis, surgical navigation, and radiotherapy planning Du et al. (2017); Hill et al. (2016); James & Dasarathy (2014). Such unified representations synthesize high-fidelity cues that single modalities lack Liu et al. (2025); Xu et al. (2026); Bernardi et al. (2026); Li et al. (2025); Zhong et al. (2023), making the synergy between fusion and downstream analysis a cornerstone of modern computer-aided diagnosis.

Despite these advancements, prevailing methodologies predominantly treat fusion and segmentation as isolated or loosely coupled stages (Sequential Paradigm) Tang et al. (2022a); Huang et al. (2022). Optimizing purely for perceptual quality metrics rather than downstream clinical utility often leads to suboptimal segmentation masks Wang et al. (2024); Xu et al. (2023); Guan et al. (2025). This **task-independent** strategy lacks an explicit feedback loop, inadvertently suppressing vital anatomical cues and yielding a "fusion for visualization's sake" dilemma that limits subsequent diagnostic efficacy Guan et al. (2023); Lu et al. (2024).

Furthermore, emerging joint frameworks designed for natural images (e.g., RGB-Infrared) struggle with unique medical imaging challenges, including intrinsic noise, modality imbalance, and fine anatomical boundaries Sun et al. (2024); Chen et al. (2025). In medicine, such general-purpose "All-in-one" models suffer from severe feature entanglement and semantic attenuation. Therefore, a specialized architecture to decouple these representational conflicts is required.

To overcome these limitations, we propose **DSS-Net**, a novel framework achieving synergistic equilibrium between high-fidelity fusion and precise segmentation. Our architecture relies on two distinct modules to reconcile the inherent conflicts between low-level visual textures and high-level semantics. Specifically, the **Spatial Complementary Interaction (SCI)** module leverages the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

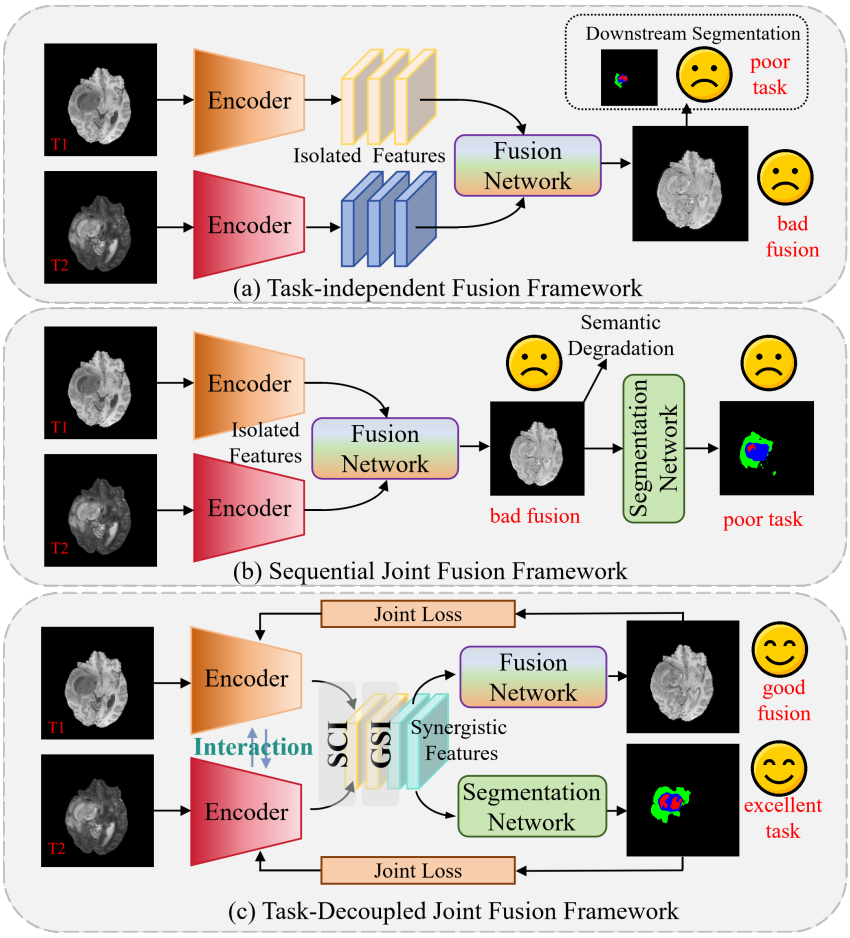

Figure 1: Evolutionary paradigms of multimodal fusion-segmentation. (a) **Task-independent Fusion**: Optimizes for generic visual metrics in isolation, remaining **agnostic** to downstream semantic requirements. (b) **Sequential Joint Fusion**: Bridges task via a unified pipeline but lacks deep interaction, resulting in **feature coupling and semantic attenuation**. (c) **Task-Decoupled Joint Fusion (Ours)**: Resolves representational conflicts by **disentangling** spatial-spectral features, achieving optimal synergy between visual realism and semantic precision.

linear-time complexity and global receptive field of State Space Models (SSMs) to achieve selective complementarity of multi-modal details, providing an enhanced structural foundation for the fusion branch. Concurrently, the **Global Spectral Integration (GSI)** module operates in the Fourier domain for cross-modal spectral fusion. Through spectral gating, GSI suppresses modality-specific noise and derives aligned, modality-invariant semantic representations for robust segmentation.

Briefly, the main contributions of this paper include:

- We design a **Spatial Complementary Interaction (SCI)** module leveraging State Space Models to achieve selective spatial complementarity, providing refined visual features for high-fidelity image fusion.
- We propose a **Global Spectral Integration (GSI)** module that executes cross-modal fusion in the Fourier domain, suppressing artifacts and deriving modality-invariant representations for precise segmentation.
- Extensive experiments demonstrate that DSS-Net achieves state-of-the-art performance, delivering superior fusion quality and highly accurate segmentation results.

2 Related Work

Deep learning has significantly advanced Multimodal Medical Image Fusion (MMIF) Shi et al. (2024); He et al. (2025); Zhu et al. (2024). Existing approaches mainly fall into two paradigms: **Task-independent Image Fusion**, focusing on pixel-level integration and visual fidelity Zhao et al. (2023b); Tang et al. (2022c); and **Joint Fusion and Segmentation Frameworks**, coupling fusion with clinical task for mutual enhancement Tang et al. (2022b); Zhong et al. (2025).

2.1 Task-independent Image Fusion

Early MMIF approaches relied on *CNNs* to capture local structures Deng et al. (2024); Liu et al. (2021); Zhao et al. (2023c); Wen et al. (2024). To overcome CNNs' limited receptive fields, *Vision Transformers (ViTs)* and hybrid architectures were later introduced for global context modeling, achieving a better balance between structural fidelity and contextual awareness Vaswani et al. (2017); Tang et al. (2022c); Cheng et al. (2025); Zhao et al. (2023b; 2024b).

Despite these architectural evolutions, mainstream methods predominantly follow a **Task-independent Fusion Framework** (Fig. 1a) Zhou et al. (2023), where image fusion and semantic segmentation are treated as isolated steps. Because the fusion network is optimized without high-level semantic feedback, the generated representations often lose critical diagnostic cues. This "task-agnostic" gap provides limited assistance for precise clinical decision-making Xu et al. (2022); Singh & Khare (2014); Ma et al. (2020).

2.2 Joint Fusion and Segmentation Frameworks

To bridge the semantic gap, joint frameworks integrate fusion and segmentation into unified pipelines Zhao et al. (2024a; 2023a). A common strategy is the **Sequential Joint Fusion Framework** (Fig. 1b), exemplified by SeAFusion Tang et al. (2022b) and SegMiF Liu et al. (2023). They cascade a fusion network with a standalone segmentation model. However, this creates a unidirectional information bottleneck. Without an effective decoupling mechanism, critical semantic cues are often degraded during fusion due to the lack of deep cross-task interaction.

Other joint models (e.g., PSFusion Tang et al. (2023), MRFS Zhang et al. (2024), AMDNet Zhong et al. (2025)) explore multi-task synergy but are primarily designed for natural images. When applied to complex medical datasets like BraTS Zhou et al. (2022), they struggle to reconcile intense representational conflicts. Lacking a specialized strategy to disentangle task-specific features, they fail to balance intricate texture preservation and precise tumor localization, resulting in a performance trade-off that limits clinical utility.

In Summary. As synthesized in Fig. 1, existing paradigms exhibit a clear evolutionary but flawed trajectory. The traditional task-independent paradigm (Fig. 1a) focuses exclusively on pixel-level aggregation for visual quality, remaining entirely agnostic to the requirements of downstream clinical task. While more recent sequential-joint frameworks (e.g., SeAFusion Tang et al. (2022b), SegMiF Liu et al. (2023)) attempt to bridge this isolation via a unified pipeline (Fig. 1b), they inherently suffer from feature coupling and a lack of deep cross-task interaction. Such a unidirectional bottleneck inevitably leads to semantic attenuation, where task-critical diagnostic cues are degraded during the fusion process, thereby limiting overall performance. To resolve these representational conflicts, we propose DSS-Net (Fig. 1c), a novel framework that explicitly disentangles spatio-spectral features to achieve an optimal synergy between visual realism and semantic precision.

To overcome these limitations, we propose **DSS-Net**, a **Task-Decoupled Joint Fusion Framework** as depicted in Fig. 1(c). Our core innovation lies in a sequential refinement strategy that progressively disentangles and aligns multi-modal features. Specifically, a **Spatial Complementary Interaction (SCI)** module is first employed to capture intricate spatial details through long-range dependency modeling. Building upon this, a **Global Spectral Integration (GSI)** module subsequently operates in the Fourier domain to achieve modality-invariant semantic alignment. By staging these operations, DSS-Net effectively reconciles representational conflicts before the features are processed by task-specific heads. Finally, a **Joint Loss** feedback loop ensures the synthesized images are optimized for both visual realism and robust clinical decision-making.

3 METHOD

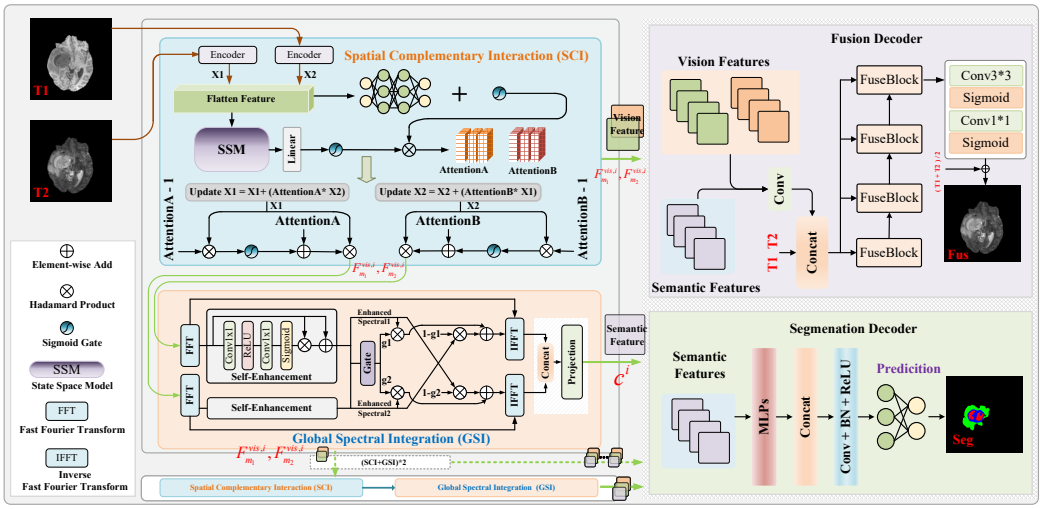

Figure 2: **Overview of the proposed DSS-Net.** (a) The *SCI module* aggregates spatial visual cues F_{vis} via Mamba-based evolution, followed by the *GSI module* which derives consistent semantic features c^i through spectral cross-gating. (b) **Dual-Path Decoding:** The *Fusion Decoder* integrates F_{vis} and c^i for image fusion, while the *Segmentation Decoder* leverages the spectral-enhanced stream c^i for precise classification.

3.1 OVERVIEW

DSS-Net (Fig. 2) is a unified framework for joint multi-modal fusion and segmentation. Its backbone follows a four-stage hierarchy, where each stage integrates an SCI and a GSI module to progressively refine spatial dependencies and enforce spectral alignment. The resulting multi-scale features are fed into dual task-specific decoders for high-fidelity fusion and precise segmentation.

Specifically, the network is organized into four progressive stages to capture hierarchical representations. At each stage, dual-modal inputs (or preceding features) are processed by a cascaded SCI–GSI pair, where SCI performs inter-modal spatial modulation and GSI conducts spectral-domain aggregation. This sequential design enables iterative refinement of visual features $F_{m_k}^{vis,i}$ and consistent semantic representations c^i. The stages are hierarchically connected, allowing deeper contextual information to be progressively accumulated. Finally, multi-scale features from all stages are bifurcated into two decoders: the Fusion Decoder reconstructs the fused image I_{fuse}, while the Segmentation Decoder produces the prediction mask S from aligned semantic features.

3.2 DECOUPLED SPATIO-SPECTRAL INTEGRATION

This core component of our framework is designed to systematically disentangle spatial visual details from global semantic cues. Instead of generic fusion, we employ a spatio-spectral decoupled strategy across four cascaded stages. Each stage orchestrates two synergistic operations: spatial detail modulation via the **Spatial Complementary Interaction (SCI)** module and spectral semantic alignment via the **Global Spectral Integration (GSI)** module. By sequentially leveraging specialized modules, the backbone progressively refines multi-modal representations for downstream task.

3.2.1 SPATIAL COMPLEMENTARY INTERACTION (SCI)

The SCI module leverages the linear sequence modeling capabilities of State Space Models (SSM) to capture global spatial dependencies and eliminate spatial redundancy between modalities. Given

the preliminary features $F^i_{m_1}, F^i_{m_2} \in \mathbb{R}^{C\times H\times W}$ extracted from the SegFormer Xie et al. (2021) encoders for modalities m_1 and m_2 (e.g., MRI T1 and T2), the SCI processes them progressively.

1) Cross-modal Hidden Evolution: To enable the model to perceive global contexts from both modalities simultaneously, we flatten the dual-stream features spatially and concatenate them along the channel dimension. This constructs a joint embedding space $\mathbf{X}_{joint} = [F^i_{m_1}; F^i_{m_2}] \in \mathbb{R}^{L\times 2C}$, where $L = H \times W$. Inside the Mamba block, the feature interaction is modeled as an input-driven Linear Time-Varying (LTV) system:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}(x_t)x(t), \\ y(t) &= \mathbf{C}(x_t)h(t). \end{aligned} \tag{1}$$

Enabled by the Selective Scan mechanism, the system matrices $\mathbf{B}$ and $\mathbf{C}$ become functions of the input. This means the model dynamically adjusts the update magnitude of the hidden state $h(t)$ based on the local saliency of the input modalities (e.g., anatomical structures in T1 and lesion highlights in T2), achieving optimal spatial information aggregation. The evolved sequence is then passed through LayerNorm, a linear projection $\mathbf{W}_p$, and a Sigmoid activation, and finally reshaped to generate two initial cross-weight maps $\mathbf{cm}_1, \mathbf{cm}_2 \in \mathbb{R}^{C\times H\times W}$.

2) Interactive Update and Discrepancy Elimination: To prevent over-smoothing during interaction, we introduce a *Global Guidance Feature* $M \in \mathbb{R}^{C\times H\times W}$, generated by passing the flattened concatenated features through a gated Multi-Layer Perceptron (MLP). The exact mixed attention representations are obtained by combining M with the SSM-derived weights: $\mathcal{A}_1 = M \odot \mathbf{cm}_1$ and $\mathcal{A}_2 = M \odot \mathbf{cm}_2$, where $\odot$ denotes the Hadamard product.

Using these attentions, we perform a Cross-Update, utilizing m_2 cues to enhance m_1, and vice versa:

$$\begin{aligned} \tilde{F}_{m_1} &= F^i_{m_1} + \mathcal{A}_1 \odot F^i_{m_2}, \\ \tilde{F}_{m_2} &= F^i_{m_2} + \mathcal{A}_2 \odot F^i_{m_1}. \end{aligned} \tag{2}$$

Since direct addition may introduce conflicting noise (e.g., local misalignment errors), we further propose a *Discrepancy Elimination* mechanism. By computing the "attention residual" between the enhanced and base features, and employing a non-linear activation as a dynamic filter threshold, we refine the final visual features:

$$\begin{aligned} F^{vis}_{m_1} &= \sigma(\tilde{F}_{m_1} \odot \mathcal{A}_1 - \tilde{F}_{m_1}) \odot \tilde{F}_{m_1} + \tilde{F}_{m_1} \odot \mathcal{A}_1, \\ F^{vis}_{m_2} &= \sigma(\tilde{F}_{m_2} \odot \mathcal{A}_2 - \tilde{F}_{m_2}) \odot \tilde{F}_{m_2} + \tilde{F}_{m_2} \odot \mathcal{A}_2. \end{aligned} \tag{3}$$

where σ is the Sigmoid function. The term $(\tilde{F} \odot \mathcal{A} - \tilde{F})$ effectively measures the discrepancy fluctuation before and after the interaction, while $\sigma(\cdot)$ suppresses activations in highly conflicting regions. Finally, the dual-stream features are aggregated as $F_{vis} = F^{vis}_{m_1} + F^{vis}_{m_2}$. This composite feature significantly enhances the contrast of lesion regions while preserving edge sharpness, providing a high-fidelity reconstruction foundation for the subsequent Fusion Decoder.

3.2.2 GLOBAL SPECTRAL INTEGRATION (GSI)

While the SCI module effectively refines local visual details, traditional spatial operations are constrained by limited receptive fields, making it challenging to capture long-range semantic consistency across modalities. The GSI module operates in the Fourier domain to achieve feature aggregation and alignment at the global energy distribution level. It takes the refined visual features $F^{vis}_{m_1}, F^{vis}_{m_2} \in \mathbb{R}^{C\times H\times W}$ from the SCI as input.

1) Spectral Decomposition: We utilize a 2D Real Fast Fourier Transform (rFFT2) to project the visual features into the spectral domain. To ensure energy conservation during the transformation, orthogonal normalization is applied:

$$\mathcal{F}_{m_k}(u, v) = \text{rFFT2}(F^{vis}_{m_k}), \quad k \in \{1, 2\}. \tag{4}$$

The complex spectral signal is decoupled into an amplitude spectrum $\mathcal{A}_{m_k} = |\mathcal{F}_{m_k}|$ and a phase spectrum $\Phi_{m_k} = \angle\mathcal{F}_{m_k}$. The amplitude spectrum captures the global intensity distribution (i.e., high-level semantics and tissue contrast), whereas the phase spectrum locks the spatial geometric structures and edge locations.

2) Spectral Self-Enhancement and Cross-Gating: Due to distinct physical imaging mechanisms, the amplitude distributions vary significantly across modalities. To achieve precise semantic alignment, we first introduce a bottleneck-based Self-Enhancement Filter to generate multiplicative enhancement weights, highlighting critical spectral components in a residual manner:

$$\hat{\mathcal{A}}_{m_k} = \mathcal{A}_{m_k} \odot (1 + \sigma(\text{Conv}_{\text{mid}}(\mathcal{A}_{m_k}))), \tag{5}$$

where σ denotes the Sigmoid function and Conv_{mid} represents the bottleneck convolution operations. Subsequently, to facilitate a "semantic soft-exchange" between modalities, we concatenate the enhanced amplitude spectra and employ a cross-gating network to learn adaptive information flow weights $\mathbf{w}_{m_1}, \mathbf{w}_{m_2}$. The fused amplitude spectra are obtained via interactive exchange:

$$\begin{aligned} \mathcal{A}_{m_1}^{fused} &= \hat{\mathcal{A}}_{m_1} \odot \mathbf{w}_{m_1} + \hat{\mathcal{A}}_{m_2} \odot (1 - \mathbf{w}_{m_1}), \\ \mathcal{A}_{m_2}^{fused} &= \hat{\mathcal{A}}_{m_2} \odot \mathbf{w}_{m_2} + \hat{\mathcal{A}}_{m_1} \odot (1 - \mathbf{w}_{m_2}). \end{aligned} \tag{6}$$

This process dynamically reallocates spectral energy on a global scale, effectively mitigating modality-specific illumination biases or artifacts while preserving common anatomical semantics.

3) Polar Reconstruction and Inverse Synthesis: To maximally retain the original spatial structural information of each modality, we recombine the fused amplitude spectra with their respective original phase spectra in the polar coordinate system. The complex tensors are then mapped back to the spatial domain using the Inverse Real FFT (irFFT2):

$$\hat{F}_{m_k} = \text{irFFT2}(\mathcal{A}_{m_k}^{fused} \cdot e^{j\Phi_{m_k}}), \quad k \in \{1, 2\}. \tag{7}$$

Finally, the deeply interacted features are concatenated along the channel dimension and projected through a 1×1 convolutional layer $\mathcal{P}$ to generate the globally consistent **semantic feature** c^i:

$$c^i = \mathcal{P}([\hat{F}_{m_1}, \hat{F}_{m_2}]). \tag{8}$$

This feature c^i not only provides the global context required by the Semantic Segmentation Decoder for pixel classification but also serves as a "semantic anchor" for the fusion network, ensuring the absolute anatomical rationality of the synthesized images.

3.3 Image Fusion Decoder

To generate high-fidelity fused images, we design a progressive decoder that integrates the visual features $\{F_m^{vis,i}\}$ from the SCI module, semantic features $\{c^i\}$ from the GSI module, and raw inputs.

1) Feature Projection and Alignment: To align channel dimensions into a unified embedding space, we employ point-wise convolutions (PConv) for dimensionality reduction:

$$f_{m,i} = \text{PConv}(F_m^{vis,i}), \quad f_{c,i} = \text{PConv}(c^i), \tag{9}$$

where $m \in \{m_1, m_2\}$.

2) Progressive Top-Down Fusion: Following a coarse-to-fine strategy, at each level $i \in \{4, 3, 2, 1\}$, the fusion block B_i aggregates the projected features, downsampled raw images $I_{raw} = [I_{m_1}, I_{m_2}]$, and upsampled feature feedback from the deeper stage:

$$\mathcal{F}_i = B_i([f_{m_1,i}, f_{m_2,i}, f_{c,i}, \text{Down}_i(I_{raw}), \text{Up}(\mathcal{F}_{i+1})]), \tag{10}$$

where $[\cdot]$ denotes concatenation. This interaction leverages deep semantic guidance to systematically supervise shallow spatial reconstruction.

3) Progressive Refinement and Residual Reconstruction: To mitigate interpolation artifacts, $\mathcal{F}_1$ undergoes a two-step progressive upsampling and refinement to yield the full-resolution $\mathcal{F}_{refined}$. To preserve the structural integrity of source modalities, we employ a residual learning paradigm. The network predicts a detail residual map I_{res} based on $\mathcal{F}_{refined}$ and the full-resolution I_{raw}:

$$I_{res} = \text{Conv}_{1\times1}([\mathcal{F}_{refined}, I_{raw}]). \tag{11}$$

Finally, I_{fuse} is generated by adding this residual to the average base intensity of the source images:

$$I_{fuse} = \text{Clamp}\left(\frac{1}{2}(I_{m_1} + I_{m_2}) + I_{res},\, 0,\, 1\right). \tag{12}$$

This ensures the network focuses on fine-grained details, maximizing structural fidelity.

3.4 SEMANTIC SEGMENTATION DECODER

To achieve efficient semantic perception, we adopt a lightweight MLP-based decoder that captures multi-scale contextual information by unifying hierarchical features.

1) Feature Embedding and Alignment: The decoder receives the hierarchical fused semantic features $\{c^i\}_{i=1}^4$ from the GSI. To resolve the discrepancy in channel counts and resolutions across stages, we first employ linear embedding layers to project each feature into a unified dimension:

$$\hat{f}_i = \text{Reshape}(\text{Linear}(c^i)). \tag{13}$$

Subsequently, all projected feature maps are upsampled to the resolution of Stage 1 via bilinear interpolation to ensure spatial consistency.

2) Global Aggregation and Prediction: The aligned feature maps are concatenated and fed into a fusion layer, where a 1×1 convolution is utilized to aggregate the multi-scale representations into a fused feature $\mathcal{M}$:

$$\mathcal{M} = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}([\hat{f}_1, \hat{f}_2, \hat{f}_3, \hat{f}_4]))). \tag{14}$$

Finally, a Dropout layer and a 1×1 classification layer are applied to generate the pixel-level segmentation mask $S \in \mathbb{R}^{N_{cls} \times H_1 \times W_1}$:

$$S = \text{Conv}_{1\times1}(\text{Dropout}(\mathcal{M})). \tag{15}$$

This design effectively integrates low-level spatial details with high-level semantic abstractions while maintaining minimal computational overhead.

3.5 LOSS FUNCTIONS

To balance fusion and segmentation, the total loss function is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{fus}, \tag{16}$$

where $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$.

1) Segmentation Loss: To address class imbalance, $\mathcal{L}_{seg}$ combines Cross-Entropy ($\mathcal{L}_{ce}$) and Dice loss ($\mathcal{L}_{dice}$):

$$\mathcal{L}_{seg} = \mathcal{L}_{ce}(S, Y) + \gamma \mathcal{L}_{dice}(S, Y), \tag{17}$$

where $\gamma = 2.0$.

2) Fusion Loss: To balance regional intensity, edge detail, and structural agreement, the fusion loss is formulated as a weighted combination:

$$\mathcal{L}_{fus} = w_{con}\mathcal{L}_{con} + w_{grad}\mathcal{L}_{grad} + w_{ssim}\mathcal{L}_{ssim}, \tag{18}$$

where w_{con}, w_{grad}, and w_{ssim} are hyperparameter trade-off weights. Specifically, $\mathcal{L}_{con}$, $\mathcal{L}_{grad}$, and $\mathcal{L}_{ssim}$ are defined as:

$$\mathcal{L}_{con} = \|I_{fuse} - \max(I_{m_1}, I_{m_2})\|_1, \tag{19}$$

$$\mathcal{L}_{grad} = \||\nabla I_{fuse}| - \max(|\nabla I_{m_1}|, |\nabla I_{m_2}|)\|_1, \tag{20}$$

$$\mathcal{L}_{ssim} = 1 - \frac{1}{2}\sum_{k=1}^{2} \text{SSIM}(I_{fuse}, I_{m_k}). \tag{21}$$

In our implementation, we set $(w_{con}, w_{grad}, w_{ssim}) = (0.5, 1.0, 1.0)$ to prioritize structural integrity and boundary sharpness over pure pixel-level intensity matching.

4 EXPERIMENT

4.1 EXPERIMENT SETTINGS

Datasets: We evaluate our method on the BraTs 2020, 2021, and 2022 datasetsMenze et al. (2014); Bakas et al. (2017); Baid et al. (2021). For our 2D framework, we extract representative axial slices from the 3D T1-weighted and T2-weighted MRI volumes to serve as source image pairs.

Table 1: Quantitative comparison of fusion performance on BraTs 2020, 2021, and 2022 datasets.

Method	BraTs 2020					BraTs 2021					BraTs 2022				
	EN↑	SD↑	MI↑	$Q^{AB/F}$ ↑	SSIM↑	EN↑	SD↑	MI↑	$Q^{AB/F}$ ↑	SSIM↑	EN↑	SD↑	MI↑	$Q^{AB/F}$ ↑	SSIM↑
SeAFusion ('22, Inf. Fusion)	2.89	81.15	1.91	0.53	0.68	2.75	81.08	1.87	0.51	0.66	2.80	81.12	1.89	0.52	0.67
PSFusion ('23, Inf. Fusion)	2.85	67.89	1.93	0.59	0.72	2.73	64.22	1.92	0.55	0.71	2.76	65.80	1.93	0.57	0.73
SegMiF ('23, ICCV)	2.71	45.33	1.97	0.58	0.76	2.62	44.10	1.95	0.53	0.77	2.65	44.25	1.94	0.54	0.77
MRFS ('24, CVPR)	2.58	87.88	1.82	0.40	0.75	2.67	87.70	1.74	0.44	0.75	2.60	87.85	1.76	0.42	0.75
AMDNet ('25, ICCV)	3.02	88.02	2.05	0.52	0.80	2.98	88.11	2.10	0.51	0.78	3.00	88.10	2.08	0.52	0.79
Ours	**3.10**	**90.81**	**2.10**	**0.60**	**0.82**	**3.08**	**90.62**	**2.15**	**0.63**	**0.82**	**3.11**	**90.88**	**2.16**	**0.62**	**0.83**

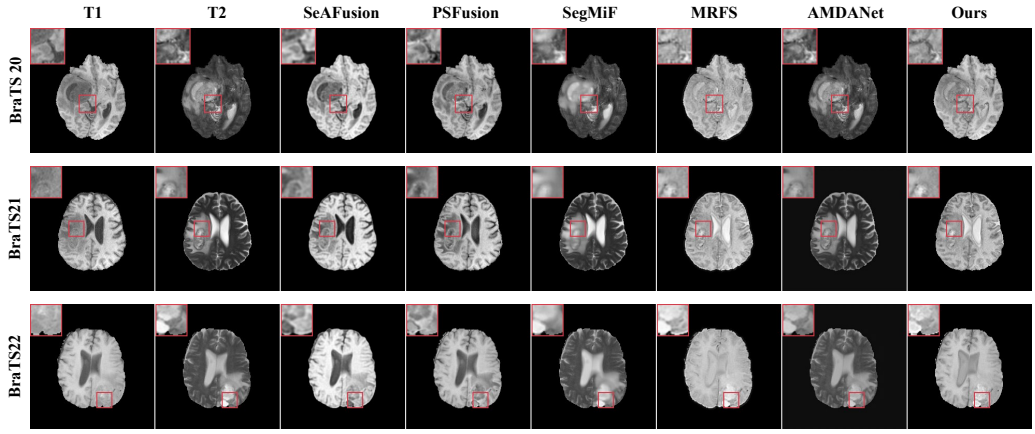

Figure 3: Qualitative comparison of image fusion results on **BraTS 2020–2022**. Columns from left: T1, T2, SeAFusion, PSFusion, SegMiF, MRFS, AMDANet, and **Ours**. Compared to other methods, **DSS-Net** achieves higher contrast and superior preservation of anatomical structures.

Evaluation Metrics: 1) Fusion: Entropy (EN) Ma et al. (2019), Standard Deviation (SD) Qu et al. (2002), Mutual Information (MI) Han et al. (2013), $Q^{AB/F}$ Xydeas & Petrović (2000), and SSIM Wang et al. (2004). **2) Segmentation:** Intersection over Union (IoU) Long et al. (2015) and Dice coefficient.

Implementation Details: The framework is implemented in PyTorch on a single RTX 3090 GPU (24 GB). We employ the Adam optimizer Kingma & Ba (2014) with an initial learning rate of 6×10^{-5}, paired with a warm-up scheduling strategy for stable convergence.

4.2 Quantitative Results

Table 1 summarizes the image fusion quantitative evaluation. Across three datasets, our method consistently achieves top performance across all five metrics (EN, SD, MI, $Q^{AB/F}$, and SSIM). Specifically, the highest SD (e.g., 90.88 on BraTS 2022) demonstrates superior visual contrast and texture, while leading $Q^{AB/F}$ and SSIM scores highlight better edge and structural preservation over SOTAs. This confirms that our dual-domain mechanism (SCI and GSI) effectively balances information retention and structural refinement.

4.3 Qualitative Results

Fig. 3 presents the qualitative comparison on BraTS. Our method produces high-contrast fused images with well-preserved subtle structures. Compared to SeAFusion and SegMiF, it achieves sharper T1/T2 interface edges and effectively suppresses artifacts. Particularly in tumor regions, our approach seamlessly integrates multi-modal complementary details for clearer textures—driven by SCI's spatial detail modeling and GSI's global cross-modal consistency.

4.4 Ablation Study

Table 2 summarizes the ablation results, validating the contribution of each component.

Table 2: Ablation study of the proposed SCI and GSI modules on the BraTS 2020 dataset.

SCI	GSI	Fusion Metrics					Seg. (%)	
		EN	SD	MI	$Q^{AB/F}$	SSIM	**mIoU**	**mDice**
		2.65	78.42	1.75	0.45	0.72	76.85	86.71
✓		2.92	85.15	1.98	0.54	0.78	79.42	88.33
	✓	2.88	84.30	1.94	0.52	0.77	78.95	88.14
✓	✓	**3.10**	**90.81**	**2.10**	**0.60**	**0.82**	**84.82**	**91.59**

Effectiveness of SCI: SCI boosts mIoU to 79.42% and $Q^{AB/F}$ to 0.54. By modeling spatial dependencies, it effectively captures fine tumor boundaries and local intensity variations.

Effectiveness of GSI: GSI raises SSIM to 0.77, MI to 1.94, and mIoU to 78.95%. Learning spectral interactions allows it to preserve global anatomical structures and suppress noise.

Dual-Domain Synergy: The full model achieves optimal performance (84.82% mIoU, 90.81 SD; Table 2). Combining SCI and GSI synergistically enhances both fusion and segmentation.

4.5 Further Analysis

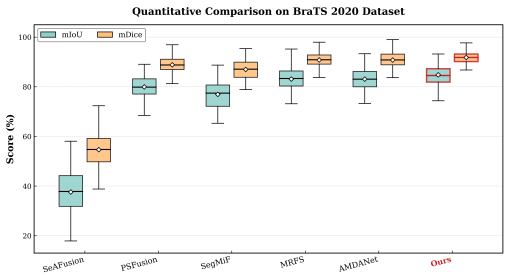

Figure 4: Downstream segmentation performance (mIoU and mDice). Red marks the best results.

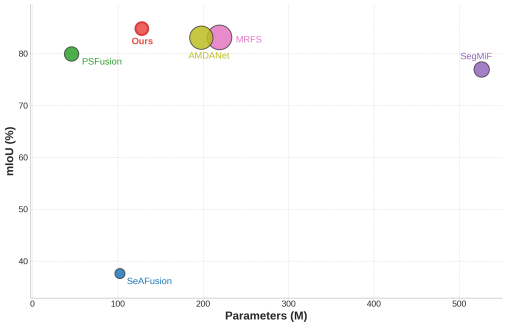

Figure 5: Efficiency comparison on BraTS 2020. **Bubble size reflects FLOPs (G)**.

Downstream Task Evaluation: Evaluated on BraTS 2020 against competitive baselines (e.g., MRFS and AMDANet), our method achieves superior performance (84.82% mIoU, 91.59% mDice; Fig. 4), confirming that task decoupling effectively preserves fine semantic features.

Model Complexity and Efficiency Analysis: As shown in Fig. 5, our method establishes a new Pareto frontier—achieving top accuracy (84.82% mIoU) with only 28.962 GFLOPs, effectively balancing high precision and efficiency for clinical deployment.

5 Conclusion

In this study, we present DSS-Net, a novel framework designed to mutually promote multi-modal image fusion and downstream task within a unified architecture. By coordinating the Spatial Complementary Interaction (SCI) and Global Spectral Integration (GSI) modules, we effectively reconcile inherent representational conflicts between low-level visual textures and high-level clinical semantics via a dual-pathway strategy. Specifically, the SCI module leverages state-space modeling to refine spatial details for high-fidelity reconstruction, while the GSI module ensures long-range semantic consistency in the Fourier domain. Extensive evaluations validate that DSS-Net not only yields visually superior fused images but also achieves state-of-the-art segmentation accuracy.

REFERENCES

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):170117, 2017.

Gwendal Bernardi, Godefroy Brisebarre, Sébastien Roman, Mohsen Ardabilian, and Emmanuel Dellandrea. A comprehensive survey on image fusion: Which approach fits which need. *Information Fusion*, 126:103594, 2026. ISSN 1566-2535.

Junyu Chen, Yihao Liu, Shuwen Wei, Zhangxing Bian, Shalini Subramanian, Aaron Carass, Jerry L Prince, and Yong Du. A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. *Med. Image Anal.*, 100:103385, 2025.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28102–28112, 2025.

Yanglin Deng, Tianyang Xu, Chunyang Cheng, Xiao-Jun Wu, and Josef Kittler. Mmdrfuse: Distilled mini-model with dynamic refresh for multi-modality image fusion. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7326–7335, 2024.

Jiao Du, Weisheng Li, and Bin Xiao. Anatomical-functional image fusion by information of interest in local laplacian filtering domain. *IEEE Transactions on Image Processing*, 26(12):5855–5866, 2017.

Yuanshen Guan, Ruikang Xu, Mingde Yao, Lizhi Wang, and Zhiwei Xiong. Mutual-guided dynamic network for image fusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 1779–1788, 2023.

Zheng Guan, Xue Wang, Wenhua Qian, Peng Liu, and Runzhuo Ma. Residual prior-driven frequency-aware network for image fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1082–1091, 2025.

Y. Han, Y. Cai, Y. Cao, and X. Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013.

Dan He, Weisheng Li, Guofen Wang, Yuping Huang, and Shiqiang Liu. Mmif-inet: Multimodal medical image fusion by invertible network. *Information Fusion*, 114:102666, 2025.

Paul Hill, Mohammed Ebrahim Al-Mualla, and David Bull. Perceptual image fusion using wavelets. *IEEE Transactions on Image Processing*, 26(3):1076–1088, 2016.

Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *Proceedings of the European Conference on Computer Vision*, pp. 539–555. Springer, 2022.

Alex Pappachen James and Belur V Dasarathy. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19:4–19, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4725–4733, 2025.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8115–8124, 2023.

Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1600–1608, 2021.

Shiqiang Liu, Weisheng Li, Dan He, Guofen Wang, and Yuping Huang. Ssefusion: Salient semantic enhancement for multimodal medical image fusion with mamba and dynamic spiking neural networks. *Information Fusion*, 119:103031, 2025. ISSN 1566-2535.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

Xudong Lu, Yuqi Jiang, Haiwen Hong, Qi Sun, and Cheng Zhuo. Dcafuse: Dual-branch diffusion-cnn complementary feature aggregation network for multi-modality image fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1524–1533, 2024.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, pp. 153–178, Jan 2019.

Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34 (10):1993–2024, 2014.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):313–315, 2002.

Yu Shi, Yu Liu, Juan Cheng, Z Jane Wang, and Xun Chen. Vdmufusion: A versatile diffusion model-based unsupervised framework for image fusion. *IEEE Transactions on Image Processing*, 2024.

Rajiv Singh and Ashish Khare. Fusion of multimodal medical images using daubechies complex wavelet transform – a multiresolution approach. *Information Fusion*, 19:49–60, 2014. ISSN 1566-2535. Special Issue on Information Fusion in Medical Image Computing and Systems.

Shanlin Sun, Kun Han, Chenyu You, Hao Tang, Deying Kong, Junayed Naushad, Xiangyi Yan, Haoyu Ma, Pooya Khosravi, James S Duncan, et al. Medical image registration via neural fields. *Med. Image Anal.*, 97:103249, 2024.

L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022a.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022b.

Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, 99:101870, 2023.

Wei Tang, Fazhi He, Yu Liu, and Yansong Duan. Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, pp. 5134–5149, 2022c.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):10944–10958, 2024.

Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, pp. 600–612, 2004.

Jinyu Wen, Asad Khan, Amei Chen, Weilong Peng, Meie Fang, C. L. Philip Chen, and Ping Li. High-quality fusion and visualization for mr-pet brain tumor images via multi-dimensional features. *IEEE Transactions on Image Processing*, 33:3550–3563, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu. RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19679–19688, 2022.

H. Xu, J. Yuan, and J. Ma. MURF: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023.

Yushen Xu, Xiaosong Li, Yuchun Wang, Xiaoqi Cheng, Huafeng Li, and Haishu Tan. Flexid-fuse: Flexible number of inputs multi-modal medical image fusion based on diffusion model. *Expert Syst. Appl.*, 296:128895, 2026. ISSN 0957-4174.

C.S. Xydeas and V. Petrović. Objective image fusion performance measure. *Electronics Letters*, pp. 308, Jan 2000.

Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26974–26983, 2024.

Guoqiang Zhao, Junjie Huang, Xiaoyun Yan, Zhaojing Wang, Junwei Tang, Yangjun Ou, Xinrong Hu, and Tao Peng. Open-vocabulary rgb-thermal semantic segmentation. In *European Conference on Computer Vision*, pp. 304–320. Springer, 2024a.

Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9380–9394, 2023a.

Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5906–5916, 2023b.

Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, and L. Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25912–25921, 2024b.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8082–8093, 2023c.

Haifeng Zhong, Fan Tang, Zhuo Chen, Hyung Jin Chang, and Yixing Gao. Amdanet: Attention-driven multi-perspective discrepancy alignment for rgb-infrared image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10645–10655, 2025.

Y. Zhong, S. Zhang, Z. Liu, X. Zhang, Z. Mo, Y. Zhang, H. Hu, W. Chen, and L. Qi. Unsupervised fusion of misaligned PAT and MRI images via mutually reinforcing cross-modality image generation and registration. *IEEE Transactions on Medical Imaging*, pp. 1702–1714, 2023.

Man Zhou, Jie Huang, Xueyang Fu, Feng Zhao, and Danfeng Hong. Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Wujie Zhou, Han Zhang, Weiqing Yan, and Weisi Lin. Mmsmcnet: Modal memory sharing and morphological complementary networks for rgb-t urban scene semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7096–7108, 2023.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7099–7108, 2024.

A APPENDIX

This Appendix provides additional technical details, hyperparameter sensitivity analyses, and qualitative/quantitative comparisons to support the main manuscript of **DSS-Net**. The content is organized as follows:

- **Section A.1: Training and Optimization Pipeline.** Formal pseudo-code detailing the end-to-end data flow and dual-task joint optimization.
- **Section A.2: Hyperparameter Sensitivity Analysis.** Comprehensive evaluations justifying our default settings across task-level (λ_1, λ_2), segmentation-level (γ), and fusion-level ($w_{con}, w_{grad}, w_{ssim}$) weights.
- **Section A.3: Extended Quantitative and Qualitative Comparisons.** Additional qualitative visualizations and multi-dataset quantitative evaluations for both image fusion and downstream segmentation.

A.1 TRAINING AND OPTIMIZATION PIPELINE

To ensure reproducibility, Algorithm 1 outlines the end-to-end training procedure of DSS-Net. The framework disentangles spatial visual features and global spectral semantics via cascaded SCI and GSI modules before feeding them into dual decoders for multi-task optimization.

Algorithm 1 Training and Optimization Pipeline of DSS-Net

Require: Multi-modal images I_{m_1}, I_{m_2}; GT masks Y; Weights $\lambda_1, \lambda_2, \gamma, w_{con}, w_{grad}, w_{ssim}$.
Ensure: Optimized parameters Θ.

for epoch = 1 to E_{max} **do**
  // 1. Feature Extraction via Cascaded Backbone
  for stage $i = 1$ to 4 **do**
    $[F_{m_1}^{vis,i}, F_{m_2}^{vis,i}] \leftarrow \text{SCI}(F_{m_1}^{i-1}, F_{m_2}^{i-1})$ {Spatial Complementary Interaction}
    $c^i \leftarrow \text{GSI}(F_{m_1}^{vis,i}, F_{m_2}^{vis,i})$ {Global Spectral Interaction}
  end for
  // 2. Dual-Path Decoding
  $I_{fuse} \leftarrow$ Fusion Decoder$(\{F_{m_k}^{vis,i}, c^i\}_{i=1}^4)$
  $S \leftarrow$ Segmentation Decoder$(\{c^i\}_{i=1}^4)$
  // 3. Loss Computation
  $\mathcal{L}_{seg} = \mathcal{L}_{ce}(S, Y) + \gamma \cdot \mathcal{L}_{dice}(S, Y)$
  $\mathcal{L}_{fus} = w_{con}\mathcal{L}_{con} + w_{grad}\mathcal{L}_{grad} + w_{ssim}\mathcal{L}_{ssim}$
  $\mathcal{L}_{total} = \lambda_1\mathcal{L}_{seg} + \lambda_2\mathcal{L}_{fus}$
  // 4. Parameter Update
  $\Theta \leftarrow \Theta - \eta \cdot \nabla_\Theta \mathcal{L}_{total}$ {Optimization via AdamW}
end for

A.2 HYPERPARAMETER SENSITIVITY ANALYSIS

All sensitivity evaluations are conducted on BraTS 2020 to justify the default configurations across task-level, segmentation-level, and fusion-level constraints.

TASK-LEVEL BALANCING WEIGHTS (λ_1, λ_2)

Weights λ_1 and λ_2 balance gradients between segmentation and fusion. As shown in Table A3, our default setting ($\lambda_1 = 1.0, \lambda_2 = 0.5$) yields optimal joint performance. Increasing λ_2 to 1.0 overemphasizes reconstruction, causing segmentation mIoU to drop to 80.45%. Conversely, reducing λ_2 to 0.1 severely impairs fusion performance (SSIM drops to 0.76).

SEGMENTATION-LEVEL WEIGHT (γ)

The weight γ controls the contribution of Dice loss ($\mathcal{L}_{dice}$) against Cross-Entropy ($\mathcal{L}_{ce}$). Fig. A6 illustrates that increasing γ from 0.0 to 2.0 substantially boosts mIoU by addressing target class

Table A3: Ablation of task-level weights λ_1 and λ_2 on BraTS 2020.

Weights		**Segmentation**		**Fusion Performance**				
λ_1	λ_2	mIoU↑	mDice↑	EN↑	SD↑	MI↑	$Q^{AB/F}$ ↑	SSIM↑
1.0	1.0	80.45	89.11	3.08	90.75	**2.12**	0.58	0.80
1.0	0.1	81.12	89.47	2.85	85.30	1.88	0.48	0.76
0.5	1.0	78.90	88.19	**3.12**	90.55	2.05	0.57	0.81
1.0	**0.5**	**84.82**	**91.59**	3.10	**90.81**	2.10	**0.60**	**0.82**

imbalance in tumor sub-regions (e.g., ET and NCR). Setting $\gamma > 2.0$ slightly decreases accuracy due to over-regularization. Thus, $\gamma = 2.0$ is chosen.

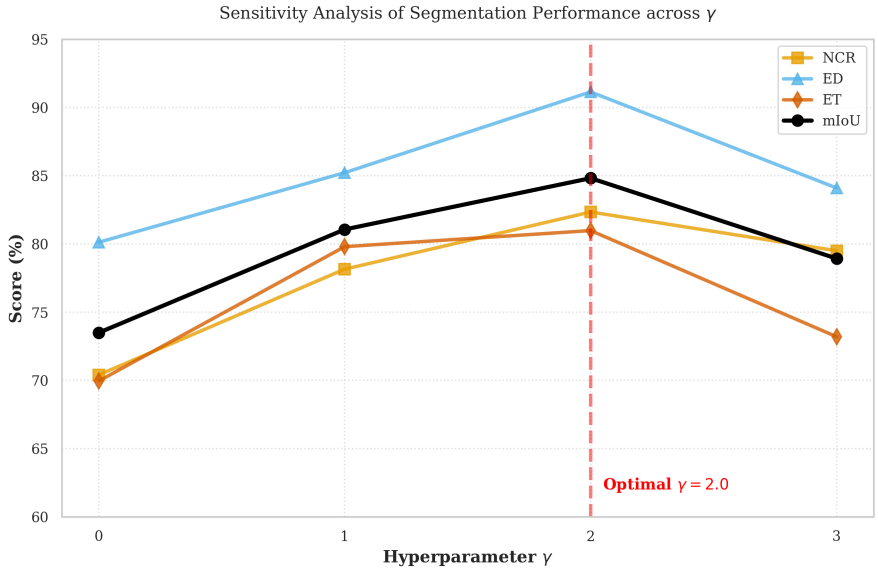

Figure A6: Sensitivity analysis of the segmentation weight γ on BraTS 2020. Peak performance across tumor sub-regions is achieved at $\gamma = 2.0$.

Fusion-level Weights $(w_{con}, w_{grad}, w_{ssim})$

We evaluate the sensitivity of fusion weights $(w_{con}, w_{grad}, w_{ssim})$ to validate model robustness. As shown in Fig. A7, metrics peak at $(0.5, 1.0, 1.0)$. Deviations lead to noticeable performance drops in structural and gradient metrics, confirming that this setting optimally balances intensity preservation and texture extraction.

A.3 Extended Comparisons

Qualitative and Quantitative Image Fusion

Qualitative results across BraTS 2020–2022 (Fig. A9) show that DSS-Net preserves sharp interfaces and fine structures without halo artifacts. Quantitatively (Table A6), DSS-Net leads in EN, MI, $Q^{AB/F}$, and SSIM across all datasets, validating the effectiveness of SCI and GSI in capturing complementary details and cross-modal consistency.

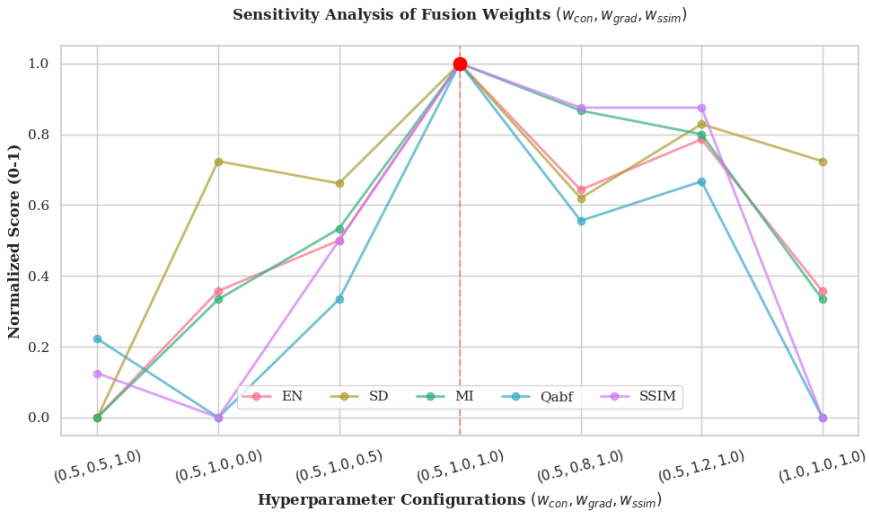

Figure A7: Sensitivity analysis of the fusion loss weights $(w_{con}, w_{grad}, w_{ssim})$ on BraTS 2020, achieving optimal trade-off at $(0.5, 1.0, 1.0)$.

Qualitative and Quantitative Segmentation

Segmentation visual comparisons (Fig. A10 and Fig. A11) confirm that spatial guidance from SCI combined with global features from GSI reduces false positives and boundary blurring at multi-class junctions (NCR, ED, and ET).

Quantitative evaluation on BraTS 2020–2022 (Table A5) shows DSS-Net consistently outperforming SOTAs, reaching up to 84.87% mIoU on BraTS 2021. Furthermore, comparing against medical segmentation baselines (Table A4) shows that DSS-Net surpasses dedicated models including nnU-Net (84.82% vs. 83.50% mIoU), demonstrating that feature disentanglement and task synergy effectively benefit downstream tasks.

Qualitative Ablation Analysis

Fig. A8 visually validates each module: removing SCI (**w/o SCI**) degrades edge sharpness, while omitting GSI (**w/o GSI**) leads to luminance imbalance. Combining both yields optimal contrast and clear anatomical boundaries.

Table A4: Segmentation performance comparison (mIoU % and mDice %) on BraTS 2020 against medical segmentation baselines.

Method	**mIoU (%)** ↑	**mDice (%)** ↑
nnU-Net	83.50	90.41
Swin-Unet	80.45	89.22
U-Mamba	82.85	90.41
BiomedParse	83.10	90.43
Ours (DSS-Net)	**84.82**	**91.59**

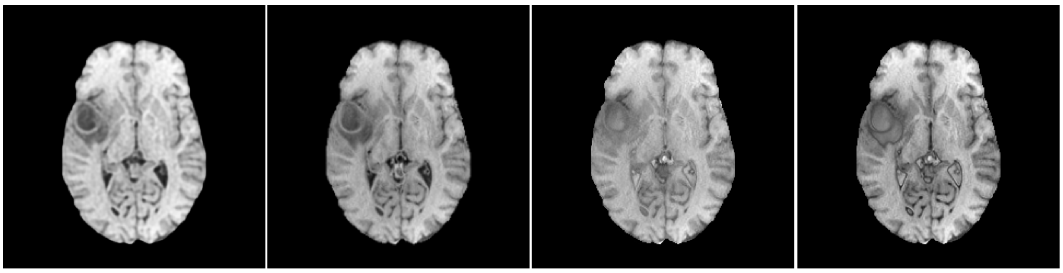

Figure A8: Qualitative ablation of SCI and GSI on image fusion over BraTS 2020: (a) Baseline, (b) w/o SCI, (c) w/o GSI, and (d) Ours.

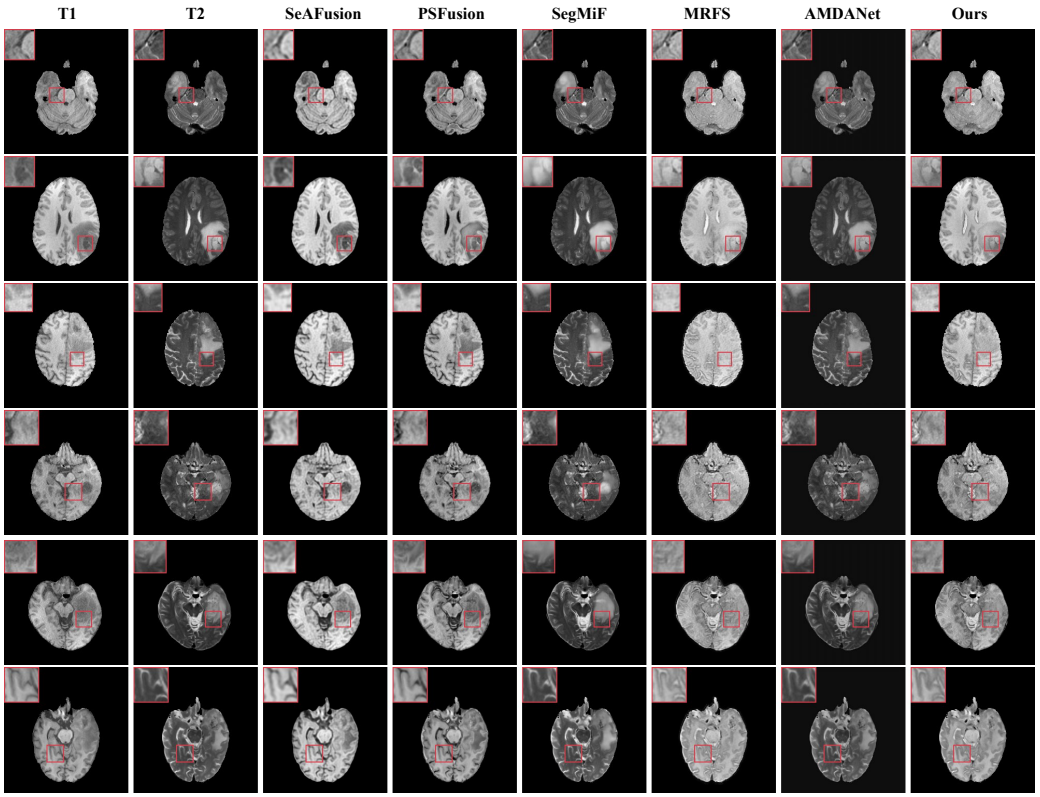

Figure A9: Qualitative comparison of image fusion results on **BraTS 2020–2022** across SOTA methods. DSS-Net achieves superior visual contrast and anatomical detail preservation.

Table A5: Quantitative comparison of segmentation performance (IoU % and Dice %) on BraTS 2020–2022.

Method	BraTS 2020					BraTS 2021					BraTS 2022				
	NCR	ED	ET	mIoU	mDice	NCR	ED	ET	mIoU	mDice	NCR	ED	ET	mIoU	mDice
SeAFusion ('22)	46.45	29.76	36.59	37.60	54.61	44.90	30.80	34.00	36.57	53.45	45.60	31.30	35.35	37.42	54.48
PSFusion ('23)	74.70	88.00	77.10	79.93	88.82	75.62	88.90	77.31	80.61	89.14	73.80	87.70	76.84	79.45	88.52
SegMiF ('23)	71.60	85.70	73.52	76.94	86.94	73.30	87.08	74.10	78.16	87.63	70.75	84.60	71.00	75.45	86.09
MRFS ('24)	79.26	91.62	78.51	83.13	90.74	77.70	90.80	78.81	82.44	90.25	77.82	90.92	78.94	82.56	90.43
AMDANet ('25)	80.33	90.08	78.79	83.07	90.72	80.60	90.28	79.70	83.53	91.10	79.00	89.81	80.00	82.94	90.61
Ours	**82.34**	**91.14**	**80.98**	**84.82**	**91.59**	**82.45**	**91.18**	**80.97**	**84.87**	**91.66**	**82.10**	**91.05**	**81.12**	**84.76**	**91.73**

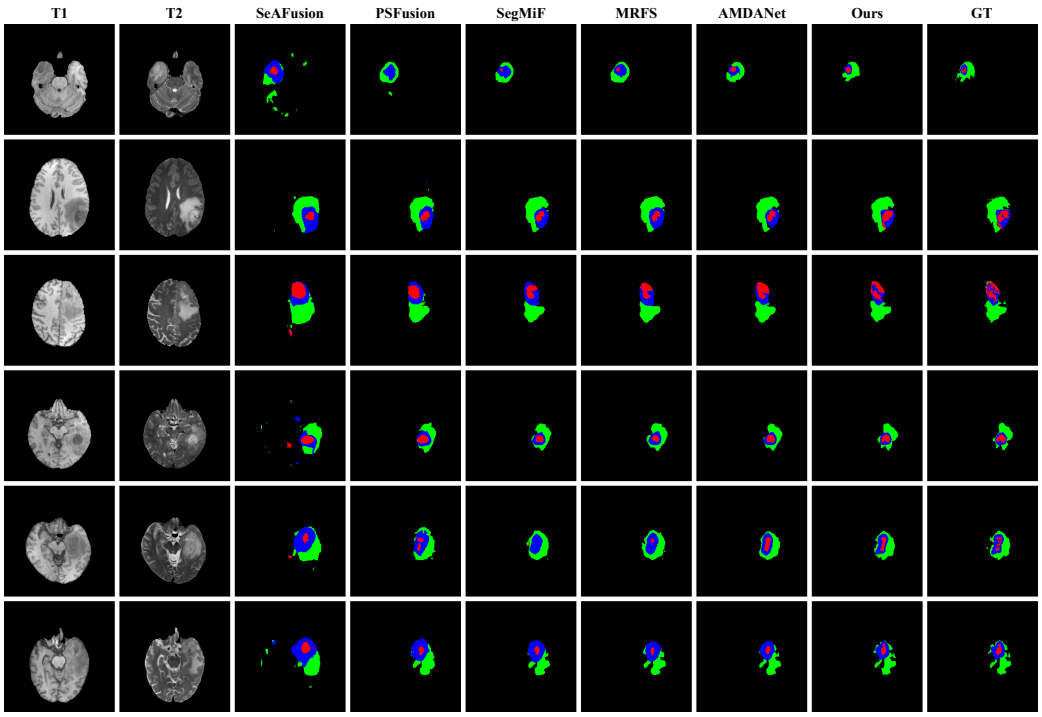

Figure A10: Qualitative segmentation results on **BraTS 2020–2022**. Color codes: *NCR* (green), *ED* (blue), *ET* (red), and *BG* (black).

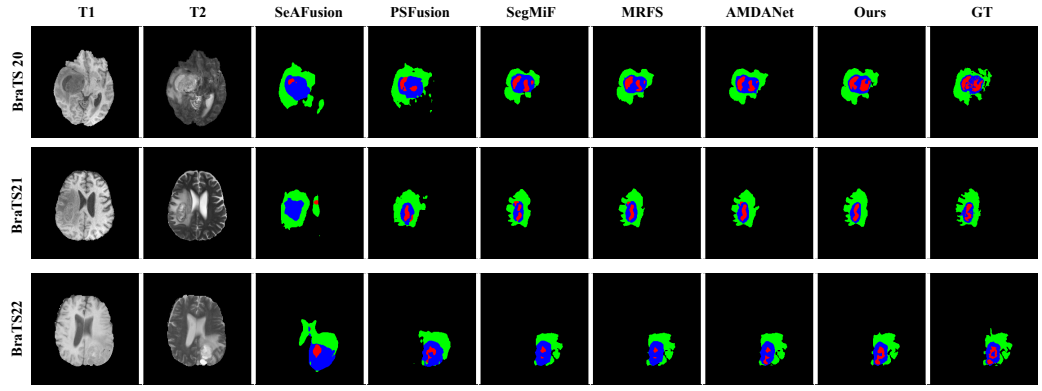

Figure A11: Extended qualitative segmentation comparisons on **BraTS 2020–2022** showing fine boundary preservation at class intersections.

Table A6: Quantitative comparison of image fusion performance (EN, SD, MI, $Q^{AB/F}$, and SSIM) on BraTS 2020–2022.

Method	**BraTS 2020**					**BraTS 2021**					**BraTS 2022**				
	EN	SD	MI	$Q^{AB/F}$	SSIM	EN	SD	MI	$Q^{AB/F}$	SSIM	EN	SD	MI	$Q^{AB/F}$	SSIM
CDDFuse ('23)	2.84	87.35	1.92	0.55	0.76	2.87	87.60	1.95	0.58	0.77	2.89	87.85	1.97	0.56	0.78
DDFM ('23)	2.92	88.10	1.98	0.56	0.77	2.95	88.45	2.01	0.59	0.78	2.98	88.70	2.03	0.58	0.79
EMMA ('24)	2.96	88.90	2.01	0.59	0.79	2.99	89.20	2.04	**0.64**	0.80	3.01	89.50	2.06	0.60	0.81
MDAFuse ('24)	2.89	89.25	1.96	0.57	0.74	2.92	89.50	1.99	0.59	0.75	2.95	89.80	2.02	0.59	0.76
UP-Fusion ('26)	3.04	**91.05**	2.06	0.58	0.80	3.02	**90.95**	2.09	0.61	0.81	3.06	**91.12**	2.11	0.61	0.82
Ours	**3.10**	90.81	**2.10**	**0.60**	**0.82**	**3.08**	90.62	**2.15**	0.63	**0.82**	**3.11**	90.88	**2.16**	**0.62**	**0.83**