Breaking Task Boundaries: A Unified Model for 3D Medical Image Fusion and Segmentation Guided by Manifold Perspective

Zeyu Wang[1], **Jiayu Wang**[1], **Haiyu Song**[1,*]

[1]College of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China
wangzeyu@dlnu.edu.cn, 202311054010@stu.dlnu.edu.cn, shy@dlnu.edu.cn

Abstract

3D medical image fusion (MIF) and segmentation (MIS) are critical and inherently synergistic tasks in medical image analysis. However, fundamentally integrating them remains highly challenging, since effective collaborative paradigms are still scarce and their optimization objectives fundamentally diverge. Moreover, existing continual learning techniques are unable to achieve truly advanced performance for both tasks using a shared weight. To address these challenges, we propose M²-CoFS, a unified model capable of jointly handling both tasks. Our core contribution is a "network-guided network learning" paradigm designed to break the task boundaries. We model the weight spaces of MIF and MIS as high-dimensional manifolds and innovatively use a lightweight neural network to implicitly construct a shared manifold. Interestingly, this network outputs unified weights for both tasks. To ensure the shared manifold retains the intrinsic geometry of both original manifolds, we embed manifold distances into the loss function of this network as a constraint. Additionally, we design a tailored three-stage training paradigm for our core contribution mentioned above. Stage I focuses on independent task optimization for high-quality weights; Stage II aims to reduce weight-space distance between tasks via our cross-task weight adaptation strategy; Our core innovation serves as Stage III. Experimental results show that M²-CoFS consistently outperforms state-of-the-art comparison models on both MIF and MIS.

Code — https://github.com/Wangjiayu0512/M2-CoFS

Introduction

3D medical image fusion (MIF) and segmentation (MIS) are two pivotal tasks in medical image analysis, providing critical support for clinical diagnosis (Yin et al. 2018; Tang et al. 2022a). MIF integrates complementary multimodal information to enhance structural detail and visual clarity (Zhao et al. 2023a). Conversely, MIS precisely delineates anatomical structures or lesions at the voxel level (Ding et al. 2021; Fang and Wang 2022), emphasizing boundary and texture features for improved accuracy.

These two tasks are inherently interdependent, naturally forming an upstream–downstream relationship. High-quality MIF provides MIS more comprehensive information, reducing segmentation difficulty and enhancing accuracy (Liu et al. 2022). Conversely, precise segmentation can guide fusion by identifying regions of interest, enabling more targeted integration (Wang et al. 2025b). Given the shared multimodal inputs, it is feasible for both tasks to benefit from shared features.

Despite their synergy, integrating MIF and MIS remains challenging for two primary issues. Firstly, existing methods address them independently, lacking inter-task collaboration and limiting overall performance. Superior fusion results can not guarantee accurate segmentation, while segmentation methods without integrated multimodal information suffer from performance limitations (Li et al. 2024; Zou et al. 2025). Additionally, existing MIF approaches focus on 2D slices (Tang, Li, and Ma 2025; Tang et al. 2024), ignoring spatial continuity and inter-slice dependencies intrinsic to 3D volumetric data, thus hindering their integration with 3D MIS methods and limiting clinical applicability. Secondly, the different optimization objectives often cause gradient conflicts in joint optimization (Liu, Chu, and Thuerey 2024; Zhou et al. 2023b; Li et al. 2025b). MIF focuses on capturing the overall complementary information from multimodal images (Zhao et al. 2023b; Liu et al. 2025b) while MIS emphasizes precise delineation of local lesion boundaries and spatial structures. To this end, RMR-Fusion (Zhang et al. 2024) is a representative method which employs a shared encoder with dual task-specific decoders and linearly combined losses. However, it still faces performance bottlenecks because the optimization conflicts have not been fundamentally resolved. This highlights a crucial question: **Can we design a collaborative framework that bridges the gap between MIF and MIS to enable true joint optimization?** To address these challenges, we propose a **M**anifold-Guided **M**ulti-modal **Co**llaborative **F**usion–**S**egmentation Network (M²-CoFS), a unified framework designed for MIF and MIS. It adopts a novel three-stage training strategy that progressively integrates task-specific weights into unified weights for both tasks. Stage I employs separate encoders and decoders to obtain high-quality task-specific weights. In Stage II, we design a cross-task weights adaptation to reduce discrepancies between task-specific weights distributions. Besides, to avoid catastrophic forgetting, Synaptic Intelligence (SI) is introduced during this adaptation (Zenke, Poole, and

*Corresponding author: Haiyu Song.

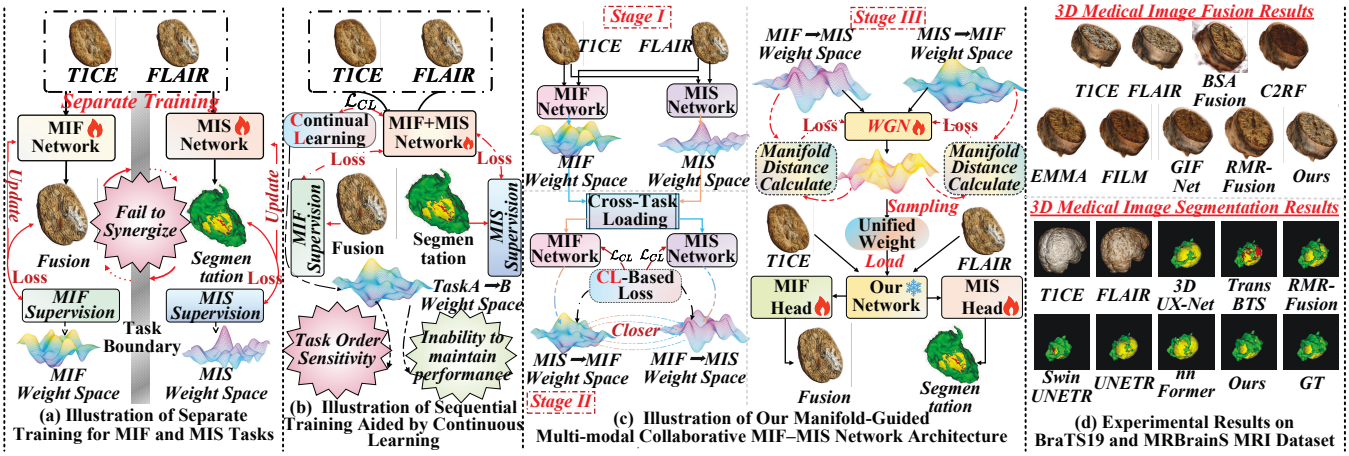

Figure 1: Workflow and performance comparison of our three-stage M²-CoFS with existing MIF–MIS Paradigms. Stage I: Independent Task Optimization; Stage II: Cross-task Weight Adaptation; Stage III: Weight Generation Network based on Implicitly Constructing a Shared Manifold.

Ganguli 2017). Nevertheless, while continual learning produces generally transferable weights, the differing task order causes them to diverge in weight spaces, limiting their collaborative potential. To bridge this gap, Stage III incorporates our core innovation: a "network-guided network learning" strategy from the manifold perspective. We regard the weight spaces of MIF and MIS as distinct manifolds in a high-dimensional space and aim to construct a shared manifold. Given the difficulty of explicit manifold construction in high-dimensions, we leverage a lightweight MLP-based weight generation network (WGN) to implicitly model the shared manifold and obtain unified task order-irrelevant weights for both tasks. To endow the WGN with the aforementioned capabilities, we embed manifold distance into loss function, preserving geometric structures of individual task manifolds while constructing the shared manifold.

Our contributions are summarized as follows:

• We design a collaborative MIF-MIS network that enables mutual promotion of both tasks, overcoming the limit of single-task approaches.

• We propose a "network-guided network learning" strategy which employs a lightweight WGN to implicitly construct a shared manifold for outputting unified task order-irrelevant weights. To preserve the geometric structure of the manifold, we tailor a manifold distance-based loss function for WGN.

• We propose a three-stage training strategy that progressively integrates weights from task-specific spaces into a precisely constructed shared space, ensuring advanced performance for both MIF and MIS.

Related Work

3D Medical Image Fusion. MIF integrates multimodal information to enhance clinical utility. Recent deep learning methods have significantly advanced this field. BSAFusion (Li et al. 2025a) and C2RF (Tang et al. 2025) achieve unaligned image fusion. EMMA (Zhao et al. 2024a) addresses the absence of ground truth by employing strong priors, while FILM (Zhao et al. 2024b) employs vision-language models to enhance cross-modal interactions. However, these 2D methods neglect the spatial continuity in volumetric data. DC2Fusion (Liu et al. 2023) addresses this by introducing deformable modules specifically designed for 3D volumes.

However, existing MIF methods primarily focus on visual quality which is not guaranteed to improve accurate segmentation performance in MIS. RMR-Fusion (Zhang et al. 2024) attempted joint optimization of MIF and MIS, but relies solely on shared encoders and linearly combined losses, leading to optimization conflicts and suboptimal performance. To this end, we propose M²-CoFS, which implicitly constructs a shared manifold to obtain unified weights, facilitating joint optimization of MIF and MIS.

3D Medical Image Segmentation. MIS provides voxel-level delineation of lesions. Deep learning models have become mainstream, including pure CNN-based methods like 3D UX-Net (Lee et al. 2023), hybrid CNN-Transformer approaches such as TransBTS (Wang et al. 2021) and Swin UNETR (Tang et al. 2022b), and Transformer-based methods like nnFormer (Zhou et al. 2023a). However, most approaches poorly utilize multimodal data as they only concatenate multimodal data or process them separately, which cause significant performance bottlenecks. To our knowledge, only RMR-Fusion attempted joint optimization but faced conflicting gradients due to divergent task objectives. To address these challenges, we design a three-stage training strategy that progressively derives unified weights for both tasks, enabling deep collaborative optimization.

Continual Learning. Continual learning tackles catastrophic forgetting when models face sequential tasks or shifting data distributions. Classic strategies fall into three categories: regularization-based methods (Kirkpatrick et al. 2017); replay approaches (Rolnick et al. 2019); and parameter-isolation techniques (Mallya and Lazebnik 2018).

Nevertheless, these methods mainly seek task compatibility rather than collaboration. In joint MIF-MIS task, existing continual learning techniques still fail to overcome op-

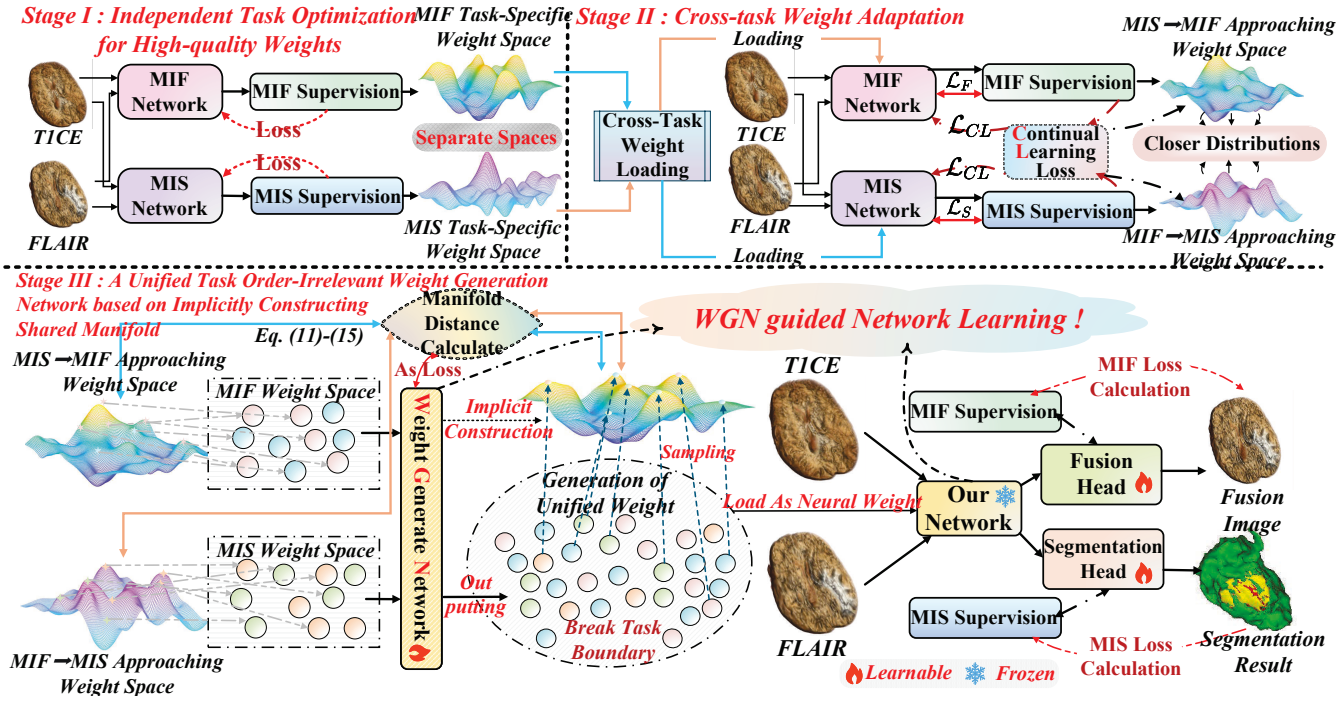

Figure 2: Schematic diagram of our model. It operates in three stages: (1) Independent Task Optimization for High-quality Weights. (2) Cross-task Weight Adaptation to Reduce Weight-space Distance between Tasks. (3) A Unified Task Order-Irrelevant Weights Generation Network based on Implicitly Constructing Shared Manifold.

timization conflicts, limiting performance. To overcome the limitations, we introduce a manifold-based "network-guided network learning" paradigm, where the weight spaces of both tasks are treated as high-dimensional manifolds, and a shared manifold is implicitly constructed to derive unified task order-irrelevant weights for joint optimization.

Method

Problem Formulation and Modeling

To achieve the collaboration of MIF and MIS tasks, we propose M²-CoFS, a unified model that jointly optimizes MIF and MIS. The inference procedure is defined in Eq. (1), which takes a pair of images $\mathcal{A}, \mathcal{B}$ and generates a high-quality fused image and an accurate segmentation mask. The architecture of our model is illustrated in Fig. 2.

$$\widehat{Y}_f, \widehat{Y}_s = \{\psi, \varphi\} \circ (\mathcal{F}(\Phi(\mathcal{A}, \mathcal{B})); \underbrace{\mathcal{G}(\theta_{\mathrm{F}}, \theta_{\mathrm{S}})}_{\theta_{\mathcal{U}}}). \quad (1)$$

Here Φ denotes a shared encoder, $\mathcal{F}$ is the feature-fusion module, ψ, φ are dedicated task heads, respectively, and $\mathcal{G}$ is a weight generation network (WGN), $\theta_{\mathcal{F}}$ and $\theta_{\mathcal{S}}$ denotes the weights of the $\mathcal{F}$ to each task, $\theta_{\mathcal{U}}$ is the unified weights.

To progressively leverage inter-task synergy, we design a three-stage training strategy: (1) task-specific independent training, (2) cross-task weight adaptation, and (3) a "network-guided network learning" strategy using WGN from a manifold perspective to construct a shared manifold, thereby generating unified task order-irrelevant weights.

Stage I: Task-Specific Weights Obtaining

In Stage I, MIF and MIS are independently trained to obtain task-specific weights θ_{F} and θ_{S}, providing a foundation for subsequent collaborative optimization. The optimization objectives are defined in Eq. (2) and (3):

$$\arg\min_{\theta_{\mathcal{F}}} \mathcal{L}_{\mathcal{F}} \left(\psi \circ \mathcal{F}_F \left(\Phi(\mathcal{A}, \mathcal{B}); \theta_{\mathcal{F}}\right)\right), \quad (2)$$

$$\arg\min_{\theta_{\mathcal{S}}} \mathcal{L}_{\mathcal{S}} \left(\varphi \circ \mathcal{F}_S \left(\Phi(\mathcal{A}, \mathcal{B}); \theta_{\mathcal{S}}\right)\right). \quad (3)$$

Specifically, the encoder Φ consists of a 3D CNN to extract local features and a Swin Transformer to capture global context, resulting in high-quality multi-scale features. Feature-fusion modules $\mathcal{F}_{\mathrm{F}}$ and $\mathcal{F}_{\mathrm{S}}$ each employs a 3D CNN to integrate multimodal features. The fusion head ψ uses stacked 3D CNN layers to enhance fusion quality, while the segmentation head φ leverages the Swin UNETR (Hatamizadeh et al. 2022a; Zou et al. 2024) architecture to refine boundaries and textures for precise segmentation. The loss functions $\mathcal{L}_F$ and $\mathcal{L}_S$ are in Subsection Loss Function.

Stage II: Cross-Task Weights Adaptation based on Continual Learning

To reduce discrepancies between task-specific weights distributions, we propose a cross-task weights adaptation mechanism. As shown in Eq. (4), the MIF's feature-fusion module initializes with the MIS's pre-trained weights $\theta_{\mathcal{S}}$ to obtain $\theta_{\mathcal{S}\to\mathcal{F}}$,

$$\theta_{\mathcal{S}\to\mathcal{F}} = \arg\min_{\theta_{\mathcal{S}}} \mathcal{L}_{\mathcal{F}} \left(\psi \circ \mathcal{F}_F \left(\Phi(\mathcal{A}, \mathcal{B}); \theta_{\mathcal{S}}\right)\right). \quad (4)$$

Conversely, MIS initializes with $\theta_{\mathcal{F}}$ and continues training for $\theta_{\mathcal{F}\to\mathcal{S}}$ as shown in Eq. (5).

$$\theta_{\mathcal{F}\to\mathcal{S}} = \arg\min_{\theta_{\mathcal{F}}} \mathcal{L}_{\mathcal{S}} \left(\varphi \circ \mathcal{F}_S \left(\Phi(\mathcal{A}, \mathcal{B}); \theta_{\mathcal{F}}\right)\right). \quad (5)$$

Through this reciprocal initialization, $\theta_{\mathcal{S}\to\mathcal{F}}$ and $\theta_{\mathcal{F}\to\mathcal{S}}$ converge in the weight space.

However, such cross-task adaptation risks the well-documented issue of catastrophic forgetting, potentially degrading performance on original task. To mitigate this, we incorporate Synaptic Intelligence (SI), allowing weights to adapt without forgetting. As shown in Eq. (6), SI quantifies the importance of each weight by allowing weights to adapt while preserving critical knowledge. For the k-th task, the importance of θ_i is defined as Ω_i :

$$\Omega_i = \frac{\sum_{t=1}^{T} \frac{\partial L^{(k)}}{\partial \theta_i} \cdot \Delta\theta_i^{(t)}}{\left(\theta_i^* - \theta_i^p\right)^2 + \xi}, \quad (6)$$

where $\Delta\theta_i^{(t)}$ denotes the weights updates at step t, and θ_i^*, θ_i^p are the post- and pre-training weights values, respectively.

Finally, as indicated in Eq. (7), we add the SI term as a regulariser into the loss function. Weights critical to the prior task are constrained during new-task training, enabling adaptation without catastrophic forgetting.

$$\mathcal{L}_{\mathcal{SI}} = \sum_i \Omega_i \left(\theta_i - \theta_i^*\right)^2. \quad (7)$$

Stage III: Network-Guided Network Learning - A Unified Task Order-Irrelevant Weights Generation Network based on Implicitly Constructing Shared Manifold

The primary goal of Stage III is to obtain unified task order-irrelevant weights $\theta_{\mathcal{U}}$, then the inference procedure delineated in Eq. (8) can be executed to generate the outputs for both tasks. Accordingly, our focus centers on acquiring $\theta_{\mathcal{U}}$.

Although Stage II brings the two tasks weights distributions closer, their sensitivity to task order continues to limit their suitability for joint optimization.

To address this limitation, we propose a "network-guided network learning" collaborative strategy from the manifold perspective (Kimmel, Sochen, and Malladi 1997; Sochen, Kimmel, and Malladi 1998; Wang et al. 2023, 2025c). The optimization objective for this stage is formulated in Eq. (8):

$$\begin{aligned} \arg\min_{\theta_{\mathcal{U}}} \big[& \mathcal{L}_{\mathcal{F}}{}'(\psi \circ \mathcal{F}_F(\Phi(\mathcal{A}, \mathcal{B}); \theta_{\mathcal{U}})) \\ & + \mathcal{L}_{\mathcal{S}}{}'(\varphi \circ \mathcal{F}_S(\Phi(\mathcal{A}, \mathcal{B}); \theta_{\mathcal{U}})) \\ & + \mathcal{L}_{\mathcal{M}}(\theta_{\mathcal{U}}, \theta_{\mathcal{F}}, \theta_{\mathcal{S}}) \big], \end{aligned} \quad (8)$$

where $\theta_{\mathcal{U}}$ denotes unified task order-irrelevant weights applicable to both tasks, generated by a lightweight WGN.

WGN acts as an implicit constructor of the shared manifold, producing unified weights $\theta_{\mathcal{U}}$ for $\mathcal{F}$ at each forward pass. These features output by this module are shared by both tasks, enabling joint training of the WGN and task-specific heads to achieve advanced performance.

Details of WGN. The design of WGN is grounded in a manifold perspective. Conceptually, we treat the $\theta_{\mathcal{F}\to\mathcal{S}}$ and $\theta_{\mathcal{S}\to\mathcal{F}}$ as discrete sampling points on respective high-dimensional manifolds $\mathcal{M}_F$ and $\mathcal{M}_S$.

$$\{\theta_{\mathcal{F}\to\mathcal{S}}\} \subset \mathcal{M}_F, \{\theta_{\mathcal{S}\to\mathcal{F}}\} \subset \mathcal{M}_S. \quad (9)$$

From this insight, since Stage II already brings the two weights distributions closer, $\mathcal{M}_F$ and $\mathcal{M}_S$ become proximate in manifold space. The goal of WGN is generating a shared manifold $\mathcal{M}_U$, from which $\theta_{\mathcal{U}}$ can be sampled. However, explicitly modeling $\mathcal{M}_U$ is impractical due to its high dimensionality and complexity. To overcome this, we propose a lightweight MLP-based WGN that implicitly constructs the shared manifold and integrates the separate gradient information into a shared weight space. The WGN receives $\theta_{\mathcal{F}\to\mathcal{S}}$ and $\theta_{\mathcal{S}\to\mathcal{F}}$, and outputs unified task order-irrelevant weights $\theta_{\mathcal{U}}$ as formulated in Eq. (10).

$$\mathcal{G}(\theta_{\mathcal{F}\to\mathcal{S}}, \theta_{\mathcal{S}\to\mathcal{F}}) = \theta_{\mathcal{U}}. \quad (10)$$

Nevertheless, training such a network remains challenging for two reasons: (1) the absence of an existing guiding paradigm, and (2) the difficulty in ensuring the geometric properties specific to each task of shared manifold . To this end, we employ manifold distance metrics to encode intrinsic geometric construction, and embed them into loss function of WGN. Obviously, computing distances on manifold surfaces becomes crucial for retaining manifold structure.

The Process of Manifold Distance Computation. First, we project the sampled points $\theta_{\mathcal{F}\to\mathcal{S}} \in R^{n_f * d}$ and $\theta_{\mathcal{S}\to\mathcal{F}} \in R^{n_s * d}$ from $\mathcal{M}_F$ and $\mathcal{M}_S$, together with the $\theta_{\mathcal{U}} \in R^{n_u * d}$ from $\mathcal{M}_U$, into a shared space and represent them as $\theta \in R^{n * d}$, where d is the number of network layers and n is the number of neurons per layer.

$$\theta = [\theta_{\mathcal{U}}; \theta_{\mathcal{F}\to\mathcal{S}}; \theta_{\mathcal{S}\to\mathcal{F}}], \theta \in \mathbb{R}^{(n_u + n_f + n_s) * d}. \quad (11)$$

Next, we construct a KNN-based graph G to preserve the local geometric structure of the manifold on θ.

$$\mathrm{G} = \mathrm{KNN}(\theta, k). \quad (12)$$

Then, we compute the distance matrices $D_{U\to F}$ and $D_{U\to S}$ using Dijkstra algorithm $\mathcal{T}$ (Dijkstra 2022).

$$D_{U\to F} = \frac{1}{n^2} \mathcal{T}(G, k, \mathrm{n_u} + i), \quad (13)$$

$$D_{U\to S} = \frac{1}{n^2} \mathcal{T}(G, k, \mathrm{n_u} + n_\mathrm{f} + i). \quad (14)$$

Finally, we add these two distance terms to form the regularizer and embed it into the loss function.

$$\mathcal{L}_{\mathcal{M}} = D_{U\to S} + D_{U\to F}. \quad (15)$$

Note that while this explicit manifold distance computation is not perfectly precise, embedding it into the loss function still preserves manifold geometry. As it maintains the relative distances, once the relative relationships are consistent, the manifold's geometric structure is retained.

Loss Function. To ensure effective optimization in each training phase, we adopt different loss functions. For MIF, the loss function is defined as Eq. (16) in Stage I:

$$\mathcal{L}_F = \mathcal{L}_{L_1} + \alpha \mathcal{L}_{\mathcal{G}}, \quad (16)$$

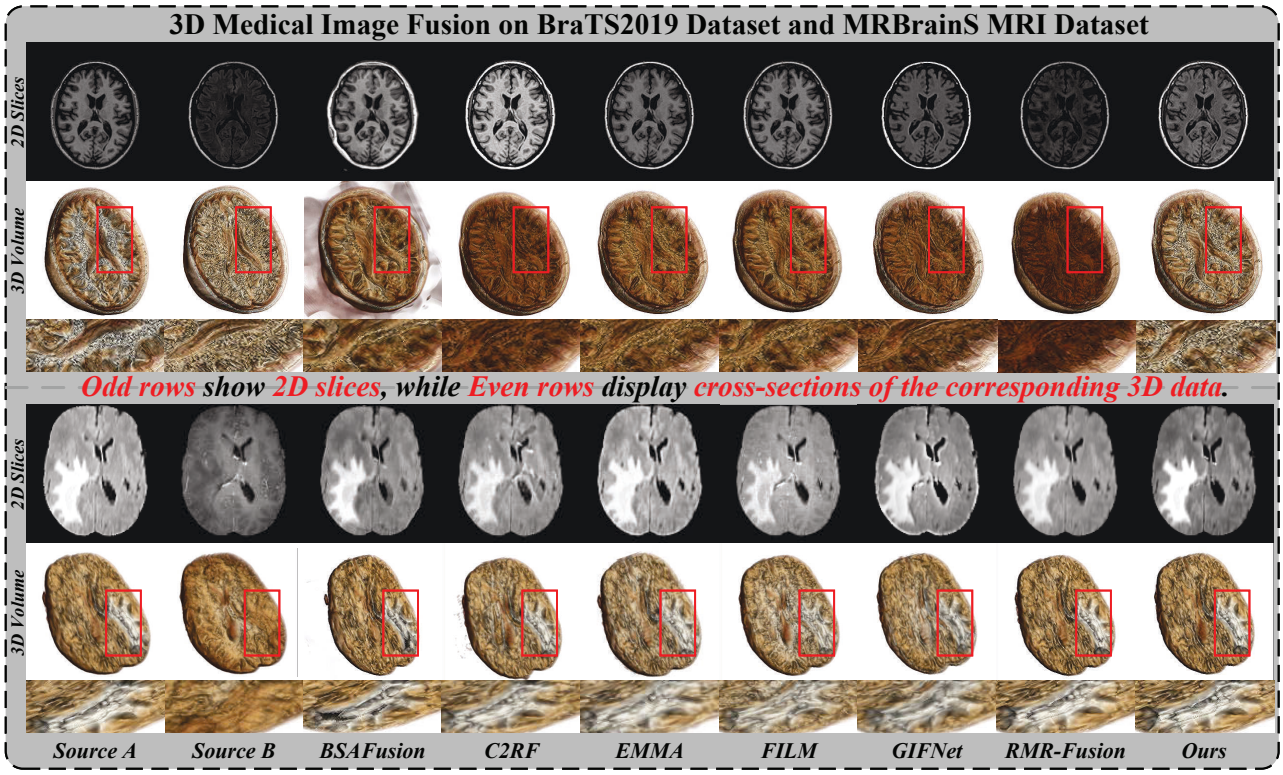

Figure 3: Qualitative comparison of various MIF models. Benefiting from MIS's precise delineation of lesions, our results retain clearer texture details and highlight the lesion area.

where $\mathcal{L}_{L_1} = ||\widehat{Y}_f - \mathcal{A}||_1 + ||\widehat{Y}_f - \mathcal{B}||_1, \mathcal{L}_{\mathcal{G}} = ||\nabla\widehat{Y}_f - \max(\nabla\mathcal{A}, \nabla\mathcal{B})||_1$, $\widehat{Y}_f$ is the fusion image. Meanwhile, the widely used Dice loss (Milletari, Navab, and Ahmadi 2016) is adopted for MIS as shown in Eq. (17):

$$\mathcal{L}_{\mathcal{S}} = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I}\widehat{Y}_{s_{i,j}}L_{i,j}}{\sum_{i=1}^{I}\widehat{Y}_{s_{i,j}} + \sum_{i=1}^{I}L_{i,j}}, \tag{17}$$

where J denotes the number of classes, I denotes the total number of pixels, $\widehat{Y}_{s_{i,j}}$ and $L_{i,j}$ represent the predicted and ground truth labels for the j-th class at the i-th voxel.

In Stage II, we adopt a cross-task weights adaptation strategy and introduce SI as a regularizer to mitigate catastrophic forgetting of task-critical information. The MIF loss function is updated as in Eq. (18):

$$\mathcal{L}_{\mathcal{F}}^{'} = \mathcal{L}_{\mathcal{F}} + \lambda\mathcal{L}_{\mathcal{SI}}, \tag{18}$$

where λ is the regularization coefficient. In the segmentation task, the loss function becomes as Eq. (19):

$$\mathcal{L}_{\mathcal{S}}^{'} = \mathcal{L}_{\mathcal{S}} + \mu\mathcal{L}_{\mathcal{SI}}, \tag{19}$$

where μ serving as its regularization parameter.

In Stage III, we embed manifold distance into loss function to constrain the shared manifold of the two tasks, and then we train both tasks jointly with WGN to obtain unified weights. The loss function is formulated as Eq. (20):

$$\mathcal{L} = \mathcal{L}_{\mathcal{F}}' + \eta\mathcal{L}_{\mathcal{S}}' + \gamma\mathcal{L}_{\mathcal{M}}, \tag{20}$$

where η and γ weight the segmentation term and the manifold-constraint term, respectively.

Experimental Results

Implementation Details

In Stage I, we set α = 0.1 for 200 epochs for both tasks; In Stage II, we set $\lambda = 0.1$, $\mu = 0.1$ for 100 epochs; Finally, we set $\eta = 2$ and $\gamma = 0.1$ in Stage III for 200 epochs. We adopt the AdamW optimizer, learning rate 3×10^{-4}. All Experiments are conducted on Ubuntu 22.04 with an Intel i9-14900K CPU, RTX 4090 GPU, and PyTorch 2.3.0. We utilize the MRBrainS MRI dataset (Mendrik et al. 2015) split into training and test sets in a 4:1 ratio, and the BraTS19 dataset (Myronenko and Hatamizadeh 2019) divided 100 random cases into training and test sets in a 9:1 ratio.

Comparison with SOTA

3D Medical Image Fusion. We select six SOTA MIF algorithms, including BSAFusion (Li et al. 2025a), C2RF (Tang et al. 2025), EMMA (Zhao et al. 2024a), FILM (Zhao et al. 2024b), GIFNet (Cheng et al. 2025), and RMR-Fusion (Zhang et al. 2024). Qualitatively, Fig. 3 presents visual results in four rows: odd rows display 2D slices from two datasets, while even rows show corresponding cross-sectional views of the resulting 3D volumes. Note that, all 2D methods perform 3D results by stacking their 2D slices. Benefiting from the guidance of MIS, our approach demonstrates superior preservation of lesion boundaries and textures. Quantitatively, we report six metrics, including Q_{MI} (Qu, Zhang, and Yan 2002), Q_{TE} (Cvejic, Canagarajah, and

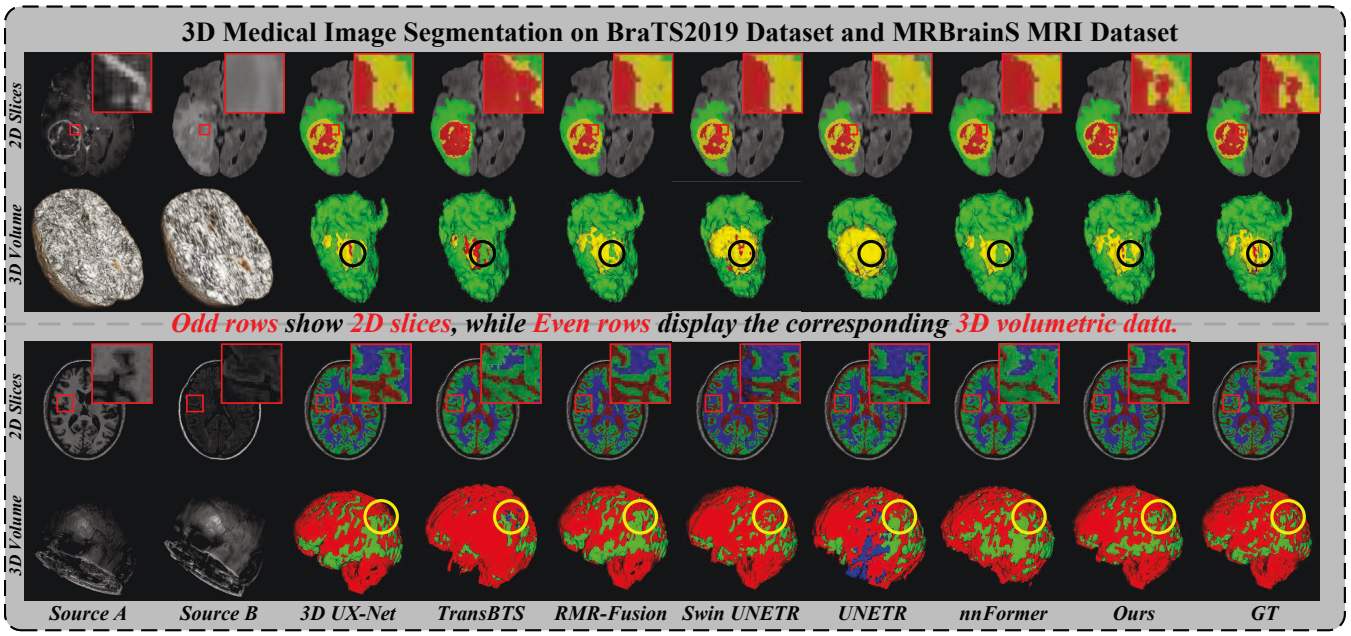

Figure 4: Qualitative comparison of various MIS models. Benefiting from the full integration and utilization of multimodal information by MIF, our results are closer to GT. Colors: For BraTS19, red/green/yellow = TC/WT/ET; while for MRBrainS MRI, red/green/blue = GM/WM/CSF.

Dataset			MRBrainS MRI Dataset						BraTS19 Dataset					
Task	Method	Pub/Year	$Q_{MI}\uparrow$	$Q_{TE}\uparrow$	$Q_{NICE}\uparrow$	$Q_G\uparrow$	$MI\uparrow$	$VIF_P\uparrow$	$Q_{MI}\uparrow$	$Q_{TE}\uparrow$	$Q_{NICE}\uparrow$	$Q_G\uparrow$	$MI\uparrow$	$VIF_P\uparrow$
MIF	BSAFusion	AAAI 25	0.7162	0.4858	0.8077	0.4208	2.8876	0.4792	0.8739	0.4174	0.8106	0.6660	3.5922	0.4852
	C2RF	IJCV 25	0.7887	0.4547	0.8082	0.6390	3.1780	0.4290	0.9356	0.4040	0.8118	0.6863	3.8924	0.5523
	EMMA	CVPR 24	0.8174	0.5286	0.8078	0.4190	3.3707	0.6096	0.9146	0.4773	0.8122	0.5589	4.0025	0.5656
	FILM	ICML 24	0.8203	0.5606	**0.8084**	0.5359	3.3730	0.5550	0.8534	0.7287	0.8116	0.4780	3.9071	0.5130
	GIFNet	CVPR 25	0.8466	0.5902	0.8083	0.6400	3.1345	0.5085	0.8656	0.7273	0.8108	0.5492	3.6522	0.4630
	RMR-Fusion	AAAI 24	0.8413	0.4570	0.8081	0.6404	3.0792	0.5943	1.1122	0.7417	0.8152	0.7077	4.5571	0.6183
	Ours	-	**0.8505**	**0.6040**	0.8083	**0.6439**	**3.4704**	**0.6318**	**1.1480**	**0.7452**	**0.8164**	**0.7077**	**4.7411**	**0.6453**
Task	Method	Pub/Year	Dice Score (%) ↑			Hausdorff Dist.(mm) ↓			Dice Score (%) ↑			Hausdorff Dist.(mm) ↓		
			CSF	GM	WM	CSF	GM	WM	ET	WT	TC	ET	WT	TC
MIS	TransBTS	MICCAI 21	0.7409	0.7123	7.7734	8.6312	6.8347	6.0558	0.7609	0.8959	0.7681	7.1735	3.1238	6.9571
	nnFormer	TIP 23	0.7787	0.7241	0.7601	6.6396	8.2773	7.1979	0.7523	0.8775	0.7751	7.4314	3.6752	6.7478
	Swin UNETR	CVPR 22	0.8143	0.8253	0.8232	5.5719	5.2416	5.3042	0.7830	0.8983	0.7815	6.5109	**3.0513**	6.5557
	UNETR	WACV 22	0.7216	0.6915	0.7321	8.3521	9.2554	8.0376	0.7531	0.8305	0.7914	7.4072	5.0859	6.2585
	3D UX-Net	ICLR 23	0.8295	0.8398	0.8635	5.1155	4.5634	4.0958	0.7851	0.8797	0.8109	6.4470	3.6095	5.6736
	RMR-Fusion	AAAI 24	**0.8323**	0.8479	0.8422	5.4981	**4.5337**	4.0124	0.8021	**0.9178**	0.8355	5.9374	3.0667	4.9352
	Ours	-	0.8312	**0.8668**	**0.8801**	**5.0642**	4.9968	**3.7973**	**0.8196**	0.9090	**0.8454**	**5.4125**	3.0782	**4.6386**

Table 1: Quantitative results averaged over two test sets for MIF and MIS; bold/underline indicate the best/second-best.

Bull 2006), Q_{NICE} (Wang, Shen, and Zhang 2005), Q_G (Xydeas and Petrovic 2000), MI (Liu et al. 2025a, 2024), and VIF_P (Wang et al. 2025a) in the upper half of Table 1 following the mainstream practice. Our model performance well, confirming that cooperative MIS enhances MIF.

3D Medical Image Segmentation. For MIS, we benchmark UNetR (Hatamizadeh et al. 2022b), TransBTS (Wang et al. 2021), Swin UNETR (Hatamizadeh et al. 2022a), nnFormer (Zhou et al. 2023a), 3D UX-Net (Lee et al. 2023), and RMR-Fusion (Zhang et al. 2024). Fig. 4 follows the same four-row layout described above. Our predictions are closest to the ground truth, benefiting from the fused multimodal information supplied by MIF. Quantitatively, Dice (Milletari, Navab, and Ahmadi 2016) and 95% Hausdorff distance (HD95) (Taha and Hanbury 2015) scores for cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM) on MRBrainS, and for enhancing tumor (ET), whole tumor (WT), and tumor core (TC) on BraTS19, are summarized in the lower half of Table 1. The improved results show that MIF and MIS indeed reinforce each other.

Ablation Study

Impact of Task Collaboration. To quantify the benefit of jointly optimizing MIF and MIS, we compare our full

Dataset		MRBrainS MRI								BraTS19							
Group	Config	$Dice\uparrow$	$HD95\downarrow$	$Q_{MI}\uparrow$	$Q_{TE}\uparrow$	$Q_{NICE}\uparrow$	$Q_G\uparrow$	$MI\uparrow$	$VIF_P\uparrow$	$Dice\uparrow$	$HD95\downarrow$	$Q_{MI}\uparrow$	$Q_{TE}\uparrow$	$Q_{NICE}\uparrow$	$Q_G\uparrow$	$MI\uparrow$	$VIF_P\uparrow$
(a1)	w/o MIF	0.850	5.013	—	—	—	—	—	—	0.792	4.751	—	—	—	—	—	—
(a2)	w/o MIS	—	—	0.840	0.600	0.804	0.643	3.432	**0.634**	—	—	0.988	0.679	0.757	0.669	4.343	0.590
(b1)	w/o Stage II	0.843	5.143	0.841	0.579	0.801	0.635	3.445	0.628	0.812	5.387	1.062	0.702	0.801	0.697	4.702	0.639
(b2)	w/o Stage III	0.714	6.234	0.752	0.543	0.758	0.628	3.089	0.594	0.729	6.842	0.983	0.614	0.758	0.667	4.381	0.615
(c1)	w/o $\mathcal{L}_{\mathcal{SI}}$	0.857	4.391	0.849	0.598	0.806	0.642	3.461	0.631	0.847	6.112	1.137	0.739	0.813	0.706	4.718	0.643
(c2)	w/o $\mathcal{L}_{\mathcal{M}}$	0.855	4.981	**0.858**	0.593	0.804	0.641	3.456	0.630	0.834	4.613	1.117	0.731	0.810	0.703	4.715	**0.647**
Default		**0.860**	**4.219**	0.851	**0.604**	**0.808**	**0.644**	**3.470**	0.632	**0.858**	**4.376**	**1.148**	**0.745**	**0.816**	**0.708**	**4.741**	0.645

Table 2: Quantitative results of ablation experiments. Bold/underline denotes the best/second best.

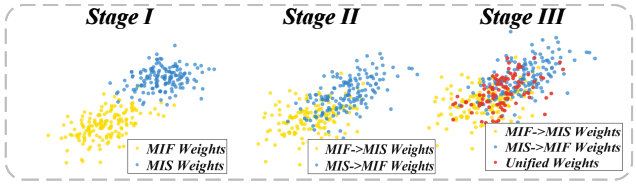

Figure 5: t-SNE visualization of weights at each stage.

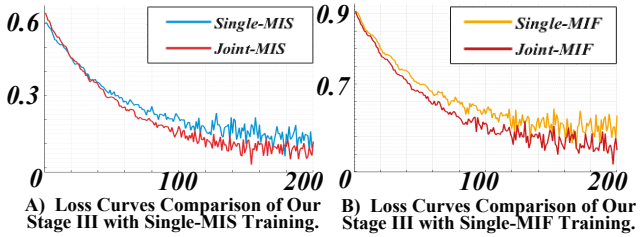

Figure 6: Visualization of loss curve.

Method	Dice↑	HD95↓	Method	Dice↑	HD95↓
TransBTS	0.808	5.752	nnFormer	0.802	5.952
TransBTS*	0.855	5.114	nnFormer*	0.823	5.309
Gain(%)	**5.71** ↑	**11.08** ↑	Gain(%)	**2.65** ↑	**10.79**↑
3D UX-Net	0.825	5.243	UNETR	0.792	6.251
3D UX-Net*	0.857	5.012	UNETR*	0.833	5.132
Gain(%)	**3.88** ↑	**4.40** ↑	Gain(%)	**5.23** ↑	**17.89**↑

Table 3: Quantitative analysis of the improvement of our method on existing MIS models.

M²-CoFS Net against models trained on only one task. As shown in Table 2(a), training exclusively for either MIS or MIF both lead to declines in performance. This confirms that without mutual guidance, each task loses the complementary information provided by the other, thereby demonstrating the critical importance of their collaboration.

Impact of our Three-Stage Training Strategy. As shown in Table 2(b), removing Stage II breaks the proximity of MIF and MIS weight distributions, causing a performance drop. Removing Stage III, which implements our core innovation of a "network-guided network learning" strategy that implicitly constructs a shared manifold to generate unified weights, causes an even larger decrease.

Impact of Regularization. The two regularizers play complementary roles. As shown in Table 2(c), removing the SI constraint results in losing important features of the previous task, whereas dropping the manifold-distance constraint damages the structure of the shared manifold. Both regularizers are indispensable for deriving unified weights.

Visualization

Weights Visualization. Fig. 5 presents the t-SNE (Maaten and Hinton 2008) visualizations of the weights at each stage. In Stage I, the MIF and MIS weights form two distant clusters, demonstrating their different objectives; in Stage II, cross-task adaptation pulls the clusters closer; in Stage III, our WGN outputs fall inside the shared region, indicating the model has learned unified task order-irrelevant weights.

Loss Curve Visualization. Fig. 6 contrasts the loss curves of our Stage III training with single-task training. The implicit shared manifold unifies separate gradients, producing faster early convergence and a lower final loss.

Broader Impact

Our key innovation is a three-stage training scheme that learns unified weights jointly optimized for MIF and MIS, and it can enhance a wide range of existing medical image segmentation networks. With all other components and settings unchanged, we replace the default backbone with UNETR, 3D-UXNet, TransBTS, and nnFormer. Table 3 lists the performance comparison. In every case the proposed strategy delivers improvements over the original models.

Conclusion

In this paper, we introduce M²-CoFS, a unified model for joint MIF and MIS. Our model consists of three sequential stages. In Stage I, two tasks are separately trained to obtain high-quality weights. In Stage II, we design cross-task weights adaptation to bring weights distributions of the two tasks closer. Stage III constitutes our core innovation, a "network guided network learning" strategy, we treat the weight spaces of MIF and MIS as high-dimensional manifolds. A lightweight WGN is designed to implicitly construct their shared manifold and sample unified task order-irrelevant weights from it. Manifold distance is embedded as a regularizer to ensure the WGN preserves the manifold's geometric structure. Finally, this network is jointly trained with task-specific heads, achieving the advanced performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [NO. 62401097]; the Natural Science Foundation of Liaoning Province (Doctoral Research Start-up Project) [NO. 2024-BS-028]; Fundamental Research Funds for Central Universities, Dalian Minzu University [No. 0854-53].

References

Cheng, C.; Xu, T.; Feng, Z.; Wu, X.; Tang, Z.; Li, H.; Zhang, Z.; Atito, S.; Awais, M.; and Kittler, J. 2025. One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28102–28112.

Cvejic, N.; Canagarajah, C.; and Bull, D. 2006. Image fusion metric based on mutual information and Tsallis entropy. *Electronics letters*, 42(11): 626–627.

Dijkstra, E. W. 2022. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: his life, work, and legacy*, 287–290.

Ding, W.; Abdel-Basset, M.; Hawash, H.; and Pedrycz, W. 2021. Multimodal infant brain segmentation by fuzzy-informed deep learning. *IEEE Transactions on Fuzzy Systems*, 30(4): 1088–1101.

Fang, L.; and Wang, X. 2022. Brain tumor segmentation based on the dual-path network of multi-modal MRI images. *Pattern Recognition*, 124: 108434.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.; and Xu, D. 2022a. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. *arXiv preprint arXiv:2201.01266*.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022b. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 574–584.

Kimmel, R.; Sochen, N.; and Malladi, R. 1997. From high energy physics to low level vision. In *Scale-Space Theory in Computer Vision: First International Conference, Scale-Space'97 Utrecht, The Netherlands, July 2–4, 1997 Proceedings 1*, 236–247. Springer.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Lee, H. H.; Bao, S.; Huo, Y.; and Landman, B. A. 2023. 3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation. In *Proceedings of the Eleventh International Conference on Learning Representations*. ICLR 2023.

Li, H.; Su, D.; Cai, Q.; and Zhang, Y. 2025a. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4725–4733.

Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; and Liu, R. 2024. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, 270–288. Springer.

Li, X.; Wang, Z.; Zou, Y.; Chen, Z.; Ma, J.; Jiang, Z.; Ma, L.; and Liu, J. 2025b. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7534–7544.

Liu, J.; Li, X.; Wang, Z.; Jiang, Z.; Zhong, W.; Fan, W.; and Xu, B. 2024. PromptFusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*.

Liu, J.; Zhang, B.; Mei, Q.; Li, X.; Zou, Y.; Jiang, Z.; Ma, L.; Liu, R.; and Fan, X. 2025a. DCEvo: Discriminative Cross-Dimensional Evolutionary Learning for Infrared and Visible Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2226–2235.

Liu, L.; Fan, X.; Zhang, C.; Dai, J.; Xie, Y.; and Liang, X. 2023. Three-dimensional medical image fusion with deformable cross-attention. In *International Conference on Neural Information Processing*, 551–563. Springer.

Liu, Q.; Chu, M.; and Thuerey, N. 2024. Config: Towards conflict-free training of physics informed neural networks. *arXiv preprint arXiv:2408.11104*.

Liu, Y.; Shi, Y.; Mu, F.; Cheng, J.; Li, C.; and Chen, X. 2022. Multimodal MRI volumetric data fusion with convolutional neural networks. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–15.

Liu, Y.; Zou, Y.; Li, X.; Zhu, X.; Han, K.; Jiang, Z.; Ma, L.; and Liu, J. 2025b. Toward a Training-Free Plug-and-Play Refinement Framework for Infrared and Visible Image Registration and Fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1268–1277.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Mallya, A.; and Lazebnik, S. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7765–7773.

Mendrik, A. M.; Vincken, K. L.; Kuijf, H. J.; Breeuwer, M.; Bouvy, W. H.; De Bresser, J.; Alansary, A.; De Bruijne, M.; Carass, A.; El-Baz, A.; et al. 2015. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Computational intelligence and neuroscience*, 2015(1): 813696.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Myronenko, A.; and Hatamizadeh, A. 2019. Robust semantic segmentation of brain tumor regions from 3D MRIs. In *International MICCAI Brainlesion Workshop*, 82–89. Springer.

Qu, G.; Zhang, D.; and Yan, P. 2002. Information measure for performance of image fusion. *Electronics letters*, 38(7): 313–315.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

Sochen, N.; Kimmel, R.; and Malladi, R. 1998. A general framework for low level vision. *IEEE transactions on image processing*, 7(3): 310–318.

Taha, A. A.; and Hanbury, A. 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1): 29.

Tang, L.; Deng, Y.; Yi, X.; Yan, Q.; Yuan, Y.; and Ma, J. 2024. DRMF: Degradation-Robust Multi-Modal Image Fusion via Composable Diffusion Prior. In *Proceedings of the ACM International Conference on Multimedia*, 8546–8555.

Tang, L.; Li, C.; and Ma, J. 2025. Mask-DiFuser: A Masked Diffusion Model for Unified Unsupervised Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Tang, L.; Yan, Q.; Xiang, X.; Fang, L.; and Ma, J. 2025. C2RF: Bridging Multi-modal Image Registration and Fusion via Commonality Mining and Contrastive Learning. *International Journal of Computer Vision*, 1–19.

Tang, W.; He, F.; Liu, Y.; and Duan, Y. 2022a. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31: 5134–5149.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022b. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740.

Wang, Q.; Shen, Y.; and Zhang, J. Q. 2005. A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena*, 200(3-4): 287–295.

Wang, W.; Chen, C.; Ding, M.; Yu, H.; Zha, S.; and Li, J. 2021. Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, 109–119. Springer.

Wang, Z.; Li, X.; Zhao, L.; Duan, H.; Wang, S.; Liu, H.; and Zhang, X. 2023. When multi-focus image fusion networks meet traditional edge-preservation technology. *International Journal of Computer Vision*, 131(10): 2529–2552.

Wang, Z.; Zhang, J.; Guan, T.; Zhou, Y.; Li, X.; Dong, M.; and Liu, J. 2025a. Efficient Rectified Flow for Image Fusion. *Advances in Neural Information Processing Systems*.

Wang, Z.; Zhang, J.; Song, H.; Ge, M.; Wang, J.; and Duan, H. 2025b. Highlight What You Want: Weakly-Supervised Instance-Level Controllable Infrared-Visible Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12637–12647.

Wang, Z.; Zhao, L.; Zhang, J.; Song, R.; Song, H.; Meng, J.; and Wang, S. 2025c. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, 4646—-4668.

Xydeas, C. S.; and Petrovic, V. 2000. Objective image fusion performance measure. *Electronics letters*, 36(4): 308–309.

Yin, M.; Liu, X.; Liu, Y.; and Chen, X. 2018. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Transactions on Instrumentation and Measurement*, 68(1): 49–64.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, 3987–3995. PMLR.

Zhang, H.; Zuo, X.; Zhou, H.; Lu, T.; and Ma, J. 2024. A robust mutual-reinforcing framework for 3d multi-modal medical image fusion based on visual-semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7087–7095.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023a. 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*, 5906–5916.

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

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023b. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8082–8093.

Zhao, Z.; Deng, L.; Bai, H.; Cui, Y.; Zhang, Z.; Zhang, Y.; Qin, H.; Chen, D.; Zhang, J.; Wang, P.; and Gool, L. V. 2024b. Image Fusion via Vision-Language Model. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; and Yu, Y. 2023a. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32: 4036–4045.

Zhou, Z.; Liu, L.; Zhao, P.; and Gong, W. 2023b. Pareto deep long-tailed recognition: A conflict-averse solution. In *The Twelfth International Conference on Learning Representations*.

Zou, Y.; Chen, Z.; Zhang, Z.; Li, X.; Ma, L.; Liu, J.; Wang, P.; and Zhang, Y. 2025. Contourlet refinement gate framework for thermal spectrum distribution regularized infrared image super-resolution. *International Journal of Computer Vision*.

Zou, Y.; Li, X.; Jiang, Z.; and Liu, J. 2024. Enhancing neural radiance fields with adaptive multi-exposure fusion: A bilevel optimization approach for novel view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7882–7890.