CMFS: CLIP-Guided Modality Interaction for Mitigating Noise in Multi-Modal Image Fusion and Segmentation

Guilin Su[1] , **Yuqing Huang**[1,2] , **Chao Yang**[1] and **Zhenyu He**[1,*]
[1]Harbin Institute of Technology, Shenzhen
[2]Peng Cheng Laboratory
{guilinsu19, domaingreen2}@gmail.com, 20b951014@stu.hit.edu.cn, zhenyuhe@hit.edu.cn

Abstract

Infrared-visible image fusion and semantic segmentation are pivotal tasks for robust scene understanding under challenging conditions such as low light. However, existing methods often struggle with high noise, modality inconsistencies, and inefficient cross-modal interactions, limiting fusion quality and segmentation accuracy. To this end, we propose CMFS, a unified framework that leverages CLIP-guided modality interaction to mitigate noise in multi-modal image fusion and segmentation. Our approach features a region-aware Modal Interaction Alignment module that combines a VMamba-based encoder with an additional shuffle layer to obtain more robust features and a CLIP-guided, regionally constrained multi-modal feature interaction block to emphasize foreground targets while suppressing low-light noise. Additionally, a Frequency-Spatial Collaboration module uses selective scanning and integrates wavelet-, spatial-, and Fourier-domain features to achieve adaptive denoising and balanced feature allocation. Furthermore, we employ a low-rank mixture-of-experts with dynamic routing to improve region-specific fusion and enhance pixel-level accuracy. Extensive experiments on several benchmarks show that, compared with state-of-the-art methods, the proposed approach demonstrates effectiveness in both image fusion quality and semantic segmentation accuracy, especially in complex environments. The source code will be released at IJCAI2025-CMFS.

1 Introduction

Infrared-visible image fusion (IVF) and multi-modal semantic segmentation (MMSS) are critical for robust scene understanding, especially under challenging conditions such as low light or adverse weather. IVF combines thermal and visible images to provide a comprehensive view, enhancing applications like night-time driving assistance and safety surveillance [Zhao *et al.*, 2023; Yi *et al.*, 2024]. Concurrently, MMSS assigns semantic labels to each pixel by leveraging information from multiple sensors, thereby improving contextual and structural scene interpretation essential for autonomous driving, robotics, and surveillance [Lv *et al.*, 2023].

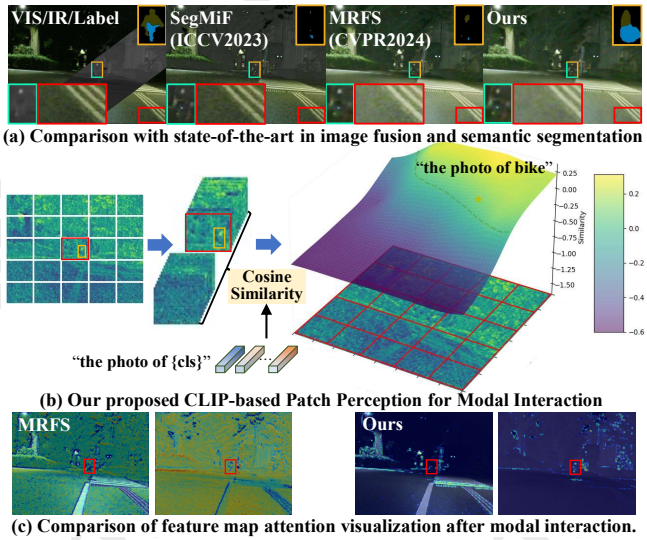

Figure 1: Visual comparison of image fusion and semantic segmentation under low-light conditions. Our method effectively perceives faint thermal objects with the guidance of CLIP, while the interaction between modal features profoundly suppresses background noise, highlights salient objects, and maintains robust performance.

Despite their importance, current approaches often treat IVF and MMSS as separate tasks, resulting in challenges related to high noise, modality inconsistencies, and inefficient cross-modal interactions. These limitations compromise both the visual quality of fused images and the accuracy of semantic segmentation. Even joint modeling attempts [Zhang *et al.*, 2024a] that use attention-based frameworks tend to be computationally intensive and sensitive to noise or imbalanced inputs. As a result, they may struggle with fine structures, elongated object shapes, and complex backgrounds, leading to incomplete scene interpretations and suboptimal fusion.

Jointly addressing IVF and MMSS remains challenging due to several factors. First, the domain gap between thermal and visible images can introduce significant noise, particularly under adverse conditions. At the same time, simple fusion strategies might propagate irrelevant or distorted information by treating all regions equally. Second, seman-

*Corresponding author

tic segmentation requires precise pixel- and region-level cues. However, noise and inconsistencies arising during fusion can hinder accurate pixel categorization. Finally, the high computational complexity inherent in many attention-based methods limits their real-time applicability. These challenges underscore the need for a robust approach to adaptively isolate inter-modal interference, preserve critical features, and efficiently model long-range dependencies.

To address these challenges, we propose a unified multi-task learning framework that simultaneously performs IVF and MMSS. Our approach begins with a Multi-modal Interaction Alignment (MIA) module, which employs a bidirectional VMamba-based encoder with an additional shuffle layer to obtain more robust features, enhancing unimodal representations and foreground saliency. This encoder partitions features and exploits CLIP's powerful classification priors using category-specific text prompts to dynamically guide region localization, as Figure 1 shows. Building on this, we introduce a regionally constrained hybrid attention (Hybrid-Att) module that selectively suppresses background noise and facilitates precise cross-modal feature interaction. Furthermore, we develop a Frequency-Space Collaboration (FSC) module to achieve joint multi-domain feature enhancement and denoising. The FSC module integrates wavelet- and spatial-domain features and leverages global Fourier-domain modeling to provide adaptive noise reduction and balanced feature allocation. Additionally, we employ a low-rank mixture-of-experts (MoE) with dynamic routing to enhance region-specific fusion and improve pixel-level accuracy while maintaining computational efficiency. The main contributions of this work can be summarized as follows:

- We propose a region-aware Modal Interaction Alignment module that combines a bidirectional VMamba-based encoder—with an additional shuffle layer and a CLIP-guided multi-modal interaction block to emphasize target regions dynamically, enhance foreground saliency, and mitigate noise.
- We develop a Frequency-Spatial Collaboration module that leverages selective scanning and integrates wavelet-, spatial-, and Fourier-domain features to achieve adaptive denoising and balanced feature allocation.
- We introduce a low-rank MoE with dynamic to improve region-specific fusion and spatial pixel accuracy.

2 Related Work

2.1 Multi-modal Learning

Multi-modal learning integrates information from diverse modalities to improve performance, overcoming the limitations of single-modal approaches, such as context sensitivity and incomplete scene understanding. In this work, we mainly focus on RGB-T tasks, where collaboration between visible and thermal images may improve robustness in challenging conditions such as low light and adverse weather.

Multi-modal Image Fusion is a fundamental task that integrates relevant and informative features from multiple modalities to generate comprehensive fused images. Autoencoders pioneered the use of deep learning techniques in this field, with methods such as DenseFuse [Li and Wu, 2018] leveraging autoencoders for feature extraction and fusion through predefined strategies. To enhance fusion effectiveness, recent approaches [Liu *et al.*, 2023; Zhang *et al.*, 2024b] incorporate high-level visual tasks to provide richer semantic guidance. TextIF [Yi *et al.*, 2024] further exploit textual cues to improve visual fidelity and perceptual robustness. However, existing multi-modal fusion methods [Zhao *et al.*, 2024c] often fail to produce visually coherent images in low-light and noisy environments, revealing significant potential for improvement.

Multi-modal Semantic Segmentation is a dense prediction task that assigns category labels to each pixel by integrating information from multiple modalities to improve accuracy and contextual understanding. Early methods like MFNet [Ha *et al.*, 2017a] used element-wise summation or concatenation for cross-modal feature fusion, often leading to redundancy and overlooking modality differences. To address these limitations, researchers introduced specialized fusion operations. For example, [Guo *et al.*, 2021] used multi-level skip connections to enhance feature flow, while [Zhang *et al.*, 2021] proposed a bridging-then-fusing strategy for multi-scale feature fusion. Attention mechanisms later became pivotal, with [Deng *et al.*, 2021] incorporating channel and spatial attention in the encoder and [Zhou *et al.*, 2022] combining cross-modal features with boundary-aware supervision for refined outputs. Recent advances emphasize exploiting modality complementarity, such as [Zhou *et al.*, 2023b]'s Transformer-based approach to leverage cross-modal correlations. However, challenges remain in fully exploring feature relationships, and interactions, and preserving spatial details.

2.2 Frequency Exploration

The Fast Fourier Transform (FFT) is a fundamental tool in frequency domain analysis, enabling efficient conversion of signals to a domain where global statistical properties are more accessible. FECNet [Huang *et al.*, 2022] leverages Fourier feature amplitudes to isolate global lightness components, improving image clarity and aesthetics. Similarly, FSDGN [Yu *et al.*, 2022] uses Fourier amplitudes as global haze indicators for image dehazing. [Kong *et al.*, 2023] introduced a frequency-domain self-attention mechanism for efficient image deblurring, and [Zhou *et al.*, 2023a] utilized the Fourier transform to model global dependencies in a novel backbone network for image degradation. While FFT has advanced many applications, its signal processing limitations highlight improvement areas. Complementing FFT, the Wavelet Transform offers multi-resolution analysis, effectively capturing localized signal variations.

2.3 State Space Models

State space models (SSMs) have gained attention for modeling long-range dependencies in sequences. The structured state-space sequence model (S4) [Gu *et al.*, 2021] was introduced to capture these dependencies with linear complexity, followed by improvements in S5 [Smith *et al.*, 2022], H3 [Fu *et al.*, 2023], and Mamba [Gu and Dao, 2023]. Notably, Mamba outperforms Transformers in linear scalability on long-sequence NLP tasks due to its data-dependent selective state space mechanism and hardware optimization. The

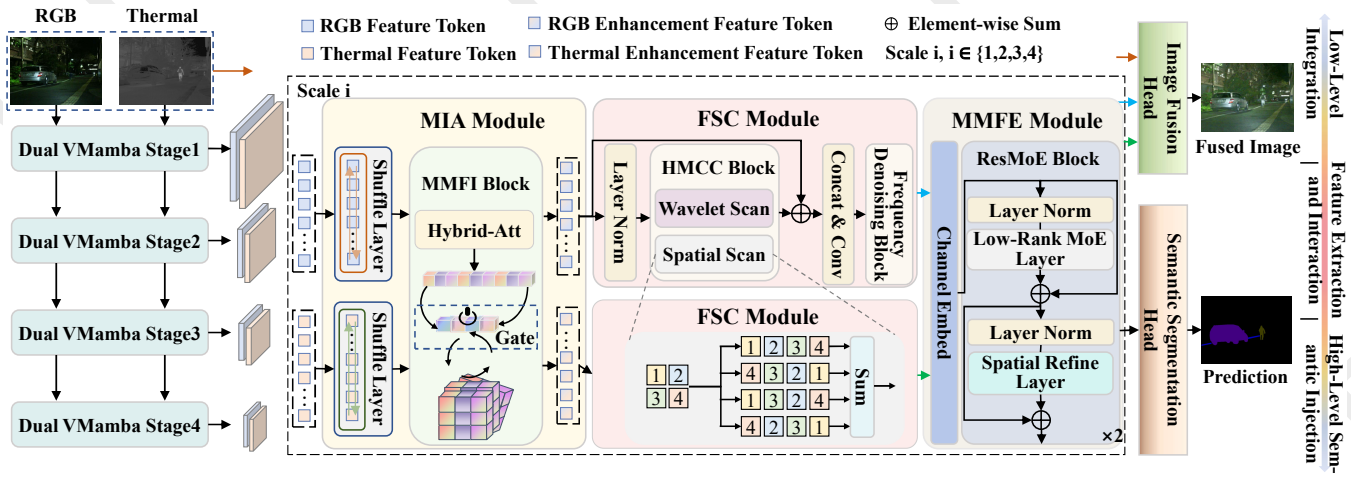

Figure 2: The overall network architecture of our CMFS. Given paired source images, multi-scale features are extracted and processed through the Modal Interaction Alignment module to enhance multi-modal interaction and target saliency. These features are then refined in the Frequency-Spatial Collaboration module, which eliminates inter-modal noise and resolves inconsistencies. The processed features are passed to the image fusion head for high-quality fusion, while the Multi-Modal Feature Enhancement module removes redundancies and further optimizes the features. Finally, the refined features are used for accurate semantic segmentation.

applicability of SSMs extends beyond NLP, with pioneering works applying Mamba to various tasks, including image classification [Zhu *et al.*, 2024; Liu *et al.*, 2024], image restoration [Guo *et al.*, 2025] and others [Li *et al.*, 2025].

3 Method

3.1 Preliminaries

State Space Models. State Space Models (SSMs) describe the dynamics of continuous systems by employing a method based on linear ordinary differential equations. They model the relationship between the input $x(t) \in \mathbb{R}$ and the output $y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^N$, which can be formulated as follows:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t), \tag{1}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are state transition matrices. The discrete variants of SSMs, such as Mamba, incorporate a discretization step parameterized by a timescale $\mathbf{\Delta}$, transforming the continuous parameters $\mathbf{A}$,$\mathbf{B}$ into their discrete counterparts $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$ through the zero-order hold (ZOH) method, expressed as:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\mathbf{\Delta A}), \\ \overline{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B}. \end{aligned} \tag{2}$$

In addition, Mamba [Gu and Dao, 2023] incorporates a selective scanning mechanism that dynamically adjusts to the input, enabling it to capture long-range contextual dependencies across different regions. This capability makes it particularly valuable for tackling complex computer vision tasks.

3.2 Overall Framework

As shown in Figure 2, CMFS is a multi-task framework that jointly tackles infrared-visible image fusion and semantic segmentation. It takes a visible image $I^{vi} \in \mathbb{R}^{H\times W\times 3}$ and thermal image $I^t \in \mathbb{R}^{H\times W\times 3}$, where H, W denote the height and width, respectively. To enhance the feature representations, CMFS progressively extracts and refines multi-modal features over n scales. At each i-th scale, VMamba-based encoders [Liu *et al.*, 2024] E_i^{vi} and E_i^t, which incorporate a bidirectional Mamba module with an additional shuffle layer to obtain more robust features, extract and refine unimodal features Φ_i^{vi} and Φ_i^t.

In extremely dark environments, both modalities often fail to capture meaningful details for objects lacking thermal radiation, resulting in severe noise and potentially introducing additional interference after interaction. To address this, the Multi-Modal Feature Interaction (MMFI) block leverages CLIP's [Radford *et al.*, 2021] category prompts and a hybrid pooling-based attention mechanism to assign region-specific weights, thereby suppressing background noise while facilitating robust inter-modal interactions. Subsequently, the Frequency-Spatial Collaboration (FSC) module further eliminates residual and amplified noise introduced during cross-modal interaction by performing long-range modeling in both the spatial and wavelet domains. Afterward, the features are sent to the Multi-Modal Feature Enhancement Module for pixel-level feature representation learning. The resulting denoised features are then fed into the fusion head to produce the final fused image F. Simultaneously, these fused features pass through a ResMoE block to extract fine-grained details, which are then fed into the segmentation head to generate the final semantic segmentation results I_s.

3.3 Multi-modal Interaction Alignment Module

RGB images provide texture information, while thermal images highlight crucial boundary cues. Fully leveraging intra-modal self-reinforcement and aligning inter-modal interactions is essential. To this end, the Multi-modal Interaction Alignment (MIA) module achieves this by combining bidirectional Mamba (BM) blocks and a MMFI block [Zhang *et al.*, 2024c] shown in Figure 3. Each BM block first updates the unimodal feature $\{\Phi^{vi}, \Phi^t\}$ to $\{\Upsilon^{vi}, \Upsilon^t\}$ through unimodal self-reinforcement and contextual modeling. The MMFI block then uses CLIP-based category prompts to guide

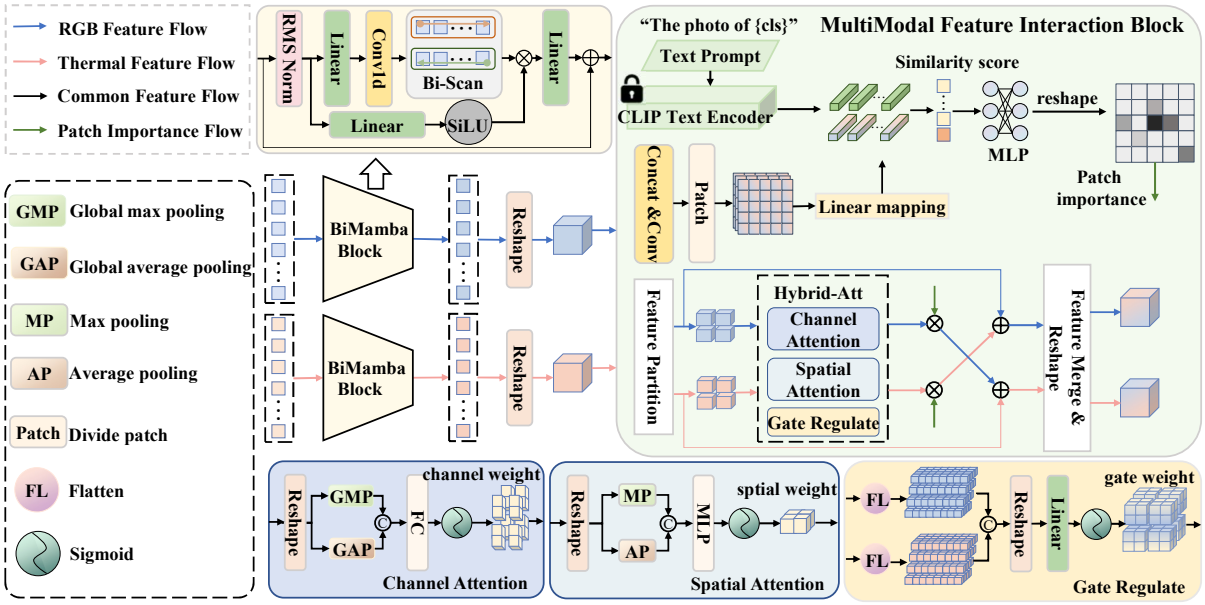

Figure 3: Structure of the Multi-modal Interaction Alignment Module. We use CLIP to achieve perception of regional importance.

region-level weighting, ensuring discriminative cross-modal interactions while suppressing background noise. Formally,

$$\{\Theta^{vi}, \Theta^{t}\} = \phi_{MMFI}(\phi_{BM}(\Phi^{vi}), \phi_{BM}(\Phi^{t})), \quad (3)$$

where Θ^{i} represents the output of related features.

Bidirectional Mamba Block. To enhance the intrinsic quality of unimodal features, we adopt a ShuffleNet-inspired approach [Zhang *et al.*, 2018] that reorders channel arrangements. This method improves feature diversity and strengthens intra-modal representations. Additionally, spatial perceptual understanding is achieved through sequential forward and backward scanning, leveraging SSM to model intra-modal relations and effectively refine and integrate critical contextual information within each modality. This process can be formulated as:

$$\{\Upsilon^{vi}, \Upsilon^{t}\} = \mathrm{SSM}(\mathrm{Shuffle}(\{\Phi^{vi}, \Phi^{t}\})). \quad (4)$$

Multi-Modal Feature Interaction (MMFI) Block. The MMFI block is designed to integrate visible and thermal features under challenging conditions. First, each unimodal feature is partitioned into non-overlapping patches to allow localized processing. Within each patch, channel attention rescales channel responses based on global pooling descriptors, while spatial attention pinpoints crucial pixel-level locations. A learnable gating mechanism then adaptively fuses signals from both modalities, preserving complementary cues while mitigating noise. To further reinforce fusion, CLIP's text encoder provides category guidance by transforming a prompt (e.g., The photo of a person) into text features. Each patch is compared with these text features, and patches with higher similarity scores, likely containing the target, receive stronger cross-modal attention. Finally, channel-spatial attention weights, gating outputs, and the CLIP-based attention map are combined to yield the updated unimodal features. The final update for each modality is defined as:

$$\begin{aligned}\Theta^{vi} &= \mathrm{Re}(\Phi^{vi} + \mathrm{Re}(\Phi^{t} \times W^{vi_c} + \Phi^{t} \times W^{vi_s}) \times W^{g} \times M) \\ \Theta^{t} &= \mathrm{Re}(\Phi^{t} + \mathrm{Re}(\Phi^{vi} \times W^{t_c} + \Phi^{vi} \times W^{t_s}) \times (1 - W^{g}) \times M),\end{aligned} \quad (5)$$

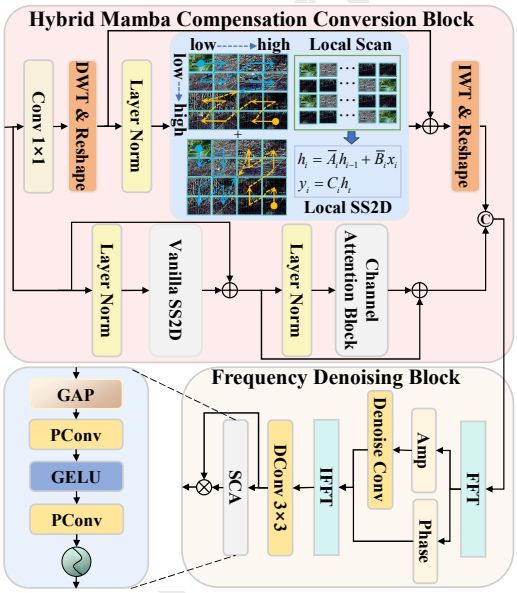

Figure 4: Overview of Frequency-Spatial Collaboration Module.

where $\mathrm{Re}(\cdot)$ indicates a reshape operation, and W_i are learnable attention weights, and M denotes the CLIP-based attention map. This synergy between localized attention, cross-modal gating, and category prompts enables robust fusion of visible and thermal streams, effectively suppressing noise and emphasizing patches most likely to contain target objects.

3.4 Frequency-Spatial Collaboration Module

Multi-modal interactions may introduce inconsistencies and noise that degrade fusion quality. As illustrated in Figure 4, the proposed FSC module addresses these by jointly modeling feature representations in both the wavelet and spatial domains, thus mitigating noise and enhancing the overall quality of the fused output. The FSC module comprises two main

Method	MFNet									PST900					FMB									
	Car	Person	Bike	Curve	CarStop	Guar.	Cone	Bump	mIoU	Hand-Drill	BackPack	Frie-Ext.	Survivor	mIoU	Car	Person	Truck	T-Lamp	T-Sign	Buil.	Vege.	Pole	mIoU
SeAFusion [Tang *et al.*, 2022]	84.2	71.1	58.7	33.1	20.1	0.0	40.4	33.9	48.8	65.6	59.6	41.1	29.5	58.9	76.2	59.6	15.1	34.4	68.0	80.1	83.5	38.4	51.9
SegFormer [Xie *et al.*, 2021]	89.5	73.2	63.8	45.9	20.8	4.14	44.8	51.5	54.7	74.3	86.4	61.1	69.3	78.1	76.5	68.4	38.7	20.9	70.6	81.4	83.8	43.9	56.3
EGFNet [Zhou *et al.*, 2022]	87.6	69.8	58.8	42.8	33.8	7.0	48.3	47.1	54.8	64.7	83.1	71.3	74.3	78.5	–	–	–	–	–	–	–	–	–
LASNet [Li *et al.*, 2023a]	84.2	67.1	56.9	41.1	39.6	18.9	48.8	40.1	54.9	77.8	86.5	82.8	75.5	84.4	73.2	58.3	33.1	32.6	68.5	80.8	83.4	41.0	55.7
SegMiF [Liu *et al.*, 2023]	87.8	71.4	63.2	47.5	31.1	0.0	48.9	50.3	56.1	66.0	81.4	76.3	75.5	79.7	78.7	65.5	42.4	35.6	71.7	80.1	85.1	35.7	58.5
MDRNet+ [Zhao *et al.*, 2024a]	87.1	69.8	60.9	47.8	34.2	8.2	50.2	55.0	56.8	63.0	76.3	63.5	71.3	74.6	75.4	67.0	27.0	41.4	68.4	79.8	82.7	45.3	55.5
GMNet [Zhou *et al.*, 2021]	86.5	73.1	61.7	44.0	42.3	14.5	48.7	47.4	57.3	85.2	83.8	73.8	78.4	84.1	–	–	–	–	–	–	–	–	–
SGFNet [Wang *et al.*, 2023]	88.4	77.6	64.3	45.8	31.0	6.0	57.1	55.0	57.6	82.8	75.8	79.9	72.7	82.1	75.0	67.2	34.6	45.8	71.4	78.2	82.7	42.8	56.0
MMSMCNet [Zhou *et al.*, 2023c]	89.2	69.1	63.5	46.4	41.9	8.8	48.8	57.6	58.1	62.4	89.2	73.3	74.7	79.8	–	–	–	–	–	–	–	–	–
CAINet [Lv *et al.*, 2023]	88.5	66.3	68.7	55.4	31.5	9.0	48.9	60.7	58.6	80.3	88.0	77.2	78.7	84.7	–	–	–	–	–	–	–	–	–
MRFS [Zhang *et al.*, 2024a]	89.4	75.4	65.0	49.0	37.2	5.4	53.1	58.8	59.1	79.7	87.4	88.0	79.6	86.9	76.2	71.3	34.4	50.1	75.8	85.4	87.0	53.6	61.2
Ours	90.5	75.6	66.3	49.2	38.5	5.5	52.7	64.4	**60.1**	78.5	88.7	88.4	84.2	**87.9**	82.5	71.2	56.3	42.2	75.9	85.4	87.3	54.0	**66.3**

Table 1: Quantitative segmentation performance on the MFNet, PST900, and FMB datasets.

components: the Hybrid Mamba Compensation Conversion (HMCC) Block and the Frequency Denoising (FD) Block.

Hybrid Mamba Compensation Conversion (HMCC) Block. The HMCC block operates in two parallel branches: one for wavelet-domain modeling and the other for spatial-domain modeling. Θ^x (with $x \in \{vi,\ t\}$) denote the output of the related features. We first modulate Θ^x with a 1×1 convolution, then apply a second-order discrete wavelet transform (DWT) to decompose the features into multiple frequency bands (LL, LH, HL, HH). A state space model with local 2D scanning [Huang *et al.*, 2024] refines these frequency bands. Finally, an inverse wavelet transform (IWT) reconstructs the wavelet-domain features [Zou *et al.*, 2024]:

$$\Theta^x_{wav} = \mathrm{IWT}\Big(\mathrm{SSM}\big(\mathrm{LN}\big(\mathrm{DWT}\big(\mathrm{Conv}_{1\times1}(\Theta^x)\big)\big)\big)\Big), \quad (6)$$

where $\mathrm{LN}(\cdot)$ denotes LayerNorm. In the spatial branch, we capture both local and global relations. Specifically, we first apply a vanilla 2D scanning within SSM to model long-range spatial correlations, then a channel attention (CA) operation to enhance local consistency. Formally,

$$\begin{aligned} \Theta^x_{sp'} &= \mathrm{SSM}\Big(\mathrm{LN}(\Theta^x)\Big) \ + \ s_1\,\Theta^x, \\ \Theta^x_{sp} &= \mathrm{CA}\Big(\mathrm{LN}(\Theta^x_{sp'})\Big) \ + \ s_2\,\Theta^x_{sp'}, \end{aligned} \quad (7)$$

where $s_1,\ s_2 \in \mathbb{R}^C$ are learnable scale factors controlling the flow of skip information. Both outputs, Θ^x_{wav} and Θ^x_{sp}, are then concatenated as follows:

$$\Theta^x_{freq-init} = \big[\Theta^x_{wav},\ \Theta^x_{sp}\big]. \quad (8)$$

Frequency Denoising (FD) Block. To further suppress mixed noise and refine the multi-domain features, we employ a Frequency Denoising (FD) block. After concatenation, we apply a Fast Fourier Transform (its degeneration-separation characteristics were validated in the supplementary materials.) to $\Theta^x_{freq-init}$, decomposing it into amplitude and phase components, Amp^x and Pha^x. A denoising convolution filters out noise in the amplitude domain, then an inverse FFT (IFFT) reconstructs the signal back to the spatial domain:

$$\Theta^x_{freq'} = \mathrm{DConv}\Big(\mathrm{IFFT}\big(\mathrm{DeConv}(\mathrm{Amp}^x),\ \mathrm{Pha}^x\big)\Big), \quad (9)$$

where $\mathrm{DConv}(\cdot)$ is a depth-wise convolution that adjusts the channel count to match the module's input dimension, and DeConv operation involves continuous pixel-wise convolution followed by the LeakyReLU activation function. Finally, we apply a simple channel attention (SCA) mechanism:

$$\Theta^x_{freq} = \mathrm{SCA}\big(\Theta^x_{freq-mid}\big) \otimes \Theta^x_{freq'}, \quad (10)$$

where $\Theta^x_{freq-mid}$ indicates an intermediate representation used for channel-attention recalibration. The result Θ^x_{freq} preserves both global structure and local detail, ensuring a robust feature representation suitable for subsequent tasks. This wavelet-spatial modeling and frequency-domain denoising fully leverage multi-domain information to suppress noise, retain crucial details, and yield semantically fused features. The denoising capability of the FSC module was further validated through the Gaussian color denoising task in the supplementary materials, demonstrating encouraging performance.

3.5 Multi-Modal Feature Enhancement Module

Semantic segmentation demands precise pixel-level classification, with different regions requiring distinct semantic cues. Prior approaches [Zhang *et al.*, 2024a] often employ self- and cross-attention mechanisms for cross-modal fusion, incurring high computational costs. To mitigate this, we propose a low-rank mixture-of-experts (MoE) mechanism that efficiently encodes fused features while eliminating redundancy. First, frequency-domain features Θ^{vi}_{freq} and Θ^{t}_{freq} are fused in a Channel Embed Block via residual convolution to produce the initial fused feature $\hat{\mathbf{o}}_{fused}$:

$$\begin{aligned} \hat{\mathbf{o}}_{fused} = \mathrm{BN}\big(&\mathrm{PConv}([\Theta^{vi}_{freq}, \Theta^{t}_{freq}]) \\ &+ \mathrm{CE}([\Theta^{vi}_{freq}, \Theta^{t}_{freq}])\big), \end{aligned} \quad (11)$$

where $[\cdot]$ is concatenation, $\mathrm{CE}(\cdot)$ is a channel embedding module composed of three convolution layers and a normalization layer, and $\mathrm{BN}(\cdot)$ denotes BatchNorm. Next, low-rank experts refine the fused features by compressing the feature space via low-rank decomposition, reducing computational complexity and filtering out redundancy. Jointly leveraging diverse subspace expressions to simplify the fused feature representation [Zamfir *et al.*, 2024]. Dynamic routing activates the most relevant experts for specific regions, enhancing semantic representations. The detailed refinement process is provided in Algorithm 1 in the supplementary materials, with the final output denoted as $\mathbf{o}_{fused}$. Finally, for object categories with elongated shapes or intricate spatial continuity, large-kernel strip convolutions (following Hou et al. [Hou *et al.*, 2024]) are applied to refine fine-grained features, thereby improving the segmentation of strip-shaped targets.

3.6 Training Objective

To generate a high-quality fused image I_f, we integrate a Gradient Residual Dense Block into the fusion head to directly produce a three-channel image with enhanced textures

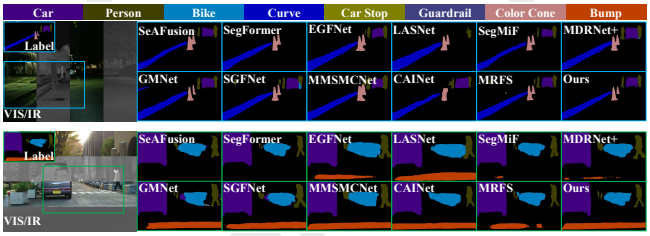

Figure 5: Qualitative segmentation results on the MFNet dataset.

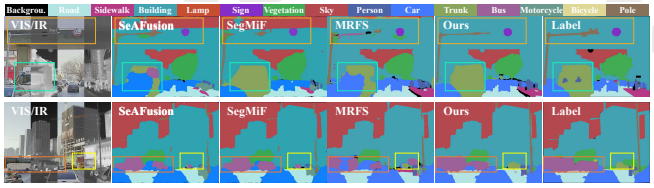

Figure 6: Qualitative segmentation results on the FMB dataset.

and without $YCbCr$ distortions. A color loss L_{color} aligns the fused Cb_F and Cr_F with the gamma-corrected channels of I_{vi}, while contrast stretching β on the thermal image improves gradient and intensity alignment. The fusion loss is defined as

$$L_{\text{fusion}} = \lambda_{\text{int}} L_1\Big(F, \max(I_{vi}, I_{\beta t})\Big) + \lambda_{\text{color}} L_{\text{color}} + \lambda_{\text{grad}} L_1\Big(\nabla F, \max(\nabla I_{vi}, \nabla I_{\beta t})\Big). \tag{12}$$

For semantic segmentation, we use the cross-entropy loss,

$$L_{\text{seg}} = -\sum P \log(I_s), \tag{13}$$

where P is the ground truth. The overall loss is

$$L = \lambda_{\text{fus}} L_{\text{fusion}} + \lambda_{\text{seg}} L_{\text{seg}}. \tag{14}$$

4 Experiments

4.1 Experimental Setting

Datasets. We evaluate the performance of CMFS on semantic segmentation and image fusion tasks using the MFNet [Ha *et al.*, 2017b], PST900 [Shivakumar *et al.*, 2020], and FMB [Liu *et al.*, 2023] datasets. Specifically, these datasets comprise 1569, 1038, and 1500 paired infrared and visible images, with resolutions of 480×640, 720×1280, and 600×800, respectively. Among these, 393, 288, and 280 image pairs are used for testing.

Implementation Details. The semantic segmentation and image fusion tasks are trained jointly for 500 epochs to enable effective multi-task learning. The training process adopts an initial learning rate of 6×10^{-5}, a batch size of 4, and the Adam optimizer with a weight decay coefficient of 0.01. During training, standard data augmentation techniques such as horizontal flipping, random scaling, and cropping are applied. All training images are cropped to a uniform size of 480×640. Refining coefficient value settings in the loss function: $\lambda_{int} = 0.5$, $\lambda_{grad} = 0.2$, $\lambda_{color} = 1.0$, $\lambda_{fus} = 0.1$, $\lambda_{seg} = 1.0$. As shown in Figure 2, we adopt a four-stage Vmamba-tiny encoder, which operates across four distinct scales. All experiments are conducted on the NVIDIA L40 GPU with 46GB of memory, coupled with an Intel Xeon Silver 4310 10-Core Processor CPU.

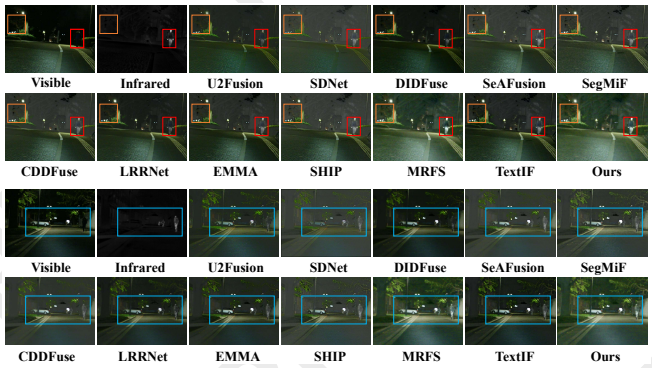

Figure 7: Visual comparison of image fusion on the MFNet dataset.

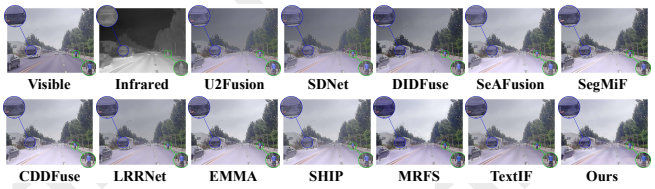

Figure 8: Visual comparison of image fusion on the FMB dataset.

4.2 Multi-modal Semantic Segmentation

To ensure a fair comparison, we primarily used data from [Zhang *et al.*, 2024a] and quantitative results reported in the respective papers, with visualizations sourced from their open-source code. Since most methods were not trained on the FMB dataset, we compared only models with publicly available optimal weights. We used mean Intersection over Union (mIoU) as the evaluation metric, which effectively reflects segmentation performance in complex scenes while mitigating the impact of class imbalance. As Figure 5 and Figure 6 shown, our method demonstrates precise segmentation of target categories in both day and night conditions, with accurate classification of edge pixels, aided by the MMFE module. It performs consistently across categories and handles ambiguous pixels effectively, even in low-light settings. For example, in Figure 5, our method delineates the contours of objects like bikes and bumps with high accuracy, achieving realistic visual results. In contrast, other methods either approximate the regions or miss features entirely. The quantitative results in Table 1 demonstrate the superiority of our method. It achieves the highest mIoU on all three datasets, with a maximum gain of 5.1% over the second-best method. To this end, we substituted the PC-Att module in the state-of-the-art MRFS model with our MMFE module, adhering strictly to its original training settings. After training for the same number of epochs on the FMB dataset, we observed a significant improvement in segmentation performance, with the mIoU increasing from **61.2** to **63.7**. Additionally, the GFLOPs were reduced from **219.13** to **177.65**, demonstrating a substantial reduction in computational complexity. Overall, quantitative and qualitative evaluations indicate that CMFS performs well in semantic segmentation.

4.3 Multi-modal Image Fusion

We evaluate the image fusion performance of CMFS on the MFNet and FMB datasets using metrics such as standard de-

Methods	Venue	MFNet Dataset			FMB Dataset		
		SD↑	EN↑	MI↑	SD↑	EN↑	MI↑
U2Fusion [Xu *et al.*, 2020]	TPAMI'20	25.273	6.002	2.385	30.783	6.678	3.046
DIDFuse [Zhao *et al.*, 2020]	IJCAI'20	35.682	6.383	2.938	37.796	6.989	3.278
SDNet [Zhang and Ma, 2021]	IJCV'21	19.623	5.808	1.902	34.832	6.606	3.173
SeAFusion [Tang *et al.*, 2022]	Inf.Fusion'22	34.011	6.422	3.479	36.161	6.754	3.883
SegMiF [Liu *et al.*, 2023]	ICCV'23	36.066	6.557	2.648	37.430	6.872	3.263
CDDFuse [Zhao *et al.*, 2023]	CVPR'23	38.062	6.496	**4.252**	37.034	6.776	4.276
LRRNet [Li *et al.*, 2023b]	TPAMI'23	32.369	6.379	3.005	26.258	6.278	3.021
EMMA [Zhao *et al.*, 2024b]	CVPR'24	40.316	6.612	3.730	36.797	6.773	4.012
SHIP [Zheng *et al.*, 2024]	CVPR'24	34.759	6.474	4.052	34.090	6.657	**5.131**
MRFS [Zhang *et al.*, 2024a]	CVPR'24	**42.333**	**7.046**	3.182	38.039	6.777	3.461
TextIF [Yi *et al.*, 2024]	CVPR'24	39.997	6.723	4.246	32.813	6.639	4.281
Ours	Ours	41.931	7.045	3.240	**43.083**	**6.898**	3.522

Table 2: Quantitative image fusion results on the MFNet and FMB datasets. The best and second-best performances are highlighted in bold and underlined, respectively.

Module	MMFI		FSC	MMFE	Fusion	mIoU
Ablation Part	BiMamba	Feature Partition & CLIP	FSC	ResMoE	Fusion Head	
Baseline					✓	63.8 (-2.5)
I		✓	✓	✓	✓	65.9 (-0.4)
II	✓		✓	✓	✓	64.4 (-1.9)
III	✓	✓		✓	✓	64.6 (-1.7)
IV	✓	✓	✓		✓	62.7 (-3.6)
V	✓	✓	✓	✓		63.2 (-3.1)
Full Model	✓	✓	✓	✓	✓	**66.3** (0.0)

Table 3: Ablation experiments of CMFS on the FMB dataset. The relative decrease in mIoU, following the ablation of each component compared to the Full Model, is highlighted in blue.

viation (SD) [Aslantas and Bendes, 2015], entropy (EN), and mutual information (MI) [Qu *et al.*, 2002] to assess image detail preservation, multi-modal information integration, and retention of complementary information. Visual results in Figure 7 and 8 show that CMFS achieves high visual fidelity, effectively removing noise even in low-light conditions. As shown in Figure 7, CMFS avoids the style bias in nighttime infrared images and effectively restores texture details like road traffic signs. Figure 8 highlights its ability to eliminate modality misalignment artifacts, ensuring sharper object boundaries. The FSC module further enhances texture, sharpens edges, and aligns the output with human perception. Quantitative results in Table 2 confirm CMFS's good performance, achieving encouraging scores across most metrics compared to state-of-the-art methods.

4.4 Ablation Studies

We conduct comprehensive ablation studies to assess the effectiveness of the specific design choices in our method, systematically evaluating six variants. **Baseline Model:** removes all components of our design, retaining only the fusion and semantic segmentation heads, and uses residual convolutional networks (Channel Embed) to fuse features from the two modalities. **Model I:** removes the Bimamba block to evaluate the impact of intra-modal feature self-reinforcement. **Model II:** omits the feature partitioning step and CLIP regional importance guidance in the MMFI module, keeping only the hybrid attention mechanism with gated correction. **Model III:** discards the FSC module to highlight the contribution of multi-domain joint modeling for denoising. **Model IV:** replaces the MMFE module with a simple single-layer 1x1 convolution to demonstrate the necessity of feature refinement and spatial modulation. **Model V:** removes the image fusion head to assess the mutual enhancement between the two tasks. Based on the controlled variable analysis presented in Table 3, it is evident that removing any individual module from the full model results in a decrease in segmentation performance, indicating that each of the designed modules contributes positively to overall segmentation accuracy. In comparison to the baseline, the **Model IV** and **V** configurations show lower performance, highlighting the importance of initial fusion via the residual network (Channel Embed Block) and the synergistic interaction between multimodal fusion and segmentation. Notably, the performance drop is most pronounced upon removal of the ResMoE Block, reinforcing the value of utilizing a MoE mechanism to refine features and eliminate redundancy, which provides a novel approach for enhancing segmentation performance. Visual comparisons in Figure 9 support these findings by introducing another extreme environment of smoke, confirming that our framework effectively meets the objectives of multi-modal segmentation and fusion tasks and demonstrates broad applicability. Additionally, the visual results in Figure 10 also substantiate the positive contribution of the designed modules to image fusion fidelity. The conclusion can be drawn by comparing the distribution and contrast of the surrounding artifacts of the person in the image. These results emphasize the effectiveness of these designs in improving image fusion and highlight the positive impact of semantic segmentation.

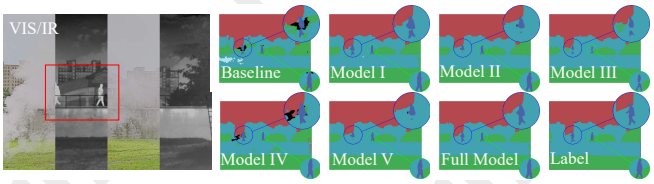

Figure 9: Qualitative segmentation of ablation studies. Our full model consistently achieves accurate segmentation of objects across different categories, even in smoke-filled environments.

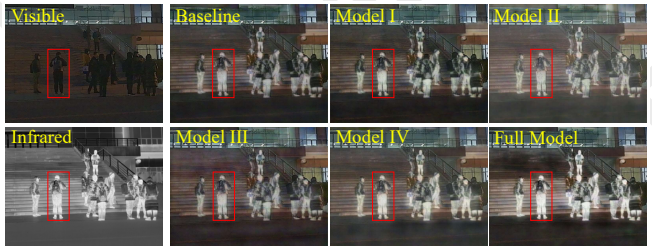

Figure 10: Qualitative fusion of ablation studies. Removing any component from the full model increases the blurriness of salient objects, such as the person in the red box.

5 Conclusion

In this work, we presented a unified framework for infrared-visible image fusion and multi-modal semantic segmentation that effectively tackles noise, modality inconsistencies, and inefficient cross-modal interactions. Our approach improves both fusion quality and segmentation accuracy through robust feature extraction, CLIP-guided alignment, and adaptive denoising, as validated by extensive experiments that outperform state-of-the-art methods.

Acknowledgements

We thank the reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (No. 62172126), and the Shenzhen Research Council (No. JCYJ20210324120202006).

References

[Aslantas and Bendes, 2015] V. Aslantas and E. Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU - International Journal of Electronics and Communications*, 69(12):1890–1896, 2015.

[Deng *et al.*, 2021] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *IROS*, pages 4467–4473, 2021.

[Fu *et al.*, 2023] Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry Hungry Hippos: Towards language modeling with state space models. In *ICLR*, 2023.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Gu *et al.*, 2021] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[Guo *et al.*, 2021] Zhifeng Guo, Xu Li, Qimin Xu, and Zhengliang Sun. Robust semantic segmentation based on rgb-thermal in variable lighting scenes. *Measurement*, 186:110176, 2021.

[Guo *et al.*, 2025] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, pages 222–241. Springer, 2025.

[Ha *et al.*, 2017a] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, pages 5108–5115, 2017.

[Ha *et al.*, 2017b] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, pages 5108–5115, 2017.

[Hou *et al.*, 2024] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE TPAMI*, 2024.

[Huang *et al.*, 2022] Jie Huang, Yajing Liu, Xueyang Fu, Man Zhou, Yang Wang, Feng Zhao, and Zhiwei Xiong. Exposure normalization and compensation for multiple-exposure correction. In *CVPR*, pages 6043–6052, 2022.

[Huang *et al.*, 2024] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024.

[Kong *et al.*, 2023] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *CVPR*, pages 5886–5895, 2023.

[Li and Wu, 2018] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE TIP*, 28(5):2614–2623, 2018.

[Li *et al.*, 2023a] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE TCSVT*, 33(3):1223–1235, March 2023.

[Li *et al.*, 2023b] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. LRRNet: A novel representation learning guided fusion framework for infrared and visible images. *IEEE TPAMI*, 45(9):11040–11052, 2023.

[Li *et al.*, 2025] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *ECCV*, pages 237–255. Springer, 2025.

[Liu *et al.*, 2023] 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 *ICCV*, 2023.

[Liu *et al.*, 2024] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[Lv *et al.*, 2023] Ying Lv, Zhi Liu, and Gongyang Li. Context-aware interaction network for rgb-t semantic segmentation. *IEEE TMM*, pages 1–13, 2023.

[Qu *et al.*, 2002] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38:313–315, 2002.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[Shivakumar *et al.*, 2020] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *ICRA*, pages 9441–9447, 2020.

[Smith *et al.*, 2022] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

[Tang *et al.*, 2022] 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, 2022.

[Wang *et al.*, 2023] Yike Wang, Gongyang Li, and Zhi Liu. Sgfnet: Semantic-guided fusion network for rgb-thermal semantic segmentation. *IEEE TCSVT*, 2023.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

[Xu *et al.*, 2020] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE TPAMI*, 2020.

[Yi *et al.*, 2024] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *CVPR*, 2024.

[Yu *et al.*, 2022] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *ECCV*, pages 181–198. Springer, 2022.

[Zamfir *et al.*, 2024] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yulun Zhang, and Radu Timofte. See more details: Efficient image super-resolution by experts mining. In *ICML*, 2024.

[Zhang and Ma, 2021] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *IJCV*, pages 1–25, 2021.

[Zhang *et al.*, 2018] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018.

[Zhang *et al.*, 2021] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *CVPR*, pages 2633–2642, 2021.

[Zhang *et al.*, 2024a] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *CVPR*, pages 26974–26983, 2024.

[Zhang *et al.*, 2024b] Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, Wenbo Huang, Yunsong Li, and Xue Yang. E2e-mfd: Towards end-to-end synchronous multimodal fusion detection. In *NeurIPS*, 2024.

[Zhang *et al.*, 2024c] Zhiwei Zhang, Yisha Liu, Weimin Xue, and Yan Zhuang. Cigf-net: Cross-modality interaction and global-feature fusion for rgb-t semantic segmentation. *IEEE TETCI*, 2024.

[Zhao *et al.*, 2020] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976. ijcai.org, 2020.

[Zhao *et al.*, 2023] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *CVPR*, pages 5906–5916, June 2023.

[Zhao *et al.*, 2024a] Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE TNNLS*, 35(7):9380–9394, 2024.

[Zhao *et al.*, 2024b] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *CVPR*, June 2024.

[Zhao *et al.*, 2024c] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-language model. In *ICML*, 2024.

[Zheng *et al.*, 2024] Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *CVPR*, pages 26374–26385, 2024.

[Zhou *et al.*, 2021] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE TIP*, 30:7790–7802, 2021.

[Zhou *et al.*, 2022] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. *AAAI*, 36(3):3571–3579, Jun. 2022.

[Zhou *et al.*, 2023a] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *ICML*, pages 42589–42601. PMLR, 2023.

[Zhou *et al.*, 2023b] Wujie Zhou, Ying Lv, Jingsheng Lei, and Lu Yu. Embedded control gate fusion and attention residual learning for rgb–thermal urban scene parsing. *IEEE T-ITS*, 24(5):4794–4803, 2023.

[Zhou *et al.*, 2023c] 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 TCSVT*, 33(12):7096–7108, 2023.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *ICML*, 2024.

[Zou *et al.*, 2024] Zhen Zou, Hu Yu, Jie Huang, and Feng Zhao. Freqmamba: Viewing mamba from a frequency perspective for image deraining. In *ACM MM*, pages 1905–1914, 2024.