Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion and Beyond

Guanyao Wu[†], Haoyu Liu[†], Hongming Fu[†], Yichuan Peng[†], Jinyuan Liu[‡], Xin Fan[†], Risheng Liu[†*]
[†]School of Software Technology, Dalian University of Technology, China
[‡]School of Mechanical Engineering, Dalian University of Technology, China
rollingplainko@gmail.com, atlantis918@hotmail.com, {xin.fan, rsliu}@dlut.edu.cn

Abstract

*Multi-modality image fusion, particularly infrared and visible, plays a crucial role in integrating diverse modalities to enhance scene understanding. Although early research prioritized visual quality, preserving fine details and adapting to downstream tasks remains challenging. Recent approaches attempt task-specific design but rarely achieve "The Best of Both Worlds" due to inconsistent optimization goals. To address these issues, we propose a novel method that leverages the semantic knowledge from the **S**egment **A**nything Model (SAM) to **G**row the quality of fusion results and **E**nable downstream task adaptability, namely **SAGE**. Specifically, we design a Semantic Persistent Attention (SPA) Module that efficiently maintains source information via the persistent repository while extracting high-level semantic priors from SAM. More importantly, to eliminate the impractical dependence on SAM during inference, we introduce a bi-level optimization-driven distillation mechanism with triplet losses, which allow the student network to effectively extract knowledge. Extensive experiments show that our method achieves a balance between high-quality visual results and downstream task adaptability while maintaining practical deployment efficiency. The code is available at* https://github.com/RollingPlain/SAGE_IVIF.

1. Introduction

In light of recent advancements in sensing technologies, multi-modality imaging [15] has gained significant attention, with applications in robotics [35], remote sensing [33], and autonomous driving [1]. Integrating infrared and visible sensors provides substantial advantages for intelligent processing [5], although both exhibit inherent limitations [16, 21]. Infrared images are robust against smoke, obstruction, and low light, while visible images excel in resolution, contrast, and texture detail. These characteristics are highly complementary, motivating their fusion to generate a comprehensive fused image that balances visual quality and downstream task requirements. Consequently, achieving this fusion efficiently remains a pressing challenge [15].

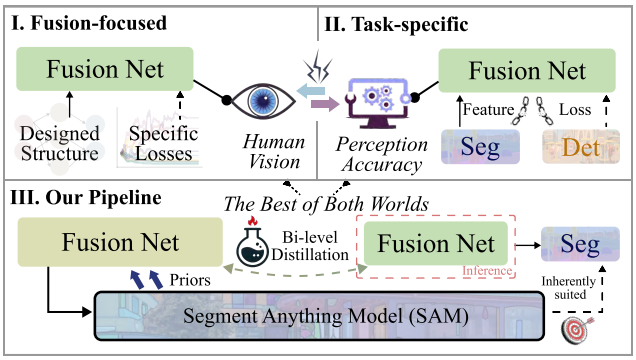

Figure 1. Differences between the proposed method and existing mainstream comparative approaches: (a) Traditional and early DL-based methods focus on the fusion visual effect. (b) Task-specific methods (e.g., TarDAL [11] & SegMiF [13]) introduce task loss and features that lead to inconsistent optimization goals, causing a conflict between visual and task accuracy. (c) Our pipeline first leverages semantic priors from SAM within a large network and then distills the knowledge into a smaller sub-network achieving practical inference feasibility while ensuring "the best of both worlds" through SAM's inherent adaptability to these tasks.

Traditional Infrared and Visible Image Fusion (IVIF) methods, based on information theory [17, 37], aim to retain as much source image information as possible but often struggle with optimizing fused image quality, especially in managing redundancy and specific scenarios. Early deep learning-based methods [8, 12, 18, 22, 32, 39, 40] focused on visualization but faced issues like edge blurring and artifacts, failing to meet downstream perception needs. Current methods [11, 13, 20, 26] focus on specific tasks, coupling fusion with downstream processing, resulting in conflicting optimization objectives and challenges in balancing visual quality with task adaptability.

*Corresponding author.

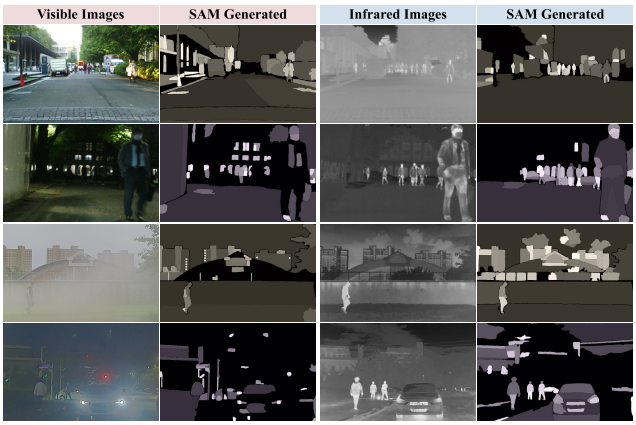

Figure 2. Demonstration of SAM's robustness in MFNet normal scenes (top) and under challenging FMB conditions (bottom).

Recent advancements in large-scale visual models have significantly improved various visual analysis tasks [30]. Among these, the Segment Anything Model [6] stands out due to its exceptional and robust capability to provide rich semantic information, making it well-suited for IVIF, as illustrated in Figure 2. Also, its segmentation capabilities align naturally with the requirements of downstream tasks in IVIF field. However, current methods that integrate SAM for low-level tasks typically require full SAM during inference, resulting in excessive practical infeasibility.

To address these issues, we propose the fusion method SAGE that fully integrates and distills semantic priors derived from SAM. As illustrated in Figure 1 III, one of our aim is to harness the advantages of SAM's semantic priors while reducing computational burden. The proposed method consists of two key components: Semantic Persistent Attention module and a bi-level distillation scheme. The former focuses on using a persistent repository throughout the process to retain source image information, while heuristically enabling the network to integrate semantic patches predicted by SAM. Furthermore, the distillation scheme is designed to efficiently exclude SAM from the inference process, thereby reducing computational complexity. By enhancing distilled information across different aspects, the triplet losses in the scheme significantly boost the performance of the fusion model. Consequently, efficient fusion is achieved using only the distilled model, delivering high-quality visual results and precise task performance without the direct involvement of SAM. In summary, our contributions can be summarized as follows:

- A novel fusion framework is proposed to leverage semantic priors from the SAM, effectively balancing visual quality and downstream task adaptability..
- The designed Semantic Persistent Attention (SPA) module utilizes a persistent repository to efficiently retain source information and simultaneously extract high-level semantic representation.
- We develop a bi-level optimization distillation scheme that transfers the information processed by SPA into the sub-network, effectively decoupling the fusion process from SAM during inference.
- Extensive experiments on multiple datasets validate the superior effectiveness of our proposed method.

2. Related Works

Learning based IVIF Methods. In recent years, deep learning based multi-modality image fusion approaches [8, 12, 32, 39, 40, 44] achieved significant progress, primarily focusing on improving fusion quality, such as enhancing image details, maintaining structural consistency, and reducing noise. These approaches, including innovations in adversarial learning, Transformer models and diffusion network, have provided new insights. However, most of these methods have not addressed how the fused images can directly contribute to downstream tasks such as object detection, semantic segmentation, and other perception-related applications. With the growing demand for more integrated solutions, this gap has sparked increasing research interest in combining fusion with various downstream tasks.

Early works began to bridge this gap, with TarDAL [11] pioneering the use of a dual-discriminator network for object detection optimization, which simultaneously considers modality differences and detection tasks during fusion. As the field has progressed, several methods [24, 38] have also focused on optimizing fusion for object detection. For semantic segmentation [36], SeAFusion [26] and SegMiF [13] have made early significant improvements, with the former enhancing fusion via task loss and the latter leveraging task-specific features. More recently, some approaches [3, 14, 19, 20, 25, 29] have integrated multi-task learning to jointly optimize detection and segmentation, improving performance on both tasks.

SAM & Its Application. The Segment Anything Model (SAM) [6] is a large-scale pre-trained foundation model that excels in zero-shot generalization, enabling it to handle a wide range of segmentation and detection tasks without the need for task-specific training. SAM demonstrates impressive performance in low-level vision applications, such as image deblurring [9], dehazing [4], super-resolution [28], and low-light image enhancement [7]. Moreover, SAM's capabilities extend to high-level tasks like object detection [23], which further strengthens its versatility. This ability to handle both low-level and high-level tasks makes SAM inherently well-suited for the needs of the IVIF field, where fusion based segmentation and detection tasks are critical. However, its impressive performance is hindered by high computational complexity, posing challenges for practical deployment. Addressing these computational burdens, while fully leveraging the semantic priors of SAM, is a primary motivation for our work.

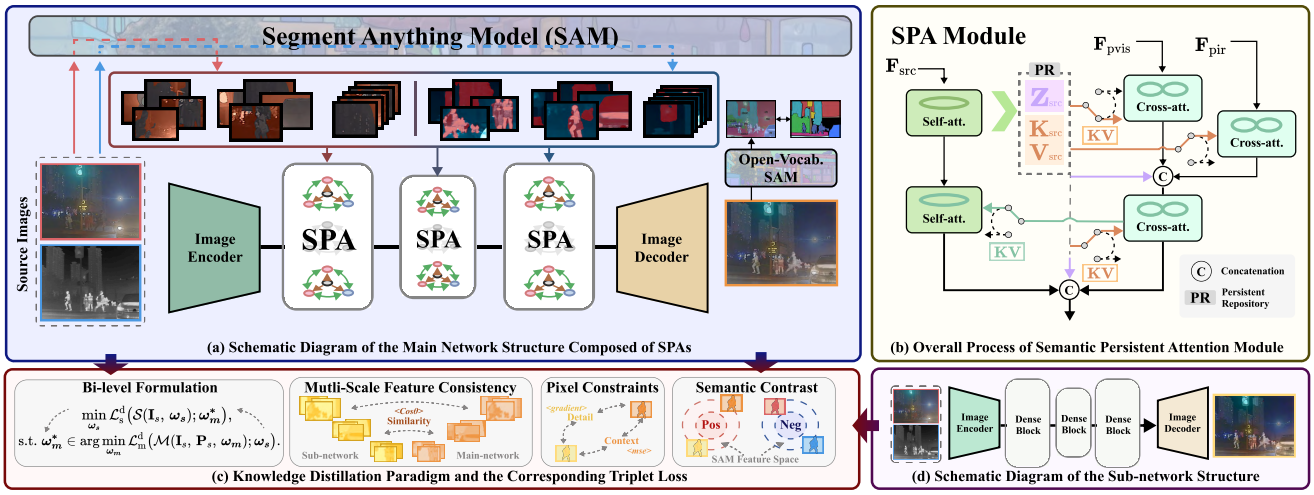

Figure 3. A overall workflow of our proposed method. (a) shows the flow structure of the main network, where the SPA module processes patches with semantic priors generated by SAM. (b) illustrates the detailed structure of the SPA module, where PR plays a key role in preserving the source and integrating the semantic information. (c) displays our distillation scheme formulation, with visualizations of the different components of the triplet loss. (d) provides a simple diagram of the sub-network, which is composed of stacked dense blocks.

3. The Proposed Method

3.1. Problem Formulation

The overall workflow of our method is illustrated in Figure 3. Our aim is to fully exploit the semantic priors of SAM during the inference stage to enhance cross-modal fusion quality. However, directly deploying a large-scale SAM model often leads to prohibitive computational overhead. To mitigate this, we adopt a knowledge distillation strategy, transferring the information encoded by the SAM-driven main network to a lightweight sub-network, thereby significantly reducing inference costs while maintaining high-quality fusion. Nevertheless, the substantial capacity gap between the SAM-enhanced main network and the compact sub-network frequently results in incomplete semantic transfer or structural inconsistencies, which hinder ideal cross-modal fusion performance. To address this issue, we propose a bi-level optimization framework that jointly optimizes both networks as a unified system, aiming to bridge the distillation gap and maintain consistent fusion guided by SAM's semantic priors.

Let $\mathbf{I}_\mathrm{s}$ denote the source visible and infrared input images, *i.e.*, $\mathbf{I}_\mathrm{s} = \{\mathbf{I}_\mathrm{vis}, \mathbf{I}_\mathrm{ir}\}$. Their corresponding semantic-prior patches generated under the guidance of SAM are denoted as $\mathbf{P}_\mathrm{s} = \{\mathbf{P}_\mathrm{vis}, \mathbf{P}_\mathrm{ir}\}$. Denote the main network by $\mathcal{M}$ with parameters $\boldsymbol{\omega}_\mathbf{m}$, and the sub-network by $\mathcal{S}$ with parameters $\boldsymbol{\omega}_\mathbf{s}$. We define the optimization process as:

$$\begin{aligned} &\min_{\boldsymbol{\omega_s}} \mathcal{L}_\mathrm{s}^\mathrm{d}\big(\mathcal{S}(\mathbf{I}_s,\, \boldsymbol{\omega_s}); \boldsymbol{\omega}_{\boldsymbol{m}}^*\big), \\ \text{s.t.} \quad &\boldsymbol{\omega}_{\boldsymbol{m}}^* \in \arg\min_{\boldsymbol{\omega_m}} \mathcal{L}_\mathrm{m}^\mathrm{d}\big(\mathcal{M}(\mathbf{I}_s,\, \mathbf{P}_s,\, \boldsymbol{\omega_m}); \boldsymbol{\omega_s}\big), \end{aligned} \tag{1}$$

where $\mathcal{L}_\mathrm{m}^\mathrm{d}$ is the distillation loss that ensures high-quality fusion with meaningful semantic cues in the main network, and $\mathcal{L}_\mathrm{s}^\mathrm{d}$ is the distillation loss guiding the sub-network to effectively mimic the main network's behavior. The fused reference image is given by: $\mathbf{I}_\mathrm{ref} = \mathcal{M}\big(\{\mathbf{I}_\mathrm{vis}, \mathbf{I}_\mathrm{ir}\}, \{\mathbf{P}_\mathrm{vis}, \mathbf{P}_\mathrm{ir}\}, \boldsymbol{\omega_m}\big)$, while the sub-network produces a fused image $\mathbf{I}_\mathrm{f} = \mathcal{S}\big(\mathbf{I}_\mathrm{vis}, \mathbf{I}_\mathrm{ir}, \boldsymbol{\omega_s}\big)$. Notably, both distillation objectives rely on outputs from the counterpart network, establishing a bidirectional dependency, which is a key feature of our distillation formulation.

Within the overall framework, the main network integrates SAM-derived semantic priors into the fusion process, while the sub-network is optimized to align with its outputs under the bi-level formulation. The following sections elaborate on two key components that support this process: the Semantic Persistent Attention (SPA) module 3.2, which integrates SAM-derived priors with source features, and the triplet-based loss scheme 3.3, which facilitates structured knowledge transfer across networks.

3.2. Semantic Persistent Attention Module

To fully exploit the semantic information provided by SAM, we propose the Semantic Persistent Attention (SPA) module, as illustrated in the flowchart in Figure 3 (b).

The key core of the SPA module is a Persistent Repository (PR), which serves as a static memory to store and maintain crucial contextual information for the fusion process. Specifically, PR stores the latent representation ($\mathbf{Z}$) of the source features $\mathbf{F}_\mathrm{src}$ and the corresponding key-value pairs ($\mathbf{K}_\mathrm{src}, \mathbf{V}_\mathrm{src}$), which provide consistent contextual support during the cross-attention operation. The core insight behind this design is that PR acts as a stable, modality-

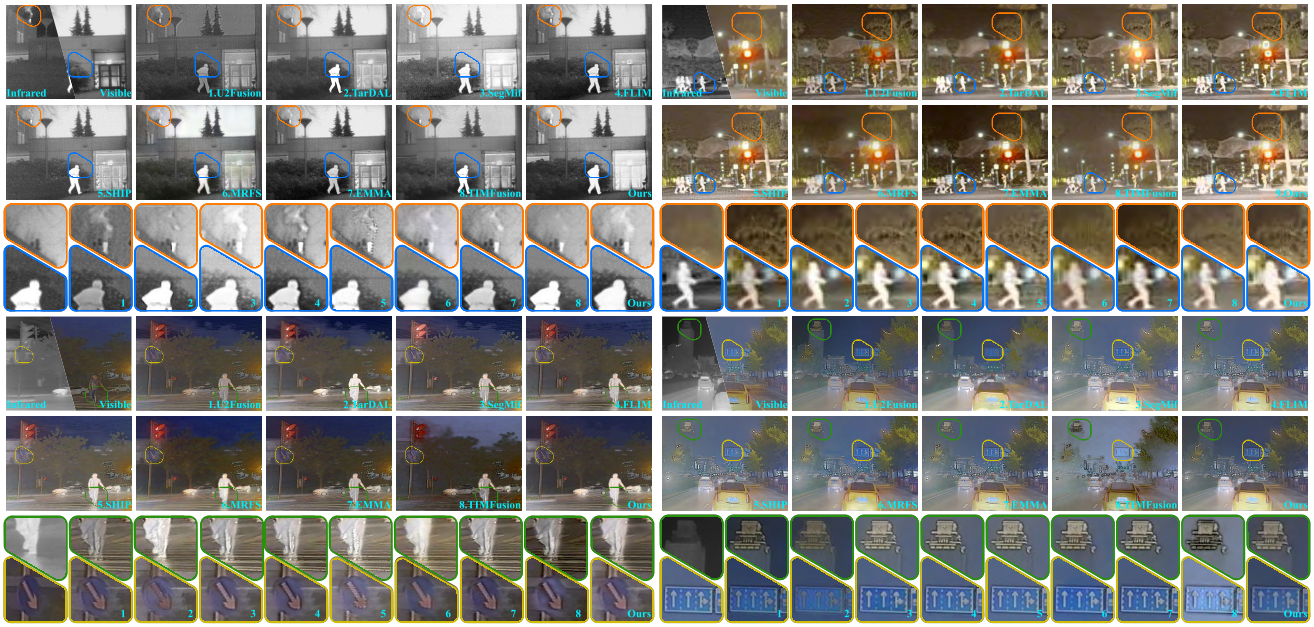

Figure 4. Qualitative demonstrations of SOTA approaches across commonly used datasets, including TNO, RoadScene, M^{3}FD and FMB.

specific information source, guiding the cross-attention mechanism to fuse the semantic patches with rich contextual details from the original source.

The visible and infrared semantic patches ($\mathbf{P}_{\text{vis}}$ and $\mathbf{P}_{\text{ir}}$) extracted from the SAM are encoded into features ($\mathbf{F}_{\text{pvis}}$ and $\mathbf{F}_{\text{pir}}$). These encoded semantic features represent limited portions of the scene and are processed through the cross-attention mechanism. By utilizing the key-value pairs ($\mathbf{K}_{\text{src}}, \mathbf{V}_{\text{src}}$) stored in PR, the cross-attention mechanism enriches the semantic queries ($\mathbf{Q}_{\text{pvis}}$ and $\mathbf{Q}_{\text{pir}}$) with the full scene context from the source image, thereby addressing the inherent incomplete scene coverage in the patches. PR ensures the fusion process stays grounded in a stable, consistent context, enabling patches to be effectively enriched with modality-specific information. This stability maintains semantic coherence while allowing flexible feature refinement via attention mechanisms.

In summary, the SPA module uses PR to guide the cross-attention mechanism, ensuring that the semantic patches from SAM are enriched with consistent and detailed modality-specific information. This approach allows the fusion of infrared and visible image modalities by combining both the high-level semantic understanding from SAM and the detailed information from the source features. The final output, $\mathbf{F}_{\text{SPA}}$, represents a semantically enriched, structurally consistent feature set, rich in semantic priors of SAM, and ready for further processing.

The main network heavily utilizes the SPA module, aiming to fully exploit the complex semantic priors provided by SAM, focus on capturing and retaining detailed semantic knowledge, thereby facilitating a powerful representation that can later be distilled into a more efficient sub-network.

3.3. Triplet Loss in Distillation Scheme

In this section, we introduce a triplet loss-driven distillation scheme that facilitates semantic knowledge transfer from the main network ($\mathcal{M}$) to the sub-network ($\mathcal{S}$) under a bi-level optimization framework. Specifically, we adopt a DARTS-style [10] training protocol, where the main network (teacher) and the sub-network (student) are updated in small alternating steps, approximating the inner-outer structure of bilevel optimization. This allows gradients to flow in a bidirectional manner: the sub-network learns from the distillation signals provided by the main network, while the main network also adapts its parameters based on the sub-network's performance and a segmentation objective. Consequently, the two networks reach a cohesive transfer with mutual compromise.

Our distillation scheme is driven by three distinct losses, each focusing on a different aspect of the fused images. Let $\mathbf{I}_{\text{ref}}$ be the fused reference image produced by the main network, and $\mathbf{I}_{\text{f}}$ be the fused output from the sub-network. We first introduce a feature alignment term:

$$\mathcal{L}_{\text{fea}} = \sum_{m=1}^{M} \left(1 - \frac{\mathbf{F}_{\text{Den}}^{m} \cdot \mathbf{F}_{\text{SPA}}^{m}}{\|\mathbf{F}_{\text{Den}}^{m}\|_2 \cdot \|\mathbf{F}_{\text{SPA}}^{m}\|_2}\right), \tag{2}$$

where $\mathbf{F}_{\text{Den}}^{m}$ and $\mathbf{F}_{\text{SPA}}^{m}$ are the feature maps of the dense and SPA blocks at the same scale m, and $\|\cdot\|_2$ is the ℓ_2-norm.

The second loss component operates at the context level, which consists of two parts: a gradient part to preserve

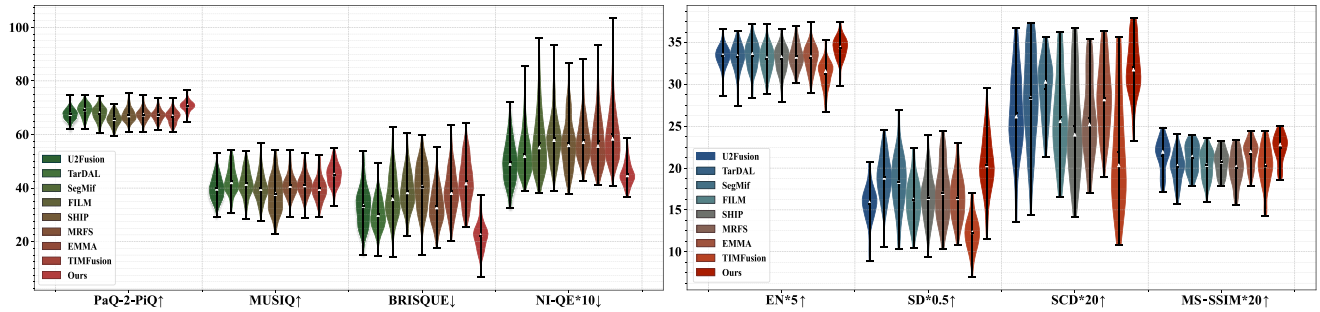

Figure 5. Quantitative comparison of fusion performance with other SOTA methods on M^{3}FD and FMB benchmarks. The violin plots illustrate the distribution of metrics, in which the the black lines and white triangles indicate the medium and mean values.

structural consistency and an MSE loss for intensity consistency. They are defined as:

$$\mathcal{L}_{\text{grad}} = \|\nabla \mathbf{I}_{\text{ref}} - \nabla \mathbf{I}_{\text{fus}}\|_1, \quad \mathcal{L}_{\text{MSE}} = \|\mathbf{I}_{\text{ref}} - \mathbf{I}_{\text{fus}}\|_2, \quad (3)$$

where ∇ denotes the Sobel operator and $\|\cdot\|_1$ is the ℓ_1 -norm. So the cont loss can be calculated as $\mathcal{L}_{\text{context}} = \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{MSE}}$. In practice, we apply these terms not only between $\mathbf{I}_{\text{ref}}$ and $\mathbf{I}_{\text{fus}}$, but also between each fused output and the original source images to ensure the reconstruction fidelity. This prevents both networks from drifting away from the source image during the distillation process.

The third loss component, contrastive semantic loss aims to ensure that the feature space of the fused image remains close to that of the reference image, while being distinct from the individual visible and infrared images archiving efficient distillation. We utilize the semantic feature space by encoder of SAM (defined as $\mathcal{S}_{\text{E}}$) to construct positive and negative pairs through the element-wise multiplication of the binary masks ($\mathbf{M}_{\text{vis}}$ and $\mathbf{M}_{\text{ir}}$) from $\mathbf{P}_{\text{vis}}$ and $\mathbf{P}_{\text{ir}}$. The contrastive semantic loss on each modality is defined as:

$$\mathcal{L}_{\text{cs}}^{\text{ir}} = \sum_{l=1}^{L} \frac{\|\mathcal{S}_{\text{E}}(\mathbf{I}_{\text{fus}} \odot \mathbf{M}_{\text{ir}}) - \mathcal{S}_{\text{E}}(\mathbf{I}_{\text{ref}} \odot \mathbf{M}_{\text{ir}})\|_2}{\|\mathcal{S}_{\text{E}}(\mathbf{I}_{\text{x}} \odot \mathbf{M}_{\text{ir}}) - \mathcal{S}_{\text{E}}(\mathbf{I}_{\text{ir}} \odot \mathbf{M}_{\text{ir}})\|_2}, \quad (4)$$

$$\mathcal{L}_{\text{cs}}^{\text{vis}} = \sum_{l=1}^{L} \frac{\|\mathcal{S}_{\text{E}}(\mathbf{I}_{\text{fus}} \odot \mathbf{M}_{\text{vis}}) - \mathcal{S}_{\text{E}}(\mathbf{I}_{\text{ref}} \odot \mathbf{M}_{\text{vis}})\|_2}{\|\mathcal{S}_{\text{E}}(\mathbf{I}_{\text{x}} \odot \mathbf{M}_{\text{vis}}) - \mathcal{S}_{\text{E}}(\mathbf{I}_{\text{vis}} \odot \mathbf{M}_{\text{vis}})\|_2}, \quad (5)$$

where l indexes the layers of $\mathcal{S}_{\text{E}}$, and $\text{x} \in \{\text{ref}, \text{fus}\}$. So the total $\mathcal{L}_{\text{cs}} = \mathcal{L}_{\text{cs}}^{\text{ir}} + \mathcal{L}_{\text{cs}}^{\text{vis}}$ can also be computed bidirectionally in the bi-level optimization scheme between two networks, ensuring that the feature spaces of both the $\mathbf{I}_{\text{ref}}$ and the $\mathbf{I}_{\text{f}}$ are easily aligned.

The total distillation loss for the sub-network, denoted as $\mathcal{L}_{\text{s}}^{\text{d}}$ in Eq. (1), is the sum of the feature loss, context loss, and contrastive semantic loss: $\mathcal{L}_{\text{s}}^{\text{d}} = \mathcal{L}_{\text{fea}} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{cs}}$.

For the main network, an additional segmentation cross-entropy loss $\mathcal{L}_{\text{seg}}$ is incorporated. This loss is computed between the segmentation predictions generated by an open-vocabulary model $\mathcal{S}_{\text{O}}$ [45, 46] and the ground truth segmentation labels. Instead of affecting the sub-network, this loss aim to prevent potential optimization conflicts. The segmentation loss $\mathcal{L}_{\text{seg}}$ is defined as:

$$\mathcal{L}_{\text{seg}} = -\sum_{c} \left[\mathbf{Y}_{\text{label}}^{c} \log(\mathbf{Y}_{\text{pred}}^{c})\right], \quad (6)$$

where $\mathbf{Y}_{\text{label}}$ represents the label segmentation map, $\mathbf{Y}_{\text{pred}}$ is the predicted segmentation map, and c is the class index. For now, the total distillation loss for the main network can be calculated as: $\mathcal{L}_{\text{m}}^{\text{d}} = \mathcal{L}_{\text{s}}^{\text{d}} + \mathcal{L}_{\text{seg}}$.

4. Experiments

4.1. Settings and Details

Five representative datasets, namely TNO [27], RoadScene [32], MFNet [2], FMB [13], and M^{3}FD [11], are utilized for both training and evaluation. For segmentation, SegFormer [31] (B3 variant) is adopted as the backbone, and the models are trained for 100 epochs. The training and testing splits are strictly in accordance with the official dataset guidelines. Adam is adopted for training, with initial learning rates of 5×10^{-4} and 2×10^{-3} for the main and sub-networks, respectively. Cosine annealing decays both to 1×10^{-5}. Prior to distillation, the models undergo a pre-training phase, followed by 5 epochs of distillation. The batch size is set to 4, and the images are randomly cropped and resized to 192×256 during training. The entire framework is implemented in PyTorch and executed on two NVIDIA GeForce RTX 4090 GPUs.

4.2. Results of Multi-modality Image Fusion

We demonstrate our fusion quality by visualization and quantitative analysis with 9 other SOTA methods in recent years, including DDFM [40], U2Fusion [32], TarDAL [11], SegMiF [13], FILM [42], SHIP [43], MRFS [36], EMMA [41] and TIMFusion [19].

Framework		Segformer - B3 [31]								X - Decoder [45]							
Method	Venue	Road	Bldg.	Sign	Veg.	Bus	Per.	Car	mIoU	Road	Bldg.	Sign	Veg.	Bus	Per.	Car	mIoU
Visible	-	84.8	81.0	64.2	84.7	46.8	61.1	77.1	49.1	79.2	67.8	18.2	73.2	52.0	56.9	82.0	50.1
Infrared	-	83.7	77.6	55.9	80.9	38.1	57.1	74.2	42.7	76.6	55.6	12.9	38.2	72.7	63.2	78.7	42.9
DDFM	ICCV'23	86.2	**83.0**	69.5	84.8	62.5	61.0	80.8	**58.2**	76.1	65.9	10.1	73.3	72.0	**66.6**	82.5	49.3
U2Fusion	TPAMI'22	86.7	82.4	68.5	84.6	50.3	61.8	80.4	57.0	**79.7**	**70.6**	22.2	72.7	75.7	65.5	**83.3**	50.4
TarDAL	CVPR'22	85.8	81.8	68.4	84.7	49.5	61.3	79.5	55.7	74.1	63.1	23.2	67.5	74.8	**66.4**	81.1	48.7
SegMiF	ICCV'23	86.3	82.2	68.0	85.0	**63.7**	60.7	80.0	57.3	79.6	68.6	14.0	73.8	74.3	65.8	**83.2**	**50.8**
FILM	ICML'24	86.9	81.9	69.1	84.0	46.1	60.6	79.4	55.7	79.2	70.3	17.9	73.7	**79.0**	64.0	82.0	50.6
SHIP	CVPR'24	84.9	82.1	68.3	83.3	61.4	60.9	80.4	57.2	76.4	66.8	**24.7**	71.2	76.3	64.6	80.9	50.4
MRFS	CVPR'24	**88.2**	82.8	**71.0**	**86.0**	61.5	**63.6**	**81.1**	56.5	73.8	65.0	**23.7**	72.3	77.2	66.2	82.6	49.6
EMMA	CVPR'24	85.2	81.2	66.8	83.3	59.0	58.9	79.9	55.8	78.6	68.4	15.0	**74.3**	77.1	64.5	82.2	50.7
TIMFusion	TPAMI'24	85.1	80.0	66.4	82.8	50.8	59.0	79.7	55.5	62.0	63.0	12.8	67.6	75.8	65.8	79.6	46.1
SAGE	Proposed	**89.0**	**83.9**	**71.4**	**86.3**	**66.7**	**64.1**	**81.7**	**61.2**	**81.0**	**70.4**	17.0	**75.2**	**77.7**	64.0	82.8	**51.1**

Table 1. Quantitative semantic segmentation results of SOTA approaches on the FMB dataset.

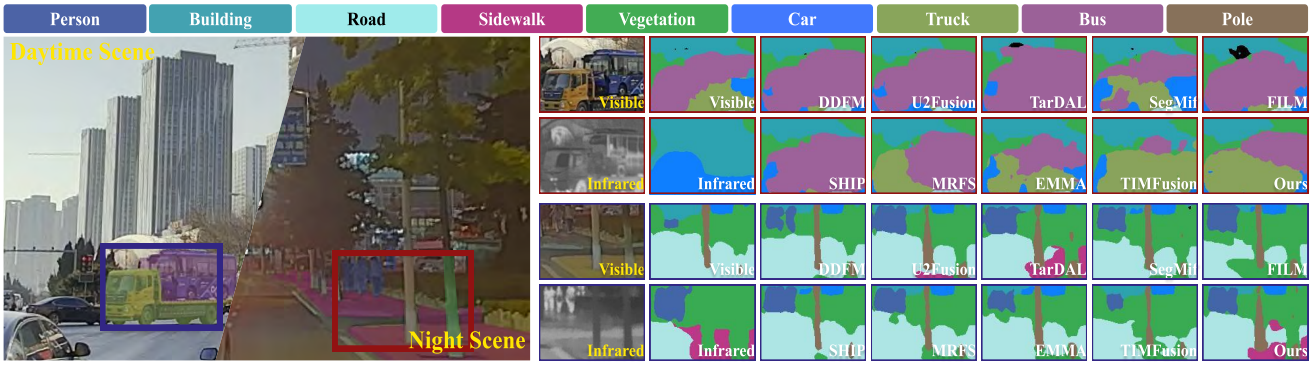

Figure 6. Qualitative segmentation results of SOTA approaches on the FMB dataset.

Qualitative Comparisons. Figure 4 presents a visual comparison of various methods. Overall, the proposed method has two main advantages. First, it effectively preserves the multi-modality information of the original images. In the TNO surveillance scene, the vegetation details from the visible image and the smoke from the chimney in the infrared image are both well retained (the image set in the top-left corner). In RoadScene, our method also achieves optimal leaf recovery. On the other hand, our approach demonstrates strong robustness to interference, as it accurately reconstructs the reflective pedestrian crossing lines on the ground and the distant building outlines in the dense fog at night (the green box in the second row). The integration of SAM enhances our method, giving it the ability to surpass other methods and achieve superior results.

Quantitative Comparisons. We also compare our numerical results with other competitors in Figure 5, based on 100 image pairs randomly selected from the FMB and M^3FD datasets. We employ four fusion evaluation metrics [21], including the Entropy (EN), Standard Deviation (SD), Sum of Correlation Differences (SCD), and Multi-Scale Structural Similarity (MS-SSIM). Additionally, we introduce four no-reference image quality assessments [34] related to image quality: BRISQUE, NIQE, MUSIQ, and PaQ-2-PiQ.

In terms of these metrics, our method demonstrates consistent superiority. High SCD and MS-SSIM values indicate that our method effectively preserves a significant amount of information from the source images, while EN reflects our rich detail and texture. The spatial, multi-scale, and patch evaluations of the NR-IQA confirm that our fusion approach has a better pixel distribution and aligns more closely with the human visual system perception.

4.3. Results of Multi-modality Segmentation

Qualitative Comparisons. We present the segmentation visualization results on the challenging new FMB dataset in Figure 6. In the daytime intersection scene, thanks to the powerful semantic priors provided by SAM, our method is the only one to completely distinguish between the truck and the bus. In the nighttime road scene, we successfully segment the sidewalk, achieving optimal performance. Additionally, we showcase the segmentation results on the MFNet dataset in Figure 7. Leveraging the advanced se-

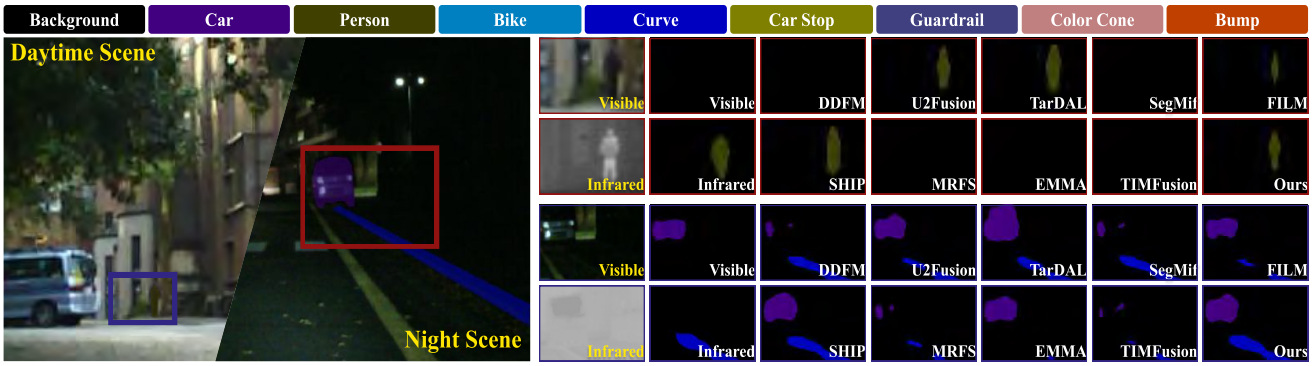

Figure 7. Qualitative segmentation results of SOTA approaches on the MFNet dataset.

Framework	Segformer - B3 [31]							
Method	Car	Per.	Bike	Curve	C.S.	C.C.	Bump	mIoU
Visible	83.3	54.7	59.6	7.46	31.0	37.4	30.8	44.6
Infrared	79.0	63.8	51.4	7.10	25.2	34.1	41.8	44.4
DDFM	86.7	62.2	63.4	30.8	**35.5**	45.7	41.1	51.7
U2Fusion	**87.3**	65.9	60.8	33.4	**36.0**	46.1	40.9	52.0
TarDAL	86.4	**67.1**	62.4	28.4	30.5	45.0	41.5	51.0
SegMiF	85.9	66.3	**64.4**	34.0	34.7	43.0	**48.0**	52.4
FILM	86.6	66.4	61.9	**36.3**	32.0	**46.7**	**46.7**	**52.7**
SHIP	86.7	63.9	59.7	33.5	35.2	44.4	42.4	51.5
MRFS	86.1	61.5	59.7	30.6	35.2	44.9	38.4	50.4
EMMA	86.1	65.6	63.8	25.4	29.1	44.9	41.3	51.7
TIMFusion	86.4	58.7	60.3	29.8	33.6	44.1	38.3	49.8
SAGE	**87.9**	**67.5**	**64.3**	**41.2**	34.1	**46.8**	**46.7**	**54.0**

Table 2. Quantitative semantic segmentation results of SOTA approaches on the MFNet dataset.

mantic information learned by SAM, our method successfully segments distant small pedestrian targets during the daytime and near-invisible lane curve at night.

Quantitative Comparisons. Table 1 presents a qualitative comparison of IoU results for two segmentation frameworks on the FMB dataset. The first framework involves re-training images generated by various fusion methods, while the second uses the output from an open-vocabulary segmentation network with prompt words, without retraining. In traditional comparisons, our method achieves a 3.0 mIoU improvement over the second-best approach, demonstrating its competitive edge across categories. Additionally, in the segmentation network that requires no training, our method also performs favorably, benefiting from the adaptability provided by SAM's semantic information. Table 2 further showcases the qualitative results on the MFNet dataset, which has lower resolution and fewer labels. Although our method shows some gaps in background categories, it achieves overall superiority in key perception categories and mean performance. Our method integrates SAM with the fusion network, achieving a breakthrough in performance for efficient segmentation.

Variants	mIoU **MFNet / FMB**	Fusion Metrics **EN / SD/ SCD / MS-SSIM**
I. The Impact of SAM		
(a) w/o SAM	50.7 / 56.5	**6.713** / 36.93 / **1.612** / **1.127**
(b) **P** from $\mathbf{Y}_{\text{label}}$	53.2 / 56.4	6.239 / 33.52 / 1.412 / 1.102
(c) Swap $\mathcal{S}^{\text{O}}$	**53.4** / 55.5	6.312 / 34.79 / 1.316 / 1.097
II. Study on the SPA Module		
(a) *w/o* **Z**	52.3 / **57.2**	6.333 / 32.81 / 1.384 / 1.002
(b) *w/o* k-v in PR	52.1 / 56.1	6.417 / 37.19 / 1.261 / 1.063
(c) *w/o* PR	53.3 / **57.2**	6.556 / 35.48 / 1.576 / 1.046
III. Discussion on the Distillation Scheme		
(a) w/o $\mathcal{L}_{\text{fea}}$	49.7 / 51.6	6.436 / 37.28 / 1.277 / 1.079
(b) w/o $\mathcal{L}_{\text{cont}}$	50.9 / 56.2	6.669 / 37.80 / 1.462 / 1.086
(c) w/o $\mathcal{L}_{\text{cs}}$	51.7 / 54.3	6.524 / **39.11** / 1.356 / 1.038
(d) Offline Dist.	50.0 / 50.4	6.221 / 32.65 / 1.581 / 0.987
SAGE	**54.0** / **61.2**	**6.872** / **40.77** / **1.604** / **1.117**

Table 3. Quantitative results on both fusion and segmentation of all variants in three ablation studies.

4.4. Ablation Studies

The Impact of SAM. The Segment Anything Module serves as a critical cornerstone in our method. To explore its impact, we derive three variants by isolating SAM from the core framework. Specifically, in variant (a), we replace the semantic patches with randomly cropped source image patches, thereby removing SAM from the main network $\mathcal{M}$. In variant (b), we assist the generation of semantic patches with segmentation labels. Additionally, we replace $\mathcal{S}_{\text{O}}$ with a conventional segmentation network [31]. Variant (a) enhances the source image information, leading to the optimal similarity metrics, SCD and MS-SSIM, as shown in Table 3. Our method fully leverages the semantic priors provided by SAM, resulting in the best foggy building contours, as illustrated in the first row of Figure 8.

Study on the SPA Module. The proposed Semantic Persistent Attention module plays a key role in integrating the semantic priors from SAM, effectively guiding the retention and enhancement of inherent modality features. To in-

Method	DDFM	U2Fusion	TarDAL	SegMiF	FILM	SHIP	MRFS	EMMA	TIMFusion	**Ours**
Time(ms)	280K	42.31	**16.57**	150.5	183.3	27.93	95.86	25.73	17.39	**10.47**
FLOPs(G)	1.34M	518.0	233.3	500.4	230.3	401.5	301.2	106.4	**100.7**	**52.06**
Params(M)	552.7	0.659	0.297	0.621	0.491	0.525	133.4	1.516	**0.158**	**0.136**

Table 4. Efficiency comparison with other state-of-the-art methods on M^3FD benchmarks.

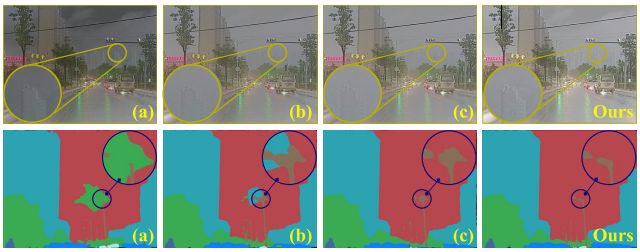

Figure 8. Visualization comparison of impact of SAM.

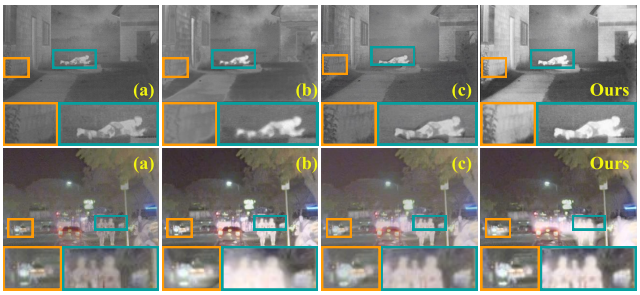

Figure 9. Visualization comparison of ablation on SPA.

vestigate its impact, we set up three variants: (a) *w/o* latent representation **Z**, (b) *w/o* key-value pairs in the Persistent Repository, and (c) *w/o* PR. The channel loss due to these modifications has been corrected. A visual comparison of these ablation variants is shown in Figure 9. It is evident that PR in the SPA module plays an essential role in maintaining critical information. The removal of any component leads to significant information loss. For instance, in (b), while the regions of interest are highlighted, the absence of source image information causes blurriness. Similarly, the absence of PR in the SPA module completely prevents the capture of beneficial semantic information, causing the network to lose focus on key areas and resulting in a low-contrast fusion results in (c).

Discussion on the Distillation Scheme. Furthermore, we discuss the distillation scheme. First, we conduct ablation studies on each component of the triple loss function, leading to variants (a)-(c) as shown in Table 3 III. Additionally, we replace the bi-level optimization distillation method with the offline distillation approach, forming variant (d). A visual comparison of these variants is shown in Figure 10. Notably, the results of the traditional online distillation are significantly inferior to our method, both in terms of visual quality and the gradient map generated by variant (b).

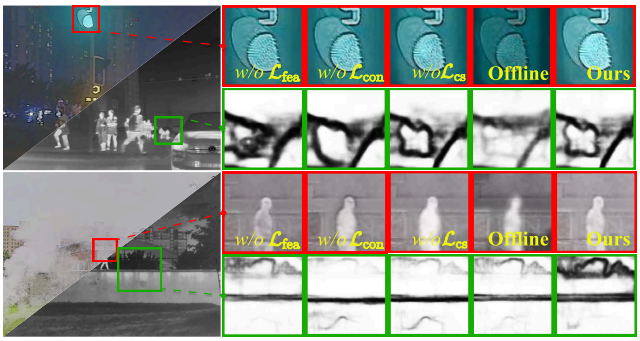

Figure 10. Visualization comparison of ablation on Distillation.

This confirms the effectiveness of the proposed distillation method when incorporating semantic information.

4.5. Computational Efficiency Analysis

In the M^3FD benchmark, we compare our method with 9 other SOTA methods in terms of time, FLOPS, and parameters. As shown in Table 4, our method demonstrates significant advantages across all aspects, particularly in time and FLOPs. Specifically, due to our distillation scheme, our sub-network is able to adjust flexibly while maintaining high efficiency, resulting in reduced FLOPs. Compared to other methods, our approach achieves a processing time of 10.47 ms and FLOPs of 52.06 G, outperforming most of the existing methods, with a parameter count of only 0.136M, showcasing computational efficiency.

Our method effectively alleviates the computational burden of SAM during inference, significantly reducing the computational overhead while preserving semantic information. This design allows our network to strike a balance between processing speed and computational resources, leading to superior efficiency in practical applications.

5. Conclusion

In this work, we propose a fusion method that utilizes semantic priors from the SAM for IVIF. We design the Semantic Persistent Attention module, which integrates semantic information while retaining source details. Additionally, we introduce a bi-level distillation scheme with triplet loss to reduce computational complexity by decoupling fusion from SAM during inference. Extensive experiments show that our method achieves superior fusion and task performance, outperforming across multiple datasets.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 62450072, U22B2052, 62302078), Central Guidance for Local Science and Technology Development Fund (Youth Science Fund Project, Category A, No. 2025JH6/101100001), the Distinguished Young Scholars Funds of Dalian (No. 2024RJ002), the China Postdoctoral Science Foundation (No. 2023M730741) and the Fundamental Research Funds for the Central Universities.

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020. 1

[2] 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 *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5108–5115. IEEE, 2017. 5

[3] Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. Multispectral image stitching via global-aware quadrature pyramid regression. *IEEE Transactions on Image Processing*, pages 4288–4302, 2024. 2

[4] Zheyan Jin, Shiqi Chen, Yueting Chen, Zhihai Xu, and Huajun Feng. Let segment anything help image dehaze. *arXiv preprint arXiv:2306.15870*, 2023. 2

[5] Harpreet Kaur, Deepika Koundal, and Virender Kadyan. Image fusion techniques: a survey. *Archives of Computational Methods in Engineering*, 28(7):4425–4447, 2021. 1

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[7] Guanlin Li, Bin Zhao, and Xuelong Li. Low-light image enhancement with sam-based structure priors and guidance. *IEEE Transactions on Multimedia*, pages 10854–10866, 2024. 2

[8] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 11040–11052, 2023. 1, 2

[9] Siwei Li, Mingxuan Liu, Yating Zhang, Shu Chen, Haoxiang Li, Zifei Dou, and Hong Chen. Sam-deblur: Let segment anything boost image deblurring. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2445–2449. IEEE, 2024. 2

[10] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *Proceedings of the International Conference on Learning Representations*, 2019. 4

[11] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. 1, 2, 5

[12] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, pages 1748–1775, 2023. 1, 2

[13] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8115–8124, 2023. 1, 2, 5

[14] Jinyuan Liu, Guanyao Wu, Zhu Liu, Long Ma, Risheng Liu, and Xin Fan. Where elegance meets precision: Towards a compact, automatic, and flexible framework for multi-modality image fusion and applications. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1110–1118, 2024. 2

[15] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025. 1

[16] Risheng Liu, Shichao Cheng, Yi He, Xin Fan, Zhouchen Lin, and Zhongxuan Luo. On the convergence of learning-based iterative methods for nonconvex inverse problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12):3027–3039, 2019. 1

[17] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Transactions on Image Processing*, 30:1261–1274, 2020. 1

[18] Risheng Liu, Zhiying Jiang, Shuzhou Yang, and Xin Fan. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Transactions on Image Processing*, 31:4922–4936, 2022. 1

[19] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024. 2, 5

[20] Zhu Liu, Jinyuan Liu, Guanyao Wu, Long Ma, Xin Fan, and Risheng Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1240–1248, 2023. 1, 2

[21] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 1, 6

[22] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 1

[23] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2

[24] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4003–4011, 2022. 2

[25] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022. 2

[26] 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. 1, 2

[27] Alexander Toet. The tno multiband image data collection. *Data in Brief*, 15:249–251, 2017. 5

[28] Chengcheng Wang, Zhiwei Hao, Yehui Tang, Jianyuan Guo, Yujie Yang, Kai Han, and Yunhe Wang. Samdiffsr: Structure-modulated diffusion model for image super-resolution. *arXiv preprint arXiv:2402.17133*, 2024. 2

[29] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3508–3515, 2022. 2

[30] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, pages 1–36, 2023. 2

[31] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 5, 6, 7

[32] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 1, 2, 5

[33] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1366–1375, 2021. 1

[34] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 6

[35] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[36] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26974–26983, 2024. 2, 5

[37] Xingchen Zhang, Ping Ye, and Gang Xiao. Vifb: A visible and infrared image fusion benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 104–105, 2020. 1

[38] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023. 2

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023. 1, 2

[40] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023. 1, 2, 5

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921, 2024. 5

[42] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, et al. Image fusion via vision-language model. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60749–60765, 2024. 5

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26384–26395, 2024. 5

[44] Huabing Zhou, Jilei Hou, Yanduo Zhang, Jiayi Ma, and Haibin Ling. Unified gradient-and intensity-discriminator generative adversarial network for image fusion. *Information Fusion*, 88:184–201, 2022. 2

[45] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 5, 6

[46] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36:19769–19782, 2023. 5