Domain Adaptation Guided Infrared and Visible Image Fusion

Tianwei Guan,[1] Haozhen Wei,[2] Yuhan Zhou,[2] Jun Ma,[2] Zecheng Xu,[4]
Zhiying Jiang,[5] Jinyuan Liu,[2] Xingyuan Li[3]*
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]School of Software Technology, Dalian University of Technology
[3]School of Computer Science, Zhejiang University
[4]School of Software Engineering, Beijing Jiaotong University
[5]College of Information Science and Technology, Dalian Maritime University
GuanTW@link.cuhk.edu.hk xingyuan_lxy@163.com

Abstract

Infrared and Visible Image Fusion (IVIF) integrates complementary information from distinct modalities to enhance image quality. However, the effectiveness declines under unseen conditions such as novel weather or scenes, due to domain shifts primarily from variations of data distribution in the visible modality, while the infrared modality remains relatively stable. To overcome domain shifts caused by the imbalance between modalities during image fusion, we propose a Domain Adaptation Guided Infrared and Visible Image Fusion method, termed DAFusion, leveraging a dual-rank domain adapter to enable fast adaptation to diverse adverse conditions during image fusion. Specifically, trainable low-rank and high-rank embedding spaces are respectively used to capture knowledge common across domains (domain-shared) and those unique to target domains (domain-specific). To leverage the dual-rank adapter more effectively, we develop a homeostatic knowledge allotment strategy to integrate the distinct types of knowledge dynamically based on the uncertainty value of target domains. Since domain adaptation typically optimizes for feature alignment across domains and emphasizes invariance rather than preserving specific cues critical for image fusion, while the fusion objective requires retaining discriminative and complementary features, a conflict between the two modules appears. To reconcile this, we further adopt a bi-level optimization framework that structurally decouples the two objectives, enabling the fusion module to steer the adaptation process while benefiting in return from domain-aligned representations. Experimental results on three benchmarks demonstrate that our method significantly outperforms state-of-the-art approaches, achieving both an enhancement in fusion quality and an improvement on subsequent high-level tasks.

Introduction

Infrared images capture thermal features, enabling the provision of information in complex conditions, but suffer from limited resolution and texture (Li et al. 2024b, 2025b). Conversely, visible images present high resolution and rich details with color, yet are degraded by severe weather and low light. Consequently, Infrared and Visible Image Fusion (IVIF) integrates information from multi-modal images to create high-quality fused images. High-level tasks, such as multi-modal saliency detection, object detection, and semantic segmentation, benefit subsequently from the more informative representations of scenes and objects in images.

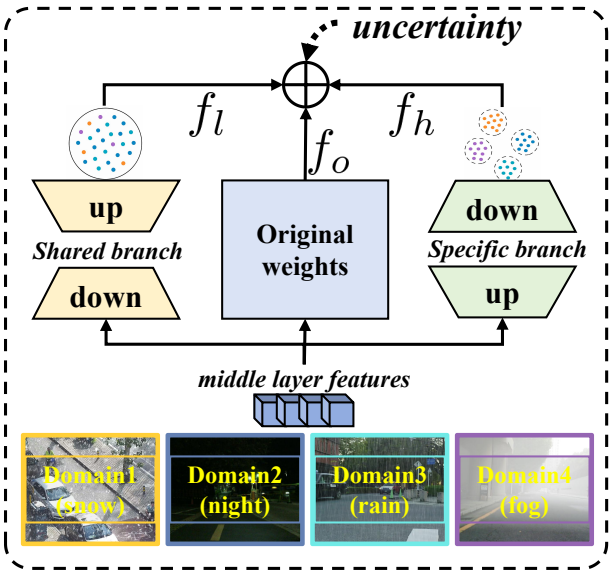

Figure 1: Architecture of the dual-rank domain adapter for visible modality adaptation across target domains. The adapter contains high-rank and low-rank branches to capture domain-shared and domain-specific knowledge, and balance them based on the uncertainty value of target domains.

Recently, numerous methods (Li et al. 2025a, 2024a) have been devised to address challenges posed by IVIF. Learning-based methods (Yang et al. 2025; Sun et al. 2022; Wang et al. 2025b) have substantially outperformed traditional methods in fusion performance. Despite the appealing fusion quality obtained, a major issue with these methods is that they can only present expected performance for images possessing similar distributions to the training dataset. This limited adaptability mainly arises from domain shifts in the visible modality, which is highly sensitive to changes in weather, illumination, and environments. In contrast, infrared images tend to remain stable across variations. Such modality im-

*Corresponding author.

balance causes fusion models to misrepresent or suppress critical information from the visible input, degrading both fusion quality and downstream task performance.

Although domain adaptation techniques have been explored in other low-level vision tasks such as image restoration and deblurring, their direct application to fusion tasks is non-trivial. Some recent works, such as TarDAL (Liu et al. 2022), incorporate adversarial learning to improve fusion robustness across different scenarios, while DCEvo (Liu et al. 2025a) enhances fused representation through dynamic feature evolution. However, these methods do not explicitly address domain shifts at the input level, nor do they optimize domain adaptation and fusion objectives in a coordinated manner. Most existing approaches still assume fixed feature distributions, making them unreliable in unseen conditions. These limitations motivate a fusion-aware adaptation strategy that jointly models domain adaptation and image fusion.

Therefore, we propose DAFusion, a Domain Adaptation Guided Infrared and Visible Image Fusion method to solve the domain adaptation challenges during image fusion. Our proposed DAFusion integrates domain adaptation and image fusion modules into a unified architecture via a bi-level optimization scheme. In particular, we pretrain the domain adapter through a teacher-student-based scheme and employ this in the visible image encoder. To alleviate error accumulation and catastrophic forgetting, we adopt a homeostatic knowledge allotment strategy to dynamically balance domain-specific and domain-shared knowledge from high-rank and low-rank adapters as shown in Figure 1. However, optimizing the domain adapter independently from the fusion objective leads to a mismatch in goals. While adaptation encourages domain invariance, fusion requires preserving modality-specific discriminative features. This misalignment results in suboptimal representations for image fusion. Based on this, we further employ the bi-level optimization scheme to decouple the optimization of domain adaptation and image fusion, enabling the visible encoder to be guided by the fusion objective at a higher level while adapting to the target domain at a lower level. This scheme allows the fusion task to steer domain adaptation in a task-driven manner, rather than treating adaptation as an independent or preprocessing step. Thereby, our method achieves a mutual promotion between image fusion and domain adaptation. Our contributions can be organized into three aspects as follows:

- To address the challenges of error accumulation and catastrophic forgetting in the adaptation process during image fusion, we develop high-rank and low-rank adapters to capture target domain-specific and domain-shared knowledge simultaneously.
- We adopt a homeostatic knowledge allotment strategy to dynamically balance domain-specific and domain-shared knowledge, ensuring efficient extraction of domain-specific knowledge with long-term domain-shared knowledge retained concurrently.
- We establish a bi-level optimization framework to build the mutual collaboration and guidance between image fusion and domain adaptation, realizing "Best of Both ".

Related Work

Learning-based Fusion Approaches

Existing image fusion methods can be broadly categorized into discriminative and generative approaches. Discriminative methods aim to learn the mapping between dual-modal inputs and the fused output. CNN-based models (Liu et al. 2021a; Ma et al. 2022; Zhao et al. 2021; Wang et al. 2022) have been widely explored in this direction. More recently, Transformer-based architectures (Liu et al. 2024b; Zhao et al. 2023b; Rao, Xu, and Wu 2023; Wang et al. 2025a; Li et al. 2023b) are also introduced to model long-range dependencies in the fusion process.

Generative methods focus on modeling the data distribution and underlying priors to generate fused images. GAN-based frameworks (Liu et al. 2021b; Li, Wu, and Kittler 2021) generate fusion results through adversarial learning. Diffusion-based models (Zhao et al. 2023c; Cao et al. 2024) further push this boundary by capturing complex priors via iterative denoising for effective image fusion. For instance, DDFM (Zhao et al. 2023c) formulates the fusion task as a conditional generation framework based on the DDPM.

Moreover, to concentrate on both downstream task performance and visual quality, task-driven fusion methods (Liu et al. 2023a; Zhao et al. 2023a; Liu et al. 2025b) have been proposed. TarDAL (Liu et al. 2022) combines a fusion module with an object detection network, improving the perception ability of the model and demonstrating an obvious advantage in challenging scenarios with efficiency.

Domain Adaptation Approaches

Domain shift is caused by discrepancies between the source and target data distributions, which leads to performance degradation when models are applied to unseen environments. To address this issue, domain adaptation (DA) techniques have been developed to improve generalization under such distribution shifts. Traditional DA methods assume access to labeled source data and unlabeled target data during training. Early approaches focus on minimizing distribution discrepancies via statistical metrics such as Maximum Mean Discrepancy (Long et al. 2015) or adversarial learning frameworks (Ganin et al. 2016), where a domain discriminator is used to encourage feature alignment across domains.

More recent efforts have explored self-training and pseudo-labeling strategies to better leverage target data, improving performance in the absence of target annotations (Zou et al. 2018; Wang et al. 2021). In addition, domain-specific normalization (Chang et al. 2019) and style transfer techniques (Hoffman et al. 2018) have been proposed to enhance feature generalization across domains.

While most works focus on offline unsupervised domain adaptation, Test-Time Adaptation (TTA) (Boudiaf et al. 2023; Wang et al. 2020) and Continual Test-Time Adaptation (CTTA) (Liu et al. 2023b; Song et al. 2023) have gained attention for adapting models on-the-fly without source data or retraining. These methods aim to incrementally adapt to streaming, unlabeled target data, offering more practical deployment in real-world dynamic environments.

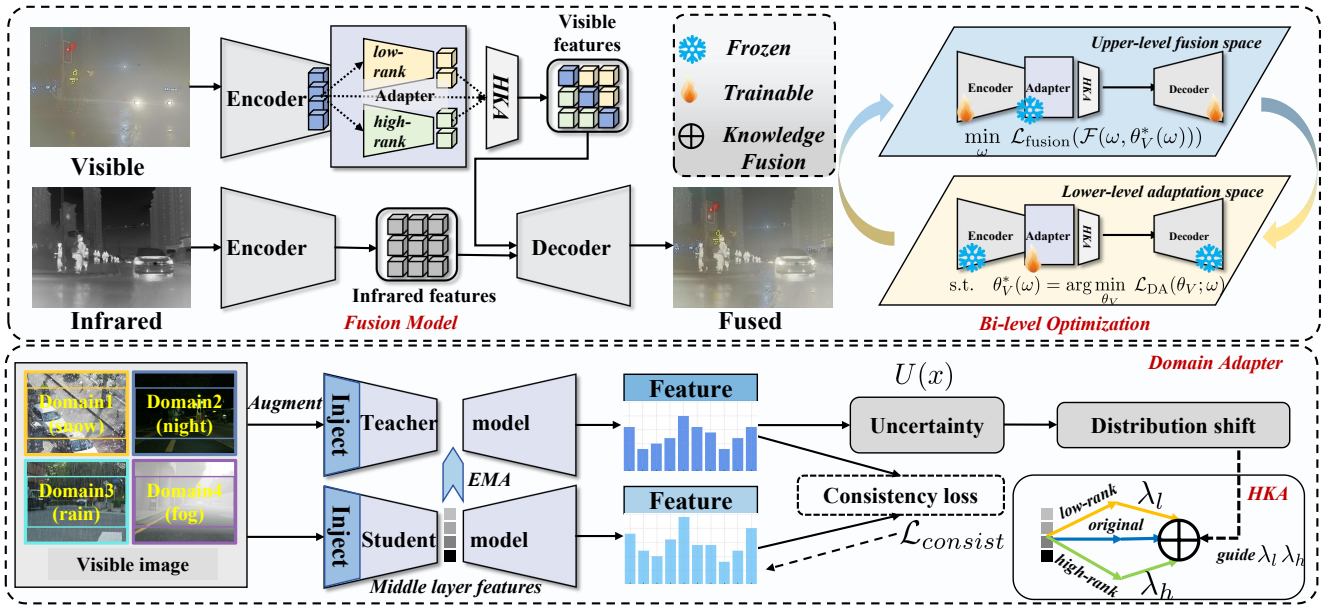

Figure 2: An overall illustration of our proposed DAFusion. The lower part depicts the pretrained domain adapter. We first inject dual-rank adapters into the linear layers of the source model. Then we construct a teacher-student-based framework to update adapters with a consistency loss as the optimization objective. The subsequent calculated distribution shift guides the balance in the HKA strategy. The upper part denotes the fusion module and the bi-level optimization framework. In this framework, the adapter and the encoder&decoder are alternately frozen to optimize mutually.

Methods

In this section, we first give an overview of the proposed Domain Adaptation Guided Infrared and Visible Image Fusion framework, **DAFusion**. We then introduce the domain adapter, the homeostatic knowledge allotment (HKA) strategy, and the bi-level optimization, as illustrated in Figure 2.

Overall Framework

As shown in Figure 2, we pretrain a domain adapter on images from 4 weather conditions: snow, night, rain, and fog through a teacher-student-based architecture, where we adopt an exponential moving average (EMA) to update the teacher model with adapters, formulated as:

$$\mathcal{T}^t = \alpha \mathcal{T}^{t-1} + (1-\alpha)\mathcal{S}^t, \tag{1}$$

$\mathcal{T}$, $\mathcal{S}$, t, and α denote the teacher model, student model, time step, and updating weight, respectively. Selecting these representative adverse weather conditions helps our adapter gain initial robustness and the ability to adapt to various real-world scenes appearing in the visible modality, thereby enhancing fusion and perception. Then the fusion module takes infrared and visible images as inputs, where the visible branch incorporates the domain adapter we pretrained earlier. To regularize the balance between domain-specific and domain-shared knowledge, an HKA strategy is designed to dynamically adjust the contributions of high-rank and low-rank adapters based on the uncertainty of target domains. Finally we employ these components in a bi-level optimization framework. The upper level optimizes the fusion quality and updates the adapter weights simultaneously indirectly, while the lower level adapts the visible encoder to target domains, so as to enhance visible images and improve fusion quality. This mutual feedback promotes both robust domain adaptation and effective multi-modal fusion.

Model Architecture

Domain Adapter While domains change across different scenarios during fusion, real-world visible images usually stem from error accumulation and catastrophic forgetting. As proven by (Liu et al. 2023b), adapters utilizing middle-layer features of varying dimensionality are effective in mitigating these issues. To be specific, a low-rank adapter reduces feature redundancy, leading to the capture of domain-shared knowledge to mitigate catastrophic forgetting. Conversely, a high-rank adapter utilizes a higher-dimensional feature representations, which more effectively align with the target data distribution, thus emphasizing the learning of domain-specific knowledge to reduce error accumulation. The above observation motivates the integration of both low-rank and high-rank adapters in the source pretrained model, with the goal of simultaneously capturing distinct domain types of knowledge. As illustrated in Figure 2, the adapter structure contains three sub-branches: the linear or Conv layer from the original network, and bottleneck structures, respectively indicating high-rank and low-rank adapters in the green branch and yellow branch. Specifically, the high-rank branch consists of an up-projection layer with parameters $W_{up}^h \in R^{d \times d_h}$, a down-projection layer with parameters

$W^h_{down} \in R^{d_h \times d}$, where $d_h \geq d$ represents the middle dimension of features in high-rank branch. When considering the projection layer for the original model, we apply a linear layer for the transformer architecture, and a Conv1×1 for the convolution network, both without any non-linear layers. On the contrary of high-rank branch, the low-rank one utilizes a down-projection layer with parameters $W^l_{down} \in R^{d_l \times d}$, an up-projection layer with parameters $W^l_{up} \in R^{d_l \times d}$ where $d_l \ll d$ implies the middle dimension of features in low-rank branch. Based on the above, we can formulate the produced features of both branches for an input feature f as:

$$f_h = W^h_{down} \cdot (W^h_{up} \cdot f); \quad f_l = W^l_{up} \cdot (W^l_{down} \cdot f). \quad (2)$$

The dual-branch bottleneck structure is linked to the output feature of the original network (f_o) through a residual connection, the fused knowledge (f_f) is formulated as:

$$f_f = f_o + \lambda_h \times f_h + \lambda_l \times f_l, \quad (3)$$

where λ_h and λ_l are scale factors acquired through the HKA strategy, which is introduced in the next section.

Homeostatic knowledge allotment (HKA) Though the dual-rank adapter contributes to learn distinct domain representations, the regulation of knowledge fusion during adaptation process is still needed to ensure the efficient capture of domain-specific knowledge with the long-term domain-shared knowledge retained. Following (Liu et al. 2023b), we calculate the uncertainty value $U(x)$ of a given input x to quantify the degree of distribution shift, formulated as:

$$U(x) = \text{NORM}\left(\left(\frac{1}{m}\sum_{i=1}^{m}\|F_i(x) - \mu\|^2\right)^{\frac{1}{2}}\right) \quad (4)$$

where m is the number of probability sets for each sample obtained in multiple forward propagation, $F_i(x)$ is the feature representation output in the i^{th} forward propagation, and μ is the average value of m times propagation. The scale factors λ of high-rank and low-rank branches are adjusted based on the uncertainty value $U(x)$, formulated as:

$$\begin{cases} \lambda_h = 1 + \mathcal{U}(x), & \lambda_l = 1 - \mathcal{U}(x), \quad \mathcal{U}(x) \geq \Theta \\ \lambda_h = 1 - \mathcal{U}(x), & \lambda_l = 1 + \mathcal{U}(x), \quad \mathcal{U}(x) < \Theta \end{cases}, \quad (5)$$

where Θ represents the threshold of uncertainty. The fusion weights for domain-specific knowledge increase with high uncertainty and decrease with low uncertainty, achieving a balance between different types of domain knowledge.

Optimization objective

Bi-level framework However, a central challenge in integrating domain adaptation and image fusion lies in the potential mismatch between their optimization objectives. Image fusion relies on preserving discriminative and modality-specific cues, especially from the visible domain, which carries rich structural and textural details. In contrast, domain adaptation typically focuses on aligning visible feature representations across domains. As a result, the domain adapter prioritizes alignment over preserving fusion-critical features, leading to suboptimal representations for the fusion task. This motivates us to rethink how to guide the domain adapter to produce enhanced visible features more beneficial to image fusion. Realizing the mutual-benefit relationship between the two modules, we decompose the complex optimization objective into two sub-objectives: the optimization of image fusion and domain adaptation.

To model this mutual-benefit relationship and achieve the "Best of Both Worlds", we introduce a bi-level optimization strategy to establish the mutual collaboration and guidance between image fusion and domain adaptation. In our scheme, the bi-level learning can be formulated as:

$$\min_{\omega} \ \mathcal{L}_{\text{fusion}}(\mathcal{F}(\omega, \theta^*_V(\omega))), \quad (6)$$

$$\text{s.t.} \quad \theta^*_V(\omega) = \arg\min_{\theta_V} \ \mathcal{L}_{\text{DA}}(\theta_V; \omega), \quad (7)$$

where ω represents fusion network parameters, $\theta^*_V(\omega)$ represents the optimal solution of adapter parameters. Specifically, the primary part is to optimize the fusion network $\mathcal{F}$ for generating enhanced images under diverse domain conditions, which facilitates robust representation for subsequent downstream tasks. In the lower-level optimization, we apply a domain adaptation objective to the visible encoder, aiming to reduce feature inconsistency across domains. In the upper-level, the fusion objective supervises the integration of aligned visible and infrared features by evaluating the fusion quality of outputs. Both levels of optimization are nested with mutual promotion, where the fusion loss implicitly supervises the domain adaptation by influencing the optimal encoder state, while the adapted encoder in turn enhances the stability and robustness of fusion. The interaction enables the network to jointly model cross-domain alignment and cross-modality integration, resulting in an effective and generalized fusion representation.

Loss function In this section, we elaborate on the concrete loss functions of our method, which can be divided into two parts for fusion quality and domain adaptability. As for the learning of the image fusion network, we give the notation of model formulation as follows. Infrared, visible and fused images are denoted as $\mathbf{I}_{ir}$, $\mathbf{I}_{vis}$ and $\mathbf{I}_{fu}$. Then we apply multiple constraints during the process of obtaining high-quality fused images. The fusion loss function is written as:

$$\mathcal{L}_{fusion} = \mathcal{L}_{SSIM} + \mathcal{L}_{deco} + \mathcal{L}_{grad} + \mathcal{L}_{int}, \quad (8)$$

where $\mathcal{L}_{SSIM}$, $\mathcal{L}_{grad}$, $\mathcal{L}_{int}$ are respectively formulated as:

$$\mathcal{L}_{SSIM} = \mathcal{L}_{SSIM}(\mathbf{I}_{ir}, \mathbf{I}_f) + \mathcal{L}_{SSIM}(\mathbf{I}_{vis}, \mathbf{I}_f), \quad (9)$$

$$\mathcal{L}_{int} = \frac{1}{HW}\left\|\mathbf{I}_f - \max(\mathbf{I}_{ir}, \mathbf{I}_{vis})\right\|, \quad (10)$$

$$\mathcal{L}_{grad} = \frac{1}{HW}\left\|\left|\nabla\mathbf{I}_f\right| - \max\left(\left|\nabla\mathbf{I}_{ir}\right|, \left|\nabla\mathbf{I}_{vis}\right|\right)\right\|. \quad (11)$$

The $\mathcal{L}_{deco}$ is proposed by (Zhao et al. 2023b) to correlate basic features from multi-modal inputs. The loss function of domain adaptation is a consistency loss, formulated as:

$$\mathcal{L}_{DA}(x) = \frac{1}{N}\sum_{i=1}^{N}\|F^T_i - F^S_i\|^2_2, \quad (12)$$

where, F^T_i and F^S_i represent the outputs of the teacher and student models at the i-th feature layer, respectively, and N is the dimensionality of the features.

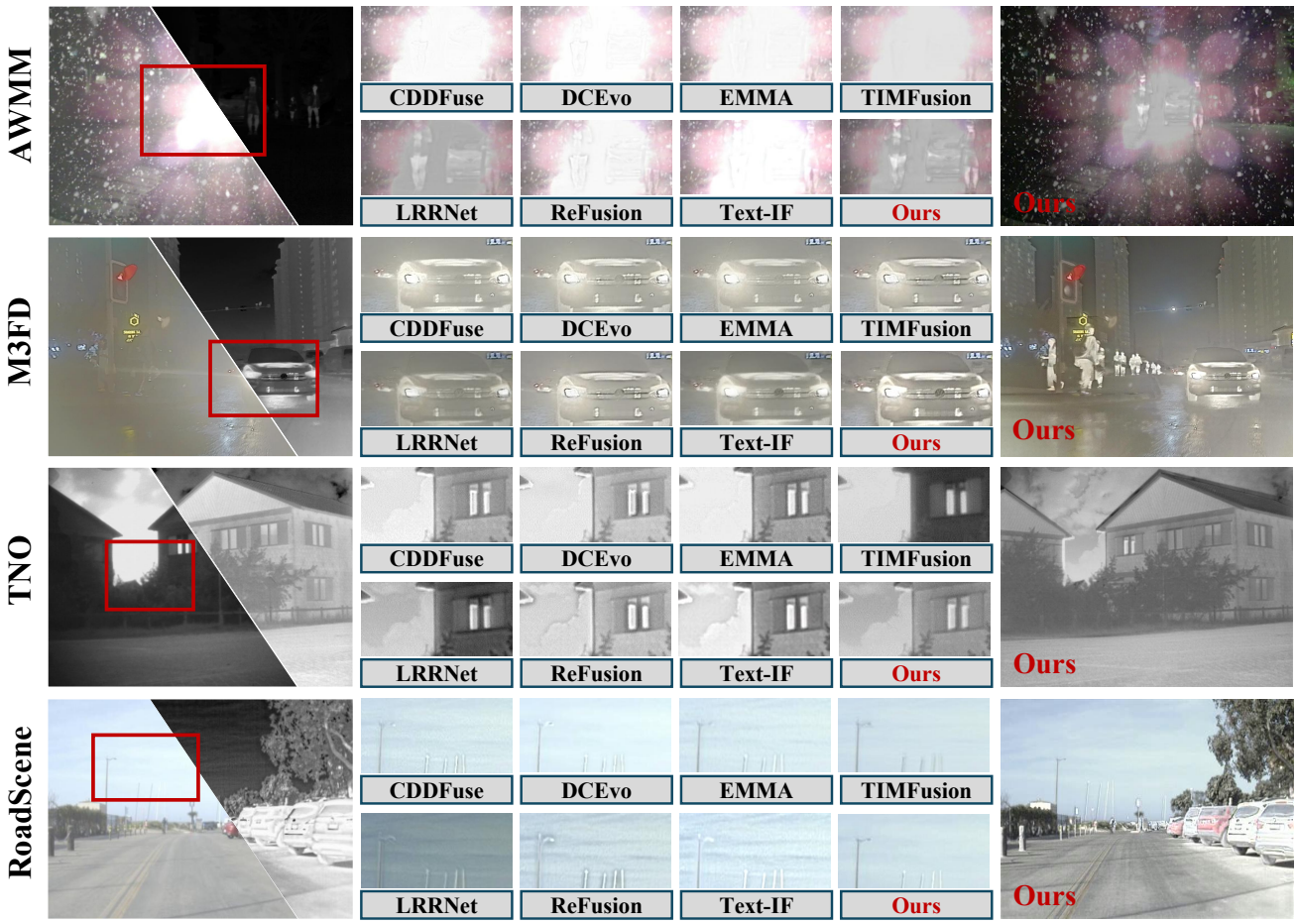

Figure 3: Qualitative comparisons of fusion quality with other fusion approaches on challenging scenarios. From top to bottom: high-brightness and snow in AWMM, nighttime and fog in M^3FD, low-quality in TNO, and daytime in RoadScene.

Experiment

To evaluate the performance of our method, we conduct experiments on five datasets, where four (M^3FD (Liu et al. 2022), RoadScene (Xu et al. 2020), TNO (Toet and Hogervorst 2012), and AWMM (Li et al. 2024c)) for IVIF, one (M^3FD) for object detection and one (FMB (Liu et al. 2023a)) for semantic segmentation. The AWMM dataset is divided into 3 parts for adapter pretraining, fusion model training, and testing. We set the learning rate to 1.5×10^{-4} for the adapter and the fusion module, $\alpha = 0.99$ for the updating weights of EMA, and $\Theta = 0.2$ for the threshold of the uncertainty value. Then the fused images of M^3FD and FMB are respectively fed to the object detection and semantic segmentation model for training and evaluation. Experiments are conducted on a NVIDIA 4080 (32GB) GPU.

Infrared and visible image fusion

We evaluate the IVIF performance of DAFusion with 11 SOTA methods, including TIM (Liu et al. 2023c), MetaFusion (Zhao et al. 2023a), Text-IF (Yi et al. 2024), CDDFuse (Zhao et al. 2023b), DCEvo (Liu et al. 2025a), BDLFusion (Liu et al. 2023d), PromptFusion (Liu et al. 2024a), LRRNet (Li et al. 2023a), EMMA (Zhao et al. 2024), ReFusion (Bai et al. 2024), and SAGE (Wu et al. 2025).

Qualitative comparisons As image fusion results displayed in Figure 3, our method exceeds other approaches, demonstrating both the remarkable capability of producing high-quality fused images and the robust adaptability to changing conditions. In the meantime of reducing noise levels, infrared objects can be clearly seen in harsh scenes, ensuring visible quality and accurate representations.

Quantitative comparisons We compare our method with 11 competing approaches on M^3FD, RoadScene, TNO, and AWMM datasets, containing 400, 221, 57, and 100 images, respectively. We adopt 4 objective metrics for analysis, and the results demonstrate that our method consistently ranks among the top two across all metrics, as shown in Table 1. Specifically, the highest CC and PSNR imply that our method can effectively achieve pixel-level fidelity. Furthermore, the significant improvements of SCD and SSIM indicate that our method integrates enough source information.

Figure 4: Qualitative comparisons of object detection performance in both smoke and daytime scenes

Methods	M^{3}FD				RoadScene				TNO				AWMM			
	CC	PSNR	SCD	SSIM	CC	PSNR	SCD	SSIM	CC	PSNR	SCD	SSIM	CC	PSNR	SCD	SSIM
TIM	0.41	60.22	1.51	0.95	0.44	59.32	1.53	0.73	0.28	59.91	1.75	0.82	0.54	60.49	1.36	0.81
Text-IF	0.44	60.12	1.56	0.97	0.43	59.69	1.59	0.91	0.29	60.07	1.78	0.87	0.56	61.83	1.36	0.90
MetaFusion	0.50	60.96	1.72	0.87	0.45	59.30	1.65	0.89	0.33	60.79	**1.90**	0.78	0.59	61.81	**1.59**	0.86
CDDFuse	0.48	59.64	1.69	1.04	0.46	59.54	1.67	0.95	0.32	60.18	1.85	0.91	0.56	61.76	1.41	0.89
DCEvo	0.45	59.60	1.52	1.03	0.42	59.13	1.57	0.94	0.30	59.21	1.81	0.89	0.54	61.12	1.28	0.94
BDLFusion	0.54	**62.16**	1.65	1.04	0.49	59.55	1.68	0.99	0.30	60.91	1.81	0.82	0.57	61.20	1.54	0.91
PromptFusion	0.44	59.93	1.54	1.02	0.45	59.71	1.64	0.96	0.31	60.36	1.81	0.90	0.56	61.99	1.36	0.89
LRRNet	0.49	61.59	1.57	0.81	0.45	59.28	1.75	0.86	0.25	59.84	1.65	0.81	0.56	61.34	1.33	0.83
EMMA	0.44	59.70	1.53	1.01	0.43	59.52	1.56	0.92	0.31	60.02	1.78	0.93	0.54	61.80	1.29	0.87
ReFusion	0.48	59.94	1.68	1.03	0.46	60.08	1.68	0.96	0.31	59.90	1.83	0.92	0.55	61.92	1.34	0.93
SAGE	0.53	60.26	1.74	1.06	0.47	59.16	1.66	0.94	0.31	59.01	1.83	0.90	0.60	61.43	1.50	0.95
Ours	**0.55**	62.02	**1.76**	**1.08**	**0.51**	**61.23**	**1.77**	**1.03**	**0.35**	**61.60**	**1.90**	**0.96**	**0.62**	**63.25**	1.54	**0.97**

Table 1: Quantitative comparisons of fusion metrics with other SOTA fusion methods on M^{3}FD, RoadScene, TNO, and AWMM datasets. **Boldface** denotes the best while underline denotes the second best results.

Downstream IVIF applications

In this section, we evaluate the practicality of our method in downstream tasks. We randomly split the M^{3}FD dataset with an 8:1:1 ratio and utilize YOLOv8s for object detection. Similarly, the FMB dataset is randomly divided in an 8:1:1 ratio using SegFormer-B1 for semantic segmentation.

We test the detection performance on six classes, shown in Table 2 (left), our method achieves the highest mAP. Figure 4 visually displays the superiority of our method, that while others miss the detection of people or misidentify trees as people, our method detects all objects precisely. In the semantic segmentation, we evaluate all methods on 14 classes, from which the results of 6 classes are selected and displayed in Table 2 (right). Our method outperforms notably in terms of mIoU of all 14 classes. The qualitative segmentation comparisons in Figure 5 also show that our method obtains the best segmentation results, precisely segmenting all classes with clear edges.

Ablation study

To understand the contribution of each component in our framework, we conduct a series of ablation experiments. Four fusion metrics and two downstream task metrics are utilized to fully assess the performance for experimental groups, displayed in Table 3. Performance of image fusion and object detection is tested on M^{3}FD dataset while that of semantic segmentation is evaluated on FMB dataset.

Study on the domain adapters. As shown in Table 3, the first five rows present our ablation experiments for our proposed domain adapter, where the 2nd to 4th rows explore the necessity for both ranks of adapters, and the 5th row proves the effectiveness of the HKA strategy for dynamical rank weights adjustment. We can achieve that the absence of any of the high-rank or low-rank adapters would lead to performance degradation. The HKA strategy is also evaluated to be valid for balancing branch weights.

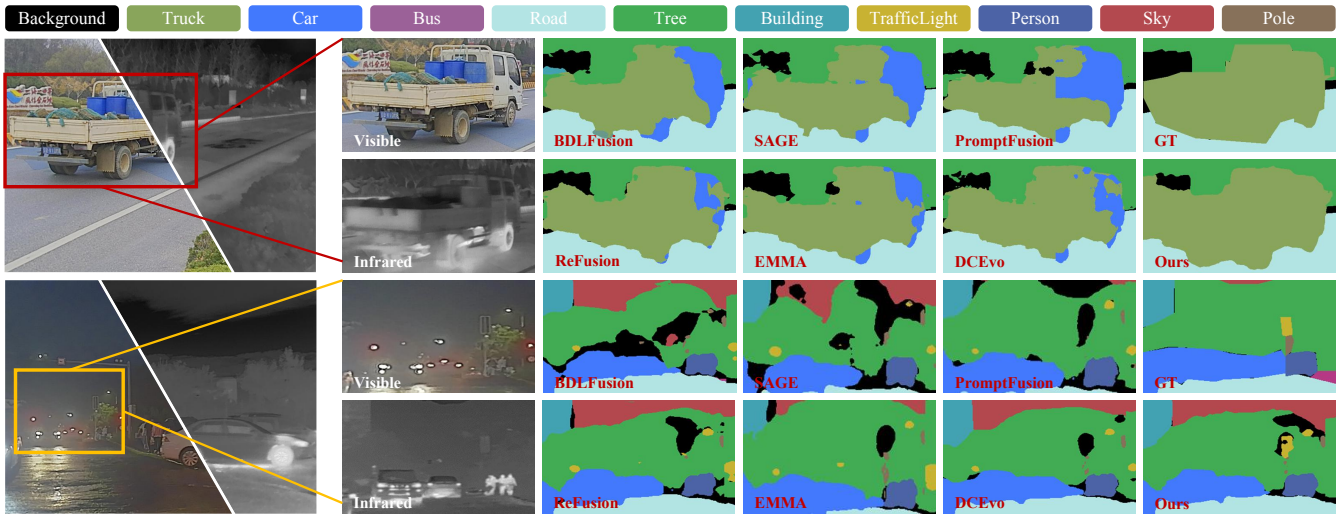

Figure 5: Qualitative comparisons of semantic segmentation performance in both daytime and nighttime scenes.

Methods	M^3FD							FMB						
	People	Car	Bus	Light	Moto	Trunk	mAP	Car	Person	Sky	Bus	Motor	Pole	mIoU
TIM	0.435	0.669	0.686	0.319	0.379	0.612	0.517	0.854	0.684	0.954	0.852	0.722	0.545	0.728
Text-IF	0.475	0.680	0.689	0.304	0.391	0.633	0.529	0.866	0.726	0.958	0.859	0.724	0.557	0.743
MetaFusion	0.449	0.653	**0.717**	0.330	0.357	0.563	0.512	0.865	0.715	0.957	0.825	0.732	0.555	0.736
CDDFuse	0.475	0.677	0.675	0.322	0.384	**0.652**	0.531	0.863	0.723	0.956	**0.877**	**0.744**	0.550	0.739
DCEvo	0.482	0.679	0.700	0.308	0.374	0.605	0.525	0.865	0.730	0.957	0.855	0.699	0.553	0.740
BDLFusion	0.487	0.672	0.670	0.299	0.399	0.624	0.525	0.863	**0.738**	**0.960**	0.857	0.690	0.558	0.739
PromptFusion	0.469	0.675	0.696	0.305	0.404	0.609	0.526	0.863	0.728	0.955	0.860	0.732	0.548	0.739
LRRNet	0.464	0.667	0.664	0.322	0.390	0.599	0.518	0.867	0.713	0.956	0.841	0.738	0.553	0.741
EMMA	0.470	0.675	0.674	0.320	0.375	0.611	0.521	0.865	0.721	0.955	0.858	0.710	0.549	0.736
ReFusion	0.478	0.679	0.672	**0.342**	0.416	0.641	0.538	0.870	0.730	0.957	0.836	0.734	0.556	0.742
SAGE	0.480	0.675	0.694	0.307	0.382	0.636	0.529	0.860	0.728	0.958	0.862	0.726	0.547	0.739
Ours	**0.489**	**0.683**	0.705	0.326	**0.423**	0.644	**0.545**	**0.884**	0.730	**0.960**	0.855	0.741	**0.559**	**0.749**

Table 2: Quantitative comparisons of downstream task performance with other existing fusion methods for object detection on M^3FD and semantic segmentation on FMB dataset. **Boldface** denotes the best while underline denotes the second best results.

Setting	CC	PSNR	SCD	SSIM	mAP	mIoU
Base	0.43	59.71	1.51	0.94	0.523	0.736
low-rank	0.47	59.82	1.55	0.98	0.527	0.738
high-rank	0.46	59.97	1.57	0.98	0.526	0.737
dual-rank	0.49	60.67	1.64	1.03	0.534	0.744
dual-rank+HKA	0.53	61.42	1.71	1.07	0.539	0.746
Ours	**0.55**	**62.02**	**1.76**	**1.08**	**0.545**	**0.749**

Table 3: Quantitative results of the ablation study. The metrics across three tasks prove every component to be effective.

Study on the bi-level optimization. To assess whether bi-level optimization plays a role in the collaboration of image fusion and domain adaptation, we compare the method with (Ours) and without bi-level optimization (dual-rank+HKA). According to the experimental results in Table 3, the incorporation of a bi-level optimization framework leads to an improvement in overall performance, achieving our goal of mutual promotion between two related modules.

The above ablation study demonstrates that every component in our framework is effective and indispensable, presenting the necessity of employing both the dual-rank domain adapter and the bi-level optimization strategy.

Conclusion

In this paper, we introduce the Domain Adaptation Guided Infrared and Visible Image Fusion method (DAFusion), obtaining significant improvement on both fusion quality and downstream task performance. Through integrating the dual-rank domain adapter and the image fusion module in a bi-level optimization framework, we achieve the mutual promotion and guidance between image fusion and domain adaptation. Extensive experiments across three benchmarks evaluate the effectiveness and robustness of our method for severe environments, surpassing existing SOTA methods.

References

Bai, H.; Zhao, Z.; Zhang, J.; Wu, Y.; Deng, L.; Cui, Y.; Jiang, B.; and Xu, S. 2024. ReFusion: Learning Image Fusion from Reconstruction with Learnable Loss Via Meta-Learning. *International Journal of Computer Vision*, 1–21.

Boudiaf, M.; Denton, T.; Van Merriënboer, B.; Dumoulin, V.; and Triantafillou, E. 2023. In search for a generalizable method for source free domain adaptation. In *International Conference on Machine Learning*, 2914–2931. PMLR.

Cao, B.; Xu, X.; Zhu, P.; Wang, Q.; and Hu, Q. 2024. Conditional Controllable Image Fusion. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 120311–120335. Curran Associates, Inc.

Chang, W.-G.; You, T.; Seo, S.; Kwak, S.; and Han, B. 2019. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 7354–7362.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, 1989–1998. Pmlr.

Li, H.; Wu, X.-J.; and Kittler, J. 2021. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73: 72–86.

Li, H.; Xu, T.; Wu, X.-J.; Lu, J.; and Kittler, J. 2023a. LRRNet: A novel representation learning guided fusion framework for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 11040–11052.

Li, X.; Li, X.; Tan, T.; Li, H.; and Ye, T. 2025a. UMC-Fuse: A Unified Multiple Complex Scenes Infrared and Visible Image Fusion Framework. *IEEE Transactions on Image Processing*.

Li, X.; Li, X.; Ye, T.; Cheng, X.; Liu, W.; and Tan, H. 2024a. Bridging the gap between multi-focus and multi-modal: a focused integration framework for multi-modal image fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1628–1637.

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

Li, X.; Liu, W.; Li, X.; Zhou, F.; Li, H.; and Nie, F. 2024c. All-weather Multi-Modality Image Fusion: Unified Framework and 100k Benchmark. *arXiv preprint arXiv:2402.02090*.

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

Li, X.; Zou, Y.; Liu, J.; Jiang, Z.; Ma, L.; Fan, X.; and Liu, R. 2023b. From text to pixels: a context-aware semantic synergy solution for infrared and visible image fusion. *arXiv preprint arXiv:2401.00421*.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. 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*, 5802–5811.

Liu, J.; Fan, X.; Jiang, J.; Liu, R.; and Luo, Z. 2021a. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 105–119.

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

Liu, J.; Lin, R.; Wu, G.; Liu, R.; Luo, Z.; and Fan, X. 2024b. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5): 1748–1775.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023a. Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation. In *International Conference on Computer Vision*.

Liu, J.; Yang, S.; Jia, P.; Zhang, R.; Lu, M.; Guo, Y.; Xue, W.; and Zhang, S. 2023b. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*.

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

Liu, R.; Liu, Z.; Liu, J.; and Fan, X. 2021b. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1600–1608.

Liu, R.; Liu, Z.; Liu, J.; Fan, X.; and Luo, Z. 2023c. A Task-guided, Implicitly-searched and Meta-initialized Deep Model for Image Fusion. *arXiv preprint arXiv:2305.15862*.

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

Liu, Z.; Liu, J.; Wu, G.; Ma, L.; Fan, X.; and Liu, R. 2023d. Bi-level Dynamic Learning for Jointly Multi-modality Image Fusion and Beyond. *IJCAI*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Rao, D.; Xu, T.; and Wu, X.-J. 2023. TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. *IEEE Transactions on Image Processing*, 1–1.

Song, J.; Lee, J.; Kweon, I. S.; and Choi, S. 2023. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11920–11929.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM international conference on multimedia*, 4003–4011.

Toet, A.; and Hogervorst, M. A. 2012. Progress in color night vision. *Optical Engineering*, 51(1): 010901–010901.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Wang, Q.; Dai, D.; Hoyer, L.; Van Gool, L.; and Fink, O. 2021. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8515–8525.

Wang, Z.; Li, X.; Duan, H.; and Zhang, X. 2022. A self-supervised residual feature learning model for multifocus image fusion. *IEEE Transactions on Image Processing*, 31: 4527–4542.

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

Wang, Z.; Zhao, L.; Zhang, J.; Song, R.; Song, H.; Meng, J.; and Wang, S. 2025b. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, 1–23.

Wu, G.; Liu, H.; Fu, H.; Peng, Y.; Liu, J.; Fan, X.; and Liu, R. 2025. Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, Z.; Zhang, Y.; Li, H.; and Liu, Y. 2025. Instruction-driven fusion of Infrared–visible images: Tailoring for diverse downstream tasks. *Information Fusion*, 121: 103148.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, W.; Xie, S.; Zhao, F.; He, Y.; and Lu, H. 2023a. 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*, 13955–13965.

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

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25912–25921.

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

Zhao, Z.; Xu, S.; Zhang, J.; Liang, C.; Zhang, C.; and Liu, J. 2021. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1186–1196.

Zou, Y.; Yu, Z.; Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, 289–305.