Progressive Modality-Adaptive Interactive Network for Multi-Modality Image Fusion

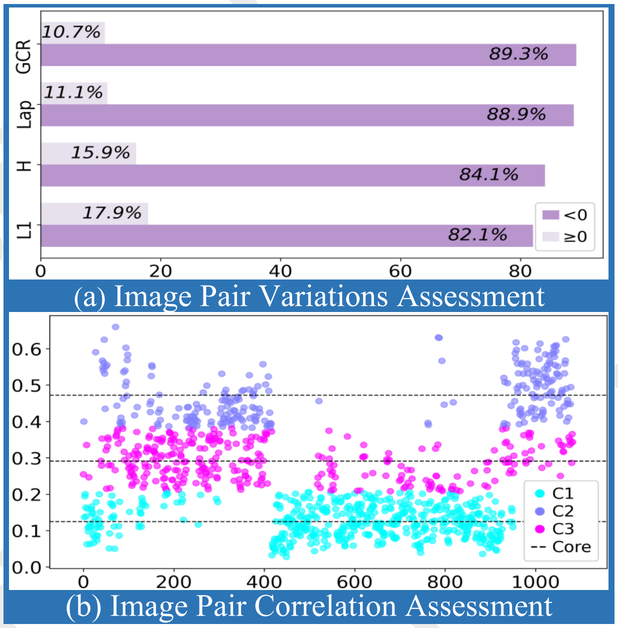
Chaowei Huang[1,2], Yaru Su[1,2], Huangbiao Xu[1,2], Xiao Ke[1,2]
[1]Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing,
College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China
[2]Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350116, China
parachutermare@gmail.com, yarusu@fzu.edu.cn, huangbiaoxu.chn@gmail.com, kex@fzu.edu.cn

Abstract

Multi-modality image fusion (MMIF) integrates features from distinct modalities to enhance visual quality and improve downstream task performance. However, existing methods often overlook the sparsity variations and dynamic correlations between infrared and visible images, potentially limiting the utilization of both modalities. To address these challenges, we propose the Progressive Modality-Adaptive Interactive Network (PoMAI), a novel framework that not only dynamically adapts to the sparsity and structural disparities of each modality but also enhances intermodal correlations, thereby optimizing fusion quality. The training process consists of two stages: in the first stage, the Neighbor-Group Matching Model (NGMM) models the high sparsity of infrared features, while the Context-Aware Modeling Network (CAMN) captures rich structural details in visible features, jointly refining modality-specific characteristics for fusion. In the second stage, the Modality-Interactive Compensation Module (MICM) refines inter-modal correlations via dynamic compensation mechanism, while freezing the first-stage modules to focus MICM solely on the compensation task. Extensive experiments on benchmark datasets demonstrate that PoMAI surpasses state-of-the-art methods in fusion quality and excels in downstream tasks.

1 Introduction

The increasing demand for image processing technologies has driven the widespread application of image fusion, particularly in complex environments [Sun *et al.*, 2022; Zhang *et al.*, 2021] where single-modality images often fail to provide sufficient information. To address this limitation, MMIF [Xu *et al.*, 2020; Zhao *et al.*, 2020] combines images from different modalities, leveraging their strengths to enhance information representation and capture fine-grained details. Within MMIF, infrared and visible image fusion (IVF) [Ma *et al.*, 2023] plays a critical role by combining the complementary information from both modalities. It leverages the infrared's advantage in harsh conditions while benefiting from the finer details provided by visible images. This fusion technique generates more comprehensive and insightful images, with significant applications [Zhang *et al.*, 2020] in object detection and surveillance. Furthermore, MMIF also encompasses medical image fusion (MIF) [James and Dasarathy, 2014], where modalities such as CT and MRI are integrated to offer more precise diagnostic insights, thereby facilitating advanced medical decision-making.



(a) Image Pair Variations Assessment

(b) Image Pair Correlation Assessment

Figure 1: Analysis of inter-modal variations and correlations: (a) indicates higher sparsity in the infrared modality, while (b) demonstrates dynamic correlation between infrared and visible modalities.

In recent years, numerous methods [Deng and Dragotti, 2020; Zhang and Ma, 2021] have been proposed in the field of MMIF. Although these methods have achieved some success, some still fail to address the distinct information distribution characteristics of different modalities. For instance, infrared images often contain sparse and localized information, while visible images provide richer, more detailed content. Furthermore, the correlation between modalities is not static but varies across different scenes, a dynamic relationship that existing methods have yet to adequately explore. This dynamic relationship, along with the modality-specific characteristics, plays a crucial role in multimodal image fusion, as evidenced by quantitative analysis.

To facilitate a deeper comparison between infrared ($\mathcal{I}$) and visible ($\mathcal{V}$) image modalities, we utilize paired images from the MSRS [Tang *et al.*, 2022] dataset. In this study, we compute several evaluation metrics to quantitatively assess the relationship between these two modalities. The visual results, as illustrated in Figure 1, consist of bar charts that depict the differences in information sparsity across various metrics, along with a scatter plot of JS divergence obtained via K-Means clustering, which captures the dynamic correlation across different scenes. As for the bar chart, it illustrates the distribution of P values across different metrics, with each bar representing a specific metric. Given an input pair $\mathcal{I}_j$ and $\mathcal{V}_j$ (representing the j-th image pair), we compute their feature values $F(\mathcal{I}_j)$ and $F(\mathcal{V}_j)$ based on various metrics, including gradient change rate (GCR) , Laplacian variance (Lap) , entropy (H) and L1 norm (L1) . The difference $F(\mathcal{I}_j) - F(\mathcal{V}_j)$ is then calculated, and the statistic P is defined as the proportion of instances where this difference is positive or negative. This provides a quantitative measure of the distribution of differences for each metric. The bar chart demonstrates that infrared images commonly exhibit higher sparsity compared to visible images, clearly highlighting the differences between the two modalities. Meanwhile, along with the scatter plot derived from the JS divergence analysis of the $\mathcal{I}_j, \mathcal{V}_j$ image pairs, K-Means clustering is applied to categorize the image pairs into three distinct groups. This clustering captures the dynamic correlation across different scenes, revealing that the correlation between the image pairs varies with the scene, highlighting how the image pairs complement each other under different conditions. These findings highlight the complex relationship between modality differences and correlations, offering key insights for designing the subsequent fusion approach.

Motivated by the critical insights into the intricate interplay between modality differences and correlations, we introduce a novel two-stage framework designed to comprehensively tackle these challenges. In the first stage, to accommodate the unique characteristics of infrared and visible images, we design two distinct modules: the Neighbor-Group Matching Model (NGMM) and the Context-Aware Modeling Network (CAMN), enabling modality-adaptive feature extraction. Unlike traditional methods that apply uniform processing across modalities, our approach tailors the processing to the intrinsic differences of each modality, enabling finer-grained feature extraction and effectively addressing the challenges posed by modality disparities. Building on this, the second stage introduces the Modality-Interactive Compensation Module (MICM), trained with frozen first-stage modules to capture deep inter-modality relationships through fine-grained interactions. The module employs a gated-weight fusion mechanism to dynamically adapt to the varying correlations between image pairs, effectively addressing the evolving interactions and ensuring the precise integration of modality-specific information.

To summarize, our contributions are as follows:

- We propose an innovative asymmetric two-stage framework to effectively capture the variations and correlations between different modalities, enhancing the robustness and adaptability of our fusion model.
- We design two distinct, modality-adaptive feature extraction modules that dynamically adapt to the varying sparsity levels and information distributions specific to infrared and visible images, ensuring a more precise and effective fusion process.
- We propose a modality-interactive compensation mechanism that adeptly captures inter-modality correlations through fine-grained interactions, enhancing fusion performance by leveraging a gated-weight fusion strategy.

2 Related Works

Recent years have witnessed significant progress in deep learning-based multimodal image fusion methods. Convolutional networks, known for their ability to capture spatial patterns, have been widely applied to image fusion tasks [Li *et al.*, 2018; Liu *et al.*, 2018]. However, convolution-based methods often struggle to capture long-range dependencies due to their inherently local receptive fields. As vision attention mechanisms [Dosovitskiy, 2020; Zamir *et al.*, 2022] have gained prominence, Transformer-based methods have emerged as powerful tools for capturing global dependencies, greatly enhancing performance in complex fusion tasks [Zhao and Nie, 2021; Tang *et al.*, 2024]. Generative Adversarial Networks (GANs) [Creswell *et al.*, 2018; Goodfellow *et al.*, 2020] use a game-theoretic interaction between a generator and a discriminator to generate high-quality fused images. Unlike convolutional and Transformer-based methods, which focus on feature extraction, GANs use adversarial training to enhance fusion, effectively preserving thermal target information and retaining key details [Li *et al.*, 2019; Chakraborty *et al.*, 2024]. Autoencoders (AE) [Michelucci, 2022; Berahmand *et al.*, 2024] have also been widely applied to multi-modal image fusion tasks [Zhao *et al.*, 2020; Li *et al.*, 2021]. AE-based methods employ an encoder-decoder architecture to map input images into low-dimensional representations, which are then fused in a high-dimensional space to extract key features and produce high-quality fused images. With the growing demand for practical applications, image fusion methods that integrate downstream tasks have seen rapid development [Liu *et al.*, 2023a; Liu *et al.*, 2023b]. By exploiting the synergy between upstream and downstream tasks, these methods effectively balance fusion quality and performance for the specific target task. While some existing methods may not fully account for the distinct information distribution characteristics of different modalities, our approach introduces a tailored processing strategy to enhance modality-specific feature extraction. By incorporating dynamic weighting and gating mechanisms, our method captures the dynamic correlations between modalities, enabling effective interaction and compensation across them. Together, these strategies effectively handle the coexistence of modality differences and correlations, ultimately enhancing the quality of the fused images.

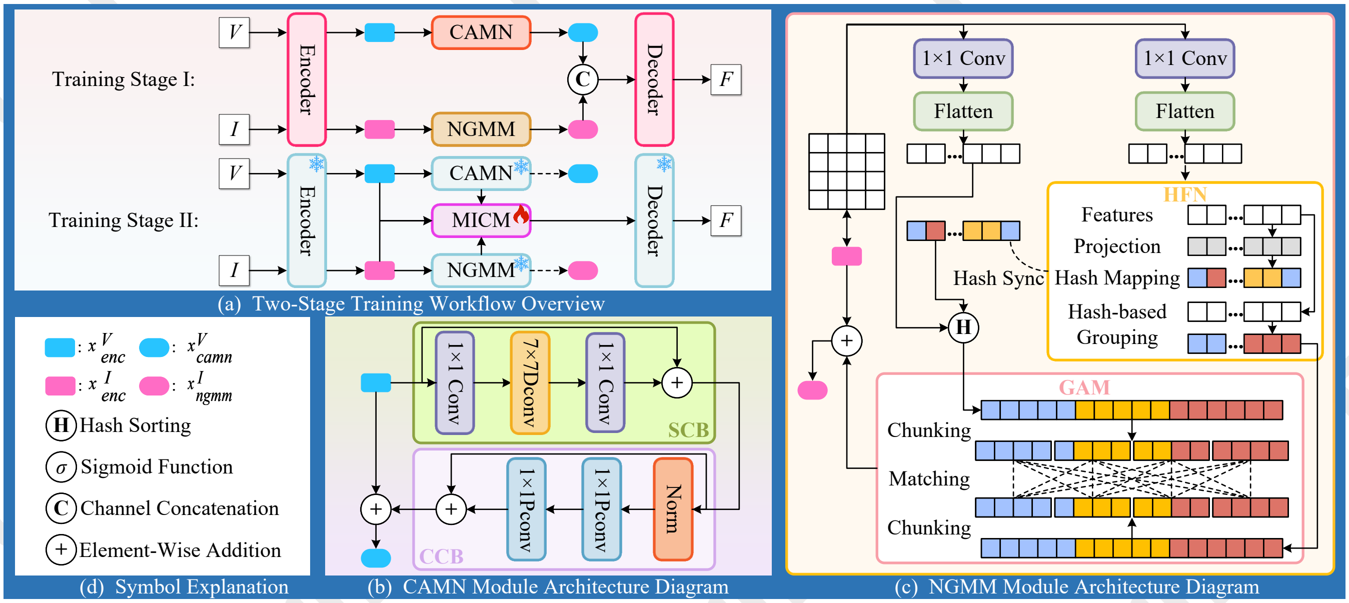


Figure 2: Overview of the proposed framework. (a) Illustration of the two-stage training workflow. (b) Detailed architecture of the Context-Aware Modeling Network (CAMN) . (c) Detailed architecture of the Neighbor-Group Matching Model (NGMM) .

3 Method

3.1 Overall Architecture

In this paper, we propose a progressive modality-adaptive interactive network (PoMAI) for MMIF, designed to address the challenges of modality-specific sparsity variations and dynamic correlations. As shown in Figure 2, PoMAI comprises five key modules and two training stages. The first stage focuses on learning modality-specific features, while the second stage builds upon the first by fine-tuning the model, with selected modules frozen, to better capture dynamic correlations. To further clarify the operation of PoMAI, we use the infrared and visible image fusion (IVF) task as a case study.

3.2 Shared Encoder

We employ a shared encoder $\mathcal{E}(\cdot)$, consisting of four cascaded Transformer blocks [Zamir *et al.*, 2022], to extract features from the input images $I_{\mathcal{I}} \in \mathbb{R}^{H\times W\times 1}, I_{\mathcal{V}} \in \mathbb{R}^{H\times W\times 3}$, yielding consistent feature representations:

$$x^{\mathcal{I}}_{enc},\, x^{\mathcal{V}}_{enc} = \mathcal{E}(I_{\mathcal{I}}), \mathcal{E}(I_{\mathcal{V}}). \tag{1}$$

The extracted features $x^{\mathcal{I}}_{enc}, x^{\mathcal{V}}_{enc} \in \mathbb{R}^{H\times W\times C}$ maintain consistent dimensionality across modalities, facilitating efficient multimodal fusion.

3.3 Neighbor-Group Matching Model

Given the feature $x^{\mathcal{I}}_{enc}$, NGMM integrates the Hash-based Feature Neighboring (HFN) and Group Attention Mechanism (GAM) to better align with the sparsity characteristics of infrared images, as shown in Figure 2.

Specifically, $x^{\mathcal{I}}_{enc}$ is processed by a dual-branch structure, where separate convolutions with identical parameters are applied in each branch. The detailed formulation is as follows:

$$x^{\mathcal{I}}_{left},\, x^{\mathcal{I}}_{right} = Conv_{1\times1}(x^{\mathcal{I}}_{enc}), Conv_{1\times1}(x^{\mathcal{I}}_{enc}). \tag{2}$$

Following the dual-branch processing, $x^{\mathcal{I}}_{left}, x^{\mathcal{I}}_{right} \in \mathbb{R}^{H\times W\times C}$ are flattened into $f_{left}, f_{right} \in \mathbb{R}^{L\times C}$, where $L = H\times W$. Leveraging the locality-sensitive hashing mechanism introduced in [Andoni *et al.*, 2015], The HFN module begins by projecting the input features $x^{\mathcal{I}}_{right}$ onto a spherical space. Subsequently, the indices $S \in \mathbb{R}^{H\times W\times 1}$, which identify the top K most relevant elements within the feature space, are generated through hash mapping. These indices S are used to perform hash-based grouping of $x^{\mathcal{I}}_{left}, x^{\mathcal{I}}_{right}$, resulting in $g_{left}, g_{right} \in \mathbb{R}^{L\times C}$. Note that the indices S are computed in the right branch and shared with the left branch. The grouping process is formalized as follows:

$$g = \bigcup_{i=1}^{K}\bigcup_{j=1}^{L}\{f_j \mid S_j = S_i\}. \tag{3}$$

Building on HFN's feature grouping, we introduce the GAM to capture the relationships among grouped features. This process ultimately produces M aggregated feature groups, where M is the total number of groups, calculated as $M = \frac{L}{K}$. The attention scores for each group are reshaped and added to $x^{\mathcal{I}}_{enc}$ via a residual connection, yielding the NGMM's output and enhancing the features.

3.4 Context-Aware Modeling Network

Unlike infrared images that exhibit high sparsity, visible images contain a wealth of details and intricate structures. To address this, we propose the Context-Aware Modeling Network (CAMN), which integrates the Spatial-Context Block

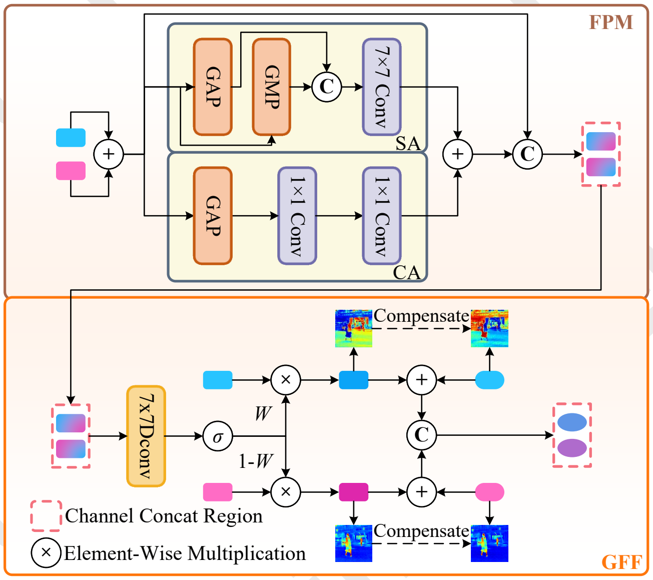


Figure 3: The MICM architecture enables interaction and compensation between infrared and visible modalities via FPM and GFF.

(SCB) and Channel-Context Block (CCB) to effectively extract and encode the rich information in visible images.

As shown in Figure 2, the processing of $x_{enc}^{\mathcal{V}}$begins with the SCB, designed to extract spatial contextual information. Within the SCB, the input feature passes through a 1×1 conv for initial processing, followed by a 7×7 dilated conv (DC) to capture spatial dependencies. A subsequent 1×1 conv refines the representation, with residual connections integrating the original features to produce $x_{scb}^{\mathcal{V}}$. The SCB module emphasizes spatial representation, formulated as follows:

$$x_{scb}^{\mathcal{V}} = Conv_{1\times1}(DC_{7\times7}(Conv_{1\times1}(x_{enc}^{\mathcal{V}}))) + x_{enc}^{\mathcal{V}}. \quad (4)$$

The output of the SCB, $x_{scb}^{\mathcal{V}}$, is fed into the CCB for channel refinement. The CCB applies a normalization layer, followed by two 1×1 pointwise conv (PC) to adjust channel dimensions. A residual connection combines $x_{scb}^{\mathcal{V}}$ with the intermediate features, yielding $x_{ccb}^{\mathcal{V}}$. The CCB module emphasizes channel-wise interaction, formulated as follows:

$$x_{ccb}^{\mathcal{V}} = PC_{1\times1}(PC_{1\times1}(Norm(x_{scb}^{\mathcal{V}}))) + x_{scb}^{\mathcal{V}}. \quad (5)$$

Finally, a residual connection combines $x_{ccb}^{\mathcal{V}}$ with $x_{enc}^{\mathcal{V}}$, helping preserve the rich information and intricate details of visible images, and yielding the final output $x_{camn}^{\mathcal{V}}$.

3.5 Modality-Interactive Compensation Module

Building on previous observations, we find that certain image pairs exhibit significant modal correlations, providing a foundation for the interaction module. Inspired by [Dai *et al.*, 2021], we propose the Modal Interaction Compensation Module (MICM), illustrated in Figure 3, which integrates the Feature Perceptual Module (FPM) and Gated Feature Fusio (GFF) to dynamically capture cross-modal correlations and compensate for modal discrepancies.

Given the features $x_{enc}^{\mathcal{I}}$ and $x_{enc}^{\mathcal{V}}$, we fuse the two modalities through element-wise addition to obtain x_{sum}. Then, x_{sum} is fed into both the channel attention (CA) and spatial attention (SA) modules. Next, the attention scores from CA and SA are combined via element-wise addition. Finally, the fused attention features are concatenated (denoted as $Cat(\cdot,\cdot)$) with x_{sum} along the channel dimension to produce the output of the FPM, x_{fpm}. This process is succinctly expressed by the following formula:

$$x_{sum} = x_{enc}^{\mathcal{I}} + x_{enc}^{\mathcal{V}}, \quad (6)$$
$$x_{fpm} = Cat(SA(x_{sum}) + CA(x_{sum}), x_{sum}). \quad (7)$$

The result feature x_{fpm} from the FPM is further processed by the GFF module. First, a 7×7 convolution is applied to x_{fpm}, and the result is passed through a sigmoid activation $\sigma(\cdot)$ to generate the gating map $\mathcal{W}$. This gating map modulates $x_{enc}^{\mathcal{I}}$ and $x_{enc}^{\mathcal{V}}$, producing weighted features. These weighted features are then added to $x_{ngmm}^{\mathcal{I}}$ and $x_{camn}^{\mathcal{V}}$, respectively. Finally, the resulting features are concatenated along the channel dimension to form x_{gff}, which also serves as the output of the MICM. This process can be described as:

$$\mathcal{W} = \sigma(\text{Dconv}_{7\times7}(x_{fpm})), \quad (8)$$
$$x_{wei}^{\mathcal{V}} = \mathcal{W} \cdot x_{enc}^{\mathcal{V}},\ x_{enh}^{\mathcal{V}} = x_{wei}^{\mathcal{V}} + x_{camn}^{\mathcal{V}}, \quad (9)$$
$$x_{wei}^{\mathcal{I}} = (1 - \mathcal{W}) \cdot x_{enc}^{\mathcal{I}},\ x_{enh}^{\mathcal{I}} = x_{wei}^{\mathcal{I}} + x_{ngmm}^{\mathcal{I}}, \quad (10)$$
$$x_{gff} = Cat(x_{enh}^{\mathcal{I}}, x_{enh}^{\mathcal{V}}). \quad (11)$$

3.6 Decoder

The decoder $\mathcal{D}(\cdot)$ is used in both stages, mirroring the encoder's structure to ensure consistent feature processing. In the first stage, it takes concatenated features from both modalities as input, while in the second stage, it processes the fused features from the MICM module, ensuring visual coherence and effective multimodal integration. Formally:

$$\text{Stage 1:} \quad I_{\mathcal{F}} = \mathcal{D}(Cat(x_{ngmm}^{\mathcal{I}}, x_{camn}^{\mathcal{V}})), \quad (12)$$
$$\text{Stage 2:} \quad I_{\mathcal{F}} = \mathcal{D}(x_{micm}). \quad (13)$$

3.7 Progressive Training

As illustrated in Figure 2, our framework employs a progressive two-stage training strategy. Specifically, in the first stage, we jointly train the encoder $\mathcal{E}(\cdot)$, NGMM, and CAMN to capture modality-specific feature from paired images $\{I_{\mathcal{I}}, I_{\mathcal{V}}\}$. These features are then decoded by $\mathcal{D}(\cdot)$ to generate a fused image that integrates the essential information from both modalities. In the second stage, we freeze the first-stage modules and introduce MICM to build the dynamic correlation between the two modalities. The resulting features, refined through enhanced cross-modal interactions, are decoded by $\mathcal{D}(\cdot)$ to produce the final fused image, reflecting the dynamic correlation between the modalities. Since the objective of both stages is to synthesize a fused image that effectively integrates information from the two modalities, we employ the same loss function $\mathcal{L}$ for consistency and training efficiency. The loss function is formulated as follows:

$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{ssim} + \alpha_2 \mathcal{L}_{mse} + \alpha_3 \mathcal{L}_{int} + \alpha_4 \mathcal{L}_{grad}. \quad (14)$$

Method	Source	TNO				RoadScene				M3FD				MSRS			
		SD	SF	AG	VIF	SD	SF	AG	VIF	SD	SF	AG	VIF	SD	SF	AG	VIF
DeFusion	ECCV'22	31.37	7.26	2.97	0.54	37.65	8.91	3.54	0.59	29.90	8.31	2.91	0.55	34.88	7.98	2.60	0.75
UMF	IJCAI'22	29.95	9.09	3.34	0.56	39.47	10.86	4.14	0.64	31.44	10.00	3.33	0.61	20.76	7.10	2.13	0.43
CDDFuse	CVPR'23	45.49	**13.56**	5.07	0.73	**55.62**	**17.18**	**6.11**	**0.65**	41.28	16.49	5.41	0.78	43.38	**11.56**	3.73	**1.05**
SegMIF	ICCV'23	45.77	13.23	5.14	**0.80**	49.47	15.12	5.77	0.63	42.41	16.03	5.42	**0.82**	41.96	11.01	3.60	0.76
BDLFusion	IJCAI'23	37.94	8.95	3.72	0.59	39.86	8.83	3.66	0.59	31.43	9.31	3.32	0.64	33.63	7.80	2.59	0.73
IGNet	ACMMM'23	39.57	9.91	4.27	0.56	43.13	11.11	4.66	0.54	42.50	13.97	5.07	0.60	33.91	9.66	3.18	0.69
LRRNet	TPAMI'23	42.78	10.86	4.30	0.55	40.90	11.60	4.29	0.48	30.13	11.63	3.95	0.56	31.76	8.47	2.64	0.54
EMMA	CVPR'24	**47.40**	12.94	**5.36**	0.67	54.44	15.09	5.75	0.64	**43.00**	**16.78**	**5.86**	0.76	**44.59**	**11.56**	**3.77**	**0.97**
CAF	IJCAI'24	36.17	11.62	4.49	0.58	45.48	14.55	5.55	0.61	36.01	14.18	4.85	0.60	26.59	9.81	3.13	0.59
MMDRFuse	ACMMM'24	30.18	8.77	3.37	0.59	31.40	7.88	3.00	0.58	29.17	11.24	3.76	0.66	37.11	9.65	3.20	0.85
ITFuse	PR'24	40.17	7.67	3.34	0.53	54.60	10.23	4.31	0.56	36.00	8.56	3.22	0.55	35.16	6.90	2.46	0.66
TIM	TPAMI'24	44.50	9.94	3.95	0.70	38.32	9.60	3.45	0.61	37.39	13.70	4.46	0.62	43.30	11.52	3.77	0.67
PoMAI(Ours)	—	**49.15**	**15.96**	**6.07**	**0.76**	**65.93**	**21.32**	**7.73**	**0.66**	**47.54**	**17.28**	**5.81**	**0.79**	**45.11**	**12.29**	**3.84**	**0.97**

Table 1: Quantitative comparison of the proposed PoMAI with 12 advanced image fusion methods, evaluated on multiple datasets for the IVF task. Bold **red** indicates the best, Bold **blue** indicates the second best.

The coefficients α_1, α_2, α_3, α_4 in the total loss function are weighting factors that balance the contributions of the different loss components to the overall objective. Each term in the overall loss function is computed as follows:

$$\mathcal{L}_{ssim} = 1 - SSIM(I_{\mathcal{F}}, I_{\mathcal{I}}) + 1 - SSIM(I_{\mathcal{F}}, I_{\mathcal{V}}), \quad (15)$$

$$\mathcal{L}_{mse} = \|I_{\mathcal{F}} - I_{\mathcal{I}}\|_2^2 + \|I_{\mathcal{F}} - I_{\mathcal{V}}\|_2^2, \quad (16)$$

$$\mathcal{L}_{int} = \frac{1}{HW}\|I_{\mathcal{F}} - max(\mathcal{I}_{\mathcal{I}}, \mathcal{I}_{\mathcal{V}})\|_1, \quad (17)$$

$$\mathcal{L}_{grad} = \| \, |\nabla I_{\mathcal{F}}| - \max(|\nabla I_{\mathcal{I}}|, |\nabla I_{\mathcal{V}}|)\|_1, \quad (18)$$

where $SSIM(\cdot, \cdot)$ represents the structural similarity index [Wang *et al.*, 2004]. The operator ∇ denotes the Sobel gradient operator, used to capture edge information. The $\max(\cdot, \cdot)$ function selects the pixel-wise maximum values between $I_{\mathcal{I}}$ and $I_{\mathcal{V}}$.

4 Infrared and Visible Image Fusion

4.1 Setup

Datasets. We conduct IVF experiments on four popular datasets: TNO [Toet and Hogervorst, 2012], RoadScene [Xu *et al.*, 2020], M3FD [Liu *et al.*, 2022], and MSRS [Tang *et al.*, 2022]. Our proposed method, PoMAI, is trained on the MSRS training set (1083 pairs) and tested on TNO (20 pairs), RoadScene (70 pairs), M3FD (100 pairs), and the MSRS test set (361 pairs). To evaluate the generalization capability of the model, no fine-tuning was performed on TNO, RoadScene, or M3FD during testing.

Metrics. We use four metrics to evaluate the fusion results: standard deviation (SD), spatial frequency (SF), average gradient (AG) and visual information fidelity (VIF). A higher value of these metrics indicates better fusion quality, with details available in [Ma *et al.*, 2019].

Implement details. Our experiments are implemented with the Pytorch framework and on a machine with four NVIDIA Tesla P100 GPUs. The training images are initially processed by randomly cropping them into 128×128 patches, ensuring a diverse set of training samples for the model. The training process is structured into two phases: the first phase comprises 80 epochs, followed by the second phase with 20 epochs, both utilizing a batch size of 16. The Adam optimizer is employed to update the parameters of each module, with an initial learning rate of 1×10^{-4} that decreases by a factor of 0.5 every 20 epochs. In Eq. 14, the values of α_1, α_2, α_3, and α_4 are set to 5, 1, 1, and 10.

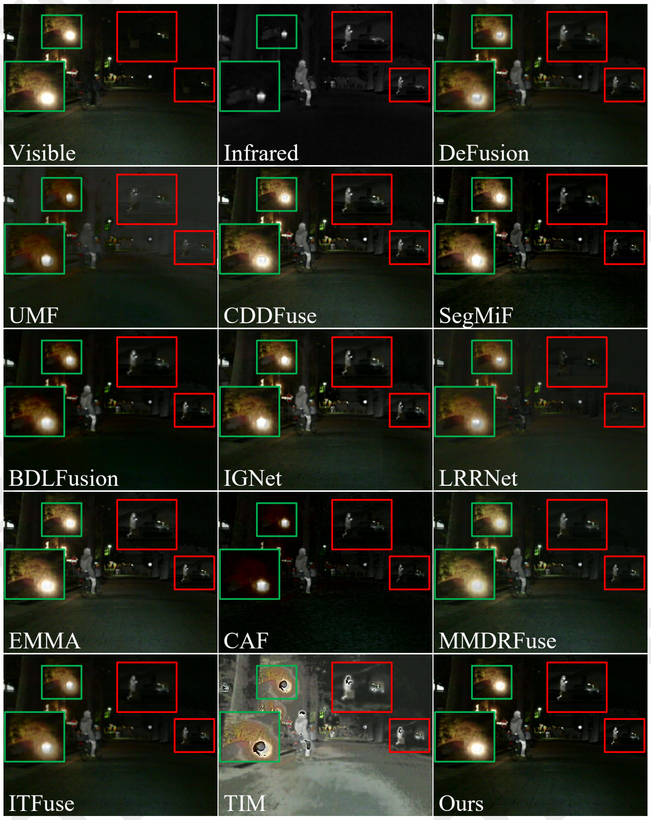


Figure 4: Qualitative comparisons of various methods on MSRS.

	Configuration	SD	SF	AG	VIF
I	w/o NGMM	44.51	11.69	3.65	0.92
II	w/o CAMN	44.48	11.72	3.69	0.91
III	w/o MICM	42.93	11.36	3.57	0.84
IV	w/o Two-Stage	43.04	11.48	3.72	0.89
V	w/o Weight-Frozen	43.67	11.33	3.70	0.92
	PoMAI (Ours)	**45.11**	**12.29**	**3.84**	**0.97**

Table 2: Ablation studies on MSRS test datasets. Two-Stage denotes using MICM in stage two. Weight-Frozen denotes freezing stage-one weights.

4.2 Comparison with SOTA Methods

We conduct comprehensive qualitative and quantitative analyses on four popular datasets, comparing our approach with 12 state-of-the-art (SOTA) methods, including DeFusion [Liang *et al.*, 2022], UMF [Wang *et al.*, 2022], CDDFuse [Zhao *et al.*, 2023], SegMiF [Liu *et al.*, 2023a], BDLFusion [Liu *et al.*, 2023b], IGNet [Li *et al.*, 2023b], LRRNet [Li *et al.*, 2023a], EMMA [Zhao *et al.*, 2024], CAF [Liu *et al.*, 2024a], MMDRFuse [Deng *et al.*, 2024], ITFuse [Tang *et al.*, 2024] and TIM [Liu *et al.*, 2024b].

Qualitative Comparisons. Figure 4 presents visual comparisons on MSRS dataset. Compared with other methods, the proposed approach better preserves the rich details from the visible modality (highlighted by the green-colored box) while simultaneously retaining the thermal information from the infrared image (highlighted by the red-colored box) to the greatest extent, achieving superior fusion results.

Quantitative Comparisons. Beyond visual comparisons, we evaluate our method quantitatively on four datasets using four metrics, as shown in Table 1. Our method achieves significant performance improvements, with superior SD values ensuring optimal contrast and competitive VIF scores indicating effective correlation preservation.

4.3 Ablation Study

We conduct a series of ablation studies on the MSRS dataset, using the same evaluation metrics as in the IVF experiments, to evaluate the individual contributions of the three proposed modules and the effectiveness of the progressive training strategy. The results are summarized in Table 2.

NGMM. To evaluate the effectiveness of the NGMM module, we conduct an ablation study by removing it from the PoMAI framework, while keeping all other components intact. This setup, denoted as Exp.I, demonstrates that the removal of the NGMM module leads to a significant performance decline across most evaluation metrics, failing to match the performance of the full PoMAI model. This highlights the crucial role of the NGMM module in enhancing system performance.

CAMN. To assess the contribution of CAMN, we perform an ablation study by removing CAMN from the PoMAI framework while retaining all other components. This experimental setup is denoted as Exp.II. The result consistently demonstrates that the removal of CAMN resulted in a notable decline in performance across the majority of evaluation metrics, highlighting its critical role in augmenting the model's overall effectiveness.

Method	Source	SD	SF	AG	VIF
DeFusion	ECCV'22	54.28	17.12	4.36	**0.63**
UMF	IJCAI'22	33.86	16.48	4.10	0.39
CDDFuse	CVPR'23	61.59	19.96	5.09	**0.66**
SegMiF	ICCV'23	63.09	19.92	4.95	0.61
BDLFusion	IJCAI'23	53.20	14.06	4.00	0.58
IGNet	ACMMM'23	38.44	13.42	4.25	0.50
LRRNet	TPAMI'23	38.20	11.94	3.15	0.35
EMMA	CVPR'24	**66.70**	19.25	5.23	0.59
CAF	IJCAI'24	52.90	21.56	**5.51**	0.55
MMDRFuse	ACMMM'24	57.26	**28.09**	4.81	0.53
TIM	TPAMI'24	38.06	9.87	2.98	0.37
PoMAI(Ours)	—	**74.96**	**30.59**	**7.55**	**0.66**

Table 3: Quantitative comparison on Harvard Medical datasets. Bold **red** indicates the best, Bold **blue** indicates the second best.

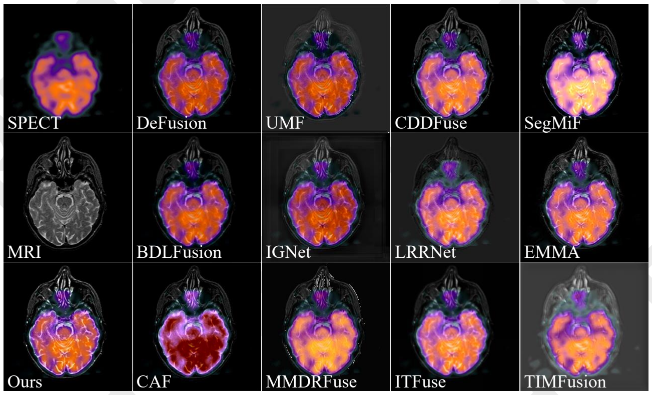


Figure 5: Qualitative comparisons of various methods on Harvard Medical dataset.

MICM and Training Strategy. Finally, to evaluate the effectiveness of incorporating the MICM module and adopting the two-stage training strategy, in which the weights of the first-stage modules (including the Shared Encoder $\mathcal{E}(\cdot)$, NGMM, CAMN and Decoder $\mathcal{D}(\cdot)$) are frozen and only the MICM module is trained, we design three ablation experiments. Specifically, Exp.III employs only the first-stage network without introducing the MICM module, equivalent to training without the two-stage strategy. Exp.IV integrates the MICM module's training process directly into the first-stage network, effectively eliminating the two-stage architecture. Finally, Exp.V introduces the MICM module while allowing the first-stage network's weights to be updated during training, thereby exploring the impact of weight optimization. The results of the three ablation experiments clearly demonstrate the necessity of incorporating the MICM module, as well as the benefits of adopting a two-stage training strategy. Specifically, by freezing the weights of the first-stage modules and progressively training the MICM module in the second stage, we observe a significant improvement in the model performance. This approach underscores the importance of both the MICM module and the progressive training strategy in achieving superior results.

Method	MMOD							MMSS									
	Peo	Car	Bus	Mot	Lam	Tru	mAP	Unl	Car	Per	Bik	Cur	CS	GR	CC	Bu	mIOU
Visible	77.74	93.38	96.13	91.96	81.19	91.69	88.68	97.61	86.09	56.40	65.35	40.73	60.96	66.74	49.64	58.73	64.70
Infrared	83.59	91.52	94.74	80.41	87.14	84.60	87.00	97.44	84.48	66.73	62.52	36.31	44.97	4.25	35.44	55.26	54.16
DeFusion	83.06	93.74	95.91	90.64	79.41	89.72	88.74	97.76	86.46	65.79	65.52	37.97	58.23	54.43	48.85	58.53	63.73
UMF	83.31	93.49	95.45	90.88	81.49	89.83	89.07	97.61	86.18	63.62	64.93	38.45	53.81	44.17	47.52	51.39	60.85
CDDFuse	82.64	93.44	94.43	90.28	82.92	91.11	89.14	97.78	86.62	65.56	64.86	40.16	60.05	61.24	49.45	60.37	65.12
SegMIF	82.43	93.09	94.14	91.11	81.32	89.45	88.59	97.75	86.65	66.00	66.03	38.35	58.38	56.01	46.79	60.07	64.00
BDLFusion	82.85	93.44	95.70	89.56	85.54	88.57	89.29	97.71	86.33	66.67	64.44	38.68	57.54	57.77	46.90	58.65	63.85
IGNet	82.64	93.09	95.31	91.23	83.38	89.02	89.11	97.64	85.70	63.91	65.27	36.38	54.68	43.17	47.00	54.11	60.88
LRRNet	81.73	93.53	95.70	91.29	81.74	89.53	88.92	97.75	86.63	63.88	64.88	41.80	58.22	53.23	48.67	58.11	63.69
EMMA	83.24	93.85	95.74	91.94	84.11	91.64	90.09	97.79	86.73	66.04	65.54	40.22	60.50	62.44	48.85	60.33	65.38
CAF	83.96	93.24	95.83	88.88	84.23	88.48	89.10	97.60	85.76	64.99	63.57	34.78	53.63	41.07	45.87	57.89	60.57
MMDRFuse	82.66	93.26	95.04	92.00	82.01	90.18	89.19	97.76	86.55	65.65	65.59	39.09	58.86	53.61	48.77	58.47	63.82
ITFuse	82.82	93.19	93.66	90.37	81.60	91.23	88.81	97.72	86.55	66.50	65.80	37.97	55.88	44.99	49.05	55.63	62.23
TIM	81.00	93.48	94.37	91.92	84.31	90.79	89.31	97.63	85.69	63.25	64.04	36.49	55.92	51.55	47.32	56.98	62.10
PoMAI(Ours)	84.81	94.03	96.09	91.62	91.55	92.70	91.80	98.05	88.94	68.23	66.66	46.21	62.09	63.72	53.16	66.49	68.17

Table 4: Quantitative comparison of the proposed PoMAI with 12 advanced image fusion methods, evaluated on multiple downstream tasks across diverse datasets. Bold **red** indicates the best, Bold **blue** indicates the second best.

5 Medical Image Fusion

Setup. We perform MIF experiments using 50 image pairs from the Harvard Medical dataset [Johnson and Becker, 2005]. Notably, we directly apply the model trained on IVF tasks to the Harvard dataset without fine-tuning. The quantitative metrics were consistent with those used in IVF task.

Comparison with SOTA Methods. We conduct comprehensive qualitative and quantitative analyses with 12 state-of-the-art (SOTA) competitors, all of which are trained exclusively on IVF-related datasets without fine-tuning on MIF. These methods are consistent with those mentioned in the IVF section. The qualitative results, as illustrated in Figure 5, demonstrate that PoMAI effectively preserves fine-grained texture details, avoids color distortion and accentuates structural information. Quantitatively, as shown in Table 3, PoMAI outperforms most existing methods across the majority of metrics, highlighting its potential for MIF tasks.

6 Downstream IVF Applications

6.1 Evaluation in Object Detection

Setup. The multi-modal object detection (MMOD) is performed on the M3FD dataset, which consists of 4,200 pairs of infrared and visible images, categorized into six labels: people, car, bus, motorcycle, truck, and lamp. The dataset is split into training, validation, and test sets with an 8:1:1 ratio. YOLOv5 [Jocher, 2020] is employed to evaluate detection performance using the mAP@0.5 metric. The training configuration consists of 100 epochs, a batch size of 8, the SGD optimizer and an initial learning rate of 1e-2.

Comparison with SOTA Methods. Table 4 indicates that PoMAI achieves competitive performance across multiple evaluation metrics, with notably higher performance in the mAP@0.5 score compared to existing methods. These results suggest that our approach demonstrates certain effectiveness in improving object detection tasks.

6.2 Evaluation in Semantic Segmentation

Setup. The multi-modal semantic segmentation (MMSS) is performed on the MSRS dataset with pixel-level semantic information of nine object categories(*e.g.*, background, car, person, bike, curve, car stop, guardrail, color cone and bump). DeepLabV3+ [Chen *et al.*, 2018] is utilized to assess semantic segmentation performance based on the intersection-over-union (IoU) metric. The training process is configured with 300 epochs, where the backbone network is frozen for the first 100 epochs. Additional settings include a batch size of 8, the SGD optimizer and an initial learning rate of 7e-3.

Comparison with SOTA Methods. Table 4 summarizes the quantitative semantic segmentation results of various fusion methods on the MSRS dataset. Our method achieves competitive performance across multiple metrics, particularly attaining the highest mIoU score. These results indicate that our approach effectively improves the performance of semantic segmentation tasks.

7 Conclusion

In this paper, we propose a progressive modality-adaptive interactive network (PoMAI) for multi-modality image fusion. We analyze the sparsity differences and dynamic correlations between infrared and visible modalities in the MSRS dataset using diverse metrics. Our analysis reveals that infrared features exhibit sparsity and localization characteristics, while visible features contain richer and more detailed information, with their cross-modal correlations dynamically evolving across different scenes. Based on these findings, we design modality-adaptive feature extraction modules and an modality-interactive compensation module to effectively address these challenges. Our experiments demonstrate PoMAI's superior performance in MMIF, with significant gains in downstream tasks. In future work, we will explore advanced feature integration methods beyond simple channel concatenation to enhance inter-modal fusion.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503, in part by the National Natural Science Foundation of China under Grant 61972097 and U21A20472, in part by the Major Scientific Research Project for Technology Promotes Police under Grant 2024YZ040001, in part by the Natural Science Foundation of Fujian Province under Grant 2021J01612 and 2020J01494, in part by the Major Science and Technology Project of Fujian Province under Grant 2021HZ022007, in part by the Fujian Collaborative Innovation Center for Big Data Application in Governments, and in part by the Fujian Engineering Research Center of Big Data Analysis and Processing.

Contribution Statement

This work was a collaborative effort by all contributing authors. Chaowei Huang and Yaru su made equal contributions to this study and are designated as co-first authors. Xiao Ke, as the corresponding author, is responsible for all communications related to this manuscript.

References

[Andoni *et al.*, 2015] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. *Advances in neural information processing systems*, 28, 2015.

[Berahmand *et al.*, 2024] Kamal Berahmand, Fatemeh Daneshfar, Elaheh Sadat Salehi, Yuefeng Li, and Yue Xu. Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review*, 57(2):28, 2024.

[Chakraborty *et al.*, 2024] Tanujit Chakraborty, Ujjwal Reddy KS, Shraddha M Naik, Madhurima Panja, and Bayapureddy Manvitha. Ten years of generative adversarial nets (gans): a survey of the state-of-the-art. *Machine Learning: Science and Technology*, 5(1):011001, 2024.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[Creswell *et al.*, 2018] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[Dai *et al.*, 2021] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3560–3569, 2021.

[Deng and Dragotti, 2020] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multimodal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3333–3348, 2020.

[Deng *et al.*, 2024] Yanglin Deng, Tianyang Xu, Chunyang Cheng, Xiao-Jun Wu, and Josef Kittler. Mmdrfuse: Distilled mini-model with dynamic refresh for multi-modality image fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7326–7335, 2024.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

[Jocher, 2020] Glenn Jocher. Yolov5, 2020. Accessed: 2024-10-8.

[Johnson and Becker, 2005] Keith A. Johnson and J. Alex Becker. Harvardmedweb, 2005. Accessed: 2024-10-8.

[Li *et al.*, 2018] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. In *2018 24th international conference on pattern recognition (ICPR)*, pages 2705–2710. IEEE, 2018.

[Li *et al.*, 2019] Qilei Li, Lu Lu, Zhen Li, Wei Wu, Zheng Liu, Gwanggil Jeon, and Xiaomin Yang. Coupled gan with relativistic discriminators for infrared and visible images fusion. *IEEE Sensors Journal*, 21(6):7458–7467, 2019.

[Li *et al.*, 2021] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.

[Li *et al.*, 2023a] 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*, 45(9):11040–11052, 2023.

[Li *et al.*, 2023b] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4471–4479, 2023.

[Liang *et al.*, 2022] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022.

[Liu *et al.*, 2018] Yu Liu, Xun Chen, Juan Cheng, Hu Peng, and Zengfu Wang. Infrared and visible image fusion with convolutional neural networks. *International Journal of Wavelets, Multiresolution and Information Processing*, 16(03):1850018, 2018.

[Liu *et al.*, 2022] 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.

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

[Liu *et al.*, 2023b] 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. *arXiv preprint arXiv:2305.06720*, 2023.

[Liu *et al.*, 2024a] 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 *IJCAI*, 2024.

[Liu *et al.*, 2024b] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and metainitialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

[Ma *et al.*, 2023] Weihong Ma, Kun Wang, Jiawei Li, Simon X Yang, Junfei Li, Lepeng Song, and Qifeng Li. Infrared and visible image fusion technology and application: A review. *Sensors*, 23(2):599, 2023.

[Michelucci, 2022] Umberto Michelucci. An introduction to autoencoders. *arXiv preprint arXiv:2201.03898*, 2022.

[Sun *et al.*, 2022] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022.

[Tang *et al.*, 2022] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[Tang *et al.*, 2024] Wei Tang, Fazhi He, and Yu Liu. Itfuse: An interactive transformer for infrared and visible image fusion. *Pattern Recognition*, 156:110822, 2024.

[Toet and Hogervorst, 2012] Alexander Toet and Maarten A Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):010901–010901, 2012.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2022] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515, 2022.

[Xu *et al.*, 2020] 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, 2020.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[Zhang and Ma, 2021] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021.

[Zhang *et al.*, 2020] Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, and Gang Xiao. Object fusion tracking based on visible and infrared images: A comprehensive review. *Information Fusion*, 63:166–187, 2020.

[Zhang *et al.*, 2021] 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.

[Zhao and Nie, 2021] Haibo Zhao and Rencan Nie. Dndt: Infrared and visible image fusion via densenet and dual-transformer. In *2021 International Conference on Information Technology and Biomedical Engineering (ICITBE)*, pages 71–75. IEEE, 2021.

[Zhao *et al.*, 2020] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[Zhao *et al.*, 2024] 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.