Probing Synergistic High-Order Interaction in Infrared and Visible Image Fusion

Naishan Zheng [1], Man Zhou [1], Jie Huang [1], Junming Hou [2], Haoying Li [3], Yuan Xu [4], Feng Zhao [1*]
[1]University of Science and Technology of China
[2]Southeast University, [3]Zhejiang University, [4]Nanyang Technology University
{nszheng,manman,hj0117}@mail.ustc.edu.cn,
junming_hou@seu.edu.cn, lhaoying@zju.edu.cn, xu.yuan@ntu.edu.sg, fzhao956@ustc.edu.cn

Abstract

Infrared and visible image fusion aims to generate a fused image by integrating and distinguishing complementary information from multiple sources. While the cross-attention mechanism with global spatial interactions appears promising, it only capture second-order spatial interactions, neglecting higher-order interactions in both spatial and channel dimensions. This limitation hampers the exploitation of synergies between multi-modalities. To bridge this gap, we introduce a Synergistic High-order Interaction Paradigm (SHIP), designed to systematically investigate the spatial fine-grained and global statistics collaborations between infrared and visible images across two fundamental dimensions: 1) ***Spatial dimension****: we construct spatial fine-grained interactions through element-wise multiplication, mathematically equivalent to global interactions, and then foster high-order formats by iteratively aggregating and evolving complementary information, enhancing both efficiency and flexibility; 2)* ***Channel dimension****: expanding on channel interactions with first-order statistics (mean), we devise high-order channel interactions to facilitate the discernment of inter-dependencies between source images based on global statistics. Harnessing high-order interactions significantly enhances our model's ability to exploit multi-modal synergies, leading to superior performance over state-of-the-art alternatives, as shown through comprehensive experiments across various benchmarks. Code is available at https://github.com/zheng980629/SHIP.*

1. Introduction

Infrared and visible image fusion strives to aggregate and discern complementary information from source images into fused images, enhancing their applicability in subsequent tasks [8, 10, 18, 32, 48, 50, 75]. Specifically, visible images are distinguished by their intricate texture details and alignment with human visual perception. In contrast, infrared images excel in capturing essential thermal radiation information, enabling the highlighting of significant targets like vehicles and pedestrians, especially in low-light environments. Consequently, there has been considerable attention on investigating synergistic correlations between various modalities to integrate complementary information.

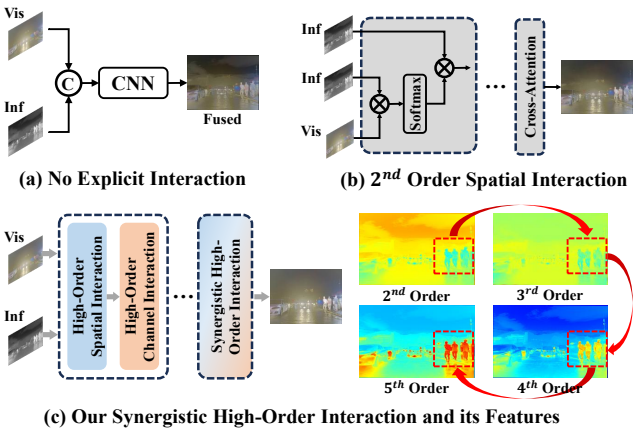

Figure 1. Comparison between previous fusion rules and our proposed paradigm. Previous works either (a) lack explicit interaction or (b) achieve only 2^{nd} order spatial interactions; and (c) our SHIP incorporates high-order spatial and channel interactions to explore synergistic correlations between modalities in spatial fine-grained details and global statistics, progressively integrating and distinguishing the complementary information.

Recently, the remarkable advancements in deep learning [19, 21, 31, 35, 47, 54, 79, 81, 86, 87] have sparked a revolution within this community. Some approaches typically begin by pre-training an encoder-decoder for feature extraction and image reconstruction. Subsequently, specific fusion rules such as concatenation [45], addition [51], weight summation [25], and maximum [88] are applied to capture correlations between source images for information

*Corresponding author.

fusion. Additionally, the design of loss functions based on image fusion properties, like saliency-based [39, 45, 62] and illumination-aware [56], provides crucial guidance for models to explore complementary information within input sources. In parallel, GAN-based methods [27, 30, 39, 42] compels the generator to investigate reciprocal information between modalities by conceptualizing image fusion as a game between the generator and discriminator. However, as shown in Fig. 1(a), these approaches do not explicitly establish synergistic correlations and inter-dependencies learning between infrared and visible images.

The emergence of Transformer [6, 20, 38], characterized by second-order spatial interactions [53] deviating from the dot-products among key, query, and value, challenges the dominance of CNNs. YDTR [59] introduces a dynamic Transformer module for acquiring local features and context information from different modalities. SwinFusion [46] develops self-attention and cross-attention mechanisms, modeling and integrating dependencies within both intra-domain and inter-domain features. PanFormer [84] designs a customized Transformer, incorporating two value terms from disparate modalities to effectively capture their collaborations. However, these transformer architectures limit their interactions to the second order in spatial dimensions, neglecting the untapped potential of high-order interactions in both spatial and channel dimensions. This limitation results in a restricted exploration of synergistic modality correlations, as illustrated in Fig. 1 (b).

Motivated by the above analysis, our objective is to model high-order interactions in spatial and channel dimensions to comprehensively explore synergies between infrared and visible modalities. Regrettably, employing a straightforward approach of cascading self-attention operations in two dimensions only captures multiple second-order interactions, simultaneously imposing an intolerable computational burden due to matrix multiplications. Therefore, *how to extend second-order interactions to arbitrary orders without introducing substantial computational overhead* is the key ingredient.

In this paper, we introduce a Synergistic High-order Interaction Paradigm (SHIP), offering an innovative approach to efficiently capture synergies in spatial fine-grained and global statistics among multiple modalities through high-order interactions. Specifically, it involves two dimensions: 1) **Spatial dimension:** we employ the frequency domain to establish spatial fine-grained correlations between infrared and visible representations through element-wise multiplication, a mathematically equivalent yet computationally efficient alternative to the costly matrix multiplication. Subsequently, we iteratively aggregate complementary information and evolve the synergistic correlations, thus fostering high-order spatial interactions and effectively excavating collaborations between modalities. 2) **Channel dimension:** building on the SE block's adaptive recalibration of feature responses using first-order statistics (mean) for first-order channel interaction, we elevate this concept into a high-order format. This extension enables the exploration of synergistic correlations grounded in global statistics from source images, providing a deeper insight into distinguishing intricate inter-dependencies among different modalities.

Our contributions are summarized as follows:

- The novel Synergistic High-order Interaction Paradigm (SHIP) in this study explores intricate high-order interactions in infrared and visible image fusion. By incorporating high-order interactions in both spatial and channel dimensions, SHIP stands as a pioneering approach, investigating synergistic correlations between modalities.
- This paradigm investigates high-order interactions involving spatial fine-grained and global statistics, collaboratively aggregating complementary information and distinguishing inter-dependencies from source modalities.
- Our experiments on multiple infrared-visible benchmarks show that the proposed framework outperforms state-of-the-art methods. Furthermore, we also demonstrate its effectiveness in the pan-sharpening task.

2. Related Work

2.1. Infrared and Visible Image Fusion

Infrared and visible image fusion aims to obtain a synthetic image effectively highlights salient objects from the source images while preserving visual quality. To align the synthesized fusion results more closely with human visual perception, various image processing techniques were introduced, including discrete wavelet [36], Laplacian pyramid [49], contourlet transform [74], sparse representation [37], low-rank representation [26], principal component analysis [9], and total variation [41]. However, these methods involved the formulation of fusion rules and intricate activity levels, constraining their applicability in complex scenarios.

Recently, explosive deep learning-based methods [17, 28, 29, 33, 34, 58, 62, 70, 78, 80, 83] have revolutionized image fusion, falling into three main categories: auto-encoder (AE)-based methods, CNN-based methods, and generative adversarial network (GAN)-based methods. AE-based methods [25, 57, 65, 66, 76] typically employ pre-trained auto-encoders for feature extraction and image reconstruction, emphasizing the design of network architectures and fusion strategies. CNN-based methods [40, 45, 45, 55, 56, 61] integrate these components within an end-to-end framework. Zhang *et al.* [72] developed intensity and gradient branches to preserve these essential properties of source images. In addition, researchers have developed various loss functions [40, 43, 45, 56] grounded in image fusion properties, providing substantial guidance during network training. For example, Ma et al. [45] designed a fusion

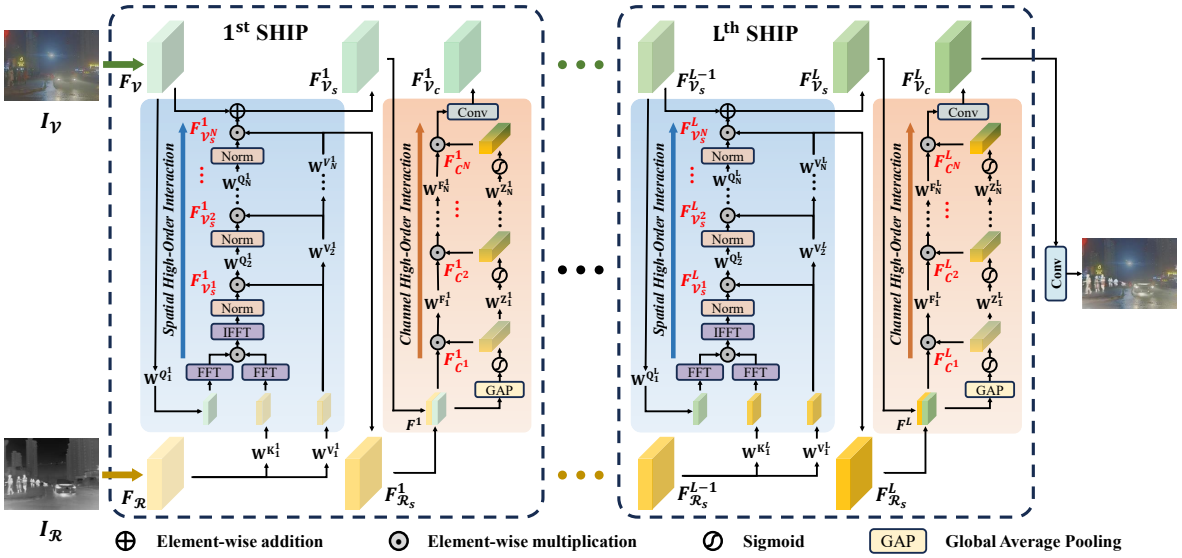

Figure 2. The detailed framework of the proposed Synergistic High-order Interaction Paradigm (SHIP) comprises alternating spatial and channel high-order interactions, executed over L iterations. Specifically, the spatial high-order interaction sufficiently excavates collaborations between two modalities and integrates spatial fine-grained complementary information through high-order modelling. Subsequently, the channel high-order interaction, rooted in the global first-order statistic (mean), further investigates global statistics, distinguishing interdependencies between visible and infrared modalities.

loss that utilizes a specialized target mask, allowing for selective fusion of target and background regions. However, due to the absence of authentic fused images for reference, researchers [27, 30, 42, 44, 82] attempted to introduce GAN into the learning paradigm. They utilized the discriminator to compel the generator to preserve more texture details highlight salient objects from source images.

2.2. High-Order Interaction Modeling

The vanilla convolution operation does not inherently capture the spatial interactions between a specific location and its neighboring region. A refined approach, known as dynamic convolution [2, 13, 22], introduces a first-order spatial interaction by generating dynamic weights adapted to the input. In Transformers [6], the self-attention mechanism facilitates a second-order spatial interaction via its key ingredient: the intrinsic matrix multiplication involving queries, keys, and values. Shifting to the channel dimension, the Squeeze-and-Excitation block [15, 63] utilizes the first-order statistic (mean) to recalibrate channel responses. Generally, these improvements only focus on capturing interactions in either the spatial or channel dimensions, rather than achieving high-order interactions in both dimensions.

3. Methods

3.1. Overview Framework

The proposed paradigm, illustrated in Fig. 2, operates as follows: given an infrared image, $I_{\mathcal{R}} \in \mathbb{R}^{H\times W\times 1}$, and a visible image, $I_{\mathcal{V}} \in \mathbb{R}^{H\times W\times 3}$, we extract corresponding shallow features using separate convolution layers for each modality, yielding $F_{\mathcal{R}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ and $F_{\mathcal{V}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$. Then, these modality-aware features undergo a series of core Synergistic High-order Interactions Paradigm (SHIP), incorporating both spatial and channel dimensions. This process explores synergies between the two modalities in spatial fine-grained details and global statistics. Finally, these features are projected back into the image space to generate the fused result, $I_{\mathcal{F}} \in \mathbb{R}^{H\times W\times 1}$. The fusion process specifically targets the Y channel in the $\mathrm{YC_bC_r}$ color space, following the approach of prior works [24, 57]. In summary, the paradigm can be formulated as follows:

$$I_{\mathcal{F}} = \mathrm{SHIP}_L(\psi(I_{\mathcal{R}}), \phi(I_{\mathcal{V}})), \tag{1}$$

where $\psi(\cdot)$ and $\phi(\cdot)$ denote feature extractors and L indicates the iteration number of our SHIP.

3.2. High-Order Spatial Interactions

Revisiting the Self-attention. The self-attention mechanism, the key ingredient of Transformer [6], fosters second-order spatial interactions through matrix multiplications among key, query, and value components. This process empowers the model to dynamically distinguish and aggregate complementary information, grounded in the query modality. For the infrared and visible image fusion, the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are derived by:

$$\mathbf{Q} = F_{\mathcal{V}}\mathbf{W}^{Q}, \quad \mathbf{K} = F_{\mathcal{R}}\mathbf{W}^{K}, \quad \mathbf{V} = F_{\mathcal{R}}\mathbf{W}^{V}, \tag{2}$$

where $\mathbf{W}^{Q}$, $\mathbf{W}^{K}$, and $\mathbf{W}^{V}$ indicate linear transformations applied to project modality-aware feature representations.

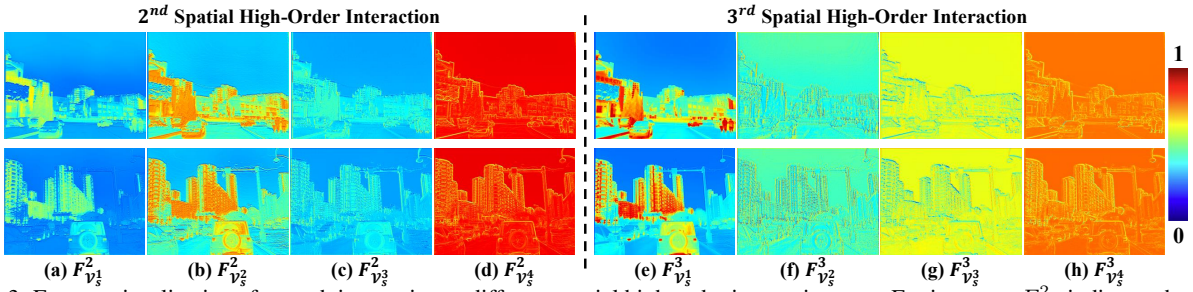

Figure 3. Feature visualization after each interaction at different spatial high-order interaction step. For instance, $F^2_{\mathcal{V}^3_s}$ indicates the feature after the third-order interaction in the 2^{nd} spatial high-order interaction. These visualizations illustrate the efficacy of high-order spatial interactions in two perspectives: (1) within each high-order interaction, feature responses escalate as the order increases, highlighting the salient objects; (2) distinct high-order interactions yield unique responses, showcasing the diversity in feature representations.

The self-attention mechanism, which captures second-order spatial interactions centered on the input $F_{\mathcal{V}}$, is realized through dot-product operations among these components:

$$\mathcal{O}_S((F_{\mathcal{V}})^2) = F^1_{\mathcal{V}_S} = \text{softmax}\left(\frac{\mathbf{Q}\otimes\mathbf{K}^T}{\sqrt{d_k}}\right)\otimes\mathbf{V} = \mathbf{A}\otimes\mathbf{V}, \tag{3}$$

where d_k represents the dimension of the key, $\otimes$ indicates the dot-product operation, $\mathbf{A} \in \mathbb{R}^{\text{HW}\times\text{HW}}$ is the correlation matrix, and $\mathcal{O}_S((F_{\mathcal{V}})^2)$ signifies the output of the self-attention module, capturing second-order spatial interactions about the input feature $F_{\mathcal{V}}$.

However, the dot product, despite its effectiveness, comes with significant computational costs, rendering it impractical for achieving high-order operations through cascading self-attention mechanisms.

Equivalently efficient form. Each element of $\mathbf{A}$ can be redefined by the inner product: $\mathbf{A}_{ij} = \langle \mathbf{q}_i, \mathbf{k}_j \rangle$, $\mathbf{q}_i \in \mathbf{Q}$, $\mathbf{k}_j \in \mathbf{K}$, and $\langle\cdot\rangle$ indicates the inner product. The convolution theorem establishes that the correlation or convolution of two signals in the spatial domain equals them Hadamard product between them in the frequency domain. To leverage this property, we incorporate the frequency domain into the self-attention mechanism, simplifying matrix multiplication to a lightweight element-wise operation. Initially, we transform the modality-aware features $F_{\mathcal{R}}$ and $F_{\mathcal{V}}$ into the frequency domain using fast Fourier transform (FFT). The correlation is computed as follows:

$$\mathbf{A} = \mathcal{F}^{-1}\left(\mathcal{F}\left(F_{\mathcal{V}}\mathbf{W}^Q\right)\odot\overline{\mathcal{F}\left(F_{\mathcal{R}}\mathbf{W}^K\right)}\right), \tag{4}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote FFT and inverse FFT, $\odot$ indicates the Hadamard product, and $\overline{\mathcal{F}(\cdot)}$ represents the conjugate transpose operation. Furthermore, the integrated features with second-order spatial interactions are obtained:

$$\mathcal{O}_S((F_{\mathcal{V}})^2) = F^1_{\mathcal{V}_S} = \text{Norm}(\mathbf{A})\odot(\text{F}_{\mathcal{R}}\mathbf{W}^{\text{V}}), \tag{5}$$

where Norm represents layer normalization applied to $\mathbf{A}$.

Delving into the High-Order Format. Recent methodologies, such as [3, 4, 73], have shown a strong preference for employing self-attention mechanisms. However, these approaches, often seen in cascading self-attention blocks, tend to generate multiple second-order interactions centered around the query feature, rather than achieving higher-order modeling. Formally, the recursive format of L cascaded self-attention can be expressed as:

$$\begin{aligned}&\mathcal{O}_S((F^{i-1}_{\mathcal{V}_S})^2) = F^i_{\mathcal{V}_S} = \text{Attention}\left(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\right),\\ &\mathbf{Q}_i = F^{i-1}_{\mathcal{V}_S}\mathbf{W}^{\mathbf{Q}_i}, \mathbf{K}_i = F_{\mathcal{R}}\mathbf{W}^{\mathbf{K}_i}, \mathbf{Q}_i = F_{\mathcal{R}}\mathbf{W}^{\mathbf{V}_i},\end{aligned} \tag{6}$$

where $1 \le i \le L$. It is apparent that this process only captures second-order interactions about the input feature $F_{\mathcal{V}_{i-1}}$ while inducing huge computational costs.

In contrast, standing on the equivalently efficient form, we progress beyond second-order interactions and extend our reach to arbitrary-order interactions (N-order) while maintaining efficiency. Concretely, for each i^{th} iteration, we extend Eq. 5 into the following high-order formulation:

$$\begin{aligned}&\mathcal{O}_S((F^{i-1}_{\mathcal{V}})^j) = F^i_{\mathcal{V}^j_s} = \text{Norm}(F^i_{\mathcal{V}^{j-1}_S}\mathbf{W}^{\mathbf{Q_j^i}})\odot(F^i_{\mathcal{R}^{j-1}_S}\mathbf{W}^{\mathbf{V_j^i}}),\\ &F^i_{\mathcal{V}^{j-1}_S} = \text{Norm}(F^i_{\mathcal{V}^{j-2}_S}\mathbf{W}^{\mathbf{Q_{j-1}^i}}), F^i_{\mathcal{R}^{j-1}_S} = F^i_{\mathcal{R}^{j-2}_S}\mathbf{W}^{\mathbf{V_{j-1}^i}},\end{aligned} \tag{7}$$

where $2 \le j \le N$. This formulation enables us to capture interactions up to the N-th order efficiently.

In general, for the traditional transformer chain with L, the sequence unfolds as follows:

$$\begin{aligned}&F_{\mathcal{V}} \to \mathcal{O}_S((F_{\mathcal{V}})^2) \to F^1_{\mathcal{V}_S} \to \mathcal{O}_S((F^1_{\mathcal{V}_S})^2) \to F^2_{\mathcal{V}_S}\ldots\\ &\to \mathcal{O}_S((F^i_{\mathcal{V}_S})^2) \to F^{i+1}_{\mathcal{V}_S}\ldots\mathcal{O}_S((F^{L-1}_{\mathcal{V}_S})^2) \to F^L_{\mathcal{V}_S}.\end{aligned} \tag{8}$$

In contrast, our high-order modeling replaces this with:

$$\begin{aligned}&F_{\mathcal{V}} \to \mathcal{O}_S((F_{\mathcal{V}})^N) \to F^1_{\mathcal{V}_S} \to \mathcal{O}_S((F^1_{\mathcal{V}_S})^N) \to F^2_{\mathcal{V}_S}\ldots\\ &\to \mathcal{O}_S((F^i_{\mathcal{V}_S})^N) \to F^{i+1}_{\mathcal{V}_S}\ldots\mathcal{O}_S((F^{L-1}_{\mathcal{V}_S})^N) \to F^L_{\mathcal{V}_S}.\end{aligned} \tag{9}$$

Indeed, this modification empowers us to capture interactions up to the N-th order within each iteration. As depicted in Fig. 3, diverse orders within each spatial high-order interaction integrate complementary information of varying

granularity. Moreover, interactions at different iterations exhibit discriminative responses, enriching the feature diversity throughout the iterative process.

3.3. High-Order Channel Interaction

Revisiting Squeeze and Excitation Block. The Squeeze-and-Excitation (SE) block [15] leverages the first-order global statistic, mean, to model channel interactions. This approach enables the SE block to explicitly capture inter-dependencies between input feature channels. For the infrared and visible image fusion, the SE block formulates the dependency between infrared and visible features from the i^{th} high-order spatial interaction as follows:

$$\begin{aligned} Z^i &= \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} F^i(x, y), \\ \mathcal{O}_C((F^i)^1) &= F_C^i = \sigma\left(\mathbf{W}^{\mathbf{Z}_1^i} Z^i\right) \cdot F^i, \end{aligned} \tag{10}$$

where $F^i = \text{concat}[F_{\mathcal{V}_S}^i, F_{\mathcal{R}_S}^i]$, Z_c represents the first-order statistic, and σ denotes the Sigmoid function. $\mathbf{W}^{\mathbf{Z}}$ includes two linear transformations and a ReLU function.

Delving into the High-Order Format. Similar to high-order spatial interactions, we extend the SE block to achieve high-order channel interactions:

$$\begin{aligned} &\mathcal{O}_C((F^i)^j) = F_{C^j}^i = \sigma\left(\mathbf{W}^{\mathbf{Z}_j^i} Z_{j-1}^i\right) \cdot (\mathbf{W}^{\mathbf{F}_j^i} F_{j-1}^i), \\ &Z_{j-1}^i = \sigma(\mathbf{W}^{\mathbf{F}_{j-1}^i} Z_{j-2}^i), F_{j-1}^i = \mathbf{W}^{\mathbf{F}_{j-1}^i} F_{j-2}^i. \end{aligned} \tag{11}$$

Finally, a convolution layer integrates F_C^i into the fused modality, yielding the integrated feature, $F_{\mathcal{V}_C}^i$.

Through N-order spatial and channel interactions conducted over L iterations, the interaction chain can be mathematically expressed as follows:

$$\begin{aligned} &F_{\mathcal{V}} \rightarrow \mathcal{O}_S((F_{\mathcal{V}})^N) \rightarrow \mathcal{O}_C((F_{\mathcal{V}_S}^1)^N) \rightarrow \mathcal{O}_S((F_{\mathcal{V}_C}^1)^N) \\ &\rightarrow \mathcal{O}_C((F_{\mathcal{V}_S}^2)^N) \rightarrow \ldots \mathcal{O}_S((F_{\mathcal{V}_C}^{L-1})^N) \rightarrow \mathcal{O}_C((F_{\mathcal{V}_S}^L)^N). \end{aligned} \tag{12}$$

We analyze channel responses at 2^{nd} channel high-order interaction along the channel dimension. Contrary to consistent responses across different orders, our high-order modeling adaptively distinguishes inter-dependencies between source modalities, as illustrated in Fig. 4.

3.4. Loss Functions

The loss function comprises the intensity and gradient terms: $\mathcal{L} = \mathcal{L}_{\text{int}} + \lambda \mathcal{L}_{\text{gra}}$. λ represents the trade-off parameter. To emphasize salient objects from visible and infrared images [30, 39], we introduce a saliency-based intensity loss, defined as follows:

$$\mathcal{L}_{\text{int}} = \left\| (\omega_{\mathcal{V}} \circ I_{\mathcal{V}} + \omega_{\mathcal{R}} \circ I_{\mathcal{R}}) - I_{\mathcal{F}} \right\|_1, \tag{13}$$

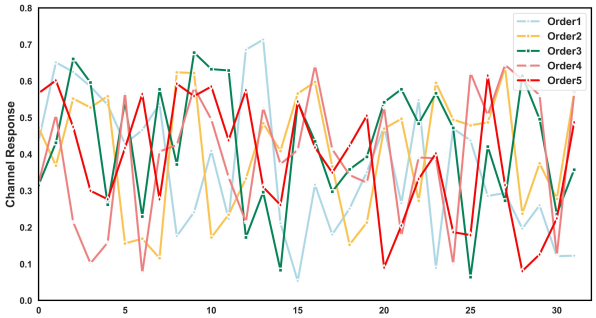

Figure 4. Channel interactions of different orders across channel indices. This observation serves as compelling evidence, suggesting that the interactions of varying orders explore diverse inter-dependencies between infrared and visible modalities.

where $\|\cdot\|_1$ denotes ℓ_1 norm. The weighted maps $\omega_{\mathcal{V}}$ and $\omega_{\mathcal{R}}$ are derived from the visible and infrared images as $\omega_{\mathcal{V}} = S_{\mathcal{V}} / (S_{\mathcal{V}} - S_{\mathcal{R}})$ and $S_{\mathcal{R}} = 1 - S_{\mathcal{V}}$, where S is the saliency matrix computed using the algorithm in [11].

To preserve crucial texture details from the source images in the fused results, we introduce a gradient loss:

$$\mathcal{L}_{\text{gra}} = \frac{1}{HW} \left\| \nabla I_{\mathcal{F}} - \max\left(\nabla I_{\mathcal{R}}, \nabla I_{\mathcal{V}}\right) \right\|_1, \tag{14}$$

where ∇ indicates the gradient operator used for texture information measurement within an image, and $max(\cdot)$ denotes the element-wise maximum operation.

4. Experiments

4.1. Experimental Settings

Datasets and metrics. To assess the effectiveness of our SHIP, we conduct comprehensive experiments on three publicly available datasets: M3FD [30], RoadScene [64], and TNO [60]. The M3FD dataset comprises 4200 paired infrared and visible images, with 3900 allocated for training and 300 for official testing. To further evaluate the generalization capability of our approach, we test our algorithm (trained on M3FD) on the RoadScene and TNO datasets. Since the latter two datasets lack a predefined split, we follow the configuration described in [72] and randomly select 25 pairs from each dataset for comparison.

A high-quality fused image should capture both salient objects and visual quality from the multi-modal images. To comprehensively measure the fusion results, we employ six metrics, including spatial frequency (SF) [7], mutual information (MI) [52], visual information fidelity (VIF) [14], average gradient (AG) [5], Q_{abf} [68], and feature mutual information (FMI) [12]. Moreover, higher values for these metrics indicate superior fusion performance.

Implementations. We implement our SHIP with PyTorch on a single NVIDIA GTX 3090 GPU. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ to update our model for $30K$, each with a batch size of 8. The initial

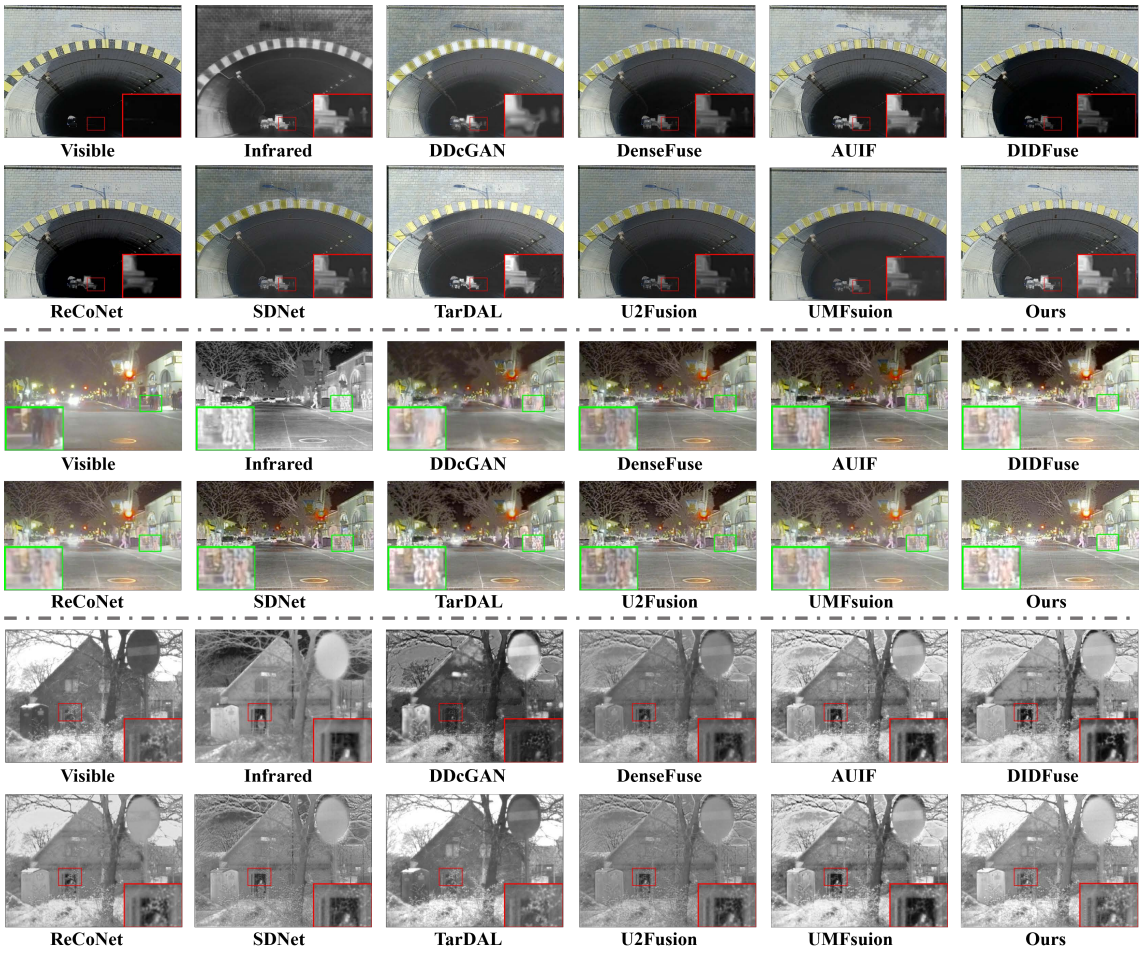

Figure 5. Qualitative results of different fusion methods on M3FD, RoadScene, and TNO datasets, respectively.

learning rate is set to 1×10^{-4} and decreases by a factor of 0.5 every $5K$ iterations. The patch size is set to 128×128.

4.2. Comparison with State-of-the-Arts

We compare our proposed SHIP with 9 state-of-the-art approaches: DDcGAN [44], DenseFuse [25], AUIF [77], DIDFuse [76], ReCoNet [16], SDNet [71], TarDAL [30], U2Fusion [64], and UMFusion [61] on three datasets.

Qualitative Comparisons. The qualitative results for three typical image pairs from various datasets are illustrated in Fig. 5. Compared to other existing methods, our proposed SHIP boasts two notable advantages. Firstly, our algorithm excels in preserving prominent objects from both infrared and visible images. As demonstrated in Fig. 5, pedestrians and tree branches in our method showcases high contrast, and distinctive contours, enhancing its suitability for visual observation (see red tangles of the first and third examples). Additionally, our results generate fusion outputs with intricate textures, aligning well with human visual perception. In contrast, visualization results show that DenseFuse, SDNet, and U2Fusion fail to effectively highlight discriminative targets, whereas DDcGAN and ReCoNet lack the ability to capture intricate textural details.

Quantitative Comparisons. Table 1 reveals the exceptional performance of our method across multiple metrics on the three datasets. The superior MI and FMI scores indicate our model's ability to effectively leverage information from both source images, showcasing its competence in transferring abundant information into the fused results. Furthermore, our approach's leading performance in SF, AG, and Q_{abf} signifies its remarkable capacity to integrate multi-modal complementary information and preserve intricate texture details. These achievements contribute to the preservation of fine-grained textures, ultimately resulting in visually appealing and detailed fused images. Moreover, the highest VIF also demonstrates that our fusion results have high-quality visual effects and small distortion, satisfying the human visual perception. These results collectively emphasize the robustness and generalization of our SHIP across various evaluation metrics, confirming its effectiveness in diverse scenarios and datasets.

Table 1. Quantitative comparison of our SHIP with 9 state-of-the-art methods on M3FD, RoadScne, and TNO datasets. The best and the second results are marked in bold and underlined.

Methods	M3FD dataset [30]					
	SF↑	MI↑	VIF↑	AG↑	Q_{abf}↑	FMI↑
DDcGAN [44]	0.059	2.540	0.768	**5.275**	0.480	0.836
DenseFuse [25]	0.037	2.930	0.762	3.303	0.503	0.863
AUIF [77]	0.046	3.049	0.819	3.922	0.503	0.845
DIDFuse [76]	0.055	3.048	0.877	4.871	0.494	0.831
ReCoNet [16]	0.041	3.049	0.818	3.929	0.486	0.845
SDNet [71]	0.053	3.231	0.678	4.730	0.528	0.846
TarDAL [30]	0.049	3.162	0.810	4.136	0.405	0.825
U2Fusion [64]	0.042	2.759	0.709	3.967	0.538	0.850
UMFusion [61]	0.034	3.087	0.709	2.928	0.397	0.855
Ours	**0.060**	**4.813**	**0.922**	5.177	**0.549**	**0.892**
Methods	**RoadScene dataset [64]**					
	SF↑	MI↑	VIF↑	AG↑	Q_{abf}↑	FMI↑
DDcGAN [44]	0.039	2.618	0.595	3.771	0.309	0.859
DenseFuse [25]	0.040	3.128	0.803	4.012	0.513	0.868
AUIF [77]	0.055	3.111	0.847	5.318	0.517	0.856
DIDFuse [76]	0.052	3.184	0.827	5.038	0.487	0.853
ReCoNet [16]	0.033	3.159	0.796	3.484	0.394	0.858
SDNet [71]	0.054	3.423	0.821	5.531	0.533	0.863
TarDAL [30]	0.050	3.464	0.787	4.389	0.448	0.852
U2Fusion [64]	0.047	2.811	0.740	4.885	0.526	0.861
UMFusion [61]	0.038	3.202	0.791	3.715	0.505	0.866
Ours	**0.058**	**3.914**	**0.905**	**5.589**	**0.550**	**0.873**
Methods	**TNO dataset [60]**					
	SF↑	MI↑	VIF↑	AG↑	Q_{abf}↑	FMI↑
DDcGAN [44]	0.048	1.847	0.674	4.831	0.349	0.858
DenseFuse [51]	0.037	2.402	0.800	3.634	0.423	**0.890**
AUIF [77]	0.051	2.271	0.815	4.712	0.426	0.879
DIDFuse [76]	0.049	2.442	0.829	4.636	0.405	0.863
ReCoNet [16]	0.029	2.426	0.827	3.323	0.357	0.878
SDNet [71]	0.050	2.186	0.762	4.844	0.421	0.883
TarDAL [30]	0.051	2.648	0.860	4.887	0.467	0.881
U2Fusion [64]	0.035	1.922	0.688	3.733	0.418	0.879
UMFusion [61]	0.036	2.247	0.717	3.166	0.398	0.888
Ours	**0.052**	**3.849**	**0.933**	**5.012**	**0.518**	**0.890**

4.3. Ablation Studies

We conduct ablation studies on the M3FD dataset to further investigate the effectiveness of our proposed SHIP under different the number of orders N and iterations L. For example, $L4N5$ denotes SHIP with 4 iterations and 5-order spatial and channel interactions.

Effect of the number of orders N: To investigate the impact of different orders of spatial and channel interactions, we conduct experiments on proposed SHIP with varying orders denoted as N. As illustrated in Fig. 6, the performance significantly increases as the number of stages increases until reaching 5. After this threshold, the performance stabilizes with slight improvements as N increases further. To strike a balance between performance and computational cost, we set $N = 5$ as the default order number. Furthermore, we present visualization results of our SHIP's feature responses with varied orders in Fig. 8. These visualizations demonstrate that the effectiveness of high-order interactions in unveiling synergistic correlations between modalities, thus highlighting salient objects.

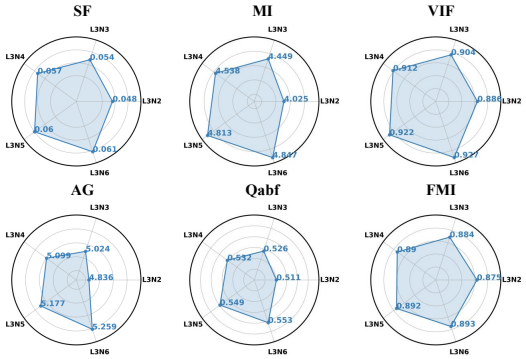

Figure 6. Ablation studies of the proposed SHIP with different numbers of order N on the M3FD dataset.

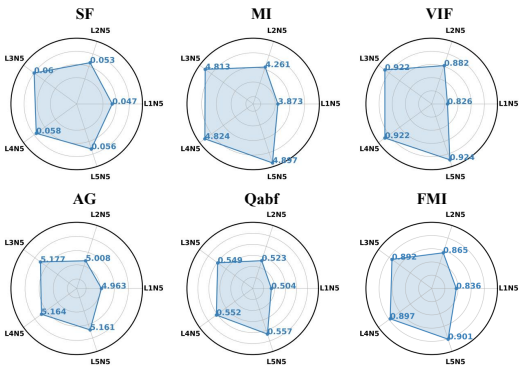

Figure 7. Ablation studies of the proposed SHIP with different numbers of block L on the M3FD dataset.

Effect of the number of iterations L: We conduct 5 experiment (L1N5, L2N5, L3N5, L4N5, and L5N5) to investigate the effect of the number of blocks on the results. From the observations in Fig. 7, it is evident that the model's performance improves considerably with an increase in the number of blocks. However, further increments in L result in a decreasing trend in SF and AG, possibly due to challenges in gradient propagation. Consequently, we employ $L = 3$ as the default block number in all experiments.

4.4. Extension on Pan-sharpening

To further demonstrate the effectiveness of our SHIP in the multi-modality image fusion task, we extend it on the panchromatic and multi-spectral image fusion task, named

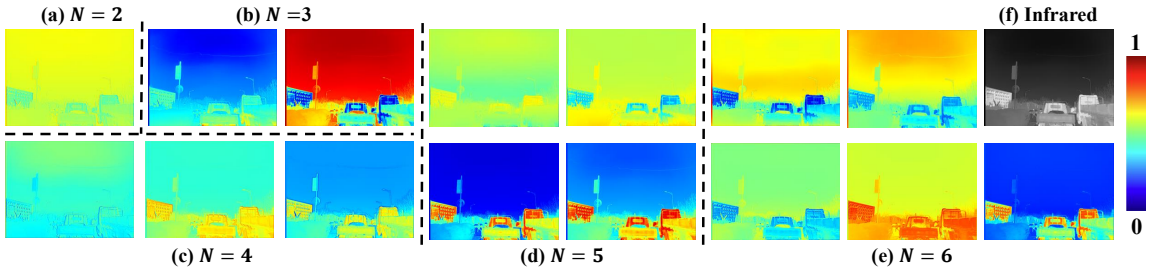

Figure 8. Visualization results of the proposed SHIP with different numbers of orders N on the M3FD dataset.

Table 2. Quantitative comparison of our SHIP with 6 state-of-the-art methods on WorldView II, GaoFen2, and WorldView III datasets.

Methods	WordView II				GaoFen2				WordView III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
GS [23]	35.638	0.918	0.042	1.877	37.226	0.903	0.031	1.674	22.561	0.547	0.122	8.243
PANNet [69]	40.818	0.963	0.026	1.056	43.066	0.969	0.018	0.858	29.684	0.907	0.085	3.426
SRPPNN [1]	41.454	0.968	0.023	0.990	47.200	0.988	0.011	0.559	30.435	0.920	0.077	3.155
GPPNN [67]	41.162	0.968	0.024	1.032	44.215	0.982	0.014	0.736	30.179	0.918	0.078	3.260
MutNet [85]	41.677	0.971	0.022	0.952	47.304	0.989	0.010	0.547	30.491	0.922	0.075	3.113
INNformer [84]	41.690	0.970	0.023	0.951	47.353	0.989	0.010	0.548	30.537	0.921	0.074	3.099
Ours	**41.736**	**0.973**	**0.022**	**0.948**	**47.458**	**0.990**	**0.010**	**0.537**	**30.615**	**0.925**	**0.074**	**3.056**

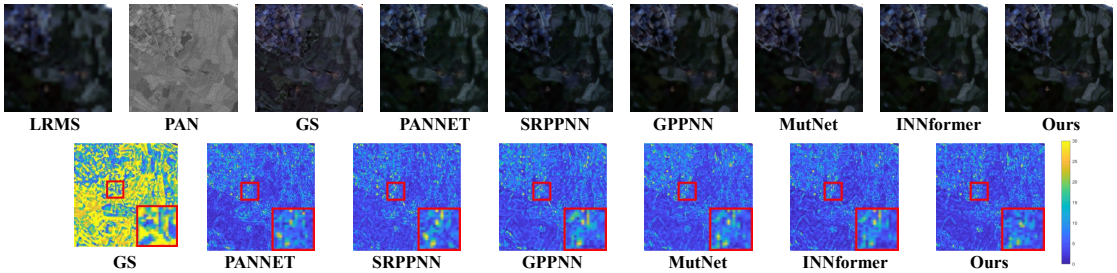

Figure 9. Visual comparison on WorldView-II dataset.

d Pan-sharpening. Following [84], we conduct extensive experiments on three widely used datasets: WorldView II, GaoFen2, and WorldView III datasets [84].

Quantitative Comparisons. Table 2 presents the evaluation metrics across three datasets, with the best and second-best values highlighted in bold and underline. Our proposed method consistently outperforms other competitive techniques across all satellite datasets. Specifically, our method surpasses 0.105 dB over the second-best INNformer on the GaoFen2 dataset. These consistent performances underscore the lower spectral distortion and superior preservation of spatial textures achieved by our SHIP.

Qualitative Comparisons. Fig. 9 visually demonstrates the comparison results, providing further confirmation of the effectiveness of our method. The last row displays the Mean Squared Error (MSE) residuals between the output pan-sharpened results and the ground truth. In comparison, our model exhibits minimal spatial and spectral distortions. The outstanding performance of our method underscores the effectiveness of the proposed synergistic high-order interaction mechanism, which integrates complementary information and enhance the visual quality of results.

5. Conclusion

In this paper, we pioneer the exploration of the Synergistic High-order Interaction paradigm (SHIP) to investigate collaborations between infrared and visible image modalities for image fusion. Our SHIP comprises both spatial and channel dimensions. The spatial high-order interaction progressively captures synergistic correlations between infrared and visible modalities, effectively integrating spatial fine-grained complementary information through high-order modeling. The channel high-order interaction, grounded global statistic, investigates and distinguishes the inter-dependencies between source modalities. Extensive experiments on multiple infrared and visible image fusion benchmarks have shown the superiority of our proposed synergistic high-order interaction paradigm.

Acknowledgments: This work was supported by the JKW Research Funds under Grant 20-163-14-LZ-001-004-01, and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

[1] Jiajun Cai and Bo Huang. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5206–5220, 2020. 8

[2] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020. 3

[3] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022. 4

[4] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. *arXiv preprint arXiv:2303.06373*, 2023. 4

[5] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015. 5

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[7] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on communications*, 43(12):2959–2965, 1995. 5

[8] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22752–22761, 2023. 1

[9] Zhizhong Fu, Xue Wang, Jin Xu, Ning Zhou, and Yufei Zhao. Infrared and visible images fusion based on rpca and nsct. *Infrared Physics & Technology*, 77:114–123, 2016. 2

[10] Hongbo Gao, Bo Cheng, Jianqiang Wang, Keqiang Li, Jianhui Zhao, and Deyi Li. Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics*, 14(9): 4224–4231, 2018. 1

[11] Sanjay Ghosh, Ruturaj G Gavaskar, and Kunal N Chaudhury. Saliency guided image detail enhancement. In *2019 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2019. 5

[12] Mohammad Bagher Akbari Haghighat, Ali Aghagolzadeh, and Hadi Seyedarabi. A non-reference image fusion metric based on mutual information of image features. *Computers & Electrical Engineering*, 37(5):744–756, 2011. 5

[13] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. Demystifying local vision transformer: Sparse connectivity, weight sharing, and dynamic weight. *arXiv preprint arXiv:2106.04263*, 2(3), 2021. 3

[14] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information fusion*, 14(2):127–135, 2013. 5

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3, 5

[16] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European Conference on Computer Vision*, pages 539–555. Springer, 2022. 6, 7

[17] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European Conference on Computer Vision*, pages 539–555. Springer, 2022. 2

[18] Deyi Ji, Haoran Wang, Hanzhe Hu, Weihao Gan, Wei Wu, and Junjie Yan. Context-aware graph convolution network for target re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1646–1654, 2021. 1

[19] Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16876–16885, 2022. 1

[20] Deyi Ji, Feng Zhao, and Hongtao Lu. Guided patch-grouping wavelet transformer with spatial congruence for ultra-high resolution segmentation. *International Joint Conference on Artificial Intelligence*, 2023. 2

[21] Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23621–23630, 2023. 1

[22] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 3

[23] Craig A Laben and Bernard V Brower. Process for enhancing the spatial resolution of multispectral imagery using pansharpening, 2000. US Patent 6,011,875. 8

[24] Zhuliang Le, Jun Huang, Han Xu, Fan Fan, Yong Ma, Xiaoguang Mei, and Jiayi Ma. Uifgan: An unsupervised continual-learning generative adversarial network for unified image fusion. *Information Fusion*, 88:305–318, 2022. 3

[25] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018. 1, 2, 6, 7

[26] Hui Li, Xiao-Jun Wu, and Josef Kittler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29:4733–4746, 2020. 2

[27] Jing Li, Hongtao Huo, Chang Li, Renhua Wang, and Qi Feng. Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Transactions on Multimedia*, 23:1383–1396, 2020. 2, 3

[28] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. 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, 2021. 2

[29] Jinyuan Liu, Yuhui Wu, Zhanbo Huang, Risheng Liu, and Xin Fan. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Processing Letters*, 28:1818–1822, 2021. 2

[30] 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. 2, 3, 5, 6, 7

[31] 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 1–28, 2023. 1

[32] 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

[33] 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. 2

[34] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1600–1608, 2021. 2

[35] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10561–10570, 2021. 1

[36] Yipeng Liu, Jing Jin, Qiang Wang, Yi Shen, and Xiaoqiu Dong. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. *Signal Processing*, 97:9–30, 2014. 2

[37] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. Image fusion with convolutional sparse representation. *IEEE signal processing letters*, 23(12):1882–1886, 2016. 2

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[39] 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. 2, 5

[40] Yongzhi Long, Haitao Jia, Yida Zhong, Yadong Jiang, and Yuming Jia. Rxdnfuse: A aggregated residual dense network for infrared and visible image fusion. *Information Fusion*, 69:128–141, 2021. 2

[41] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016. 2

[42] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48:11–26, 2019. 2, 3

[43] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020. 2

[44] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 3, 6, 7

[45] Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao. Stdfusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. 1, 2

[46] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 2

[47] 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

[48] Long Ma, Dian Jin, Nan An, Jinyuan Liu, Xin Fan, Zhongxuan Luo, and Risheng Liu. Bilevel fast scene adaptation for low-light image enhancement. *International Journal of Computer Vision*, pages 1–19, 2023. 1

[49] Run Mao, Xian Song Fu, Ping-juan Niu, Hui Quan Wang, Jie Pan, Shu Shu Li, and Lei Liu. Multi-directional laplacian pyramid image fusion algorithm. In *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pages 568–572. IEEE, 2018. 2

[50] Nirmala Paramanandham and Kishore Rajendiran. Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications. *Infrared Physics & Technology*, 88:13–22, 2018. 1

[51] K. R. Prabhakar, V. S. Srikar, and R. V. Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *IEEE International Conference on Computer Vision*, pages 4724–4732, 2017. 1, 7

[52] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):1, 2002. 5

[53] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions.

Advances in Neural Information Processing Systems, 35: 10353–10366, 2022. 2

[54] 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. 1

[55] 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. 2

[56] 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. 2

[57] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023. 2, 3

[58] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, page 101870, 2023. 2

[59] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 2022. 2

[60] Alexander Toet. The tno multiband image data collection. *Data in brief*, 15:249–251, 2017. 5, 7

[61] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022. 2, 6, 7

[62] Di Wang, Jinyuan Liu, Risheng Liu, and Xin Fan. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98: 101828, 2023. 2

[63] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[64] 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. 5, 6, 7

[65] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7:824–836, 2021. 2

[66] Meilong Xu, Linfeng Tang, Hao Zhang, and Jiayi Ma. Infrared and visible image fusion via parallel scene and texture learning. *Pattern Recognition*, 132:108929, 2022. 2

[67] 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. 8

[68] Costas S Xydeas, Vladimir Petrovic, et al. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000. 5

[69] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017. 8

[70] Wei Yu, Qi Zhu, Naishan Zheng, Jie Huang, Man Zhou, and Feng Zhao. Learning non-uniform-sampling for ultra-high-definition image enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1412–1421, 2023. 2

[71] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129:2761–2785, 2021. 6, 7

[72] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *AAAI Conference on Artificial Intelligence*, pages 12797–12804, 2020. 2, 5

[73] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022. 4

[74] Qiong Zhang and Xavier Maldague. An adaptive fusion approach for infrared and visible images based on nsct and compressed sensing. *Infrared Physics & Technology*, 74:11–20, 2016. 2

[75] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2153–2162, 2023. 1

[76] 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. 2, 6, 7

[77] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. 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, 2021. 6, 7

[78] Naishan Zheng, Jie Huang, Feng Zhao, Xueyang Fu, and Feng Wu. Unsupervised underexposed image enhancement via self-illuminated and perceptual guidance. *IEEE Transactions on Multimedia*, 2022. 2

[79] Naishan Zheng, Jie Huang, Qi Zhu, Man Zhou, Feng Zhao, and Zheng-Jun Zha. Enhancement by your aesthetic: An intelligible unsupervised personalized enhancer for low-light images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6521–6529, 2022. 1

[80] Naishan Zheng, Jie Huang, Man Zhou, Zizheng Yang, Qi Zhu, and Feng Zhao. Learning semantic degradation-aware guidance for recognition-driven unsupervised low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3678–3686, 2023. 2

[81] Naishan Zheng, Man Zhou, Yanmeng Dong, Xiangyu Rui, Jie Huang, Chongyi Li, and Feng Zhao. Empowering low-light image enhancer through customized learnable priors.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12559–12569, 2023. 1

[82] Huabing Zhou, Wei Wu, Yanduo Zhang, Jiayi Ma, and Haibin Ling. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. *IEEE Transactions on Multimedia*, 2021. 3

[83] 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

[84] Man Zhou, Jie Huang, Yanchi Fang, Xueyang Fu, and Aiping Liu. Pan-sharpening with customized transformer and invertible neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3553–3561, 2022. 2, 8

[85] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. Mutual information-driven pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1798–1808, 2022. 8

[86] Qi Zhu, Man Zhou, Naishan Zheng, Chongyi Li, Jie Huang, and Feng Zhao. Exploring temporal frequency spectrum in deep video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12428–12437, 2023. 1

[87] Qi Zhu, Jie Huang, Naishan Zheng, Hongzhi Gao, Chongyi Li, Yuan Xu, Feng Zhao, et al. Fouridown: Factoring downsampling into shuffling and superposing. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[88] Zhiqin Zhu, Hongpeng Yin, Yi Chai, Yanxia Li, and Guanqiu Qi. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Information Sciences*, 432:516–529, 2018. 1