FHFusion: Frequency Heterogeneity Network for Infrared and Visible Image Fusion

Zhou Fu[1], Min Li[1,2,3,*], Chen Chen[4], Xiaoyi Lv[4], and Cheng Chen[4]
[1]Xinjiang University, School of Computer Science and Technology, Urumqi, China
[2]The First People's Hospital of Kashi Prefecture, Kashgar 844099, China
[3]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[4]Xinjiang University, College of Software, Urumqi, China
1815407326@qq.com, MinLi927@163.com, 1343432873@qq.com, xjuwawj01@163.com, chenchenoptics@gmail.com
*Corresponding Author: Min Li Email: MinLi927@163.com

***Abstract*—Infrared and visible image fusion aims to integrate complementary information from multiple modalities. Recently, frequency-domain-based fusion methods have gained attention due to their advantages in edge texture recognition and global information integration. However, existing frequency-domain fusion methods fail to fully utilize the rich information embedded within frequency representations and the cross-modal complementarity among different spectral components. Therefore, we propose FHFusion. Specifically, our approach leverages the semantic and stylistic heterogeneity inherent in the frequency domain of infrared and visible images. Through spectral decoupling mechanisms, we achieve differentiated information fusion that effectively exploits cross-modal frequency domain complementary information. Further, we design an Amplitude Selection Module (ASM) and a Phase Discrimination Module (PDM) adaptively fusing energy distribution differences in the amplitude spectrum and structural complementary information in the phase spectrum, respectively. Experiments on the M3FD, LLVIP, and TNO datasets show that FHFusion achieves competitive or superior performance against representative state-of-the-art methods in both subjective and objective evaluations.**

***Index Terms*—Infrared and visible image fusion, frequency-domain heterogeneity, adaptive frequency-domain fusion**

I. Introduction

Infrared and visible image fusion integrates the rich texture details of visible images with the thermal target detection capability of infrared images, and has been widely applied in military surveillance and target detection [1]. With the development of deep learning, image fusion techniques have achieved significant progress. In particular, frequency-domain-based fusion methods have attracted increasing attention due to their ability to capture edge texture features and global structural information through spectral decomposition [2], [3]. Existing approaches can be broadly categorized into traditional frequency-domain transform-based methods and deep learning-based frequency-domain fusion methods. Traditional methods based on wavelet or Fourier transforms are limited by manually designed features and cannot fully exploit frequency-domain information [4], [5]. Although deep learning-based frequency-domain fusion methods improve fusion performance, their modeling of frequency-domain heterogeneity remains insufficient.

Existing methods exhibit clear limitations. SHIP leverages frequency-domain operations to enhance global interaction modeling, but its frequency-domain processing mainly serves interaction computation without explicitly distinguishing the roles of different spectral components [6]. SFMFusion improves global modeling via spatial–frequency enhancement but relies primarily on holistic spectral enhancement [7]. FISCNet integrates phase information in the frequency domain via direct addition, which fails to fully exploit the structural representation potential of phase spectra [8].

To address these issues, we propose FHFusion, which reframes infrared and visible image fusion by explicitly considering frequency-domain heterogeneity as a core modeling factor. By explicitly modeling the differentiated roles of spectral components in semantic expression and structural representation, FHFusion constructs a spectral decoupling-based fusion mechanism that enables targeted and role-aware cross-modal information integration. The main contributions are summarized as follows:

- We systematically investigate semantic and stylistic heterogeneity in the frequency-domain of infrared and visible images, and propose a spectral decoupling fusion strategy to enable differentiated exploitation of cross-modal spectral complementarity.
- An Amplitude Selection Module (ASM) and a Phase Discrimination Module (PDM) are designed to adaptively filter energy distribution information and enhance structural and semantic preservation, respectively.
- Experimental results on multiple public benchmark datasets demonstrate that FHFusion significantly outperforms representative state-of-the-art methods in both subjective and objective evaluations.

II. Related Work

Existing frequency-domain-based infrared and visible image fusion methods suffer from limited capability in exploiting frequency-domain information, and fail to fully utilize the rich representations in the frequency domain as well as the complementary characteristics between modalities. Traditional frequency-domain fusion approaches rely on fixed transform kernels and predefined fusion rules, which hinder adaptive

processing and ignore cross-modal frequency correlations between infrared and visible images, thereby limiting the effective exploitation of frequency-domain information [9], [10]. Although deep learning-based frequency-domain fusion approaches have achieved notable improvements in feature extraction and fusion performance, they still suffer from notable limitations [11], [12]. In SHIP, frequency-domain operations are primarily introduced to facilitate global interaction modeling, with the main focus remaining on feature-domain interaction rather than explicit differentiation of spectral components [6]. FISCNet integrates phase information in the frequency domain via direct addition, which fails to fully exploit the structural representation potential of phase spectra [8]. In addition, SFMFusion models frequency-domain information mainly through holistic spectral enhancement, which limits its ability to capture fine-grained and component-specific spectral characteristics [7]. Overall, these approaches lack sufficient understanding of the frequency-domain heterogeneity inherent in infrared and visible images, resulting in insufficient exploitation of spectral complementary information and potentially causing structural distortions in fused results. To address these limitations, we propose FHFusion, which explicitly models frequency-domain semantic and stylistic heterogeneity through differentiated modeling and complementary fusion of amplitude and phase spectra, enabling more faithful cross-modal structural and semantic information integration.

III. Method

To address the aforementioned limitations, this study proposes an FHFusion framework for infrared and visible image fusion. The proposed approach takes as inputs an infrared image $I_r \in \mathbb{R}^{1\times H\times W}$ and a visible image $I_v \in \mathbb{R}^{3\times H\times W}$. The visible image is first converted from the RGB color space to the YCbCr color space, and the luminance channel I_v^Y is extracted as the input image. As shown in Fig. 1, the feature maps of the visible and infrared images are processed through the Spatial-Frequency Domain Hierarchical Architecture (SDHA) and the Spatial Interaction Module (SIM), which is adapted from the work of Zheng *et al.* [8], for three iterations ($N = 3$), thereby enabling deep feature integration. Additionally, as illustrated in Fig. 2(a), the fused representation is progressively refined across three iterations with respect to different types of information, leading to a more stable fusion result. During the image reconstruction stage, the fused features are mapped back to the image space via convolutional layers, producing the fused luminance channel I_f^Y. Finally, I_f^Y is combined with the original chrominance channels (Cb and Cr) of the visible image to generate the final fused image I_f. Moreover, Fig. 2(b) and Fig. 2(c) further indicate that, through deeper iterative frequency-integration within SDHA, the salient targets in infrared images and the textural details in visible images are gradually emphasized.

A. Spatial-Frequency Domain Hierarchical Architecture

The SDHA is designed to achieve refined representation of frequency-domain information. Initially, the infrared and visible images are processed through independent convolutional layers to obtain feature representations $F_r \in \mathbb{R}^{16\times H\times W}$ and $F_v \in \mathbb{R}^{16\times H\times W}$, respectively. Subsequently, the Fast Fourier Transform (FFT) is applied to map these features into the frequency domain [20]:

$$A_r, P_r = \mathcal{F}(F_r), \qquad A_v, P_v = \mathcal{F}(F_v), \tag{1}$$

where $\mathcal{F}$ denotes the Fourier transform operation, and A and P represent the amplitude and phase spectra of the source images, respectively. Subsequently, to address frequency-domain heterogeneity between infrared and visible images, different spectral components are decoupled. Through the targeted design of discriminative spectral information processing mechanisms, refined frequency-domain representation is achieved. Specifically, amplitude-related energy information and phase-related structural information are processed separately:

$$A_{\text{fused}} = M_A(A_r, A_v), \qquad P_{\text{fused}} = M_P(P_r, P_v), \tag{2}$$

where M_A and M_P represent the ASM module and PDM module, respectively. Next, the complex signal is reconstructed from the fused amplitude and phase spectra using the standard amplitude–phase formulation:

$$\tilde{F}_f = A_{\text{fused}} \cdot \big(\cos(P_{\text{fused}}) + j\sin(P_{\text{fused}})\big), \tag{3}$$

where $\tilde{F}_f$ denotes the generated fused image features, thereby ensuring the consistency and integrity of the amplitude and phase information during frequency-domain fusion [20].Finally, the inverse Fast Fourier Transform (IFFT) is applied to $\tilde{F}_f$ to transform it back to the spatial domain [21].

B. Amplitude Selection Module

Considering that thermal targets in infrared images exhibit localized energy concentration, whereas textural information in visible images demonstrates a more spatially distributed energy pattern, the Amplitude Selection Module (ASM) aims to achieve optimized amplitude fusion through a frequency-adaptive selection mechanism rather than simple convolutional aggregation.

Initially, modality-specific amplitude features are projected into a unified embedding space via independent 1×1 convolutional encoders and then concatenated along the channel dimension:

$$A_{\text{concat}} = \text{Concat}\big(C_1(A_r),\, C_2(A_v)\big), \tag{4}$$

where A_r and A_v denote the amplitude spectra of infrared and visible modalities, respectively, $C_1(\cdot)$ and $C_2(\cdot)$ represent independent 1×1 convolutional encoding operations, and A_{concat} is the resulting multimodal amplitude representation constructed via channel-wise concatenation.

Subsequently, a cascaded convolutional transformation is employed to extract complex energy distribution patterns and inter-channel correlations within the amplitude spectrum:

$$F_{\text{amp}} = C_4\big(\text{LeakyReLU}(C_3(A_{\text{concat}}))\big), \tag{5}$$

where $C_3(\cdot)$ and $C_4(\cdot)$ are sequential 1×1 convolutional layers with independent learnable parameters. This hierarchical transformation enables the network to capture both local energy aggregation characteristics and cross-modal interaction cues in the frequency domain.

To further alleviate the limitation of insufficient global frequency information modeling in existing methods, a global context-aware attention weight generation mechanism is introduced:

$$\alpha_{\text{amp}} = \sigma\big(C_6(\text{ReLU}(C_5(\text{FGAP}(F_{\text{amp}}))))\big), \tag{6}$$

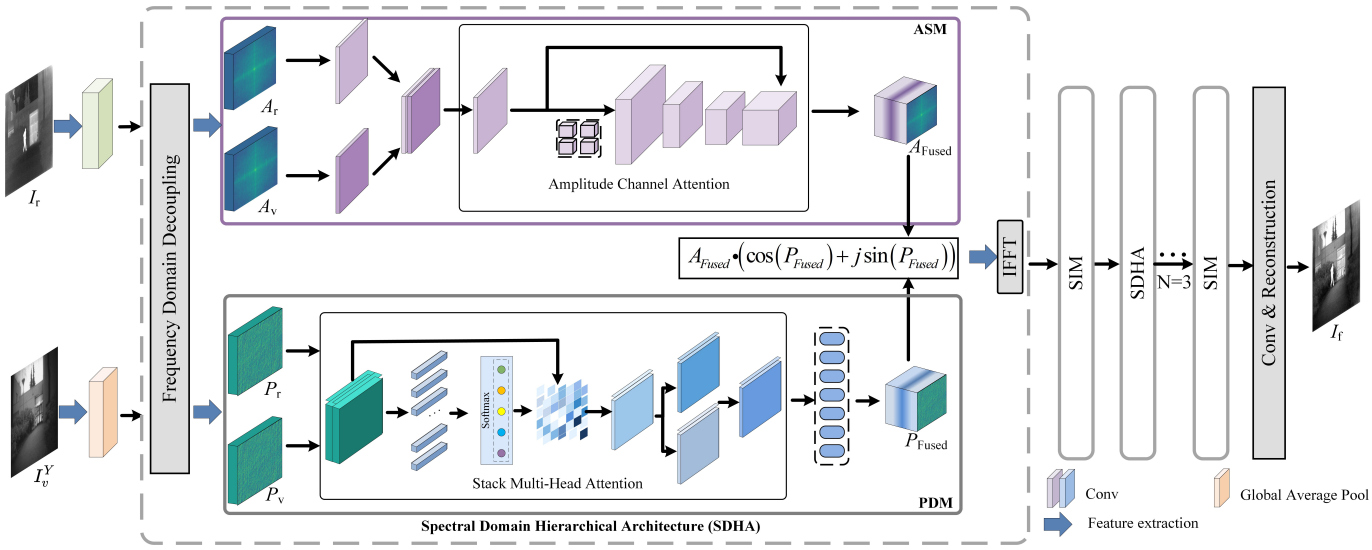

Fig. 1. Overall architecture of the FHFusion network. The SDHA and SIM [8] perform $N = 3$ iterations to achieve deep interaction between frequency and spatial domains.

TABLE I
QUANTITATIVE COMPARISONS OF FHFUSION WITH SEVEN REPRESENTATIVE METHODS ACROSS THE M3FD DATASET [13], TNO DATASET [14] AND LLVIP DATASET [15]. THE BEST RESULTS ARE SHOWN IN BOLD.

Method	M3FD Dataset						TNO Dataset						LLVIP Dataset					
	EN↑	SD↑	SF↑	AG↑	MI↑	VIFF↑	EN↑	SD↑	SF↑	AG↑	MI↑	VIFF↑	EN↑	SD↑	SF↑	AG↑	MI↑	VIFF↑
CDDFuse [16]	6.956	38.439	15.353	5.031	2.811	0.760	6.936	44.291	12.092	4.424	2.066	0.742	6.834	40.613	8.536	2.091	**2.921**	0.908
TarDAL [17]	7.085	22.489	14.847	1.327	2.394	0.830	6.913	35.015	8.510	3.765	1.818	0.686	6.147	27.623	4.517	1.331	1.877	0.676
SFMFusion [7]	6.777	34.103	12.053	3.569	2.588	0.883	6.586	37.704	9.990	3.617	1.794	0.435	7.549	54.256	12.260	2.865	2.449	0.910
DCEvo [18]	6.937	37.070	15.041	4.901	**3.215**	0.784	6.756	37.164	10.474	3.835	2.331	0.770	6.897	40.582	9.049	2.290	2.685	0.911
PSFusion [19]	7.383	49.313	21.831	7.234	1.884	0.784	**7.115**	46.153	13.806	5.416	1.581	0.709	7.292	62.640	13.441	3.551	2.189	**1.124**
FISCNet [8]	6.996	38.508	17.731	6.022	2.749	0.777	6.783	39.521	12.317	4.763	2.132	0.687	7.538	51.544	14.748	4.107	2.145	0.867
SHIP [6]	6.626	30.516	12.176	3.784	3.161	**0.924**	6.749	37.577	11.712	4.457	**2.619**	0.720	7.427	49.136	13.203	3.594	2.634	0.877
Ours	**7.446**	**63.040**	**25.513**	**8.324**	3.142	0.898	6.959	**56.691**	**15.146**	**5.654**	2.332	**0.808**	**7.553**	**67.348**	**19.157**	**5.062**	2.698	0.912

where FGAP denotes Global Average Pooling, $C_5(\cdot)$ and $C_6(\cdot)$ are 1×1 convolutional layers used for channel-wise dependency modeling, and $\sigma(\cdot)$ is the Sigmoid activation function.

Finally, the learned attention weights are applied to perform adaptive re-weighting of the amplitude features:

$$A_{\text{fused}} = F_{\text{amp}} \odot \alpha_{\text{amp}}, \tag{7}$$

where $\odot$ denotes element-wise multiplication. The attention coefficients α_{amp} dynamically emphasize informative frequency components while suppressing less relevant ones, thereby achieving effective and discriminative integration of multimodal amplitude information.

C. Phase Discrimination Module

Considering the heterogeneous structural semantic characteristics present in the phase spectra of source images, the Phase Discrimination Module (PDM) introduces a channel-wise adaptive aggregation mechanism to selectively capture and integrate key phase-related structural information.

Initially, the phase spectra of infrared and visible modalities are concatenated along the channel dimension to form a multimodal phase representation:

$$P_{\text{concat}} = \text{Concat}(P_r, P_v), \tag{8}$$

where P_r and P_v denote the phase spectra of infrared and visible images, respectively, and P_{concat} represents the channel-wise concatenated phase feature map.

Subsequently, global phase context information is extracted to facilitate adaptive discrimination:

$$Z = \text{ReLU}\Big(C_7\big(\text{FGAP}(\sum P_{\text{concat}})\big)\Big), \tag{9}$$

where FGAP denotes Global Average Pooling, $C_7(\cdot)$ is a 1×1 convolutional layer for global phase feature embedding, and Z represents the activated global phase descriptor.

To enable modality-specific phase structure discrimination, a multi-perspective feature representation strategy is adopted, in which independent attention heads are assigned to each modality:

$$A_i = C_i^{\text{attn}}(Z), \qquad i \in \{1, 2\}, \tag{10}$$

where $C_i^{\text{attn}}(\cdot)$ denotes an independent 1×1 convolutional attention mapping for the i-th modality. Specifically, C_1^{attn} is designed to emphasize salient edge and contour structures in the infrared phase spectrum, while C_2^{attn} focuses on prominent textural patterns in the visible phase spectrum.

The modality-specific attention responses are then concatenated and normalized using a Softmax operation to obtain adaptive fusion weights:

$$\alpha_{\text{pha}} = \text{Softmax}\big(\text{Concat}(A_1, A_2)\big), \tag{11}$$

where $\alpha_{\text{pha}} = \{\alpha_{\text{pha}}^1, \alpha_{\text{pha}}^2\}$ denotes the normalized attention coefficients corresponding to different modalities.

Finally, the phase features are fused in a weighted manner based on the learned attention coefficients:

$$P_{\text{fused}} = C_8\Big(\sum_{i=1}^{2} \alpha_{\text{pha}}^{i} \odot P_i\Big), \tag{12}$$

where $\odot$ denotes element-wise multiplication, α_{pha}^{i} represents the attention weight associated with the i-th modality, and $C_8(\cdot)$ is a post-processing module composed of two convolutional layers. This module further refines the fused phase representation while preserving the structural integrity of phase information. By assigning independent attention heads to different modalities, the PDM effectively alleviates mutual interference arising from shared modeling and enables precise discrimination and integration of heterogeneous phase structures.

D. Loss Function

FHFusion employs a multi-objective loss function to guide the training process, ensuring that the fused images simultaneously preserve the saliency of infrared targets while maintaining the rich textural information of visible images. The total loss function is formulated as a weighted combination of three complementary components [22]:

$$\mathcal{L} = \lambda_p \mathcal{L}_{\text{pixel}} + \lambda_m \mathcal{L}_{\text{max}} + \lambda_g \mathcal{L}_{\text{gra}} \tag{13}$$

where the weight coefficients are set as $\lambda_p = 3$, $\lambda_m = 10$, and $\lambda_g = 10$, and $\mathcal{L}_{\text{pixel}}$, $\mathcal{L}_{\text{max}}$ and $\mathcal{L}_{\text{gra}}$ denote the pixel-level fusion loss, maximum-intensity fusion loss, and gradient-based loss, respectively.

The pixel-level fusion loss is constructed based on a region-adaptive weight map to enforce local consistency between the fused image and the source images at the pixel level [23]:

$$\mathcal{L}_{\text{pixel}} = \|\omega_v \circ I_v + \omega_r \circ I_r - I_f\|_1 \tag{14}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm, ω_v and ω_r are the weight maps corresponding to the visible and infrared source images, respectively. The weights are computed as $\omega_v = S_v/(S_v+S_r)$ and $\omega_r = 1-\omega_v$, where S_v and S_r denote the saliency maps of the visible and infrared images.

The maximum-intensity fusion loss is designed to preserve salient infrared targets by encouraging the fused image to retain the strongest response from the source images [24]:

$$\mathcal{L}_{\text{max}} = \frac{1}{HW}\|I_f - \max(I_r, I_v)\|_1 \tag{15}$$

The gradient loss focuses on preserving fine-grained details and textural structures in the fused image and is defined as [22]:

$$\mathcal{L}_{\text{gra}} = \frac{1}{HW}\|\nabla I_f - \max(\nabla I_r, \nabla I_v)\|_1 \tag{16}$$

where ∇ denotes the gradient operator implemented using the Sobel operator. This loss term enhances edge and texture preservation, ensuring that fine structural details from the visible image are effectively retained in the fused result.

IV. Experimental Configurations

For training, we randomly select 500 image pairs from the M3FD dataset and generate 24,090 training samples of 128×128 resolution through random cropping augmentation. Additionally, we construct the test set using 200 image pairs from the M3FD dataset and 25 image pairs from the TNO dataset. For the LLVIP dataset, we directly use the paired test samples for quantitative and qualitative evaluation.We employ the Adam optimizer with hyperparameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model is trained for 20,000 iterations with a batch size of 16 and an initial learning rate of 1×10^{-4}. We adopt a step-wise decay schedule where the learning rate is reduced by 50% every 5,000 iterations.

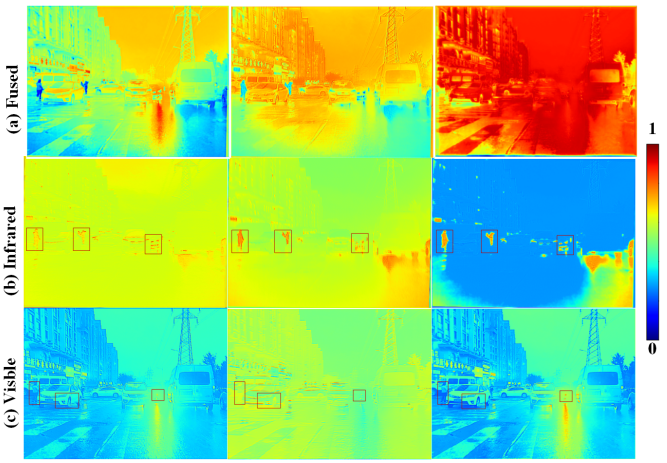

Fig. 2. Visualization of feature representations for infrared and visible images through three iterations of the SDHA.

A. Quantitative Comparison

We quantitatively evaluate FHFusion on the M3FD, TNO, and LLVIP dataset, with results reported in Table I. On the M3FD dataset, FHFusion achieves the best overall performance across most evaluation metrics, demonstrating clear advantages in contrast enhancement, texture preservation, and perceptual quality. On the TNO dataset, our method ranks first in SD, SF, AG, and VIFF, and remains competitive in MI, suggesting good adaptability across diverse scenes. For the LLVIP dataset, FHFusion attains the best results in EN, SD, SF, and AG, while achieving second-best performance in MI and VIFF, validating its robustness and perceptual fidelity in low-light pedestrian scenarios.

B. Qualitative Comparison

Fig. 3 presents qualitative comparisons on the M3FD, TNO, and LLVIP dataset. As shown in Fig. 3(a), compared with existing methods that suffer from texture degradation or weakened target edges, FHFusion better preserves fine visible details while maintaining salient infrared structures. In Fig. 3(b), FHFusion achieves a more balanced fusion by simultaneously retaining thermal radiation characteristics and fine-grained visible textures. For challenging high-contrast scenarios in Fig. 3(c), FHFusion produces clearer target–background separation and more stable structural boundaries, avoiding excessive smoothing artifacts observed in other methods.

C. Efficiency and Infrared-Visible Object Detection

We further evaluate FHFusion from the perspectives of computational cost and downstream object detection performance to assess its practical applicability. For efficiency analysis, we compare different fusion methods on infrared and visible image pairs with a resolution of 620×448. As shown in Table II,

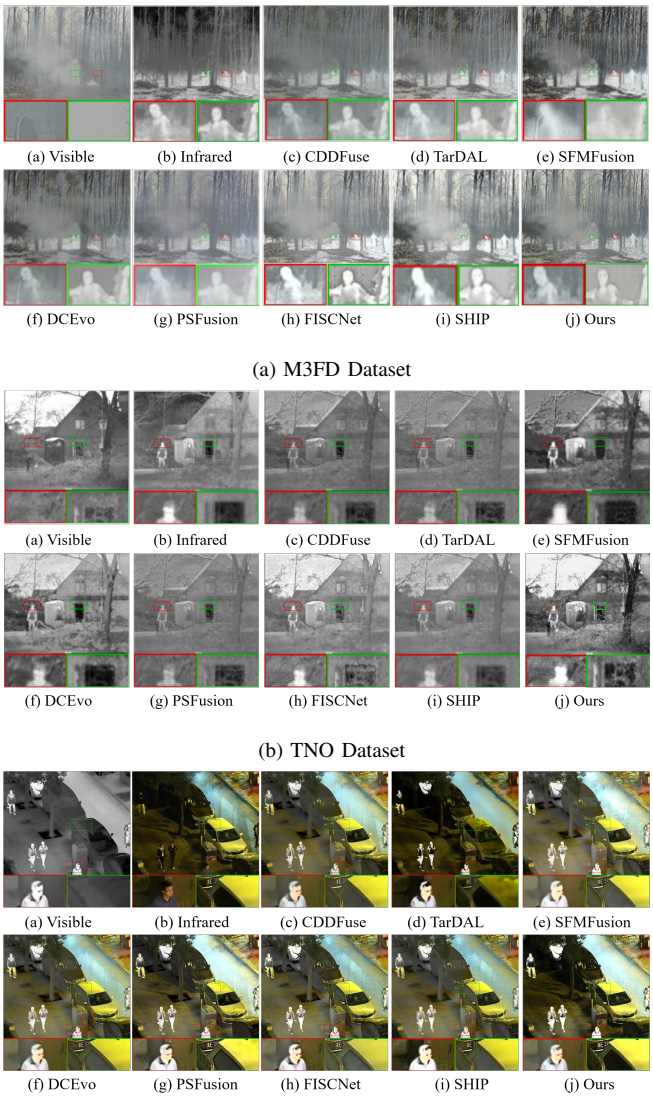

Fig. 3. Qualitative comparisons of FHFusion versus other methods on (a) M3FD, (b) TNO, and (c) LLVIP dataset.

PSFusion and DCEvo introduce relatively high computational overhead in terms of parameter count and FLOPs. In contrast, FHFusion achieves a lightweight design with 0.33 M parameters and 101.82 G FLOPs, indicating reduced computational cost while maintaining competitive performance.

To further investigate the impact of fusion quality on downstream tasks, we evaluate infrared-visible object detection performance using the CenterNet [25] detector on the M3FD dataset. The experimental comparison includes seven representative fusion methods. As reported in Table III, FHFusion achieves competitive detection performance across key categories such as *People* and *Car*, and attains an mAP score only 0.1% lower than the best-performing method, showing that the proposed fusion strategy remains competitive for downstream detection tasks.

TABLE II
EFFICIENCY ANALYSIS WITH DIFFERENT METHODS.

Metric	TarDAL	DCEvo	CDDFuse	PSFusion	SHIP	SFMFusion	Ours
Parameters (M)	0.30	2.00	1.19	45.91	0.55	1.61	0.33
FLOPs (G)	91.14	912.66	547.74	180.86	156.84	171.41	101.82

TABLE III
QUANTITATIVE RESULTS OF OBJECT DETECTION WITH DIFFERENT METHODS ON THE M3FD DATASET. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN BOLD AND UNDERLINED, RESPECTIVELY.

Method	Lamp	Car	Bus	Motor	Truck	People	mAP
CDDFuse [16]	0.285	0.720	0.445	0.325	0.385	0.525	0.448
TarDAL [17]	0.231	0.632	0.419	0.295	0.327	0.505	0.401
SFMFusion [7]	0.247	0.634	0.459	0.331	0.349	0.449	0.412
DCEvo [18]	0.295	0.735	0.460	0.341	0.431	0.540	0.467
PSFusion [19]	**0.461**	0.729	0.485	0.342	0.447	**0.589**	**0.509**
FISCNet [8]	0.324	0.742	0.491	0.362	0.460	0.572	0.492
SHIP [6]	0.361	0.711	0.495	0.359	0.469	0.568	0.494
Ours	0.370	**0.753**	**0.501**	**0.365**	**0.478**	0.579	0.508

V. ABLATION STUDIES

To evaluate the contribution of different components in FHFusion, we conduct comprehensive ablation studies on the M3FD dataset. The results of module ablations, iteration analysis, and loss component ablations are summarized in Table IV. In all cases, removing any component or simplifying the design leads to consistent performance degradation, demonstrating the effectiveness of the proposed architecture.

Specifically, removing SDHA, ASM, PDM, or SIM degrades fusion quality in different aspects, indicating that frequency-domain hierarchical modeling, adaptive amplitude selection, phase discrimination, and spatial–frequency interaction are all essential. The performance improves with increasing iteration number N and stabilizes at $N = 3$, while additional iterations provide marginal gains. Moreover, loss ablation results show that the complete loss function achieves the best overall performance, confirming that pixel-level, intensity-based, and gradient-based losses jointly contribute to balanced target saliency preservation and texture retention.

VI. CONCLUSIONS

This paper presents FHFusion, a novel approach that systematically exploits frequency-domain heterogeneity between infrared and visible images via spectral decoupling strategies. Furthermore, the carefully designed ASM module enables adaptive amplitude information selection and the PDM module achieves precise phase information discrimination, establishing a refined frequency-domain fusion framework. Experimental results demonstrate that FHFusion achieves superior performance in both quantitative fusion metrics and visual quality assessment, especially in preserving salient target features and textural details. Future research will focus on extending this framework to other multimodal image fusion tasks.

TABLE IV
ABLATION RESULTS ON THE M3FD DATASET. PANEL A: MODULE ABLATIONS. PANEL B: IMPACT OF ITERATION NUMBER. PANEL C: LOSS COMPONENT ABLATIONS. THE BEST RESULTS ARE SHOWN IN BOLD.

Setting	EN↑	SD↑	SF↑	AG↑	MI↑	VIFF↑
Panel A: Module ablations						
w/o SDHA	7.096	50.910	17.459	5.119	2.341	0.720
w/o ASM	7.281	55.023	22.030	7.198	2.822	0.859
w/o PDM	6.967	39.864	15.514	5.269	2.849	0.751
w/o SIM	6.929	37.372	15.858	5.467	2.909	0.811
Ours	**7.446**	**63.040**	**25.513**	**8.324**	**3.142**	**0.898**
Panel B: Impact of iteration number N						
$N = 1$	6.988	39.472	18.027	5.991	2.668	0.699
$N = 2$	7.431	57.284	22.652	7.480	2.726	0.840
$N = 3$ (Ours)	7.446	63.040	25.513	8.324	**3.142**	0.898
$N = 4$	**7.452**	**63.984**	**26.123**	**8.598**	2.984	**0.902**
Panel C: Loss component ablations						
$\mathcal{L}$	**7.446**	**63.040**	**25.513**	**8.324**	**3.142**	**0.898**
w/o $\mathcal{L}_{pixel}$	7.318	60.434	23.686	8.251	2.537	0.864
w/o $\mathcal{L}_{max}$	7.192	42.023	23.356	6.586	1.281	0.640
w/o $\mathcal{L}_{gra}$	7.396	54.910	21.459	7.119	2.541	0.753

ACKNOWLEDGMENTS

This work was supported in part by the Science and Technology Supporting Xinjiang Program (Directive Project) under Grant No. 2025E02057, and in part by the Key Research Project of Medical and Health System Reform under Grant No. CHEATG2502070503.

REFERENCES

[1] Min Li, Enguang Zuo, Feng Li, Cheng Chen, Chaoxun Guo, Pei Liu, Yunling Wang, Xiaoyi Lv, and Chen Chen, "Dcfusion: Difference correlation-driven fusion mechanism of infrared and visible images," *Pattern Recognition*, vol. 158, pp. 111002, 2025.

[2] Xingchen Zhang, Ping Ye, and Gang Xiao, "VIFB: A visible and infrared image fusion benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 104–105.

[3] Kening Cui, Jinyong Cheng, and Yan Pan, "Fdafusion: A joint frequency decomposition and adaptive fusion network for nighttime infrared and visible image fusion," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[4] Hui Li, Xianbiao Qi, and Wuyuan Xie, "Fast infrared and visible image fusion with structural decomposition," *Knowledge-Based Systems*, vol. 204, pp. 106182, 2020.

[5] Zhiqiang Zhou, Bo Wang, Sun Li, and Mingjie Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters," *Information Fusion*, vol. 30, pp. 15–26, 2016.

[6] Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao, "Probing synergistic high-order interaction in infrared and visible image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26384–26395.

[7] Hui Sun, Long Lv, Pingping Zhang, Tongdan Tang, Feng Tian, Weibing Sun, and Huchuan Lu, "Spatial-frequency enhanced Mamba for multi-modal image fusion," *arXiv preprint arXiv:2511.06593*, 2025.

[8] Naishan Zheng, Man Zhou, Jie Huang, and Feng Zhao, "Frequency integration and spatial compensation network for infrared and visible image fusion," *Information Fusion*, vol. 109, pp. 102359, 2024.

[9] Chenwu Wang, Junsheng Wu, Aiqing Fang, Zhixiang Zhu, Pei Wang, and Hao Chen, "An efficient frequency domain fusion network of infrared and visible images," *Engineering Applications of Artificial Intelligence*, vol. 133, pp. 108013, 2024.

[10] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao, "Frequency and spatial dual guidance for image dehazing," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 181–198.

[11] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong, "Deep Fourier-based exposure correction network with spatial-frequency interaction," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 163–180.

[12] Aowei Yu, Yueyang Li, and Weimiao Feng, "Moefuse: Degradation-aware fusion of infrared and visible images using mixture of experts," in *International Joint Conference on Neural Networks, IJCNN 2025, Rome, Italy, June 30 - July 5, 2025*. 2025, pp. 1–8, IEEE.

[13] 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*, vol. 83, pp. 79–92, 2022.

[14] Alexander Toet, "The TNO multiband image data collection," *Data in Brief*, vol. 15, pp. 249, 2017.

[15] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3496–3504.

[16] 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 (CVPR)*, 2023, pp. 5906–5916.

[17] 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 (CVPR)*, 2022, pp. 5802–5811.

[18] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan, "DCEvo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 2226–2235.

[19] 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*, vol. 99, pp. 101870, 2023.

[20] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian, "A Fourier-based framework for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14383–14392.

[21] Yanchao Yang and Stefano Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4085–4095.

[22] Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma, "C2RF: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning," *International Journal of Computer Vision*, 2025.

[23] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.

[24] 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.

[25] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.