CFSFusion: A Cross-Domain Interaction and Frequency-Spatial Collaborative Network for Infrared and Visible Image Fusion

Zhiyu Xiao[1,2], Qiang Tong[1,2], Yuli Chen[1,2,3], Xiulei Liu[1,2], Siyu Zhu[1,2,*]
[1]Laboratory of Data Science and Information Studies, Beijing Information Science and Technology University, Beijing, China
[2]School of Computer Science, Beijing Information Science and Technology University, Beijing, China
[3]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, China
2603361579@qq.com, tongq85@bistu.edu.cn, chenyuli@bupt.edu.cn, liuxiulei@bistu.edu.cn, zhusiyu@bistu.edu.cn
*Corresponding Author: Siyu Zhu Email: zhusiyu@bistu.edu.cn

***Abstract*—Fusing infrared and visible images is a key technique for intelligent perception in complex environments, such as nighttime surveillance and autonomous driving. Existing methods often combine frequency and spatial features without sufficient interaction, resulting in limited structure–saliency integration, degraded detail fidelity, and less prominent object representation. Moreover, high-frequency details crucial for textures and contours are commonly enhanced in a non-selective manner, which can introduce noise and blurred edges. To address these issues, we propose CFSFusion, a collaborative framework that integrates frequency-informed and spatial features. Specifically, we design a Cross-Domain Interaction Module (CDIM) that enables structure–saliency guided interaction between frequency-informed representations and spatial features. In addition, a structure-aware high-frequency enhancement mechanism (StructHF) is introduced to selectively emphasize key details, while an improved Multi-Scale Enhancement Channel Attention (MECA) further reinforces salient structures and suppresses noise. Extensive experiments on public datasets demonstrate that our method achieves high-quality fusion across multiple metrics and shows strong robustness in downstream object detection tasks. We release the code[1] to facilitate future research.**

***Index Terms*—Infrared and visible image fusion, cross-domain interaction, multimodal fusion, object detection**

I. INTRODUCTION

Obtaining high-quality perceptual images in complex environments is essential for applications such as nighttime perception [1], surveillance [2], and autonomous driving [3]. Single-modality images often fail to achieve both robustness and fine-detail representation [4]. Infrared images provide strong thermal saliency but lack structural details, while visible images contain rich textures but degrade under poor illumination [5]. Therefore, infrared–visible image fusion, which balances target saliency and structural information, has become an important problem in multimodal perception.

Image fusion integrates complementary information from multiple sources for perception tasks such as object detection [6] and segmentation [7]. Infrared–visible fusion methods are typically divided into spatial-domain and frequency-domain approaches. Spatial-domain methods use CNNs or attention mechanisms but often suffer from blurred high-frequency details [8], while frequency-domain methods preserve texture through frequency decomposition [9] but may fail to model global structure–saliency relationships. Recent spatial–frequency fusion methods, including sequential fusion [10] and parallel architectures [11], improve performance but still lack effective cross-domain coordination. Consequently, spatial and frequency features are often modeled separately, and non-selective high-frequency enhancement may introduce noise and false contours.

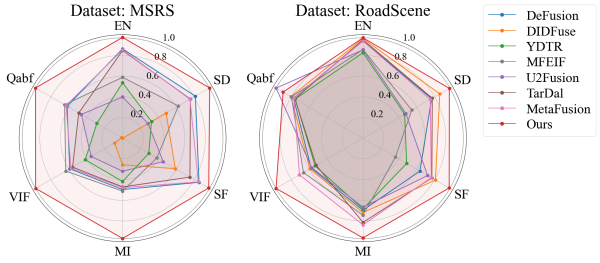

Fig. 1. Our method (highlighted in red) achieves the state-of-the-art performance on MSRS and RoadScene in six metrics.

To address these issues, we propose a collaborative fusion framework integrating frequency-informed and spatial features to capture global structural cues and fine-grained details. The framework includes a Cross-Domain Interaction Module (CDIM) for structure–saliency guided interaction between frequency-informed and spatial features, and a structure-aware high-frequency enhancement mechanism (StructHF) based on a Feature Pyramid Network (FPN) and an Edge Gradient Module (EGM) to enhance task-relevant high-frequency components without amplifying noise. In addition, a Multi-Scale Enhancement Channel Attention (MECA) reinforces salient structures and suppresses redundant responses. Experiments on three public datasets demonstrate the effectiveness of the proposed method. As shown in Fig. 1, our method achieves

[1]The code is available at: https://github.com/xxx260/CFSFusion.

leading performance on MSRS and RoadScene and competitive results on M3FD, while improving downstream object detection performance.

Overall, the contributions of our work are as follows:

- We propose a frequency–spatial collaborative fusion framework with a bidirectional structure–saliency guidance mechanism between frequency and spatial domains.
- We design a structure-aware high-frequency enhancement with edge guidance and multi-scale enhancement channel attention to strengthen target textures and boundaries.
- Experiments on three public datasets (M3FD, MSRS, and RoadScene) demonstrate that the proposed method outperforms existing approaches in fusion quality and object detection, showing strong generalization ability.

II. RELATED WORKS

A. Traditional Image Fusion

Early infrared and visible image fusion relies on traditional methods using handcrafted features and fusion rules, including pixel-level, saliency, and optimization approaches. Li et al. [12] propose pixel-level fusion with low computational complexity but reduced contrast and blurred details. Yan et al. [13] develop a spectral wavelet-based multi-scale decomposition method to enhance detail representation. Xiang et al. [14] combine adaptive dual-channel PCNN with NSCT to improve edge and texture preservation. Bin et al. [15] propose an approximate sparse representation method to reduce complexity while maintaining fusion accuracy. Meng et al. [16] introduce a fusion strategy based on saliency maps and keypoints, whose performance depends on saliency detection accuracy. Overall, these traditional methods struggle to preserve structural fidelity and target saliency, thereby limiting fusion performance.

B. Single-Domain and Multi-Domain Image Fusion

With the rise of deep learning, more methods focus on fusing spatial and frequency information. Zhao et al. [8] propose CDDFuse, a dual-branch network using CNNs and attention to extract salient spatial features, but fine details remain weak. Frequency-domain methods extract components via Fast Fourier Transform (FFT) [17] to preserve structure and texture. Wang et al. [9] propose a directional lifting wavelet fusion network that separately processes low- and high-frequency coefficients, but its reliance on frequency-domain features limits spatial detail and structure–saliency modeling. Zheng et al. [10] propose FISC, which processes frequency and spatial information sequentially, risking feature loss in one domain. Hu et al. [11] introduce SFDFusion, which fuses spatial and frequency features in parallel but still lacks deep cross-domain interaction. Differing from prior approaches, our method introduces bidirectional structure–saliency interaction to improve high-frequency detail enhancement and edge clarity.

III. METHODS

In this paper, we propose the CFSFusion framework for multi-modal image fusion, as shown in Fig. 2, consisting of FFM, MSFM, and CDIM.

A. Frequency Fusion Module

High-frequency information is important for fusion, but spatial-domain methods lack frequency discrimination. We propose the Frequency Fusion Module (FFM), which decomposes images using the Fourier Transform and integrates infrared and visible frequency information via adaptive fusion and multi-scale channel attention. Given an input image $I \in \mathbb{R}^{1\times H\times W}$, it is mapped to the complex frequency domain using FFT:

$$\mathcal{F} = \mathrm{FFT}(I), \quad \tilde{\mathcal{F}} = A \cdot e^{jP} \tag{1}$$

where $\mathcal{F} \in \mathbb{C}^{1\times H\times W}$ is the complex spectrum, $\tilde{\mathcal{F}}$ is its polar form, $A = |\mathcal{F}|$ is the amplitude spectrum carrying texture and edge information, $P = \angle\mathcal{F}$ is the phase spectrum encoding the spatial structure, e is the natural constant, and j is the imaginary unit with $j^2 = -1$.

Compute the frequency-domain components separately for the infrared image I_{ir} and the visible image I_{vi} :

$$A_{ir}, P_{ir} = \mathrm{FFT}(I_{ir}), \quad A_{vi}, P_{vi} = \mathrm{FFT}(I_{vi}) \tag{2}$$

where A_{ir}, A_{vi} are amplitudes and P_{ir}, P_{vi} are phases.

For amplitude and phase, we design AmpFuse and PhaFuse, both using MECA for weighted fusion, selectively fusing infrared and visible frequency information to preserve structural consistency. This can be expressed as:

$$A_f = \mathrm{AmpFuse}(A_{ir}, A_{vi}) = \mathrm{MECA}([A_{ir}, A_{vi}]) \tag{3}$$

where A_f is the fused amplitude, $\mathrm{MECA}(\cdot)$ denotes the MECA module, and $[A_{ir}, A_{vi}]$ denotes channel concatenation.

$$P_f = \mathrm{PhaFuse}(P_{ir}, P_{vi}) = \mathrm{MECA}([P_{ir}, P_{vi}]) \tag{4}$$

where P_f is the fused phase spectrum, and $[P_{ir}, P_{vi}]$ denotes concatenation of infrared and visible phase spectra. This enhances key high-frequency components while suppressing noise and misalignment. The fused spectrum is obtained as:

$$\mathcal{F}_f = A_f \cdot e^{jP_f} \tag{5}$$

where $\mathcal{F}_f$ is the fused complex spectrum composed of the fused amplitude A_f and fused phase P_f.

Finally, the fused frequency domain result is reconstructed into spatial features using the IFFT:

$$F_f = \mathrm{IFFT}(\mathcal{F}_f) = \mathrm{IFFT}(A_f \cdot e^{jP_f}) \tag{6}$$

where F_f denotes the fused image in the frequency domain.

This module performs frequency-domain fusion to preserve high-frequency details, compensate for spatial fusion limitations, and support cross-domain interaction and structural completion.

B. Multi-Scale Spatial Fusion Module

We design a Multi-Scale Spatial Fusion Module (MSFM) for structure-aware high-frequency enhancement (StructHF). MSFM integrates a FPN and an EGM to model multi-scale structures and edge-aware high-frequency cues. FPN captures hierarchical structures, while EGM guides detail enhancement and noise suppression. Specifically, for the input infrared

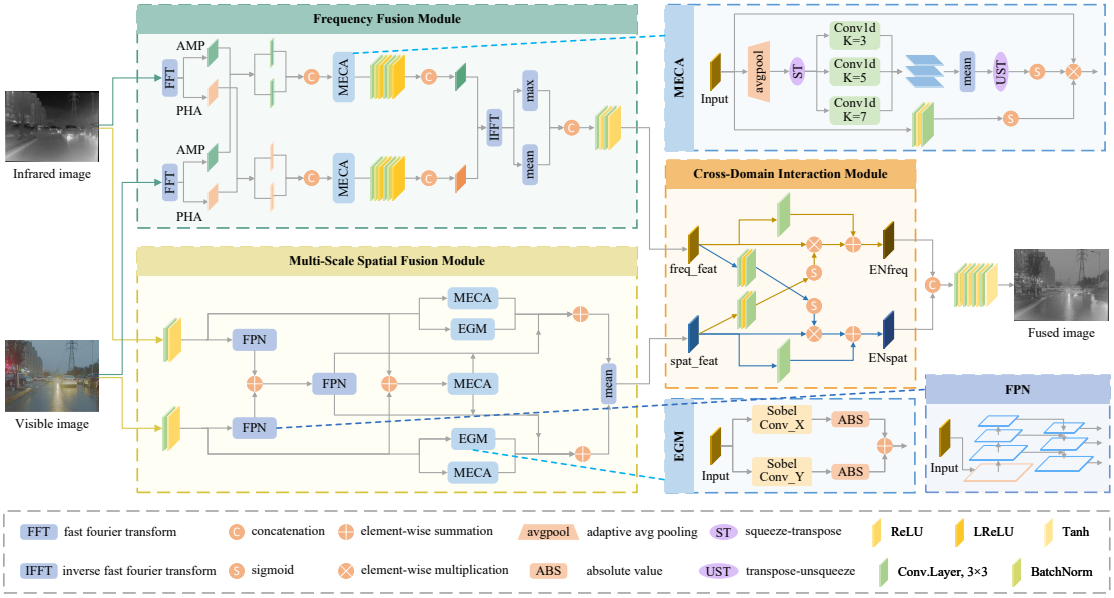

Fig. 2. Overview of the proposed CFSFusion framework. It consists of a Frequency Fusion Module, a Cross-Domain Interaction Module, and a Multi-Scale Spatial Fusion Module, which implements the structure-aware high-frequency enhancement by integrating a Feature Pyramid Network with an Edge Gradient Module. An improved Multi-Scale Enhancement Channel Attention is used to reinforce salient structures.

image $I_{ir} \in \mathbb{R}^{1\times H\times W}$ and visible image $I_{vi} \in \mathbb{R}^{1\times H\times W}$, preliminary feature representations are generated through convolutional embedding:

$$X_{ir} = Conv_{3\times3}(I_{ir}), \quad X_{vi} = Conv_{3\times3}(I_{vi}) \tag{7}$$

where $Conv_{3\times3}(\cdot)$ denotes a 3×3 convolution, and X_{ir}, X_{vi} are the embedded features.

Then, MECA enhances single-modal and fused features, highlighting salient regions and suppressing noise:

$$X_{ir}^{att} = \mathrm{MECA}(X_{ir}) + \mathrm{MECA}(X_{ir} + X_{vi}) \tag{8}$$

$$X_{vi}^{att} = \mathrm{MECA}(X_{vi}) + \mathrm{MECA}(X_{ir} + X_{vi}) \tag{9}$$

where X_{ir}^{att} is the enhanced infrared feature, and X_{vi}^{att} is the enhanced visible feature. To enhance target contours, an EGM uses Sobel operators in the x/y directions, followed by absolute summation:

$$\nabla X_{ir} = |S_x * X_{ir}| + |S_y * X_{ir}| \tag{10}$$

$$\nabla X_{vi} = |S_x * X_{vi}| + |S_y * X_{vi}| \tag{11}$$

where S_x, S_y are Sobel kernels in the x and y directions, $*$ denotes convolution, and ∇X_{ir}, ∇X_{vi} are the gradient feature of infrared and visible.

Next, multi-scale convolutions fuse infrared and visible features to model spatial structure:

$$C_3 = Conv_{3\times3}(Conv_{3\times3}(X_{ir}) + Conv_{3\times3}(X_{vi})) \tag{12}$$

where C_3 is the multi-scale fused feature.

Finally, fused multi-scale, attention-enhanced, and gradient features are averaged across channels to produce unified spatial structural features:

$$F_{s,ir} = X_{ir}^{att} + \nabla X_{ir} + C_3, \quad F_{s,vi} = X_{vi}^{att} + \nabla X_{vi} + C_3 \tag{13}$$

$$F_s = \mathrm{Mean}(F_{s,ir}, F_{s,vi}) \tag{14}$$

where $F_{s,ir}$, $F_{s,vi}$ are structure-enhanced infrared and visible features, $\mathrm{Mean}(\cdot)$ denotes the element-wise mean, and F_s is the spatial feature guiding frequency features.

Edge-gradient guidance enhances contours, and multi-scale channel attention highlights salient regions and suppresses noise. The spatial feature F_s guides frequency features for detail-preserving fusion.

C. Cross-Domain Interaction Module

To enhance coordination between frequency-informed and spatial representations, we propose the Cross-Domain Interaction Module (CDIM). Although reconstructed into the spatial domain via inverse Fourier transform, these features retain global spectral priors and frequency-selective responses, referred to as *frequency-informed spatial features*. CDIM performs lightweight region-level structure–saliency calibration instead of heavy cross-attention, as fusion prioritizes cross-domain consistency over token-wise correspondence.

Specifically, CDIM first applies lightweight linear projections to the frequency-informed features F_f and spatial features F_s to align channel dimensions:

$$F'_{freq\_proj} = \mathrm{Proj}(F_f), \quad F'_{spatial\_proj} = \mathrm{Proj}(F_s) \tag{15}$$

where $\mathrm{Proj}(\cdot)$ denotes a linear projection operator.

CDIM consists of two complementary interaction branches. In the frequency-to-spatial branch ($\mathcal{B}_{F\to S}$), a spatial guidance map is generated from frequency-informed features:

$$\mathcal{G}_{F\to S} = \sigma(\mathrm{Conv}_{F\to S}(F'_{freq\_proj})) \tag{16}$$

where $\sigma(\cdot)$ denotes the Sigmoid function. The Sigmoid activation enables smooth, bounded region-level modulation, which

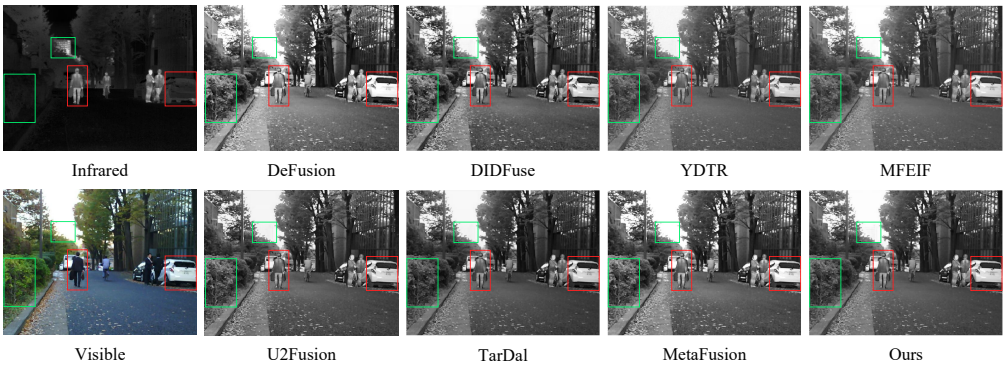

Fig. 3. Comparison of fusion results on the image "01379D" from the MSRS dataset using different methods.

is sufficient for enforcing structural consistency. The calibrated spatial features are obtained by:

$$F''_{spatial} = F'_{spatial\_proj} \odot \mathcal{G}_{F\to S} \tag{17}$$

Conversely, the spatial-to-frequency branch ($\mathcal{B}_{S\to F}$) uses spatial structural cues to suppress redundant frequency-informed responses:

$$\mathcal{G}_{S\to F} = \sigma(\text{Conv}_{S\to F}(F'_{spatial\_proj})) \tag{18}$$

$$F''_{freq} = F'_{freq\_proj} \odot \mathcal{G}_{S\to F} \tag{19}$$

By performing bidirectional, low-cost structure–saliency calibration rather than explicit correspondence modeling, CDIM outputs mutually refined frequency-informed and spatial features, providing an efficient foundation for high-quality fusion and robust downstream perception.

IV. EXPERIMENTS

A. Experimental Setting

Datasets. Experiments are conducted on three public infrared–visible fusion datasets: M3FD [18], MSRS [19], and RoadScene [20]. M3FD contains 4200 registered pairs (1024×768) with detection labels, split into training and testing sets with an 8:2 ratio. MSRS includes 1444 pairs (640×480) with an official split of 1083 training and 361 testing pairs. RoadScene contains 221 registered pairs captured in real traffic scenarios.

Evaluation Metrics. We evaluate fusion performance using Entropy (EN), Standard Deviation (SD), Spatial Frequency (SF), Mutual Information (MI), Visual Information Fidelity (VIF), and Gradient-based Fusion Performance (Qabf) [5], where higher values indicate better performance. Mean Average Precision (mAP) is used for object detection.

Implementation Details. Experiments are conducted on a workstation with two RTX 3090 GPUs. The model is trained on MSRS and tested on M3FD, MSRS, and RoadScene. The batch size is 4 and the model is trained for 50 epochs using Adam with a learning rate of 0.0005 ($\beta_1 = 0.9$, $\beta_2 = 0.999$).

TABLE I
QUANTITATIVE COMPARISON RESULTS ACROSS THREE DATASETS. **BOLDFACE** AND UNDERLINE SHOW THE BEST AND SECOND-BEST VALUES, RESPECTIVELY.

Dataset: M3FD Infrared-Visible Fusion Dataset						
Methods	EN	SD	SF	MI	VIF	Qabf
DeFusion	7.159	35.267	13.386	3.825	0.624	0.521
DIDFuse	<u>7.163</u>	**46.438**	**13.721**	3.960	0.668	0.512
YDTR	6.474	28.359	11.001	3.825	0.626	0.459
MFEIF	6.701	30.012	9.173	3.890	0.640	0.457
U2Fusion	6.847	33.103	13.018	3.715	0.643	**0.566**
TarDal	7.157	42.891	13.482	3.895	0.602	0.461
MetaFusion	**7.203**	<u>43.125</u>	13.476	<u>4.038</u>	<u>0.689</u>	0.484
Ours	6.851	35.828	<u>13.492</u>	**4.727**	**0.787**	<u>0.563</u>
Dataset: MSRS Infrared-Visible Fusion Dataset						
Methods	EN	SD	SF	MI	VIF	Qabf
DeFusion	<u>6.521</u>	<u>39.264</u>	<u>11.562</u>	2.968	0.724	0.536
DIDFuse	4.230	30.002	9.598	2.356	0.311	0.201
YDTR	5.653	25.401	7.401	2.769	0.581	0.353
MFEIF	5.792	33.901	8.068	<u>3.008</u>	<u>0.761</u>	<u>0.551</u>
U2Fusion	5.289	24.302	8.598	2.514	0.527	0.448
TarDal	6.481	37.689	10.812	2.903	0.698	0.464
MetaFusion	6.498	37.561	11.503	2.921	0.713	0.542
Ours	**6.816**	**43.764**	**12.368**	**4.183**	**1.038**	**0.729**
Dataset: RoadScene Infrared-Visible Fusion Dataset						
Methods	EN	SD	SF	MI	VIF	Qabf
DeFusion	7.327	47.826	12.241	3.804	0.575	0.453
DIDFuse	<u>7.402</u>	<u>51.588</u>	<u>14.156</u>	3.871	0.583	0.463
YDTR	6.951	35.436	10.613	3.763	0.549	0.447
MFEIF	7.053	38.711	9.237	3.951	0.625	0.461
U2Fusion	7.042	35.987	13.198	3.712	0.571	**0.518**
TarDal	7.363	48.132	13.714	4.160	0.542	0.448
MetaFusion	7.321	47.214	13.648	<u>4.231</u>	<u>0.651</u>	0.452
Ours	**7.426**	**55.964**	**15.845**	**4.614**	**0.804**	<u>0.492</u>

B. Infrared and Visible Image Fusion Performance

Quantitative Analysis. To validate the effectiveness of our method, we conduct a systematic comparison on three mainstream infrared-visible image fusion datasets, using multiple evaluation metrics and seven representative fusion methods: DeFusion [21], DIDFuse [22], YDTR [23], MFEIF [24], U2Fusion [20], TarDal [18], and MetaFusion [25]. Detailed quantitative results are shown in Table I.

On MSRS and RoadScene, our method achieves leading or competitive results across fusion metrics, as shown in Fig. 1, demonstrating effective information preservation and

TABLE II
MODEL EFFICIENCY COMPARISON. **BOLDFACE** AND UNDERLINE SHOW THE BEST AND SECOND-BEST VALUES, RESPECTIVELY.

Methods	Time(s)	Params/M	FLOPs/G
DeFusion	0.050	0.30	119.03
DIDFuse	0.046	0.25	115.21
YDTR	0.034	**0.13**	95.94
MFEIF	0.073	0.41	114.21
U2Fusion	0.251	0.69	367.12
TarDal	<u>0.030</u>	0.32	<u>92.11</u>
MetaFusion	0.150	1.20	449.87
Ours	**0.020**	<u>0.14</u>	**29.38**

TABLE III
OBJECT DETECTION METRICS (MAP @0.5) ON THE M3FD DATASET. **BOLDFACE** AND UNDERLINE SHOW THE BEST AND SECOND-BEST VALUES, RESPECTIVELY.

Methods	People	Car	Bus	Motor	Lamp	Truck	All
DeFusion	0.561	0.874	0.823	0.657	0.615	0.846	0.754
DIDFuse	0.472	0.869	0.827	0.613	0.541	0.683	0.715
YDTR	0.612	0.895	0.821	0.741	<u>0.682</u>	0.866	0.769
MFEIF	0.593	<u>0.901</u>	0.752	0.761	0.673	0.869	0.768
U2Fusion	0.589	**0.904**	0.775	0.752	0.554	0.851	0.758
TarDal	0.684	0.878	0.881	<u>0.781</u>	0.557	<u>0.873</u>	0.770
MetaFusion	<u>0.686</u>	0.893	<u>0.893</u>	0.775	0.551	**0.874**	<u>0.791</u>
Ours	**0.784**	0.897	**0.905**	**0.787**	**0.868**	0.851	**0.812**

detail representation. On M3FD, although some metrics show minor differences, the overall performance remains among the best, with notable improvements in MI and VIF, indicating better information integrity and visual consistency. To assess practical deployability, we compare inference time, model size, and computational cost, as reported in Table II. Under identical evaluation settings, our method demonstrates competitive efficiency with a compact model size, achieving a favorable balance between performance and computational overhead.

Qualitative Analysis. For visual comparison, a representative MSRS image pair is fused using multiple methods, as shown in Fig. 3. Red boxes indicate foreground targets, and green boxes highlight background textures.

Overall, our method preserves clear target contours and infrared saliency while maintaining visible structural details. This is mainly due to the cross-domain interaction module and structure-aware high-frequency enhancement, which preserve structural consistency and fine textures. Compared with other methods, DeFusion introduces over-bright backgrounds due to spectral contamination, while MetaFusion produces cleaner backgrounds but reduces texture sharpness. In contrast, our method achieves a better balance between foreground saliency and background texture without excessive enhancement or detail suppression.

C. Object Detection Performance

To evaluate downstream perception performance, we conduct object detection experiments on the M3FD dataset using YOLOv8 detectors trained on infrared and visible modalities. Fused images are directly used for inference without retraining. This protocol evaluates perceptual compatibility and cross-domain generalization rather than upper-bound performance from task-specific retraining. Performance improvements under this inference-only setting indicate better consistency with features learned by pre-trained detectors.

As shown in Table III, our method achieves competitive and stable detection performance across multiple object categories. This suggests that the proposed frequency–spatial collaborative fusion preserves discriminative cues effectively and enhances target visibility, thereby supporting more reliable recognition and localization without explicit detection-oriented optimization. Fig. 4 presents qualitative detection results on scenes with complex backgrounds and numerous small targets. Compared with existing methods, our fused images better

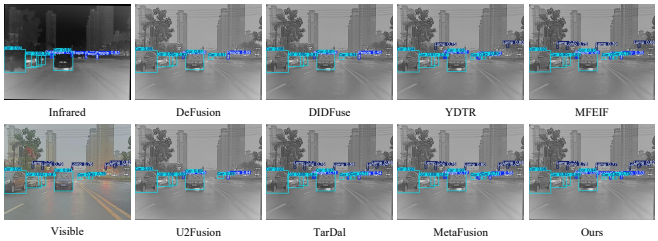

Fig. 4. A comparison of the detection performance after fusion on image "00129" from the M3FD dataset.

maintain target-related structures and edge information, reducing missed detections while avoiding background interference. These results demonstrate that the proposed fusion strategy produces perceptually compatible representations beneficial for downstream object detection tasks.

D. Ablation Study

We conduct ablation experiments on MSRS to evaluate each module, as shown in Table IV, where "w/o" indicates module removal or replacement. Removing CDIM weakens frequency–spatial interaction and reduces information retention and target saliency, with drops in MI and VIF. Removing StructHF degrades detail sharpness and edge clarity, reflected by declines in SF and Qabf. Replacing MECA with single convolution attention limits multi-scale context modeling, causing slight performance degradation with declines in SD and VIF. These results show that the three modules jointly improve information fusion, detail representation, and visual fidelity.

To further illustrate the impact of each module, we provide visual comparisons in Fig. 5 using image "01516D". The yellow-boxed regions highlight the advantages of our method in multiple areas. For nearby pedestrians and vehicles, the fused image preserves infrared saliency while maintaining visible textures. For distant targets and buildings, contours and boundaries remain clear even in complex backgrounds. In detailed regions such as ground tiles, the results show sharp structures without noticeable artifacts. In contrast, ablated variants show degraded saliency, blurred edges, and reduced texture contrast. Overall, our method produces higher visual quality and structural fidelity across different scenes.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS ON MSRS DATASET. **BOLDFACE** AND UNDERLINE SHOW THE BEST AND SECOND-BEST VALUES, RESPECTIVELY.

	EN	SD	SF	MI	VIF	Qabf
w/o CDIM	6.726	42.957	11.570	3.514	0.823	0.680
w/o StructHF	6.746	42.831	11.657	3.694	<u>0.874</u>	0.656
w/o MECA	<u>6.761</u>	<u>43.353</u>	<u>12.059</u>	<u>4.076</u>	0.851	<u>0.704</u>
Ours	**6.816**	**43.764**	**12.368**	**4.183**	**1.038**	**0.729**

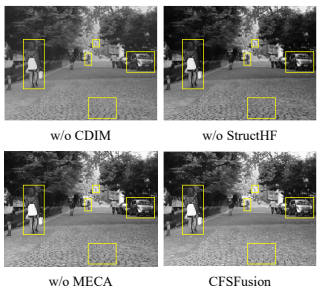

Fig. 5. Visual comparison of ablation experiments.

V. CONCLUSION

In this paper, we propose CFSFusion, an infrared–visible image fusion framework designed to address limitations in frequency–spatial feature interaction and detail representation. The proposed CDIM enables structure–saliency guided feature interaction between frequency-informed and spatial representations, while StructHF selectively enhances task-relevant details through multi-scale modeling and edge-gradient guidance. In addition, the improved MECA facilitates effective cross-scale channel modeling to reinforce salient structures. Extensive experiments on public datasets demonstrate that CFSFusion achieves high-quality and competitive fusion performance, and consistently improves detection accuracy and target localization in downstream object detection tasks.

ACKNOWLEDGMENT

This work is supported by the Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan (NO.20251027).

REFERENCES

[1] Z. Zhou, M. Dong, X. Xie, and Z. Gao, "Fusion of infrared and visible images for night-vision context enhancement," *Applied optics*, vol. 55, no. 23, pp. 6480–6490, 2016.

[2] N. Paramanandham and K. Rajendiran, "Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications," *Infrared Physics & Technology*, vol. 88, pp. 13–22, 2018.

[3] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li, "Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 4224–4231, 2018.

[4] K. Song, Y. Zhao, L. Huang, Y. Yan, and Q. Meng, "Rgb-t image analysis technology and application: A survey," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105919, 2023.

[5] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information fusion*, vol. 45, pp. 153–178, 2019.

[6] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 35–43.

[7] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[8] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. 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*, 2023, pp. 5906–5916.

[9] C. Wang, J. Wu, A. Fang, Z. Zhu, P. Wang, and H. Chen, "An efficient frequency domain fusion network of infrared and visible images," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108013, 2024.

[10] N. Zheng, M. Zhou, J. Huang, and F. Zhao, "Frequency integration and spatial compensation network for infrared and visible image fusion," *Information Fusion*, vol. 109, p. 102359, 2024.

[11] K. Hu, Q. Zhang, M. Yuan, and Y. Zhang, "Sfdfusion: an efficient spatial-frequency domain fusion network for infrared and visible image fusion," *arXiv preprint arXiv:2410.22837*, 2024.

[12] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *information Fusion*, vol. 33, pp. 100–112, 2017.

[13] X. Yan, H. Qin, J. Li, H. Zhou, and J.-g. Zong, "Infrared and visible image fusion with spectral graph wavelet transform," *Journal of the Optical Society of America A*, vol. 32, no. 9, pp. 1643–1652, 2015.

[14] T. Xiang, L. Yan, and R. Gao, "A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain," *Infrared Physics & Technology*, vol. 69, pp. 53–61, 2015.

[15] Y. Bin, Y. Chao, and H. Guoyu, "Efficient image fusion with approximate sparse representation," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 14, no. 04, p. 1650024, 2016.

[16] F. Meng, B. Guo, M. Song, and X. Zhang, "Image fusion with saliency map and interest points," *Neurocomputing*, vol. 177, pp. 1–8, 2016.

[17] Y. Tatsunami and M. Taki, "Fft-based dynamic token mixer for vision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 328–15 336.

[18] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. 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*, 2022, pp. 5802–5811.

[19] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Piafusion: A progressive infrared and visible image fusion network based on illumination aware," *Information Fusion*, vol. 83, pp. 79–92, 2022.

[20] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 502–518, 2020.

[21] P. Liang, J. Jiang, X. Liu, and J. Ma, "Fusion from decomposition: A self-supervised decomposition approach for image fusion," in *European conference on computer vision*. Springer, 2022, pp. 719–735.

[22] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "Didfuse: Deep image decomposition for infrared and visible image fusion," *arXiv preprint arXiv:2003.09210*, 2020.

[23] W. Tang, F. He, and Y. Liu, "Ydtr: Infrared and visible image fusion via y-shape dynamic transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 5413–5428, 2022.

[24] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. 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*, vol. 32, no. 1, pp. 105–119, 2021.

[25] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 955–13 965.