Lightweight Multi-Scale Cross-Fusion Network Model for Infrared Small Target Detection

Zengpei Zhang[1], Xuekui Zhang[1], Min Li[2], Ruixuan Zhang[1], Jing Liu[1], Dexin Zhang[1], Hui Zhang[1,*]
[1]Tianjin University of Science and Technology, Tianjin, China
[2]Suzhou Nuclear Power Research Institute Co., Ltd., Suzhou, China
{zpzhang2026, 19201249508, minli2506}@163.com, {zhangrx, jing.liu, dxzhang, zhanghui2022}@tust.edu.cn
[*]Corresponding Author

***Abstract*—Infrared small target detection (IRSTD) is challenging due to extremely small target sizes, low signal-to-noise ratios, and strong background clutter. Existing deep models often improve accuracy at the cost of prohibitive computation, whereas lightweight designs tend to lose weak target cues during downsampling. This paper presents LWMCNet, a lightweight multi-scale cross-fusion network for IRSTD. LWMCNet integrates a Lightweight Multi-Scale Feature Attention (LMFA) module with a Cascaded Feature Enhancement (CFE) strategy to preserve weak target information, and employs Residual Feature Fusion (RFF) together with Adaptive Fusion Upsampling (AFU) to strengthen cross-scale interaction and decoder reconstruction. On the IRSTD-1K dataset, LWMCNet achieves 70.25% IoU with only 0.21M parameters and 0.44G FLOPs at 256×256 input resolution. Moreover, it runs at 29.70 FPS on an NVIDIA Jetson NX edge platform, demonstrating its practicality for resource-constrained real-time deployment. The code is available at https://github.com/zpzhang-star/LWMCNet.**

***Index Terms*—Infrared Small Target Detection, Cascade Feature Enhancement, Lightweight**

I. Introduction

IRSTD has widespread applications, such as in infrared early warning and military surveillance. However, due to the imaging mechanism, infrared small targets often exhibit low contrast, blurred details, and a low signal-to-noise ratio, making them easily obscured by complex backgrounds and noise. This significantly increases the difficulty of IRSTD.

Recent advances in deep learning have led to numerous deep learning-based detection methods, such as ACM [1], DNANet [2], and UIUNet [3], which mainly focus on improving multiscale feature learning and interaction between low- and high-level features to mitigate information loss. Deep learning methods for IRSTD face a fundamental trade-off between accuracy and efficiency. High-accuracy algorithms, like SeRankDet [4], SCTransNet [5], and TCI-Former [6], have significant computational complexity and large parameter counts. This makes them unsuitable for resource-constrained, real-time applications. Conversely, lightweight networks, such as SpirDet [7], ALCNet [8], and LWNet [9], are computationally efficient but often fail to achieve sufficient detection accuracy for real-time performance.

To address this critical trade-off, we propose a novel Lightweight Multi-Scale Cross-Fusion Network (LWMCNet) designed to significantly reduce computational load and parameter count while preserving high detection accuracy.

In summary, the primary contributions of this paper are as follows:

- We propose LWMCNet, an ultra-lightweight multi-scale cross-fusion network for IRSTD that achieves a favorable accuracy–efficiency trade-off (0.21M parameters and 0.44G FLOPs at 256×256) and real-time throughput on an embedded Jetson NX platform (29.70 FPS).
- To mitigate the loss of weak target cues during downsampling, we design a Lightweight Multi-Scale Feature Attention (LMFA) module with a Cascaded Feature Enhancement (CFE) strategy that propagates low-level information across encoder stages and suppresses background interference.
- To strengthen cross-scale interaction and decoder reconstruction for small-target segmentation, we introduce a Residual Feature Fusion (RFF) module and an Adaptive Fusion Upsampling (AFU) decoder to refine fused features and preserve target boundaries and details.

II. METHODOLOGY

We define cross-fusion as the integration of feature maps from adjacent spatial scales. Unlike standard skip connections that directly concatenate features, our RFF module implements cross-fusion via channel-wise cross attention, selectively exchanging complementary information between adjacent scales while preserving spatial details through a residual pathway.

Our novelty lies in task-specific architectural synergy for IRSTD challenges. Unlike generic module stacking, the CFE and RFF modules prevent feature vanishing of tiny targets in deep layers—a critical bottleneck in lightweight models. By integrating WTConv for high-frequency details and SPDConv to replace lossy pooling, we ensure precise feature retention within a budget of just 0.21M parameters.

We present the novel LWMCNet, with its architecture illustrated in Fig. 1. The model adopts an efficient U-shaped architecture inspired by UNet [10], featuring:

- Encoder: Initial layer: RVEA (residual vision-expansion attention) and EMA [11] (Efficient Multi-Scale Attention) modules extract low-level features. Feature hierarchy: Three cascaded LMFA (Lightweight Multi-Scale Feature Attention) modules progressively extract features (channel dimensions: 8→16→32→64). Bottleneck: Unified LMFA module maintains 64-channel features.

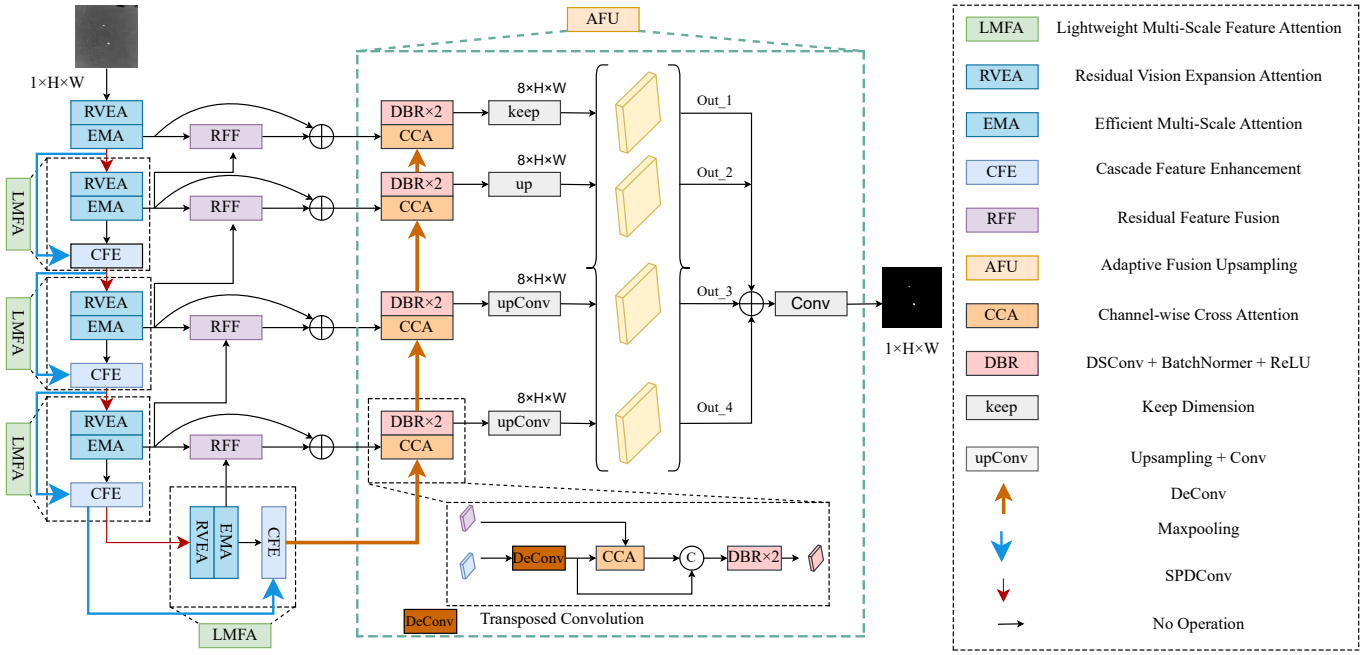

Fig. 1. Overview of the proposed LWMCNet for IRSTD. AFU is an Adaptive Fusion Upsampling Module.

- Cross-connection: RFF (Residual Feature Fusion) module enhances feature interaction through skip connections
- Decoder: Four AFU (Adaptive Fusion Upsampling) modules form symmetric structure (32→16→8→8 channels)

The network integrates multi-scale features for global context, while unlike UNet [10], it accesses adjacent features directly, removes redundancy, and reduces parameters.

A. *Lightweight Multi-Scale Feature Attention*

Infrared small targets typically appear as point-like objects occupying only a few pixels. Relying on a simple backbone for feature extraction can thus lead to significant feature loss during downsampling. To address this, we introduce the LMFA module, which combines RVEA, EMA, and CFE(Cascade Feature Enhancement) strategies (Fig. 2).

First, the architecture integrates Depthwise Separable Convolution (DSConv) [12] for localized feature extraction and Wavelet Convolution (WTConv) [13] for extended receptive fields to maintain model efficiency. The WTConv layer leverages wavelet transforms to achieve 40% larger receptive fields than standard convolutions while reducing parameters by 35%, which is crucial for multi-scale feature capture. Within the RVEA module processing $C \times H \times W$ input tensors, DSConv first extracts spatial features with BatchNorm-LeakyReLU for non-linear enhancement. WTConv hierarchically decomposes features via Haar wavelets, preserving high-frequency texture details critical for small target detection. It achieves exponential receptive field expansion with logarithmic parameter growth while decoupling targets from low-frequency background clutter across multiple frequency bands and enhancing global context to highlight target anomalies. This design is particularly effective for infrared backgrounds dominated by slowly varying low-frequency clutter with limited high-frequency interference, yielding notable performance gains.

Second, channel attention (CA) and spatial attention (SA) mechanisms [14] are used to adaptively modulate the importance of features, helping the network focus on salient attributes. A residual branch concurrently processes the input features, using a 1×1 convolution for dimension alignment while preserving global information. Features from both branches are then aggregated via a ReLU activation to generate the output. Finally, the EMA mechanism captures intricate multi-scale details and texture information of small targets. To

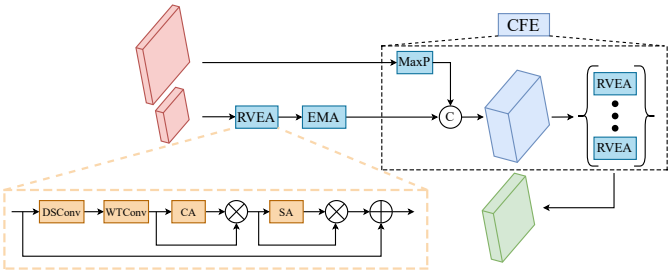

Fig. 2. This LMFA model is mainly composed of RVEA, EMA and CFE.

reduce computational overhead, the EMA module restructures channel dimensions and divides them into sub-feature groups, ensuring even distribution of spatial features. EMA encodes global information to dynamically recalibrate channel weights, which suppresses background noise and enhances responses for small targets. Additionally, cross-dimensional interactions aggregate outputs from parallel branches, capturing pixel-level relationships.

Last, SPDConv [15] is employed during the downsampling stage as a substitute for pooling layers, leveraging learnable convolutional operations to more effectively preserve the detailed information of small targets.

B. Cascaded Feature Enhancement

Infrared small targets are difficult to detect due to low SNR, weak brightness, and cluttered backgrounds. A simple stacking of feature extraction modules risks a loss of target features to background noise in deeper network layers. To address this, the proposed CFE architecture, as shown in Fig. 2, systematically integrates features across layers. This approach progressively incorporates features from adjacent layers during encoding to enhance information flow. Specifically, each encoder stage concatenates its output with features from the preceding layer, which are downsampled via max pooling to align spatial dimensions. This strategy effectively aggregates target information from multiple scales and preserves critical low-level details. This mechanism enables comprehensive multi-scale representation learning, progressively building more discriminative features while maintaining the integrity of spatial details.

C. Residual Feature Fusion Module

Deep semantic features lack fine details and spatial information, while high-resolution shallow features provide texture cues.To enhance these critical details for accurate small target contour segmentation, we introduce the RFF module (Fig. 3).

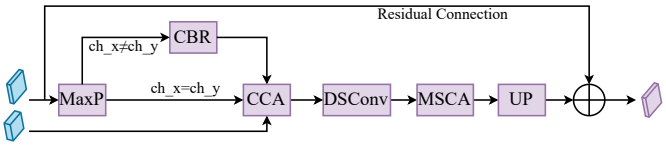

Fig. 3. This RFF model is mainly composed of CCA, DSConv and MSCA.

This module integrates two input features from adjacent layers in the encoder stage: $C \times H \times W$ and $C \times H/2 \times W/2$. Initially, the input $C \times H \times W$ is downsampled to $C \times H/2 \times W/2$ through Maxpooling. Subsequently, the two input features are fused via channel-wise cross attention (CCA) [5], and when channel numbers are misaligned, a 1×1 convolution ensures alignment. The fused features are subsequently refined through DSConv and the multi-spectral channel attention (MSCA) module [16], enabling the extraction and enhancement of critical frequency components. Thereafter, the enhanced features are upsampled to the original spatial dimensions of $C \times H \times W$ and combined element-wise with the residual features, yielding the optimized output $C \times H \times W$. Integrating features across layers enhances the model's ability to capture and refine target edge information. As a core component, the MSCA module uses the discrete cosine transform (DCT) to build multi-frequency channel responses, which compensates for information ignored by global average pooling (GAP) and improves the detection of low-contrast small targets. This process enhances high-frequency components in channel attention, preserving subtle target features.

D. Adaptive Fusion Upsampling Module

In the decoder, we introduce the AFU module to refine features. The AFU module replaces conventional upsampling with transposed convolution, which dynamically modulates feature generation via parameter learning to enhance detection accuracy. The input feature map is first expanded by transposed convolution. Then, CCA fuses this upsampled map with the output from the RFF module. The fused map is concatenated and refined by two DSConv layers to produce the final output. Finally, four feature maps are generated in the decoder, resized to $8 \times H \times W$ using keep, up, and upConv operations. These maps are fused by element-wise summation, with a 1×1 convolution refining the channel dimension to yield the final output. CCA compresses spatial information via global average pooling, models encoder-decoder channel-wise dependencies, and generates attention masks for adaptive recalibration—bridging the semantic gap to mitigate conflicts from direct concatenation and thereby improving small target detection while reducing false alarms.

III. EXPERIMENTS AND RESULTS

A. Experimental Settings

For our evaluation, we used two public datasets: SIRST [1] and IRSTD-1K [22]. For SIRST, 50% of the images were for training and the other half for testing. For IRSTD-1K, we allocated 80% of the images for training and 20% for testing. All experiments are conducted with an input image size of 256×256.

To quantitatively evaluate the performance of the proposed model, we employ three standard metrics: pixel-level Intersection over Union (IoU), Probability of Detection (Pd), and False Alarm rate (Fa).

- **Intersection over Union (IoU)**: Measures pixel-level overlap, $IoU = A_i/A_u$, where A_i and A_u are intersection and union areas.
- **Probability of Detection (Pd)**: Target-level metric, $Pd = N_{pred}/N_{all}$, where N_{pred} is correctly detected targets and N_{all} is total actual targets. Detection criterion: Euclidean distance between predicted and ground truth centers < 3 pixels (standard IRSTD protocol).
- **False Alarm Rate (Fa)**: Pixel-level ratio, $Fa = P_{false}/P_{all}$, where P_{false} is false positives and P_{all} is total pixels. Reported in 10^{-6} units for small-value comparison.

All experiments are performed on a NVIDIA GeForce RTX 3090 GPU (24GB), using PyTorch. The model was trained with the SoftIoU loss function and optimized by the Adam optimizer for 1000 epochs with a batch size of 16.

B. Experimental Comparison

This section compares LWMCNet with state-of-the-art IRSTD methods. As shown in Table I, existing methods face a trade-off: high-accuracy models achieve better results but with

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON SIRST AND IRSTD-1K DATASETS. THE BEST-PERFORMING VALUES IN THE HEAVYWEIGHT, SEMI-LIGHTWEIGHT, AND ULTRA-LIGHTWEIGHT CATEGORIES ARE HIGHLIGHTED IN THEIR RESPECTIVE COLORS.

Category	Method	Param(M)	FLOPs(G)	SIRST			IRSTD-1K		
				IoU(%)↑	Pd(%)↑	Fa(10^{-6})↓	IoU(%)↑	Pd(%)↑	Fa(10^{-6})↓
Heavy	SeRankDet [4] (2024)	108.89	140.71	**81.27**	**100.0**	**1.86**	**73.66**	**94.28**	**8.37**
	UIUNet [3] (2022)	50.54	54.43	76.25	95.06	24.35	64.37	92.93	18.62
	HintU [17] (2024)	31.04	54.66	75.58	94.68	34.23	63.25	92.60	26.57
	SCTransNet [5] (2024)	11.19	10.12	77.50	96.95	13.92	68.03	93.27	10.74
Semi	FC3Net [18] (2024)	6.90	10.57	74.22	**99.12**	**6.57**	64.98	92.93	**15.73**
	DNANet [2] (2022)	4.70	14.26	75.55	95.06	20.85	64.23	90.90	17.40
	ARFC-WAHNet [19] (2025)	3.46	5.86	74.29	96.96	33.08	64.92	**93.94**	26.61
	CSPENet [20] (2025)	1.44	9.49	**79.83**	96.96	15.18	**66.79**	93.27	23.53
Ultra	ALCNet [8] (2021)	0.43	0.38	70.83	94.30	36.15	60.60	92.98	58.80
	ACM [1] (2021)	0.40	0.40	68.93	91.63	15.23	59.23	**93.27**	65.28
	RDIAN [21] (2022)	0.22	3.72	75.56	95.44	19.41	59.63	87.88	54.58
	Ours (LWMCNet)	**0.21**	**0.44**	**76.62**	**96.96**	**11.59**	**70.25**	92.93	**17.46**

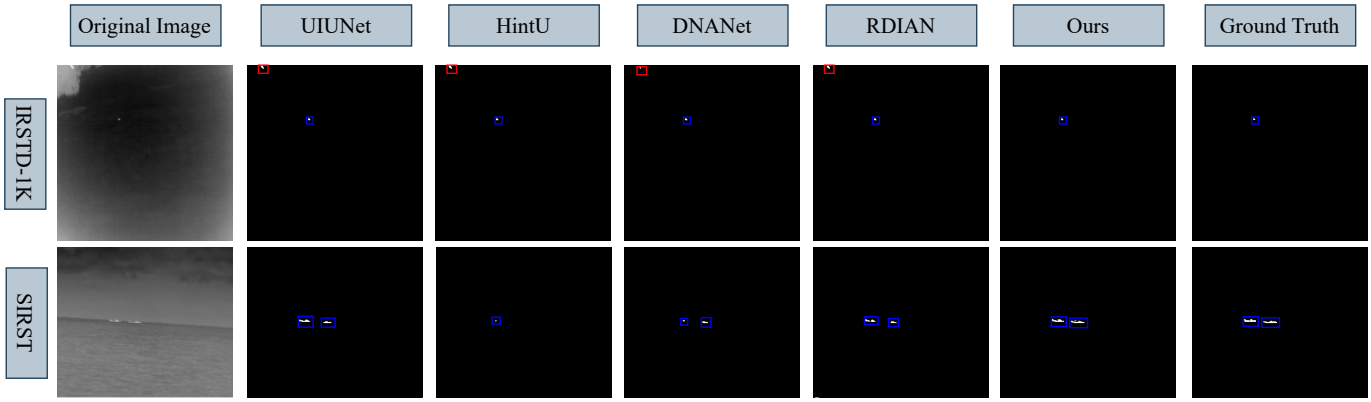

Fig. 4. Visual results obtained from various IRSTD methods on the SIRST and IRSTD-1K datasets are presented. The blue and red rectangle boxes represent correctly detected targets and false alarms, respectively.

TABLE II

ABLATION STUDY ON THE CONTRIBUTION OF LMFA, RFF, AND AFU MODULES ON THE IRSTD-1K DATASET. THE SYMBOL '✓' INDICATES THE MODULE IS UTILIZED, WHILE '✗' INDICATES IT IS NOT UTILIZED.

Method	LMFA	RFF	AFU	IoU(%)↑	Pd(%)↑	Fa($\times 10^{-6}$)↓
Exp1	✗	✗	✗	60.68	91.92	45.66
Exp2	✓	✗	✗	66.33	92.26	26.78
Exp3	✗	✓	✗	61.69	93.27	46.84
Exp4	✗	✗	✓	63.28	93.27	43.69
Exp5	✗	✓	✓	64.20	91.92	42.57
Exp6	✓	✓	✗	66.32	93.93	34.24
Exp7	✓	✗	✓	68.72	**93.94**	22.40
Exp8	✓	✓	✓	**70.25**	92.93	**17.46**

heavy overhead, while lightweight ones often underperform in complex cases. For fair benchmarking under deployment constraints, we group the models into three categories based on their parameter count and FLOPs($\wedge$ means and, $\vee$ means or):

- Heavyweight models (parameters $>$10M $\wedge$ FLOPs $>$10G)
- Semi-lightweight models (parameters $<$10M $\vee$ FLOPs $<$10G)
- Ultra-lightweight models (parameters $<$1M $\wedge$ FLOPs $<$1G)

Designed for ultra-lightweight applications, LWMCNet has only 0.21M parameters and 0.44G FLOPs. Located at the top-left corner of the Pareto front in Fig. 5, LWMCNet achieves an optimal efficiency–accuracy trade-off with 70.25% IoU on IRSTD-1K, demonstrating performance close to that of heavyweight models at an extremely low computational cost. It surpasses mainstream lightweight methods while yielding performance close to that of heavyweight models, making it suitable for resource-constrained edge platforms such as

drone-based infrared surveillance. Visual results in Fig. 4 further validate its accurate localization and strong background clutter suppression.

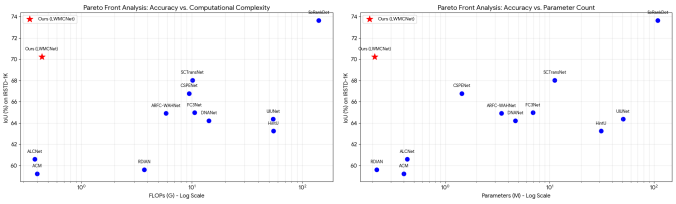

Fig. 5. Pareto-front analysis of detection accuracy versus model complexity on the IRSTD-1K dataset. Left: IoU (%) vs. FLOPs (G). Right: IoU (%) vs. Parameters (M). The x-axis is presented in log scale to clearly visualize the performance across different model scales. Our LWMCNet occupies the top-left corner, demonstrating a superior trade-off between accuracy and efficiency.

C. Ablation Experiments

1) Module Ablation Study: To verify the contribution of each component, we conduct extensive ablation experiments on the IRSTD-1K dataset using UNet as the baseline (Exp1), with results summarized in Table II. Individual module evaluation (Exp2-4) demonstrates that each component consistently enhances detection performance, with the LMFA module yielding the most significant gain, improving IoU from 60.68% to 66.33%. This highlights the necessity of multi-scale feature attention in capturing dim targets. Synergy analysis (Exp5-7) reveals that integrating RFF and AFU with LMFA substantially outperforms their independent applications, confirming the architectural compatibility of our design. Notably, the robust feature extraction provided by LMFA serves as a foundational stage, ensuring that spatial details fused by the RFF module remain discriminative and intact. Finally, the full integration (Exp8) achieves the optimal balance, reaching a peak IoU of 70.25%.

2) Effect of CFE and RVEA Quantity in CFE: As illustrated in Fig. 2, the LMFA module consists of three sub-modules: RVEA, EMA, and CFE, where CFE integrates multiple RVEA modules. The effectiveness of the CFE strategy is validated through two comparative experiments. First, as shown in Table III, CFE improves IoU and reduces Fa compared to using only RVEA and EMA. Second, we analyze the impact of RVEA module count within CFE to determine its optimal configuration, as shown in Table IV. The results indicate that integrating three RVEA modules yields the best performance, whereas further aggregation may degrade feature extraction due to redundancy or background interference. The inherent challenges of IRSTD also limit deeper extraction, as target features are easily obscured by the background.

TABLE III
VERIFY THE VALIDITY OF CFE ON THE IRSTD-1K DATASET.

Method	IoU(%)↑	Pd(%)↑	Fa($\times 10^{-6}$)↓
LMFA without CFE	68.16	**93.94**	24.43
LMFA with CFE	**70.25**	92.93	**17.46**

TABLE IV
EFFECT OF THE NUMBER OF RVEA MODULES IN THE CFE ON THE IRSTD-1K DATASET (256×256).

Number	Params (M)↓	FLOPs (G)↓	IoU (%)↑	Pd P_d(%)↑	Fa ($\times 10^{-6}$) ↓
$m = 1$	**0.17**	**0.38**	66.81	92.93	42.47
$m = 2$	0.19	0.41	67.39	90.91	24.41
$\mathbf{m = 3}$	0.21	0.44	**70.25**	**92.93**	**17.46**
$m = 4$	0.24	0.47	68.29	92.26	22.66

D. Efficiency Experiments

1) Efficiency Analysis: To evaluate the real-time performance of the proposed LWMCNet, we benchmark its inference speed on an NVIDIA GeForce RTX 3090 GPU (24GB). All FPS results are measured with batch size 1 after warm-up and averaged over multiple runs on the same hardware settings, excluding data loading and disk I/O. As summarized in Table V, while our method does not achieve the highest raw FPS among the compared state-of-the-art competitors, it comfortably satisfies the real-time processing requirements ($\sim$60 FPS at 256×256 on RTX 3090 GPU) while delivering superior detection accuracy (70.25% IoU).

Interestingly, its inference speed is lower than expected given its minimal parameters and FLOPs. To investigate this bottleneck, we conduct a component-wise latency analysis in Table VI. The results pinpoint the 'WTConv' and 'Attention (CA, SA)' modules within LMFA as the primary contributors to computational latency. This suggests a justifiable trade-off: a modest reduction in throughput yields substantial gains in sophisticated feature aggregation and target discrimination.

TABLE V
REAL-TIME PROCESSING CAPABILITIES OF DIFFERENT MODELS ON THE IRSTD-1K DATASET. THE IMAGE SIZE IS SET TO 256×256, AND FPS IS MEASURED ON NVIDIA GEFORCE RTX 3090 (24GB).

Model	FPS↑	IoU(%) ↑	Pd(%) ↑	Fa($\times 10^{-6}$) ↓
DNANet [2]	89.47	64.23	90.90	17.40
RDIAN [21]	**465.71**	59.63	87.88	54.58
UIUNet [3]	73.48	64.37	92.93	18.62
SCTransNet [5]	57.20	68.03	**93.27**	**10.74**
Ours	<u>59.50</u>	**70.25**	<u>92.93</u>	17.46

2) Deployment Potential on Edge Platforms: To evaluate the practical utility of LWMCNet in real-world scenarios, we conduct a comprehensive deployment benchmark on NVIDIA Jetson NX, a representative resource-constrained edge platform. As summarized in Table VII, LWMCNet demonstrates exceptional portability and scaling efficiency across hardware architectures and input resolutions. At 256×256 resolution, it achieves 29.70 FPS on Jetson NX—nearly double the speed of state-of-the-art methods such as DNANet (16.32 FPS) and UIUNet (13.46 FPS). This framerate enables fluid, real-time target tracking in embedded infrared surveillance systems.

TABLE VI

COMPONENT-WISE LATENCY AND COMPLEXITY ANALYSIS ON THE IRSTD-1K DATASET (256×256). THIS ABLATION STUDY VALIDATES THE IMPACT OF WTCONV, ATTENTION (CA, SA), AND RFF MODULES ON MODEL EFFICIENCY. '✓' INDICATES THAT THIS COMPONENT IS USED, WHILE '✗' INDICATES THAT THIS COMPONENT IS NOT USED.

Components			**Metrics**		
WTConv	CA, SA	RFF	Params(M)↓	FLOPs(G)↓	FPS↑
✓	✓	✓	0.21	0.44	59.50
✗	✓	✓	0.18	0.39	80.50
✓	✗	✓	0.20	0.43	62.37
✓	✓	✗	0.18	0.41	50.49
✗	✗	✓	0.17	0.38	110.48
✗	✓	✗	0.14	0.35	83.43
✓	✗	✗	0.17	0.40	68.27

The advantage becomes more pronounced at 512×512 resolution. Under this demanding configuration, LWMCNet sustains 9.77 FPS, whereas DNANet and UIUNet drop to approximately 4 FPS—well below the threshold for effective real-time monitoring. While RDIAN yields higher raw FPS, its detection performance is substantially inferior, with IoU 10.62% lower than ours (59.63% vs. 70.25%). Similarly, although SCTransNet achieves a competitive false alarm rate, its IoU and P_d remain lower than those of our architecture. These comparisons underscore that LWMCNet provides a superior accuracy-to-latency trade-off: it is lightweight enough to maintain practical throughput on embedded hardware without sacrificing the detection sensitivity required for small, dim infrared targets. This consistent performance lead across both server-grade GPUs and edge devices confirms the robustness of LWMCNet for onboard deployment in complex environments.

TABLE VII

EFFICIENCY COMPARISON ACROSS DIFFERENT INPUT RESOLUTIONS AND HARDWARE PLATFORMS. RESULTS HIGHLIGHT THE BALANCE BETWEEN ACCURACY AND REAL-TIME THROUGHPUT.

Model	**FPS (RTX 3090)** ↑		**FPS (Jetson NX)**↑		**IoU**	**Pd**	**Fa**
	256^2	512^2	256^2	512^2	**(%)** ↑	**(%)** ↑	$(\times 10^{-6})$ ↓
DNANet [2]	89.47	44.86	16.32	4.57	64.23	90.90	17.40
RDIAN [21]	**465.71**	**191.39**	**67.25**	**19.72**	59.63	87.88	54.58
UIUNet [3]	73.48	39.07	13.46	4.36	64.37	92.93	18.62
SCTransNet [5]	57.20	44.87	24.29	8.33	68.03	**93.27**	**10.74**
Ours	59.50	49.62	<u>29.70</u>	<u>9.77</u>	**70.25**	<u>92.93</u>	17.46

IV. CONCLUSION

This paper proposes a LWMCNet model to address the trade-off between detection accuracy and computational efficiency. By introducing the LMFA module and the CFE strategy, it effectively suppresses the disappearance of weak target features in deep networks; at the same time, by combining RFF and AFU, it achieves robust cross-layer interaction and precise feature reconstruction.

Experiments on SIRST and IRSTD-1K show that LWMCNet achieves 76.62% and 70.25% IoU, respectively, reaching new performance levels. With only 0.21M parameters and 0.44G FLOPs (at 256×256 resolution), it attains 29.70 FPS on NVIDIA Jetson NX, demonstrating efficient multi-scale fusion under constrained hardware conditions.

REFERENCES

[1] Y Dai, Y Wu, F Zhou, and K Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proceedings of WACV*, 2021, pp. 950–959.

[2] B Li, C Xiao, L Wang, Y Wang, Z Lin, M Li, W An, and Y Guo, "Dense nested attention network for infrared small target detection," *IEEE TIP*, vol. 32, pp. 1745–1758, 2022.

[3] X Wu, D Hong, and J Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE TIP*, vol. 32, pp. 364–376, 2022.

[4] Y Dai, P Pan, Y Qian, Yand Li, X Li, J Yang, and H Wang, "Pick of the bunch: Detecting infrared small targets beyond hit-miss trade-offs via selective rank-aware attention," *IEEE TGRS*, vol. 62, pp. 1–15, 2024.

[5] S Yuan, H Qin, X Yan, N Akhtar, and A Mian, "Sctransnet: Spatial-channel cross transformer network for infrared small target detection," *IEEE TGRS*, 2024.

[6] T Chen, Z Tan, Q Chu, Y Wu, B Liu, and N Yu, "Tci-former: Thermal conduction-inspired transformer for infrared small target detection," in *Proceedings of AAAI*, 2024, vol. 38, pp. 1201–1209.

[7] Q Mao, Q Li, B Wang, Y Zhang, T Dai, and C Chen, "Spirdet: towards efficient, accurate and lightweight infrared small target detector," *IEEE TGRS*, vol. 62, pp. 1–12, 2024.

[8] Y Dai, Y Wu, F Zhou, and K Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE TGRS*, vol. 59, no. 11, pp. 9813–9824, 2021.

[9] R Kou, C Wang, Y Yu, Z Peng, M Yang, F Huang, and Q Fu, "Lw-irstnet: Lightweight infrared small target segmentation network and application deployment," *IEEE TGRS*, vol. 61, pp. 1–13, 2023.

[10] O Ronneberger, P Fischer, and T Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI*, 2015, pp. 234–241.

[11] Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang, "Efficient multi-scale attention module with cross-spatial learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[12] F Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of CVPR*, 2017, pp. 1251–1258.

[13] S Finder, R Amoyal, E Treister, and O Freifeld, "Wavelet convolutions for large receptive fields," in *Proceedings of ECCV*, 2025, pp. 363–380.

[14] S Woo, J Park, J Lee, and I Kweon, "Cbam: Convolutional block attention module," in *Proceedings of ECCV*, 2018, pp. 3–19.

[15] R Sunkara and T Luo, "No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects," in *Proceedings of ECML-PKDD*, 2022, pp. 443–459.

[16] Z Qin, P Zhang, F Wu, and X Li, "Fcanet: Frequency channel attention networks," in *Proceedings of ICCV*, 2021, pp. 783–792.

[17] W Quan, Wand Zhao, W Wang, H Xie, and M Wang, Fand Wei, "Lost in unet: Improving infrared small target detection by underappreciated local features," *arXiv:2406.13445*, 2024.

[18] M Zhang, K Yue, J Zhang, Y Li, and X Gao, "Exploring feature compensation and cross-level correlation for infrared small target detection," in *Proceedings of MM*, 2022, pp. 1857–1865.

[19] X Cui, J Luo, J Deng, K Li, X Qiu, and Z Peng, "Arfc-wahnet: Adaptive receptive field convolution and wavelet-attentive hierarchical network for infrared small target detection," *arXiv:2505.10595*, 2025.

[20] J Deng, K Li, X Cui, J Li, C Long, T Pu, and Z Peng, "Cspenet: Contour-aware and saliency priors embedding network for infrared small target detection," *arXiv:2505.09943*, 2025.

[21] H Sun, J Bai, F Yang, and X Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset irdst," *IEEE TGRS*, vol. 61, pp. 1–13, 2023.

[22] M Zhang, R Zhang, Y Yang, H Bai, J Zhang, and J Guo, "Isnet: Shape matters for infrared small target detection," in *Proceedings of CVPR*, 2022, pp. 877–886.