RHNet: Residue Channel Prior-guided High-Low Feature Collaborative Network for Infrared Small Target Detection

1st Ziheng Fang
Department of Computer Science and Engineering
East China University of Science and Technology
Shanghai, China
20002321@mail.ecust.edu.cn

2nd Qian Zhang*
Department of Computer Science and Engineering
East China University of Science and Technology
Shanghai, China
qianzhang@ecust.edu.cn

3rd Yunfei Tong
Department of Computer Science and Engineering
East China University of Science and Technology
Shanghai, China
10182415@mail.ecust.edu.cn

4th Zhe Wang*
Department of Computer Science and Engineering
East China University of Science and Technology
Shanghai, China
wangzhe@ecust.edu.cn

***Abstract*—Following the detection-by-segmentation paradigm, U-Net and its variants have attained competitive performance in Infrared Small Target Detection (IRSTD) benchmarks. However, as typical CNNs relying on pure spatial-domain convolution stacking, they face limitations in utilizing differentiated high-low feature (semantic-spatial) collaboration and lack a global residue channel view. Such neglect of high-low feature synergy, coupled with the absence of residue channel-aware learning and redundant computations from excessive convolution stacking, induces representation bias that impairs detection reliability. This paper proposes a Residue Channel Prior-guided High-Low Feature Collaborative Network (RHNet), where Multi-Scale Feature Enhancement (MSFE) module and Prior-Guided Feature Boost (PGFB) module are advised to tackle the aforementioned issues. The low-level MSFE-D integrates bidirectional multi-scale and dilated convolutions to capture fine-grained contour details. The high-level MSFE-S employs saliency kernels to extract discriminative semantic information. The PGFB module leverages residue channel prior to guide feature optimization, compensating for the lack of global residue channel perspective and replacing redundant convolutions to suppress false responses. Extensive experiments on IRSTD-1K, NUAA-SIRST, and NUDT-SIRST demonstrate that RHNet outperforms 26 state-of-the-art methods with an IoU of 70.29% on IRSTD-1K, while maintaining an ultra-lightweight architecture (3.79M parameters).**

***Index Terms*—Multi-scale representation, residue channel prior, feature extraction, infrared small target detection**

*Corresponding authors: Qian Zhang (qianzhang@ecust.edu.cn), Zhe Wang (wangzhe@ecust.edu.cn).

This work was supported by the Natural Science Foundation of China under Grant No. 62476087 and 62306115, Shanghai Municipal Education Commission's Initiative on Artificial Intelligence-Driven Reform of Scientific Research Paradigms and Empowerment of Discipline Leapfrogging, and the Natural Science Foundation of Shanghai under the 2024 Shanghai Action Plan for Science, Technology and Innovation (No. 24ZR1416800).

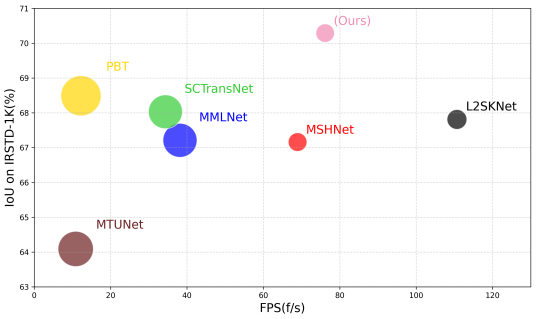

Fig. 1. Visualization of the detection performance (IoU), Frames Per Second (FPS) as well as Floating-Point Operations Per Second (area of circles) of some deep learning-based methods.

I. Introduction

Infrared small target detection (IRSTD) serves as a critical technology with extensive applications across military, aerospace exploration, and forest security domains. Owing to the fact that visible images are highly sensitive to light intensity fluctuations, IRSTD has gained remarkable research interest. However, IRSTD remains an intractable challenge due to inherent target characteristics: small targets typically occupy less than 50 image pixels [1], lacking distinct grayscale features, and exhibit substantial size variations across scenarios. These traits make targets easily obscured by complex backgrounds such as cloud clutter [2], ground textures, and noise, severely hindering the accuracy and robustness of detection algorithms. Addressing this dilemma is crucial to meeting the escalating demands for reliable target detection in practical applications.

Conventional model-driven approaches primarily resort to

filtering [3], local contrast [4], and low-rank representation [5] techniques to achieve background suppression. In addition, limited-data image synthesis has also been systematically reviewed to support data-scarce visual tasks [6]. These learning-free methods feature favorable computational accessibility, yet they have inherent limitations in practical applications. Specifically, filtering and local contrast methods depend on hand-crafted features, which makes it challenging to capture the intricate characteristics of complex background scenes comprehensively. For low-rank representation methods, though they exhibit adaptability to infrared images with low signal-to-noise ratios, their practical deployment is hindered by relatively high computational complexity and the tendency to generate false alarm detections.

CNN-based deep learning methods dominate IRSTD, with most studies adopting U-Net-inspired encoder-decoder architectures under the detection-by-segmentation paradigm. ISNet [7] refines the calculation of local pixel grayscale variations to effectively capture the edge and shape features of small targets, demonstrating superior detection performance compared to traditional CNNs in low signal-to-noise ratio (SNR) scenarios. DNANet [8] enhances the model's adaptability to targets with dynamic sizes by facilitating repeated interaction and information complementation across cross-layer features. ACMNet [2] introduces an asymmetric context modulation strategy, achieving the collaborative optimization of "background suppression and target enhancement". Beyond vision tasks, multi-omics data fusion has also been widely used for disease-related gene discovery [9].

Additionally, leveraging the Gaussian distribution characteristics inherent to infrared small targets, several studies focus on optimizing deep kernels in convolutional networks. L^2SKNet [10] designs a learnable saliency kernel that subtracts surrounding pixel values from the central pixel, while Yang $et\ al.$ [11] propose a pinwheel-shaped convolution kernel to sample features from four directions, thereby enhancing discriminative feature extraction for small targets.

Despite the impressive performance and significant progress of CNN-based methods, two critical limitations remain insufficiently addressed. 1) U-Net architectures fail to adequately exploit differentiated high-low feature characteristics [12]. As low-level features emphasize target detail structures while high-level features prioritize semantic discrimination between targets and backgrounds, yet existing methods adopt homogeneous extraction strategies and neglect effective fusion resulting in low-level features lacking sufficient receptive field size and multi-directionality and high-level features failing to fully exert semantic extraction advantages. 2) CNN methods over-rely on convolution stacking for feature enhancement, ignoring the effective integration of prior information, efficient fusion of multi-level features and guidance from knowledge based on the residue channel characteristics of image channels. Redundant convolutions not only fail to significantly improve feature representation but also incur unnecessary computational overhead, compromising the models' detection efficiency and practical applicability.

To address the aforementioned issues, this paper proposes a Residue Channel Prior-guided High Low Feature Collaborative Network (RHNet). The network integrates Multi-Scale Feature Enhancement (MSFE) modules for differentiated high and low-level feature extraction, complemented by Prior-Guided Feature Boost (PGFB) modules that leverage residue channel prior to guide feature optimization. The MSFE module is equipped with a differentiated convolution kernel tailored to the needs of high and low-level feature extraction. The low-level MSFE encoder integrates custom-designed bidirectional multi-scale convolution kernels and dilated convolution kernels to capture high-resolution target contour information, effectively addressing the limitations of traditional convolutions in directional perception and scale coverage. For the high-level MSFE encoder, we adopt learnable kernels based on the Gaussian distribution characteristics of infrared small targets to accurately extract target semantic information. The PGFB module abandons simple convolutional layer stacking and introduces residue channel prior for feature extraction assistance. This prior information is derived from the multi-channel feature maps output by the MSFE module. The PGFB module first performs multi-scale convolution on the prior information map and then uses it to guide the concatenation map of upsampled features and the peer encoder features, achieving precise feature optimization. As shown in Fig. 1, by leveraging the aforementioned strategies, our RHNet achieves an IoU of 70.29% on the IRSTD-1k dataset while maintaining relatively high inference speed. Furthermore, it significantly reduces parameter costs and FLOPs compared to mainstream methods. In summary, the contributions of this paper are as follows.

- **Network design:** This paper proposes a novel neural network that exploits multi-receptive field perception for high-low feature characteristics and residue channel prior.
- **Differentiated convolution design for MSFE:** This paper proposes a MSFE module which adopts two types of learnable kernels for low/high levels to address ineffective feature collaboration.
- **PGFB module:** This paper proposes a PGFB module which eliminates redundant convolution stacking with residue channel prior and enhances feature optimization accuracy.

II. Proposed Method

A. Overall Structure

As is illustrated in Fig. 2, our RHNet contains a feature extraction subnetwork and a prior-guided feature boost subnetwork. First, the input image is fed into the feature extraction subnetwork, which is constructed by stacking four Multi-Scale Feature Enhancement-Dual (MSFE-D) modules and one Multi-Scale Feature Enhancement-Single (MSFE-S) module. This subnetwork leverages multi-scale convolution to effectively amplify the contrast between targets of various sizes and complex backgrounds, and it adopts distinct learnable kernel strategies tailored to high-level and low-level positions. Specifically, MSFE-D is designed for low-level

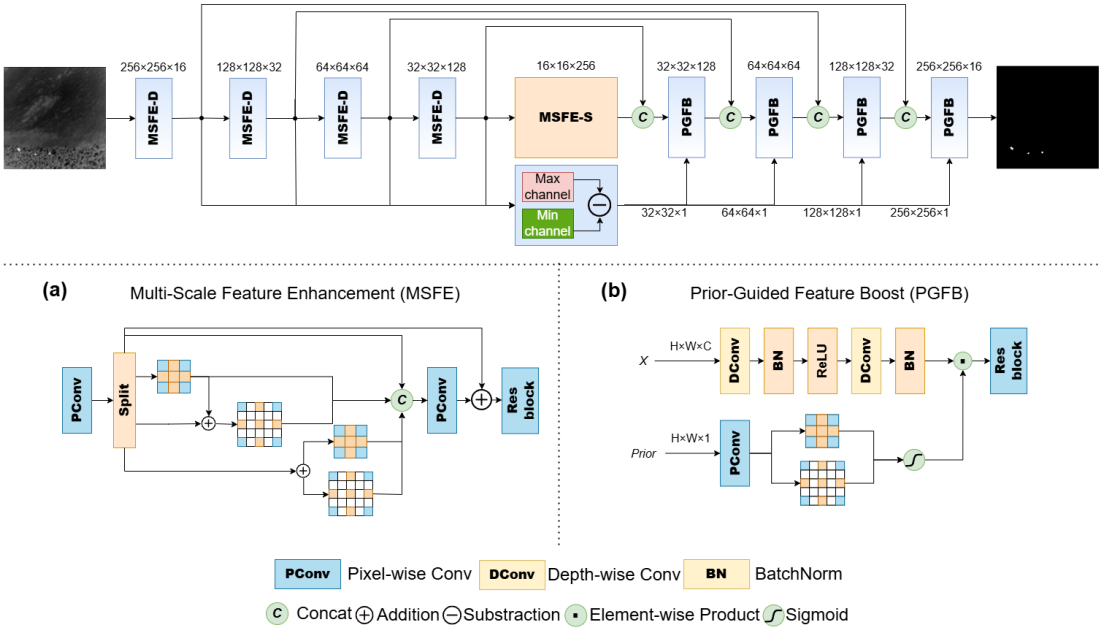

Fig. 2. Pipeline of our RHNet containing Multi-Scale Feature Enhancement (a,MSFE) and Prior-Guided Feature Boost (b,PGFB).

processing while MSFE-S serves high-level feature extraction. Second, the multi-channel feature maps output by the feature extraction subnetwork are processed to generate a single-channel residue channel prior map. Third, the prior-guided feature boost subnetwork, which is composed of four Prior-Guided Feature Boost (PGFB) modules, takes this prior map and the concatenation of upsampled decoder features and the peer encoder features as inputs. It uses the prior map to guide feature optimization throughout the decoding process. After sequential enhancement via four PGFB modules, this subnetwork outputs the final prediction map.

B. Multi-Scale Feature Enhancement Module

Fig. 2.a illustrates the detailed structure of the MSFE module. Firstly, the module enhances the channel dimension of the input feature map via 1×1 convolution, and then partitions the feature map X with enhanced channels into four feature subsets $\{X_0, X_1, X_2, X_3\}$ along the channel dimension. Among them, X_0 is directly connected to the output layer and outputs as Y_0; the remaining feature subsets X_1, X_2, and X_3 are fed into their respective independent branches for feature processing, and finally output the corresponding features Y_1, Y_2, and Y_3. The specific operation formulas of each branch are as follows:

$$Y_i = \begin{cases} X_i, & i = 0 \\ \text{Conv}(X_i), & i = 1 \\ \text{Dilated Conv}(X_i + Y_{i-1}), & i = 2 \\ \text{Mixed Conv}(X_i + Y_{i-1}), & i = 3 \end{cases} \tag{1}$$

where Conv denotes the standard convolution operation, Dilated Conv represents the dilated convolution operation and Mixed Conv is a hybrid computation strategy that sums the outputs of standard convolution and dilated convolution. X_i denotes the i-th feature subset obtained by channel partitioning, and Y_i denotes the output feature of the i-th branch. The three convolution operations involved in the formulas have different specifications: Conv(s,d), Dilated Conv(s,d), and the standard convolution component in Mixed Conv shares the same specification as Conv(s,d), while the dilated convolution component shares the same specification as Dilated Conv(s,d), where "s" denotes the kernel size and "d" represents the dilation rate. Subsequently, the output features of all branches are concatenated along the channel dimension to form the feature $Y = \{Y_0, Y_1, Y_2, Y_3\}$, which is then passed through a 1×1 convolution for information mixing. A residual block is conducted on Y and X to obtain a contrast-enhanced feature map, using the following equation:

$$F_{\text{out}} = \text{Res\_block}(\text{Pconv}(Y) + X) \tag{2}$$

where Pconv refers to the 1×1 convolutional block. The Res_Block($\cdot$) consists of a 3×3 convolution, channel attention, and spatial attention.

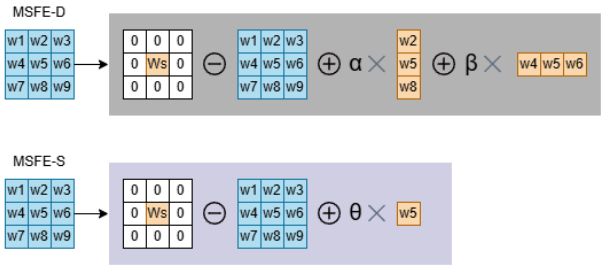

Fig. 3. Visualization of the decomposition of the custom-designed learnable kernels. α, β and θ are weighted parameters

The MSFE module is equipped with different learnable kernels based on its hierarchical position in the network, thereby deriving MSFE-D for low-level features and MSFE-S for high-level features.

Since MSFE-D operates at the low level, it focuses on capturing the detailed structure of the target. Specifically, it computes the Central Sum Weight by aggregating all weight elements of the 3×3 convolution kernel, and then generates the corresponding Central Sum Weight convolution feature map. By calculating the difference between this map and the original convolution feature map of the kernel, the gray-scale disparity between the target region and the background is perceived, which enhances local contrast and extracts salient features at the target center. Subsequently, the result is fused with the strip convolution feature maps W_{col}, W_{row} using weighted parameters to perceive the target's detailed structure from two directions.

$$W_{\text{finald}} = W'_{\text{sum}} - W' + \alpha \cdot W_{\text{col}} + \beta \cdot W_{\text{row}} \tag{3}$$

where W_{finald} denotes the final fused feature map of MSFE-D, W'_{sum} is the convolution feature map generated by the Central Sum Weight. W' is the original convolution feature map of the kernel, α and β are trainable weighted parameters that balance the contributions of W_{col} and W_{row}, W_{col} is the strip convolution feature map extracted by the middle column of the original kernel, and W_{row} is the strip convolution feature map extracted by the central row of the original kernel.

In contrast, MSFE-S operates at the high level and focuses on extracting pixel-wise semantic information. We apply the LLSKM [10] inside the MSFE-S. After obtaining the salient features of the target center in the same manner, MSFE-S fuses this result with the 1×1 convolution feature map W_{center} from the central position of the original kernel. This weighted fusion allows MSFE-S to capture the semantic information of the target pixels.

$$W_{\text{finals}} = W'_{\text{sum}} - W' + \theta \cdot W_{\text{center}} \tag{4}$$

where W_{finals} denotes the final fused feature map of MSFE-S, W'_{sum} is the convolution feature map generated by the Central Sum Weight. W' is the original convolution feature map of the kernel, θ is a trainable weighted parameter that modulates the contribution of W_{center}. W_{center} is a feature map extracted via a 1×1 convolution constructed from the central position of the original kernel, and it is specifically designed to capture pixel-wise semantic information of infrared small targets.

C. Prior-Guided Feature Boost Module

As illustrated in Fig. 2.b, the PGFB module accepts two core inputs: one is a single-channel prior map $Prior$ from the peer MSFE module; the other is the input feature map X, which is generated by concatenating the upsampled feature from the lower decoder and the feature from the peer encoder. For the input X, a serial processing pipeline is adopted to achieve feature refinement and dimension adaptation, yielding the intermediate feature X'

$$X' = \text{BN}(\text{Conv}_{3\times3}(\text{ReLu}(\text{BN}(\text{Conv}_{3\times3}(X))))) \tag{5}$$

For the prior map $Prior$, the number of channels is first adjusted via 1×1 Pixel-wise Convolution to match the channel dimension of the intermediate feature X'. Subsequently, multi-scale convolution is applied to extract multi-scale prior information, followed by the Sigmoid activation function to generate the channel attention weight W. The calculation formula is as follows:

$$W = \text{Sigmoid}(\text{Mixed Conv}(\text{Conv}_{1\times1}(Prior))) \tag{6}$$

where Mixed Conv is the same as its counterpart in the MSFE module, which is obtained by summing the results of the standard convolution and the dilated convolution.

Finally, the intermediate feature X' is element-wise multiplied by the attention weight W, and the result is fed into a residual block for feature enhancement to obtain the output F_{out} of the PGFB module. This realizes the accurate guidance of prior knowledge on feature learning.

$$F_{\text{out}} = \text{Res\_block}(X' \odot W) \tag{7}$$

The Res_Block$(\cdot)$ consists of a 3×3 convolution, channel attention, and spatial attention.

To optimize the network performance by quantifying the similarity between the final prediction map and the ground truth, we adopt the scale and location sensitive (SLS) loss [13] $\mathcal{L}_{\text{SLS}}$ that combines scale-sensitive loss $\mathcal{L}_S$ and location-sensitive loss $\mathcal{L}_L$.

III. Experiments And Results

A. Experimental Settings

We conduct our experiments on three widely-adopted public datasets: IRSTD-1k [7], NUDT-SIRST [8], and SIRST [2], which contain 1001, 427, and 1327 IRIs, respectively. Following existing works [7], [14], the images in NUAA-SIRST and NUDT-SIRST are divided equally into training sets and testing sets, while the images in IRSTD-1k are divided into training sets and testing sets with a 4:1 ratio.

Our Experiments are constructed under the PyTorch framework and run on a single RTX 2080ti GPU with 12GB of memory. Then, we augment the training data by flipping and rotation. Following existing works, we resize the input image to 256×256 for both training and testing. The model is trained for 800 epochs with a batch size of 4, utilizing the AdaGrad optimizer. The initial learning rate is set to 0.05 and is adjusted using the CosineAnnealingLR Learning Rate Scheduler with a period of 50 epochs, allowing the learning rate to decline smoothly to 0.045 over each period.

B. Ablation Study

In this section, we conduct ablation studies to validate the effectiveness of our proposed modules on the SIRST and NUDT-SIRST datasets. The results are reported in Table I.

Impact of MSFE: As is depicted by Table I, the integration of the MSFE module brings substantial performance gains to the baseline model, with remarkable improvements in IoU and significant reduction in F_a. By contrast, removing the MSFE module leads to a noticeable drop in IoU, highlighting

TABLE I
ABLATION STUDY OF KEY COMPONENTS

MSFE	PGFB	SIRST			NUDT-SIRST		
		IoU↑	Pd↑	Fa↓	IoU↑	Pd↑	Fa↓
		75.81	99.08	38.14	77.36	95.45	17.37
✓		78.39	100	9.75	86.45	98.73	5.26
	✓	78.08	100	10.82	85.93	98.84	5.10
✓	✓	**79.36**	**100**	**7.62**	**87.01**	**99.15**	**3.01**

the critical role of leveraging differentiated high-low feature characteristics. For low-level features, the module injects sufficient scale coverage and directional perception into receptive fields; for high-level features, it capitalizes on the Gaussian distribution inherent to infrared small targets to refine semantic information extraction. This tailored design effectively enhances the network's ability to capture target details and distinguish target from background, laying a solid foundation for detection accuracy.

Impact of PGFB: The performance of our basic model improves after adding the PGFB module, with a noticeable increase in P_d. When the PGFB module is excluded, all of the three evaluation metrics degrade evidently. By incorporating residue channel prior to guide feature optimization at each network level, the module replaces redundant convolution stacking, which not only reduces unnecessary computation but also strengthens the integration of multi-scale and cross-layer information. This mechanism enables the network to focus more on target-related features, suppressing false alarms and ensuring stable and reliable detection performance.

C. Quantitative Analysis

In this section, we conduct a quantitative comparison of our method with 26 SOTA methods. We compare our method with traditional, CNN-based and hybrid methods on the three publicly available datasets. For a fair comparison, all predicted maps are collected from publicly available results provided by the authors or computed by running the source code released by the authors.

As shown in Table II, RHNet yields superior performance on IRSTD-1k and SIRST, attaining the highest IoU and P_d. On NUDT-SIRST, it achieves the highest P_d alongside the lowest F_a. Across all nine evaluation metrics, our RHNet delivers six top performances and one second-best result. Notably, it achieves a 100% probability of detection on SIRST. While RHNet lags slightly behind SCTransNet in terms of IoU on the NUDT-SIRST dataset, our method outperforms it on the other two datasets. In comparison to HDNet, although our method exhibits a marginally higher F_a, our RHNet outperforms it in IoU and P_d. Additionally, to validate the superiority of our proposed high-low level learnable kernel update strategy, we conduct comparative experiments on the IRSTD-1k dataset. The high-low level strategy in our framework is replaced by fixed-weight kernels or learnable kernels with different update strategies. Results of different strategies are shown in Table III. Specifically, the fixed-weight kernel adopts a 3×3 configuration with a central value of 1 and surrounding values of -1/8. LLSKM [10] shares the same update strategy as MSFE-S. Pinwheel [11] is a convolutional approach that extracts feature maps via four strip-shaped convolutions. Experimental results demonstrate that our proposed high-low level learnable kernel strategy achieves the best performance.

D. Computational Analysis

Table IV reports the computational efficiency of recent model in terms of quantitative metrics. Our RHNet not only achieves superior detection performance, but also remains lightweight with 3.79M parameters and 5.99M FLOPs, while maintaining a high inference speed of 76.18 frames per second. This efficiency gain is attributed to our PGFB module, which leverages prior knowledge to replace complex computational operations, thus effectively reducing the model complexity and boosting the inference speed.

E. Visualization Analysis

As is depicted by Fig. 4, we conduct visualization experiments on five images sampled from the three datasets. These test images cover a wide range of complex background scenarios, where infrared small targets are faint and highly blended with the background. While state-of-the-art models such as SCTransNet and L^2SKNet yield favorable IoU values, they still suffer from noticeable false alarm issues. For instance, SCTransNet misclassifies complex background textures as targets, leading to false positives. In contrast, our RHNet exhibits superior performance in small target detection and false alarm reduction, excelling at accurately segmenting multiscale targets while suppressing background noise. This can be attributed to our MSFE module, which excels at suppressing background noise and accurately extracting target features. The PGFB module further consolidates and enhances these performance gains.

IV. CONCLUSIONS

This paper proposes RHNet, a prior-guided network addressing ineffective feature collaboration and redundant computations in IRSTD. Its core contributions lie in three aspects: the differentiated MSFE module resolving high-low feature collaboration issues, the PGFB module integrating residue channel prior to reduce overhead and enhance representation, and achieving an excellent accuracy-efficiency balance with state-of-the-art performance on the three datasets and ultra-lightweight architecture. Future work will focus on adaptive parameter adjustment for complex backgrounds, extending to video-based tasks, and optimizing the prior generation mechanism to further suppress distractors. RHNet offers valuable insight for lightweight small target detection with broad practical application prospects.

TABLE II

COMPARISONS WITH DIFFERENT SOTA DETECTORS ON IRSTD-1K, SIRST, NUDT-SIRST WITH IOU(%), PD(%),FA(10^{-6}). SIZE DENOTES SIZE OF INPUT IMAGE. METHODS ARE CATEGORIZED AS : TRAD(TRADITIONAL METHODS), CNN(DEEP CONVOLUTION METHODS), AND HYBRID(HYBIRD CNN AND OTHER METHODS)

Method	Size	Type	Year	IRSTD-1k			SIRST			NUDT-SIRST		
				IoU↑	Pd↑	Fa↓	IoU↑	Pd↑	Fa↓	IoU↑	Pd↑	Fa↓
Max-Median [15]	256	Trad	1999	6.70	65.21	59.73	6.02	84.34	774.30	4.20	58.41	36.89
Top-Hat [3]	256	Trad	2010	10.06	75.11	1432.00	1.51	79.74	16456.00	20.72	78.41	166.70
IPI [5]	256	Trad	2013	27.92	81.37	16.18	1.09	87.05	30467.00	17.76	74.49	41.23
RIPT [16]	256	Trad	2017	14.11	77.55	28.31	16.79	69.76	59.33	29.44	91.85	344.30
NRAM [17]	256	Trad	2018	15.25	70.68	16.93	15.25	70.68	16.93	6.93	56.40	19.27
PSTNN [18]	256	Trad	2019	24.57	71.99	35.26	30.30	72.80	48.99	14.85	66.13	44.17
MSLSTIPT [19]	256	Trad	2021	11.43	79.03	1524.00	1.08	0.05	8.18	8.34	47.40	888.10
TLLCM [4]	256	Trad	2020	3.31	77.39	6738.00	4.24	88.37	6243.00	2.18	62.01	1608.00
WSLCM [20]	256	Trad	2021	3.45	72.44	6619.00	6.39	88.74	4462.00	2.28	56.82	1309.00
MDvsFA [21]	256	CNN	2019	37.34	83.71	88.52	60.30	89.35	56.35	35.86	85.22	95.37
ALCNet [22]	256	CNN	2021	65.68	89.25	27.71	73.74	97.25	26.79	72.89	96.19	30.40
ACMNet [2]	256	CNN	2021	60.33	93.27	68.49	69.44	92.02	22.71	64.86	96.72	28.59
ISNet [7]	256	CNN	2022	61.85	90.24	31.56	70.49	95.06	67.98	81.24	97.78	6.34
DNANet [8]	256	CNN	2022	65.71	91.84	17.61	77.76	96.33	2.31	79.98	96.93	12.78
RKformer [23]	512	Hybrid	2022	64.12	93.27	18.65	77.24	99.11	1.58	-	-	-
ISTDU-Net [24]	512	CNN	2022	65.01	93.94	26.44	75.93	96.20	38.90	91.76	98.52	3.77
UIU-Net [25]	256	CNN	2023	68.69	91.25	13.48	77.53	92.40	9.33	75.91	96.83	18.61
RDIAN [26]	256	CNN	2023	59.94	87.21	33.31	70.74	95.06	48.16	82.42	96.72	14.85
MTUNet [1]	256	Hybrid	2023	64.09	90.48	12.15	74.85	99.08	7.09	77.98	96.08	17.51
MSHNet [13]	256	CNN	2024	67.16	93.88	15.03	73.50	97.25	31.05	80.55	97.99	11.77
GCI-Net [27]	512	CNN	2024	67.75	93.89	12.84	78.81	99.34	2.11	-	-	-
SCTransNet [28]	256	Hybrid	2024	68.03	93.27	10.74	77.50	96.95	13.92	**94.09**	98.62	4.29
PBT [29]	256	Hybrid	2024	68.49	92.52	8.88	78.39	99.08	2.13	83.89	97.23	4.23
L^2SKNet [10]	256	CNN	2025	67.81	90.24	17.46	73.43	98.17	20.82	93.58	97.57	5.33
MMLNet [30]	256	CNN	2025	67.21	94.28	14.00	78.71	98.88	25.71	81.81	98.43	11.77
HDNet [14]	256	Hybrid	2025	70.26	**94.56**	**4.33**	79.17	**100.00**	**0.53**	85.17	98.52	2.78
Ours	256	CNN	-	**70.29**	**94.56**	8.29	**79.36**	**100.00**	7.62	87.01	**99.15**	**2.45**

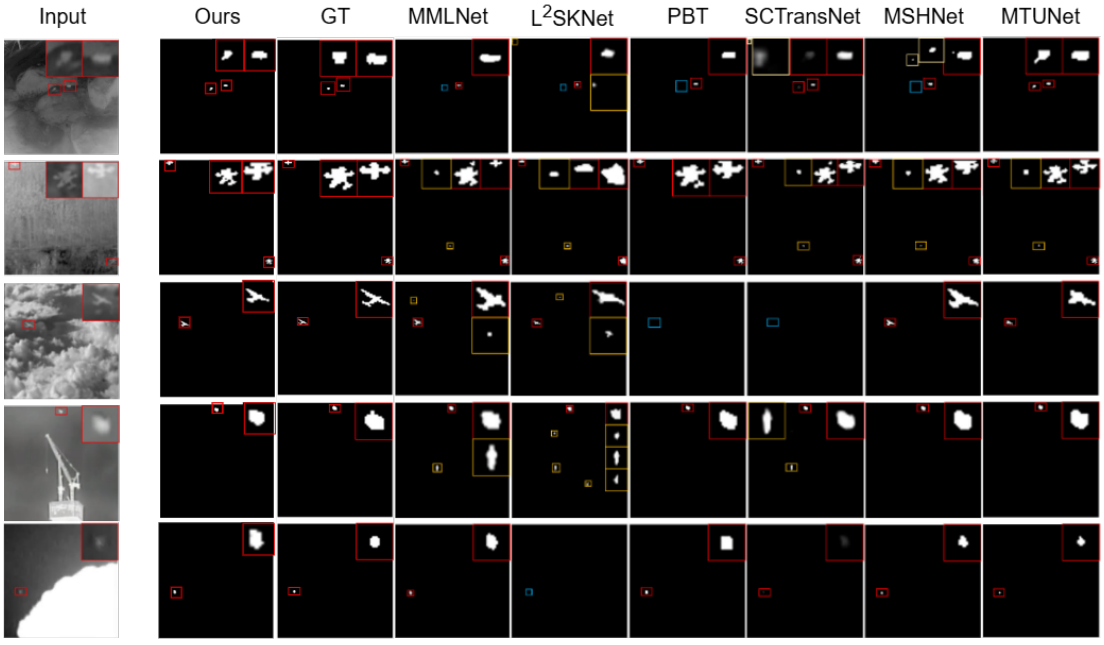

Fig. 4. Visualization results. The correctly detected targets, false alarms, and missed targets are framed by red, yellow, and blue bounding boxes, respectively.

TABLE III
COMPARISONS WITH DIFFERENT LEARNABLE KERNEL UPDATE STRATEGIES ON IRSTD-1K WITH IOU(%), PD(%),FA(10^{-6})

strategy	IoU↑	Pd↑	Fa↓
fixed weight	68.12	94.22	11.69
LLSKM	68.99	93.54	14.87
Pinwheel	66.47	93.54	18.14
Ours	**70.29**	**94.56**	**8.27**

TABLE IV
COMPARISONS OF MODEL COMPLEXITY WITH IOU(%), PARAMS(M), FLOPS(G), FPS(F/S)

Method	IoU(%)↑	Params(M)↓	FLOPs(G)↓	FPS(f/s)↑
UIU-Net [25]	68.69	50.54	54.43	9.87
RDIAN [26]	59.94	0.217	3.72	180.23
MTUNet [1]	64.09	12.75	21.94	10.87
MSHNet [13]	67.16	4.07	6.11	68.96
GCI-Net [27]	67.75	0.71	55.85	-
SCTransNet [28]	68.03	11.19	20.24	34.36
PBT [29]	68.49	26.29	28.53	12.25
L^2SKNet [10]	67.81	0.899	6.89	110.68
MMLNet [30]	67.21	3.58	20.41	38.17
HDNet [14]	70.26	3.84	5.96	78.81
Ours	70.29	3.79	5.99	76.18

REFERENCES

[1] T. Wu, B. Li, Y. Luo, Y. Wang, C. Xiao, T. Liu, J. Yang, W. An, and Y. Guo, "Mtu-net: Multilevel transunet for space-based infrared tiny ship detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[2] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 950–959.

[3] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognition*, vol. 43, no. 6, pp. 2145–2156, 2010.

[4] J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, and Q. Zhao, "A local contrast method for infrared small-target detection utilizing a tri-layer window," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 10, pp. 1822–1826, 2019.

[5] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 4996–5009, 2013.

[6] M. Yang and Z. Wang, "Image synthesis under limited data: A survey and taxonomy," *International Journal of Computer Vision*, vol. 133, no. 6, pp. 3689–3726, 2025.

[7] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "Isnet: Shape matters for infrared small target detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 877–886.

[8] 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 Transactions on Image Processing*, vol. 32, pp. 1745–1758, 2022.

[9] H. Yang, Z. Yang, L. Zhang, Y. Yang, D. Zhou, D. Li, J. Zhang, and Z. Wang, "Trans-driver: A deep learning approach for cancer driver gene discovery with multi-omics data," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 22, no. 6, pp. 3065–3076, 2025.

[10] F. Wu, A. Liu, T. Zhang, L. Zhang, J. Luo, and Z. Peng, "Saliency at the helm: Steering infrared small target detection with learnable kernels," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025.

[11] J. Yang, S. Liu, J. Wu, X. Su, N. Hai, and X. Huang, "Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9202–9210.

[12] X. Zhang, X. Zhang, S.-Y. Cao, B. Yu, C. Zhang, and H.-L. Shen, "Mrf3net: An infrared small target detection network using multireceptive field perception and effective feature fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[13] Q. Liu, R. Liu, B. Zheng, H. Wang, and Y. Fu, "Infrared small target detection with scale and location sensitivity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 490–17 499.

[14] M. Xu, C. Yu, Z. Li, H. Tang, Y. Hu, and L. Nie, "Hdnet: A hybrid domain network with multi-scale high-frequency information enhancement for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[15] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," in *Signal and Data Processing of Small Targets 1999*, vol. 3809. SPIE, 1999, pp. 74–83.

[16] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 10, no. 8, pp. 3752–3767, 2017.

[17] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, "Infrared small target detection via non-convex rank approximation minimization joint l 2, 1 norm," *Remote Sensing*, vol. 10, no. 11, p. 1821, 2018.

[18] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sensing*, vol. 11, no. 4, p. 382, 2019.

[19] Y. Sun, J. Yang, and W. An, "Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 3737–3752, 2020.

[20] J. Han, S. Moradi, I. Faramarzi, H. Zhang, Q. Zhao, X. Zhang, and N. Li, "Infrared small target detection based on the weighted strengthened local contrast measure," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 9, pp. 1670–1674, 2020.

[21] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8509–8518.

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

[23] M. Zhang, H. Bai, J. Zhang, R. Zhang, C. Wang, J. Guo, and X. Gao, "Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1730–1738.

[24] Q. Hou, L. Zhang, F. Tan, Y. Xi, H. Zheng, and N. Li, "Istdu-net: Infrared small-target detection u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

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

[26] 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[27] M. Zhang, K. Yue, B. Li, J. Guo, Y. Li, and X. Gao, "Single-frame infrared small target detection via gaussian curvature inspired network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[28] S. Yuan, H. Qin, X. Yan, N. Akhtar, and A. Mian, "Sctransnet: Spatial-channel cross transformer network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[29] H. Yang, T. Mu, Z. Dong, Z. Zhang, B. Wang, W. Ke, Q. Yang, and Z. He, "Pbt: Progressive background-aware transformer for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[30] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multi-branch mutual-guiding learning for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.