European Journal of AI, Computing & Informatics

Vol. 2 No. 1 2026

Article **Open Access**

ESE-Net: Edge-Shape Enhancement Network for Infrared Small Target Detection

Sen Yang [1] and Guanxun Cui [1,*]

[1] Chongqing University of Technology, Chongqing, China
* Correspondence: Guanxun Cui, Chongqing University of Technology, Chongqing, China

Received: 03 December 2025
Revised: 14 December 2025
Accepted: 06 January 2026
Published: 10 January 2026

Abstract: Infrared small target detection (ISTD) aims to segment small targets from infrared images and is widely applied in military and industrial fields. Although recent deep learning-based methods have achieved remarkable performance, they often fail when targets are indistinguishable from complex backgrounds. This is mainly due to the limited use of spatial domain features, which cannot capture subtle boundary cues, making precise segmentation challenging. To address this, we propose an Edge-Shape Enhanced Network (ESE-Net), which reinforces edge feature representations to improve target discrimination in complex infrared scenes. First, we design a Multiscale Spatial Edge Attention (MSEA) module to strengthen target edges by perceiving directional gradient changes. To suppress background noise while highlighting target boundaries, we introduce an Edge Guidance Module (EGM) that extracts edge features in the frequency domain via a wavelet transform and performs reversible down sampling, discarding low-frequency components before fusing with spatial features. Furthermore, a Multiscale Group Convolution Module (MGCM) is integrated in deep layers to preserve target details and mitigate the risk of small target loss. Experiments on the NUAA-SIRST and IRSTD-1K datasets demonstrate the effectiveness of our method.

Keywords: infrared small target detection; edge enhancement; spatial-frequency fusion; semantic segmentation

1. Introduction

Infrared small target detection (IRSTD) is a binary segmentation task that aims to generate a binary mask as output. Unlike optical imaging, infrared imaging relies on thermal radiation, often resulting in blurred target edges due to subtle radiometric transitions between targets and backgrounds. Compared with generic image segmentation, IRSTD faces additional challenges, such as extremely small target sizes, low target-background contrast, and heavy background noise.

Due to their low contrast and similarity to surrounding regions, infrared small targets can be easily overlooked by the human visual system. As illustrated in the top row of Figure 1, these targets exhibit dim textures and weak contrast, making their texture features poorly distinguishable. In contrast, shape features-particularly target edges-are more visually prominent and provide discriminative cues for distinguishing targets from complex backgrounds. Inspired by this observation, we argue that enhancing edge representations to emphasize shape information is crucial to mitigate the similarity between targets and backgrounds. This motivates the following two research questions: how can shape features be effectively emphasized by enhancing edge representations

with sufficient spatial and directional information and how can multi-receptive-field perception be enhanced to increase model adaptability in complex environments?

Figure 1. Small infrared targets in complex backgrounds.

The inherent similarity between infrared small targets and their backgrounds makes IRSTD more difficult than general detection or segmentation tasks [1-3]. In recent years, this challenge has attracted growing attention. With the rise of deep learning and the availability of public datasets, many researchers have proposed sophisticated methods to tackle IRSTD. Convolutional neural networks (CNNs) have become the dominant paradigm, offering robust feature learning through adversarial training, multi-level feature fusion, and integration of edge priors [1-11].

In addition, several recent studies reported in Infrared Physics & Technology have explored complementary directions, such as multi-perception of target features, coordinate-based detection strategies, and robust optimization frameworks for small-target detection, further highlighting the significance of edge, scale, and frequency-domain modeling for IRSTD [12-14].

Despite their success, conventional CNN-based methods often exhibit weak shape bias, limiting their ability to capture detailed edge structures and leading to high false alarm rates [15]. To address this issue, we propose a shape-biased CNN architecture that emphasizes edge enhancement to incorporate shape information explicitly.

Traditional edge detection techniques (e.g., Sobel, Prewitt, Laplacian) utilize fixed differential operators to extract edges. While effective to some extent, these methods lack adaptability to diverse edge geometries due to their limited scale and directional sensitivity. Recent edge-oriented small-target methods in *Infrared Physics & Technology* also emphasize the role of structural cues, such as edge-dilation segmentation and multiscale local saliency for maritime targets [16].To overcome this limitation, we introduce a novel Multi-Scale Edge Attention (MSEA) module, which employs fixed convolution kernels with multiple scales and directions. Combined with a selective fusion strategy, MSEA enables precise edge enhancement while maintaining spatial diversity and directionality, resulting in more accurate target-boundary perception.

Although MSEA strengthens structural details, it may also amplify background noise due to the spatial domain's limited ability to distinguish structural edges from noisy textures. To address this issue, we propose an Edge-Guided Module (EGM) based on the Discrete Wavelet Transform (DWT). EGM extracts high-frequency components from the input image and leverages structural priors to suppress non-structural interference, enabling accurate edge extraction while reducing background noise.

Moreover, to prevent small targets from vanishing in deeper layers of the network, we introduce a Multi-Scale Group Convolution Module (MGCM). Positioned in the deeper stages of the network, MGCM enhances the preservation of small target features by capturing contextual information at multiple receptive fields.

Based on the MSEA, EGM, and MGCM modules, we construct an end-to-end edge shape enhancement network, termed ESE-Net, which explicitly strengthens edge information and improves detection performance in complex backgrounds.

The main contributions of this work are summarized as follows:

We propose a novel spatial-frequency joint edge enhancement framework. The proposed EGM complements the MSEA by guiding edge feature extraction in the frequency domain, thereby improving detection performance under complex background conditions.

We introduce the MGCM in the deeper network layers to preserve small target features and suppress irrelevant high-frequency components introduced by DWT, further enhancing the robustness of the network.

Extensive experiments on two public datasets, NUAA-SIRST and IRSTD-1K, demonstrate that our proposed method significantly outperforms existing state-of-the-art IRSTD approaches.

2. Proposed Method

2.1. Overview of the Proposed ESE-Net

The blue region represents the encoder, composed of four stages for feature extraction, with feature maps denoted as X_{S_i} (i=1,2,3,4). The green part consists of block-wise DWT and the Edge Guidance Module (EGM), with the frequency spectrum output from EGM denoted as X_{freq}. The orange section corresponds to the decoder, where feature maps during upsampling are denoted as X_{D_i} (i=1,2,3,4). The red boxes indicate infrared small targets. The pink module in the center is the Multi-scale Group Convolution Module (MGCM), and the yellow module on the left is the Multi-scale Edge Attention Module (MSEA).

The overall architecture of the proposed ESE-Net is illustrated in Figure 2. To address the challenges of blurred and indistinct target boundaries in infrared imagery, ESE-Net enhances edge representations by integrating spatial- and frequency-domain information. Similar dual-domain strategies have also been discussed in recent studies on IRSTD, highlighting the importance of combining spatial and spectral information.

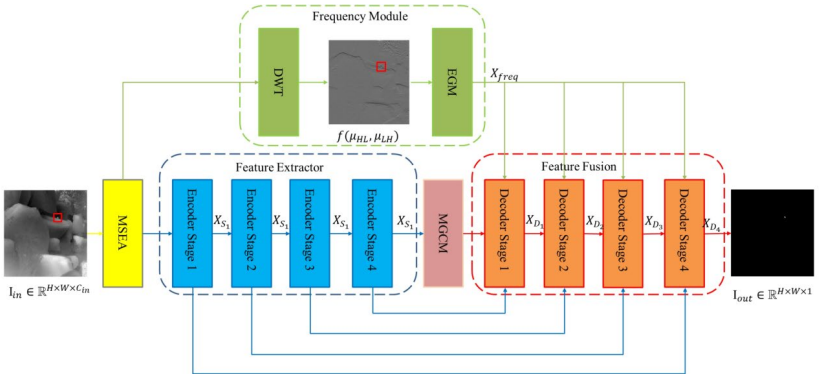

Figure 2. Overall architecture of the proposed model.

The input image is first processed by the Multi-Scale Edge Attention (MSEA) module, which employs fixed convolutional kernels at multiple scales to extract edge features from different receptive fields. In parallel, a frequency-domain branch captures high-frequency structural details via Haar wavelet transform. Unlike conventional methods that suppress high-frequency components as noise, we retain the horizontal and vertical subbands-where small targets are typically concentrated-while discarding the low-frequency background and diagonal components, thereby preserving structural target information.

To further refine the frequency-domain edge features, we introduce the Edge-Guided Module (EGM). This module takes multi-scale subbands as input and generates corresponding multi-level edge guidance maps. These guidance maps are spatially matched with the decoder features at different stages and fused accordingly, allowing explicit structural information to be injected during decoding. Through this multi-scale

edge-guided fusion strategy, EGM enhances salient high-frequency structures while suppressing background noise, facilitating more accurate target boundary perception.

The backbone of the network consists of an encoder-decoder structure. The encoder extracts hierarchical features across four stages, and the decoder progressively restores spatial resolution. At the center of the network, the Multi-Scale Group Convolution Module (MGCM) is employed to capture contextual information across multiple receptive fields. This helps to preserve small target features that may otherwise vanish in deeper layers of the network.

By jointly leveraging spatial and frequency cues, this dual-domain fusion architecture enables the network to suppress background interference while highlighting subtle target features, thereby significantly improving detection performance in complex infrared scenes.

2.2. Edge-Shape Enhancement Modules

2.2.1. Edge-Guided Module

As illustrated in Figure 3, the EGM integrates wavelet-based frequency features with CNN-derived spatial features. Specifically, given a frequency-domain input $X_{\text{freq}} \in \mathrm{R}^{M\times N\times 2}$ (from DWT) and a decoder feature map $X_{S_i} \in \mathrm{R}^{M\times N\times C}$, the following operations are applied:

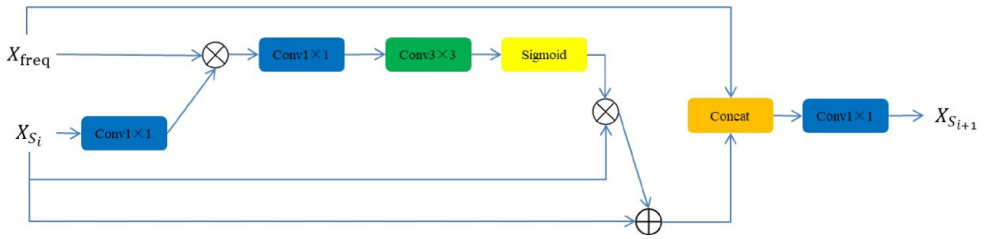

Figure 3. The proposed EGM integrates spatial and frequency information to enhance fused textures, edges, and fine details.

X_{freq} is fed into the MGCM module to generate an edge feature map $G \in \mathrm{R}^{M\times N\times 8}$. G is processed by a 3×3 convolution to obtain $w(G) \in \mathrm{R}^{M\times N\times C/2}$. X_{S_i} is also passed through a 3×3 convolution for channel reduction to match $C/2$, followed by element-wise multiplication with $w(G)$. The resulting fused map is further refined via a 3×3 convolution and a Sigmoid activation δ to obtain a guidance map $G' \in \mathrm{R}^{M\times N\times C}$. G' is added to X_{S_i} to yield a residual-guided feature map, which is concatenated with $w(G)$ and passed through a 1×1 convolution to produce the updated decoder output $X_{S_{i+1}}$.

During the upsampling process in the decoder, substantial noise is often introduced, as shown in the middle column of Figure 4.

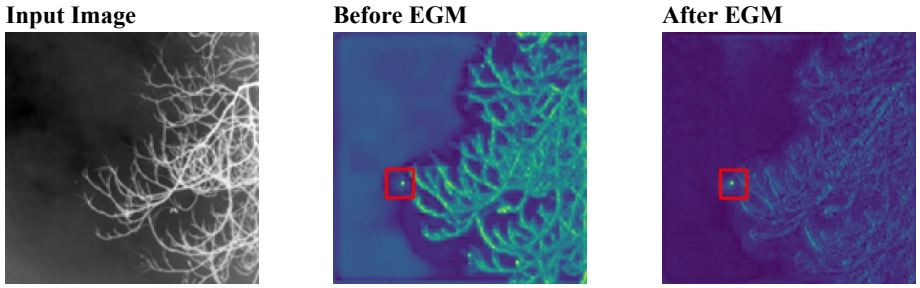

Figure 4. Feature map visualization in the decoder. The multi-level fusion of multi-scale edge guidance maps generated by the EGM effectively enhances target edge information while significantly suppressing background interference, thereby improving the efficiency of feature encoding.

Prior studies have demonstrated that decoder-stage feature fusion effectively mitigates such interference. Inspired by this, we propose an Edge-Guided Module (EGM) that leverages hierarchical feature fusion and cross-domain interaction to suppress noise and enhance target edge details.

The entire process is mathematically formulated as:

$$Q=\delta\left(\alpha\left[\alpha(w(X_{\text{freq}}))\otimes\alpha(X_{S_i})\right]\right)$$

$$X_{S_{i+1}}=\alpha\left(\Pi\left[\beta(\alpha(w(X_{\text{freq}}))),X_{S_i}+Q\otimes X_{S_i}\right]\right)$$

Here, α and β denote 3×3 and 1×1 convolutions respectively, $\otimes$ represents element-wise multiplication, "+" denotes addition, and Π indicates concatenation.

By hierarchically integrating multi-scale edge guidance maps across decoder stages, the EGM enriches feature representations with fine-grained textures and boundaries, while effectively suppressing background clutter. As visualized in Figure 4 (right column), this strategy substantially improves target edge clarity and overall detection accuracy.

2.2.2. Multi-Scale and Multi-Directional Edge Attention (MSEA)

To enhance edge representation and suppress background interference, we propose the Multi-Scale and Multi-Directional Edge Attention (MSEA) module. This module employs fixed directional convolutional kernels at multiple scales and orientations to emphasize target edges while suppressing irrelevant textures-an approach well-suited for small target detection in infrared imagery. Recent edge-guided approaches for maritime and aerial IRSTD tasks also support the effectiveness of multi-scale edge priors. MSEA utilizes three groups of fixed kernels with receptive fields of 7×7, 5×5, and 3×3, respectively. Each group contains eight directional kernels (e.g., horizontal, vertical, and diagonal), allowing edge extraction from multiple orientations. The directional response is computed using:

$$\mathrm{K}_{s\times s}^{(j)}=\begin{bmatrix}\kappa_1 & \cdots & \kappa_2 & \cdots & \kappa_3\\ \cdots & 0 & \cdots & 0 & \cdots\\ \kappa_4 & \cdots & 1 & \cdots & \kappa_5\\ \cdots & 0 & \cdots & 0 & \cdots\\ \kappa_6 & \cdots & \kappa_7 & \cdots & \kappa_8\end{bmatrix},\ \kappa_j=-1$$

$$\mathrm{G}^{(j)}=\mathrm{K}_{s\times s}^{(j)}*\mathrm{I},\ j=1,\ldots,8$$

$$\mathrm{O}^{(i)}=\mathrm{G}^{(i)}\odot\mathrm{G}^{(i+4)},\ i=1,2,3,4$$

$$\mathrm{M}_{s\times s}=\sigma\left(\sum_{i=1}^{4}\mathrm{O}^{(i)}\right)$$

$$\mathbf{M}_{\text{edge}}=\max(\mathrm{M}_{3\times3},\mathrm{M}_{5\times5},\mathrm{M}_{7\times7})$$

Here, $\sigma(\cdot)$ denotes the Sigmoid activation. Directional kernels $\mathrm{K}^{(j)}$ extract gradient responses $\mathrm{G}^{(j)}$ along eight orientations. For each symmetric pair of directions, their element-wise product $\mathrm{O}^{(i)}$ is computed to enhance bidirectional consistency. The responses across all four pairs are summed and passed through a sigmoid to generate the scale-specific edge map $\mathrm{M}_{s\times s}$.

Multi-scale fusion is achieved by taking the maximum activation across different kernel sizes (3 × 3, 5 × 5, 7 × 7), resulting in a consolidated edge attention map $\mathbf{M}_{\text{edge}}$. This map emphasizes salient edges with consistent multi-directional support while suppressing noisy or spurious gradients.

Additionally, the opposing-direction multiplication introduces local contrast enhancement, yielding directional contrast maps (denoted as $\mathrm{O}^{(i)}$) for each scale. These are fused and normalized to produce a final attention map bounded in [0,1], serving as guidance for edge-aware feature modulation in the subsequent processing stages.

2.2.3. Multi-Receptive Field Perception

Infrared images often exhibit complex and cluttered backgrounds, making small targets difficult to detect using limited receptive fields. This can lead to high false alarm rates due to insufficient contextual awareness. While methods like UIU-Net, DNANet, and RDIAN have explored deeper or wider networks to address this, deeper models incur higher computational costs, and width-based strategies often struggle to suppress background-like textures [17]. Some recent works in *Infrared Physics & Technology* have also emphasized multi-perception and robust optimization as promising solutions for enhancing receptive field adaptability [18].

To overcome these limitations, we introduce a Multi-scale Group Convolution Module (MGCM) (see Figure 5), designed to enhance receptive field diversity while maintaining efficiency. Inspired by channel-splitting and grouped convolution designs, the MGCM divides the input feature map [19-21].

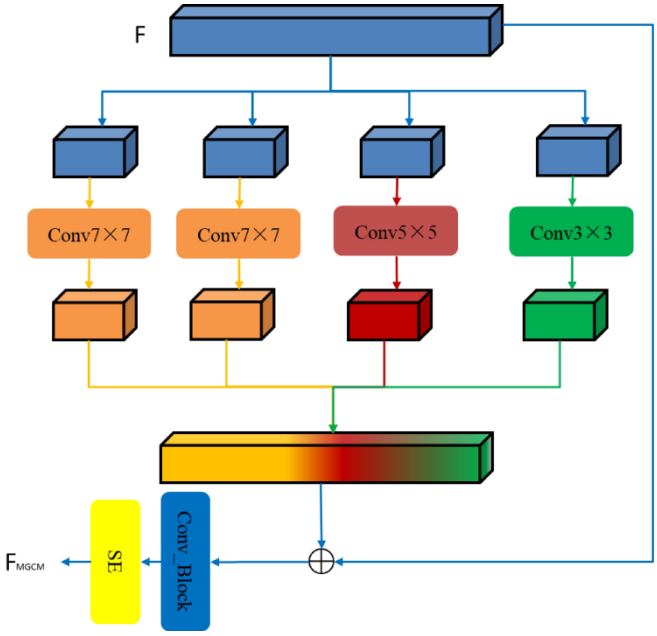

Figure 5. The structure of MGCM.

F into four branches, each processed with convolutions of varying kernel sizes to capture multi-scale context. The outputs are concatenated with the original input to preserve both local and global information. The resulting fused tensor is then passed through a Conv_Block, which consists of a 1 × 1 convolution followed by batch normalization and ReLU activation, to reduce channel redundancy and enhance feature interaction. Finally, a Squeeze-and-Excitation (SE) module is applied to adaptively recalibrate channel-wise feature responses, further boosting discriminative capability [22].

This design enables the model to effectively perceive multi-scale contextual cues, reducing false positives while retaining essential target features. Compared with conventional receptive-field enlargement, the proposed MGCM provides a lightweight yet effective way to improve target discrimination in cluttered infrared scenes [18].

3. Experiments

3.1. Experimental Setup

Datasets: This study uses the NUAA-SIRST and IRSTD-1K datasets for training, validation, and testing. NUAA-SIRST consists of 427 infrared images of varying sizes. To avoid overlap between the training, validation, and testing sets, only one representative image is selected from each infrared sequence. Due to the limited availability of infrared

sequences, the NUAA-SIRST dataset includes infrared images at a wavelength of 950 nm, in addition to short-wave and mid-wave infrared images [23]. Many of the targets are very faint and are hidden in complex backgrounds with significant clutter. Detecting these targets is challenging, even for humans, and requires a deep semantic understanding of the entire scene and focused search efforts. On the other hand, IRSTD-1K is a more challenging dataset, containing 1000 real infrared images, each with a size of 512 × 512 pixels. IRSTD-1K includes a variety of small targets, such as drones, animals, ships, and vehicles, which can be captured from long distances at various positions [24]. This dataset covers numerous scenes, with backgrounds including oceans, rivers, fields, mountains, urban environments, and clouds, all of which contain significant clutter and noise. IRSTD-1K serves as a comprehensive benchmark for evaluating ISTD methods. For each dataset, 80% of the images are used as the training set, and 20% are used as the test set.

Evaluation Metrics: We use Intersection over Union (IoU) and False Alarm Rate (Fa) as pixel-level evaluation metrics, and Detection Probability (Pd) to evaluate target-level performance. Different metrics reveal different aspects of the detector's performance. Fand Pfocus emphasize recall and false positives, while IoU considers both aspects simultaneously. Their definitions are as follows:

$$\text{IoU}=\frac{TP}{T+P-TP}$$

$$P_d=\frac{\text{Number of correctly predicted targets}}{\text{Number of all targets}}$$

$$F_a=\frac{\text{Number of falsely predicted pixels}}{\text{Number of all pixels}}$$

3.2. Implementation Details

The proposed method is implemented using the PyTorch framework. Based on existing works, the input size of the detector is set to 256 × 256. We train the different models using an RTX4090 GPU with the AdaGrad optimizer. The batch size is set to 4, and the learning rate is set to 0.05 [25].

3.3. Comparison with Existing Methods

1) Quantitative Comparison. We compare our proposed ESE-Net with both traditional and deep learning-based infrared small target detection (ISTD) methods. Traditional methods include:

Filtering-based: Top-Hat, Max-Median

Local contrast-based: WSLCM, TLLCM

Low-rank models: IPI, NRAM, RIPT, PSTNN, MSLSTIPT

Deep learning-based methods include: MDvsFA, ALCNet, ISNet, and DNANet. All deep learning methods were retrained using official implementations to ensure fair evaluation on NUAA-SIRST and IRSTD-1K [23-31].

Table 1. presents the quantitative comparison. ESE-Net consistently ranks among the top three methods across all metrics and datasets, often achieving the best or second-best performance. Traditional methods generally perform poorly due to their limited hand-crafted priors, especially on cluttered or low-contrast scenes. Deep learning-based approaches significantly outperform traditional methods but still struggle with low target-background contrast, limiting their IoU performance.

Table 1. Quantitative comparison of different IR small target detection methods on IRSTD-1k and NUAA-SIRST datasets.

Method	**Description**	**IRSTD-1k**			**NUAA-SIRST**		
3-8		IoU↑	P_d↑	F_a↓	IoU↑	P_d↑	F_a↓
Top-Hat [23]	Filtering	10.06	75.11	1432	7.143	79.84	10.12

Max-Median [24]	Local Contrast	6.998	65.21	59.73	4.172	69.20	55.33
WSLCM [25]		3.452	72.44	6619	1.158	77.95	5446
TLLCM [26]		3.311	77.39	6738	1.029	79.09	5899
IPI [27]	Low Rank	27.92	81.37	16.18	25.67	85.55	11.47
NRAM [28]		15.25	70.68	16.93	12.16	74.52	13.85
RIPT [29]		14.11	77.55	28.31	11.05	79.08	22.61
PSTNN [30]		24.57	71.99	35.26	22.40	77.95	29.11
MSLSTIPT [31]		11.43	79.03	1524	10.30	82.13	1131
MDvsFA [3]	Deep Learning	49.50	82.11	80.33	60.30	89.35	56.35
ALCNet [5]		62.05	92.19	31.56	74.31	97.34	20.21
ISNet [1]		68.77	95.56	15.39	80.02	99.18	4.92
DNANet [7]		65.71	91.84	17.61	77.54	98.10	2.510
ESE-Net (Ours)		70.14	94.90	9.034	79.64	99.22	8.549

Our method excels in both pixel-level metrics (IoU, Fa) and object-level metric (Pd). The Edge Guidance Module (EGM) enhances pixel-level segmentation accuracy by preserving spatial boundary details through wavelet-based edge extraction, while the Multi-Scale Grouped Convolution Module (MGCM) improves semantic abstraction and reduces false positives in complex scenes. The integration of these complementary modules allows the decoder to simultaneously capture high-level semantics and fine edge structures, resulting in improved overall detection performance.

Moreover, our Multi-Grained Convolution Module (MGCM) module further refines feature representation by suppressing irrelevant high-frequency noise while retaining small target cues. On IRSTD-1K, which contains more challenging real-world scenarios, our method outperforms others by a large margin, especially in reducing false alarms.

Furthermore, the ROC curves evaluated on the IRSTD-1k dataset are presented in Figure 6, where our method achieves the highest AUC score. This result clearly demonstrates the effectiveness of the proposed approach in distinguishing targets from background, particularly in low false positive rate regions where the model maintains high detection sensitivity.

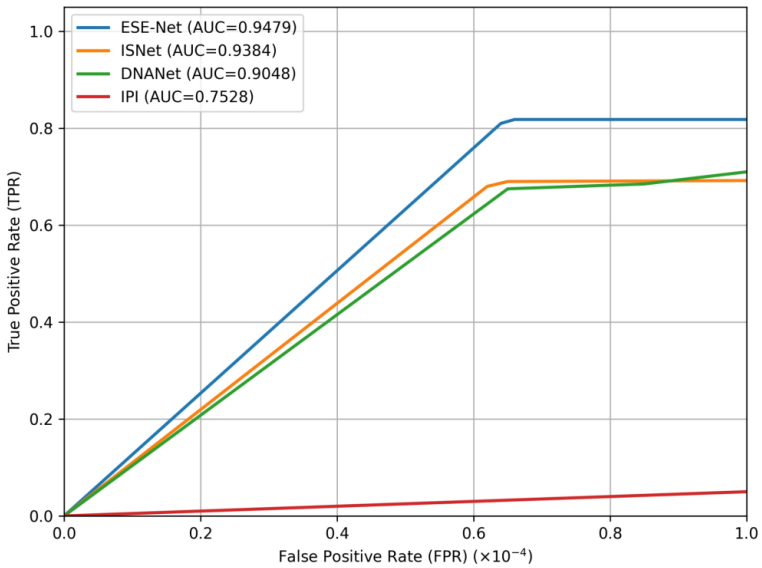

Figure 6. ROC curves of our CSRNet and other approaches on IRSTD-1k.

2) Visual Comparison. Figure 7 illustrates qualitative comparisons. Detecting small infrared targets is visually difficult due to their background similarity. As shown, many state-of-the-art methods either miss targets or introduce false positives, particularly in low-contrast, cluttered scenes. In contrast, ESE-Net demonstrates strong robustness by

accurately segmenting true targets while suppressing noise and clutter, even when targets are visually indistinguishable.

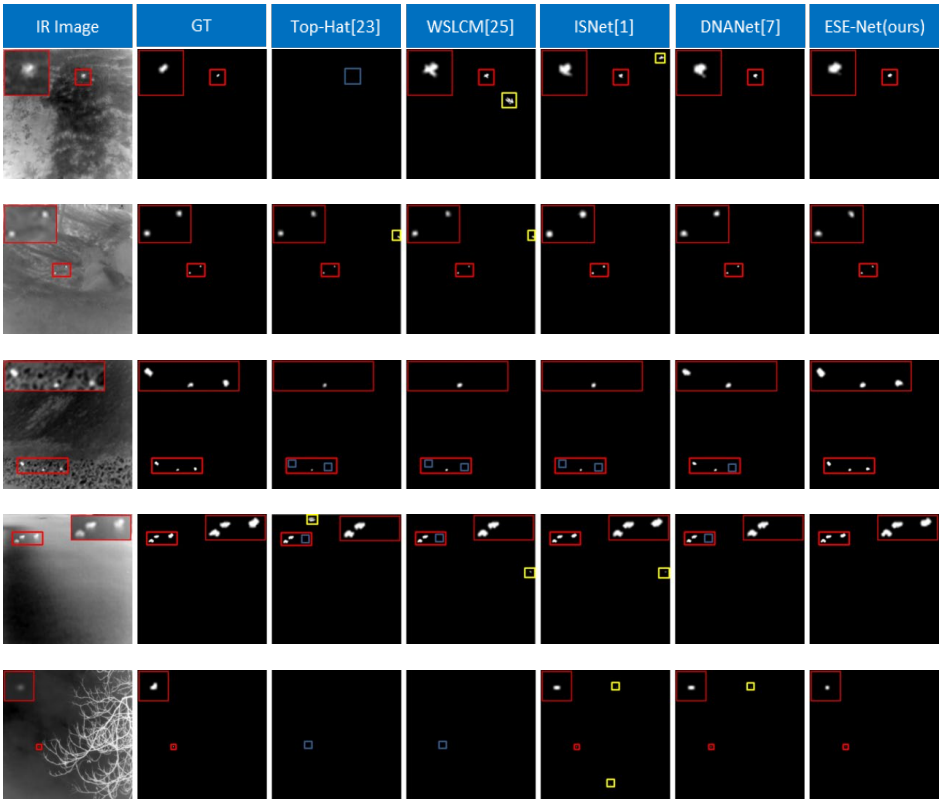

Figure 7. Visual comparison of detection results on several infrared images. Correctly detected targets, missed targets, and false alarms are framed by red, blue, and yellow boxes, respectively. A close-up view of the target is shown in image corners.

In the last two rows of Figure 7, the majority of existing methods fail to correctly recognize the true target, or mistakenly identify adjacent structures as targets. In contrast, our network-guided by the Edge Guidance Module (EGM) and refined by the MGCM module-effectively suppresses irrelevant interferences through frequency-domain feature extraction. As a result, it successfully distinguishes targets from complex backgrounds.

Importantly, we have designed a Multi-Scale Spatial Edge Attention (MSEA) module that works in tandem with the EGM. EGM employs wavelet-based or similar frequency-domain transforms to extract edge features, followed by reversible downsampling to discard low-frequency interference before spatial fusion. This spatial-frequency interactive attention mechanism integrates multi-scale spatial information and frequency-domain edge cues. This approach significantly enhances edge detection quality and suppresses noise in complex backgrounds.

3.4. Ablation Studies

To comprehensively evaluate the contribution of each component in our proposed network and their synergistic effects, we conducted a series of ablation experiments. These experiments include module-level comparisons, edge enhancement analysis, and an in-depth investigation into the number of edge-guided feature fusion stages.

1)Multiscale Design and Edge Enhancement in MSEA. To enhance the edge representation ability of the network, we design a Multi-Scale Edge Attention (MSEA) module that applies fixed convolutional filters at multiple scales and directions. This design aims to capture edge features under diverse target sizes and background

conditions. We conduct ablation experiments to validate the necessity of the multiscale design and evaluate the effectiveness of MSEA in enhancing structural edge information.

As shown in Table 2, we conducted experiments to explore the impact of kernel sizes in the parameter-fixed convolution layers within the MSEA module. Specifically, as the kernel size increases, the IoU score gradually decreases, while the detection probability (Pd) increases, and the false alarm rate (Fa) first decreases and then increases. The inconsistency among these metrics indicates that a single-scale convolution kernel lacks robustness in detection performance, which can be attributed to the diverse size distribution of infrared small targets. Therefore, we adopt a multiscale design, enabling the network to achieve optimal performance across all evaluation metrics.

Table 2. Result of Different Scales of The Patch In MSEA.

# scale of the patch	IoU (%) ↑	Pd (%) ↑	Fa (%) ↓
3	67.43	93.11	15.43
5	67.21	93.73	12.84
7	66.84	93.87	16.61
3,5,7	70.14	94.90	9.03

In addition to quantitative analysis, we provide a visual comparison of edge enhancement effects in Figure 8, the edge maps generated by our proposed method exhibit sharper and more continuous contours compared to traditional edge detectors. This clear enhancement of edge structure provides more precise boundary information, which is beneficial for the network to learn discriminative edge features.

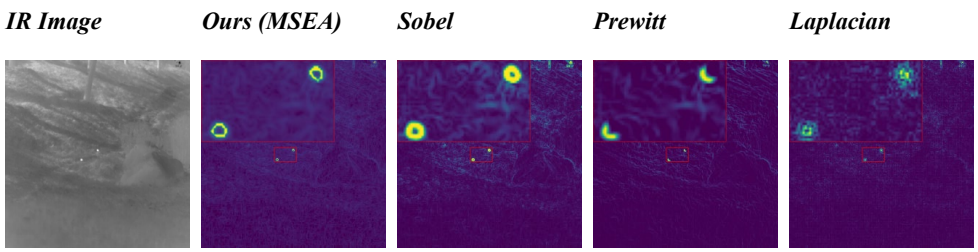

Figure 8. Comparison of edge maps generated by different methods. The proposed MSEA module achieves more precise and continuous edge enhancement compared to traditional operators.

2) Effect of Edge-Guided Feature Fusion at Different Decoder Stages.To further investigate the effectiveness of the proposed EGM module, we conduct experiments that vary the number of decoder stages fused with the edge guidance map. Specifically, we test the integration of 1 to 4 decoder stages (denoted as X_{D_1} to X_{D_4}), while keeping all other settings unchanged to ensure fair comparison.

1-stage fusion: Edge guidance is introduced only at the top decoder stage (Decoder Stage 4, X_{D_4}).

2-stage fusion: Edge information is fused with Decoder Stage 3 and 4(X_{D_3}), X_{D_4}).

3-stage fusion: Further integration at Decoder Stage 2 (X_{D_2}).

4-stage fusion: Full-scale fusion across all decoder stages (X_{D_1}-X_{D_4}).

The performance comparison is presented in Table 3.

Table 3. Ablation Study on the Number of Edge-Guided Feature Fusion Stages.

# Fusion Stages	IoU (%) ↑	Pd (%) ↑	Fa (%) ↓
1	66.85	91.02	22.37
2	68.42	92.73	12.84
3	69.58	94.10	10.27
4	70.14	94.90	9.03

These results clearly demonstrate that edge information is most effective when applied consistently throughout the decoder, rather than at a single or partial stage. Therefore, we adopt 4-stage fusion as the default configuration in our final model.

3) Effectiveness of Individual Modules in ESE-Net. We evaluate the contribution of the Multi-Scale Edge Aggregation Module (MSEA), Multi-Grained Convolution Module (MGCM), and Edge-Guided Module (EGM) through module-wise ablation. Table 4 summarizes the results.

Table 4. Ablation Study of MSEA, MGCM, and EGM.

MSEA	MGCM	EGM	IoU	Pd	Fa ($\times 10^{-6}$)
			63.28	87.43	21.54
✓			66.75	89.32	17.42
	✓		65.93	88.70	18.89
		✓	66.20	89.04	17.96
✓	✓		68.92	92.37	13.51
✓		✓	69.40	93.20	12.78
	✓	✓	69.17	92.88	13.05
✓	✓	✓	70.14	94.90	9.03

The results clearly show that each module contributes positively to performance. When all three modules are integrated, the network achieves the best results in terms of IoU, Pd, and Fa.

As observed, the detection performance steadily improves with an increasing number of fusion stages. IoU increases from 66.85% to 70.14%, and Pd reaches 94.90%. Notably, the false alarm rate (Fa) drops significantly from 22.37 to 9.03, indicating that multi-level edge-guided fusion enhances both semantic understanding and edge preservation, improving target localization while effectively suppressing background noise.

However, we also observe diminishing returns beyond the third stage, suggesting that excessive fusion may introduce feature redundancy or lead to overfitting. Therefore, a trade-off between accuracy and model complexity must be considered in network design.

4. Conclusion

In this paper, we propose a novel Edge-Shape Enhanced Network (ESE-Net) for infrared small target detection, which leverages edge information to enhance the discriminability of small targets in complex infrared scenes. Specifically, we design a Multiscale Spatial Edge Attention (MSEA) module to enhance edge contours by capturing multiscale directional gradients. To further highlight target boundaries and suppress background interference, we introduce an Edge Guidance Module (EGM) that utilizes frequency-domain edge cues via wavelet transform and selective fusion. Additionally, a Multiscale Group Convolution Module (MGCM) is deployed to preserve fine target details and improve robustness against target omission. Extensive experiments on NUAA-SIRST and IRSTD-1K datasets validate the effectiveness of our proposed approach, achieving superior performance compared with existing methods. Future work will explore lightweight extensions of the network and real-time deployment on edge devices.

References

1. 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.
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. doi: 10.1109/wacv48630.2021.00099

3. F. Hashmi, M. U. Hassan, M. U. Zubair, K. Ahmed, T. Aziz, and R. M. Choudhry, "Near-miss detection metrics: An approach to enable sensing technologies for proactive construction safety management," *Buildings*, vol. 14, no. 4, p. 1005, 2024. doi: 10.3390/buildings14041005
4. B. Zhao, C. Wang, Q. Fu, and Z. Han, "A novel pattern for infrared small target detection with generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4481-4492, 2020. doi: 10.1109/tgrs.2020.3012981
5. 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. doi: 10.1109/tgrs.2020.3044958
6. T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 4, pp. 4250-4261, 2023. doi: 10.1109/taes.2023.3238703
7. B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, and Y. Guo, "Dense nested attention network for infrared small target detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 1745-1758, 2022. doi: 10.1109/tip.2022.3199107
8. K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-13, 2022. doi: 10.1109/tgrs.2022.3163410
9. M. Qi, L. Liu, S. Zhuang, Y. Liu, K. Li, Y. Yang, and X. Li, "FTC-Net: Fusion of transformer and CNN features for infrared small target detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8613-8623, 2022. doi: 10.1109/jstars.2022.3210707
10. F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small and dim target detection with transformer under complex backgrounds," *IEEE Transactions on Image Processing*, vol. 32, pp. 5921-5932, 2023. doi: 10.1109/tip.2023.3326396
11. 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*, October, 2022, pp. 1730-1738. doi: 10.1145/3503161.3547817
12. C. Liu, F. Xie, H. Zhang, Z. Jiang, and Y. Zheng, "Infrared small target detection based on multi-perception of target features," *Infrared Physics & Technology*, vol. 135, p. 104927, 2023. doi: 10.1016/j.infrared.2023.104927
13. L. Fan, Y. Wang, G. Hu, F. Li, Y. Dong, H. Zheng, and X. Ding, "Diffusion-based continuous feature representation for infrared small-dim target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-17, 2024. doi: 10.1109/tgrs.2024.3395478
14. L. Xu, Y. Wei, H. Zhang, and S. Shang, "Robust and fast infrared small target detection based on pareto frontier optimization," *Infrared Physics & Technology*, vol. 123, p. 104192, 2022. doi: 10.1016/j.infrared.2022.104192
15. M. Zhang, N. Wang, Y. Li, and X. Gao, "Neural probabilistic graphical model for face sketch synthesis," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2623-2637, 2019. doi: 10.1109/tnnls.2019.2933590
16. E. Zhao, L. Dong, and H. Dai, "Infrared maritime target detection based on edge dilation segmentation and multiscale local saliency of image details," *Infrared Physics & Technology*, vol. 133, p. 104852, 2023. doi: 10.1016/j.infrared.2023.104852
17. 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. doi: 10.1109/tip.2022.3228497
18. 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. doi: 10.1109/tgrs.2023.3235150
19. N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," In *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116-131.
20. C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390-391.
21. K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1580-1589.
22. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141. doi: 10.1109/cvpr.2018.00745
23. J. F. Rivest, and R. Fortin, "Detection of dim targets in digital infrared imagery by morphological image processing," *Optical Engineering*, vol. 35, no. 7, pp. 1886-1893, 1996.
24. 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*, October, 1999, pp. 74-83.
25. 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. doi: 10.1109/lgrs.2020.3004978
26. 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. doi: 10.1109/lgrs.2019.2954578

27. 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. doi: 10.1109/tip.2013.2281420
28. 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.
29. 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. doi: 10.1109/jstars.2017.2700023
30. 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. doi: 10.3390/rs11040382
31. 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. doi: 10.1109/tgrs.2020.3022069