Electronic Research Archive

ERA, 34(9): 6177–6218.
DOI: 10.3934/era.2026272
Received: 09 May 2026
Revised: 19 June 2026
Accepted: 06 July 2026
Published: 16 July 2026

https://www.aimspress.com/journal/ERA

Research article

BSG-Mamba: A dual-branch network based on background-suppression sparse graph-enhanced Mamba for infrared small target detection

Xinying Huang[1,2], Jianhua Song[1,2,*], Xinrong Fu[1], Yu Chen[1] and Hao Liu[1]

[1] College of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China
[2] Key Laboratory of Light Field Manipulation and System Integration Applications in Fujian Province, Minnan Normal University, Zhangzhou 363000, China

* **Correspondence:** Email: songjianhua@mnnu.edu.cn.

Abstract: Infrared small target detection faces numerous challenges due to weak target features and complex backgrounds. Existing methods have drawbacks such as mixed background and target signals, imbalanced local and global feature modeling, and insufficient learning of direction-specific context. To address these issues, this paper proposes a dual-branch network based on background-suppression sparse graph-enhanced Mamba (BSG-Mamba). We build this network on a U-shaped encoder-decoder architecture. We design a background-suppression branch to initially separate the target from the background and combine it with a target-enhancing gating (TEG) mechanism to enhance target features. On this basis, we effectively capture non-local feature correlations in small targets by embedding a sparse graph encoder (SG-encoder), compensating for the shortcomings of traditional convolutional segmentation. At the deep encoder position, it constructs a Mamba module with multi-state space model (SSM) decoupling and adaptive fusion. We configure independent SSM parameters for four-direction scanning and introduce learnable fusion weights to achieve unified modeling of long-range dependencies and direction-specific features. Experimental results on the NUAA-SIRST and NUDT-SIRST datasets show that the core detection metrics of the model outperform existing state-of-the-art (SOTA) methods, achieving detection rates of 97.74% and 98.90%, respectively, on the two datasets, while maintaining a lightweight network parameter count and balancing detection accuracy and computational efficiency.

Keywords: infrared small target detection; background suppression; Unet; sparse graph; Mamba

1. Introduction

Infrared small target detection (IRSTD) leverages thermal radiation differences to identify targets in complex scenes, playing a pivotal role in maritime surveillance [1], military reconnaissance [2,3], and autonomous systems [4,5]. Unlike general object detection, IRSTD is fundamentally constrained by the low signal to clutter ratio (SCR) and limited spatial resolution of infrared imagery, where target signatures are frequently submerged in heavy noise and heterogeneous backgrounds. Furthermore, the inherent lack of structural and textural features in dim small targets, often occupying only a few pixels, poses significant hurdles for reliable localization and precise contour segmentation. These challenges, exacerbated by weak thermal contrast, necessitate robust algorithms to mitigate high false alarm rates and missed detections. Specifically, IRSTD faces three core challenges: extremely low SCR, where dim targets are often submerged in heavy background clutters and sensor noise; lack of discriminative structural and textural features, as small targets typically occupy only a few pixels; and strong interference from edges and textures in heterogeneous scenes, which easily leads to false positives and missed detections [6].

Early IRSTD research primarily utilized traditional techniques including filter-based methods [7–9], local contrast-based methods [10,11], human visual system models [12], and low-rank sparse representations [13]. Although these methods perform adequately in uniform backgrounds, their heavy dependence on handcrafted priors and manual hyperparameters limits their robustness in complex scenes containing significant noise interference.

The advent of deep learning has shifted the focus toward data-driven feature extraction, predominantly within U-shaped frameworks inspired by the success of U-Net in medical imaging [14]. Liu et al. pioneered this transition by proposing the first convolutional neural network (CNN)-based IRSTD method [15]. To enhance context integration, Dai et al. developed the asymmetric context module [16], while Zhang et al. introduced AGPCNet [17] to improve global correlation through attention-guided mechanisms. However, these approaches often struggle with weak target loss or feature fragmentation caused by block-based strategies. To address multi-scale representation, UIU-Net [18] employs a nested U-shaped structure, and DNA-Net [19] implements dense nested interactions. Despite their effectiveness, these models frequently face challenges regarding limited generalization or high computational complexity. Recently, transformer-based architectures have been introduced to model long-range dependencies. For instance, U-Transformer [20] integrates Swin Transformer to alleviate semantic information deficiency, while hybrid designs like the Vision Transformer and Convolutional Neural Network (ViT-CNN) encoder [21] and SCTransNet [22] further strengthen global context modeling. APTNet [23] adopts an adaptive partial transformer and dual residual attention to enhance target features while reducing partial computational consumption. STASPPNet [24] combines Swin Transformer with multi-scale atrous pooling to detect ultra-small infrared targets. Nevertheless, the substantial computational overhead and structural complexity of these transformer-based models often necessitate meticulous parameter fine-tuning and may lead to the smoothing of critical local details.

Human visual perception studies indicate that effective target recognition depends on both target signatures and background characteristics. However, most IRSTD methods focus on isolated feature extraction or background modeling, failing to exploit the synergy between the two. For instance, Zhang et al. incorporated edge information via Sobel operators [25], but its robustness remains limited in dynamic scenarios. A critical challenge involves decoupling salient target features from complex

background interference without excessive labeled data or training. While feature extraction has advanced, many approaches still rely on implicit suppression, and explicit background modeling has not yet fully met practical requirements [26].

Recently, state space models (SSMs) such as Vision Mamba have gained prominence for their long range dependency modeling and linear computational complexity [27,28]. Although Mamba variants have succeeded in medical imaging [29,30] and autonomous driving [31], their direct application to IRSTD remains problematic. Specifically, the global modeling mechanism of Mamba often overlooks fine-grained local context essential for identifying small targets. Furthermore, standard multi-directional scanning strategies typically share parameters, which limits their ability to perform specific context modeling for distinct directional features in complex backgrounds. Additionally, the high structural complexity of Mamba models necessitates further research to optimize their efficiency and sensitivity for resource constrained IRSTD applications. When combining these issues with the inherent defects of conventional CNN and transformer-based methods, we find that mainstream solutions generally struggle with unsatisfactory background suppression, heavy computational overhead, and inadequate capture of fine-grained target features. It is therefore imperative to explore a new technical paradigm to tackle these practical dilemmas.

To remedy the above problems, our framework abandons the single-branch implicit background suppression adopted by most existing methods. We construct a dedicated dual-branch structure for explicit background modeling and target feature learning, and adopt lightweight SSMs to replace resource-hungry transformer modules. Graph convolution is also embedded to capture non-local feature correlations and fit the sparse property of infrared small targets.

To put the above design into practice, this work realizes parallel processing of background information and target features, effectively enhancing the identification ability of small targets in complex backgrounds. The model combines multi-directional scanning state space modeling and graph convolution, achieving unified modeling of long-range dependencies and local context, and greatly improving detection robustness and feature discrimination under noisy and complicated scenes. The contributions of this work are as follows:

1) We propose a dual-branch collaborative mechanism that combines background suppression with target enhancement. By explicitly learning background attention maps and foreground enhancement weights, this mechanism dynamically suppresses background interference during encoding-decoding to enhance multi-scale features, effectively addressing the challenge of target-background separation in complex scenes.

2) We design a sparse graph encoder (SG-encoder) that integrates a graph convolutional module into the encoder architecture. It constructs graph-structured relationships among feature points to capture non-local and irregular features, overcoming the limitations of conventional convolutions in representing small target structures under complex backgrounds.

3) We present a multi-SSM decoupled Mamba model with adaptive fusion. It equips the four scanning directions with fully independent SSM core parameters and introduces learnable directional fusion weights, achieving lightweight modeling with directional specificity. This addresses the lack of direction-aware context modeling in existing global modeling approaches.

Experimental results on public datasets, including NUAA-SIRST [16] and NUDT-SIRST [19], demonstrate the superiority of the proposed method. Compared with more advanced (SOTA) methods in recent years, the method proposed in this paper has better robust performance and detection performance under complex backgrounds and multi-scale tasks.

2. Related work

2.1. Background suppression-based infrared small target detection

In IRSTD, complex background clutter often causes false alarms and missed detections. Therefore, effective background suppression has always been a core issue for improving detection performance.

Early traditional methods exploited statistical differences between target and background in the time, spatial, and transform domains for suppression. LCM enhanced small target salience by boosting local contrast responses [32]. Top-Hat filtering filtering used morphological operations to estimate and subtract the background [7]. Although these methods were computationally efficient, they relied on fixed, manually defined features, making them unsuitable for non-uniform backgrounds and bright structural edges. Han et al. proposed a three-layer estimation window and closest-mean principle to suppress high-brightness backgrounds with single-scale computation, but their method remains model-driven, relying on handcrafted parameters and lacking the adaptive feature learning capability of deep networks.

With the adoption of deep learning, data-driven methods have shown greater adaptability. One common approach frames the problem as pixel-level regression. Dai et al. used an encoder-decoder to map the original image to a target-background segmentation map. Another approach focused on designing a dedicated background estimation branch [16]. Li et al. designed an interference suppression module that predicts weight maps of non-target regions to suppress background responses [19]. Wang et al. introduced a reverse attention mechanism to iteratively erase the foreground and refine the background estimate [33].

To address these limitations, we propose an explicit dual-branch background suppression architecture. Unlike implicit learning approaches, we construct a parallel background estimation branch to learn and output attention maps for both global and local backgrounds. By adaptively fusing the original features with these background attention maps, we dynamically suppress background interference and enhance target response during feedforward processing.

2.2. Graph network-based infrared small target detection

With the continuous advancement of deep learning, researchers have begun to explore the application of graph neural networks (GNNs) with traditional CNNs to further enhance the models' ability to capture non-local contextual information [34]. GNNs propagate and aggregate information through the adjacency relationships between nodes to perform feature learning. By modeling dependencies between pixels at a distance, GNNs can effectively capture non-local contextual information. Models based on GNNs include graph convolutional networks (GCNs) [35], graph attention mechanism (GAT) [36], graph reasoning networks [37], and visual image network (ViG) [38].

ViG divides an image into patches, treats them as graph nodes, uses K-nearest neighbor (KNN) to construct an adjacency matrix, and extracts features via graph convolutions. However, KNN computations limit its efficiency. Munir et al. improved efficiency by introducing an adaptive clustering mechanism that avoids KNN [39]. Lin et al. proposed BPGA, which suppresses noise using mobile SAM prior and enhances target features via visual graph convolutions [40]. Shen et al. proposed GCLNet, which uses multi-scale graph inference and local feature learning modules to capture both global and local context [41].

To address the sparse distribution of small infrared targets, we propose a graph convolutional module based on local sliding windows and dynamic edge weights. Unlike KNN-based methods, our approach uses local sliding windows to construct the graph structure and introduces a dynamic edge weight mechanism, avoiding complex KNN computations.

2.3. Mamba-based infrared small target detection

To address the computational efficiency bottlenecks of transformers [42], structured SSMs have emerged as an alternative [43]. SSMs use state variables to characterize system evolution and model time series dynamics. Their recursive structure enables effective capture of long-range dependencies. Structured state space sequence model (S4) efficiently handles long sequences with near-linear time complexity [27]. S6 introduces input-dependent dynamic parameter tuning [28]. Vision Mamba extends Selective state space sequence model (S6) to computer vision, achieving global context modeling without self-attention via a bidirectional S6 model with spatial embeddings [44].

Mamba has been widely applied to various vision tasks. Ma et al. combined the U-Net with Mamba for medical image segmentation [29]. Guo et al. proposed MambaIR, integrating channel attention and local enhancement [45]. Chen et al. proposed RSMamba with a dynamic activation strategy [46]. Zhang et al. proposed MOU-Mamba, which uses MO-SS2D to eliminate redundant information and LG-SS2D to fuse local and global features [47].

However, existing Mamba-based methods lack multi-directional scanning modeling mechanisms and have limited local perception capabilities. To address these issues, we design a four-directional, fully independent selective SSM, enabling each direction to independently learn target-specific long-range dependencies. We also introduce a learnable adaptive fusion mechanism that dynamically balances the four directions via softmax-normalized weights. Additionally, we explicitly integrate a deep separable convolutional local enhancement branch to overcome SSMs' limited local perception, establishing a global-local collaborative modeling capability.

2.4. Enlightenment from cross-domain visual detection research

Recent advances in artificial intelligence, including foundation models and interpretable learning frameworks, bring new perspectives for visual anomaly detection and feature representation. Though these methods are proposed for hyperspectral image analysis, their ideas on representation learning, model interpretability, and target-background decoupling provide meaningful references for infrared small target detection.

Large-scale pre-trained foundation models exhibit powerful universal representation ability. As a typical spectral remote sensing foundation model, SpectralGPT [48] adopts a 3D generative transformer structure. Trained on massive spectral data, it captures spatial-spectral correlation via self-supervised learning and achieves excellent performance on various downstream tasks, offering valuable experience for high-quality feature modeling.

Interpretable learning that combines model-driven priors and data-driven networks improves the explainability of deep models. LRR-Net [49] integrates low-rank representation priors into a deep unfolding network, transforming fixed regularization terms into learnable parameters. It combines the interpretability of traditional models and the strong fitting capability of deep learning, and it realizes effective separation between target and background.

Inspired by the above research, we adopt the idea of target-background decoupling and build an independent background suppression branch to improve the interpretability of our detection framework. We also employ lightweight SSMs and sparse graph convolution to strengthen feature representation while ensuring practical running efficiency.

3. Methodology

3.1. Model structure

As shown in Figure 1, background-suppression sparse graph-enhanced Mamba (BSG-Mamba) employs a U-shaped encoder-decoder architecture featuring a dual-branch mechanism for concurrent background suppression and target enhancement. The background branch extracts complex clutter to generate suppression weights, which are fused into the second layer of the main encoder to refine initial target responses. The five-layer encoder incorporates three spatially adaptive stages with embedded graph convolutions to strengthen semantic correlations. In deeper layers, direction-specific adaptive fusion Mamba modules capture long-range dependencies through four-directional state space modeling. Subsequently, the decoder integrates multi-level features via skip connections and utilizes target-aware gating to dynamically adjust feature weights for precise target-background discrimination. By synergizing background decoupling, state space modeling, and graph convolutions, BSG-Mamba significantly improves the detection of dim-small targets in cluttered infrared scenarios.

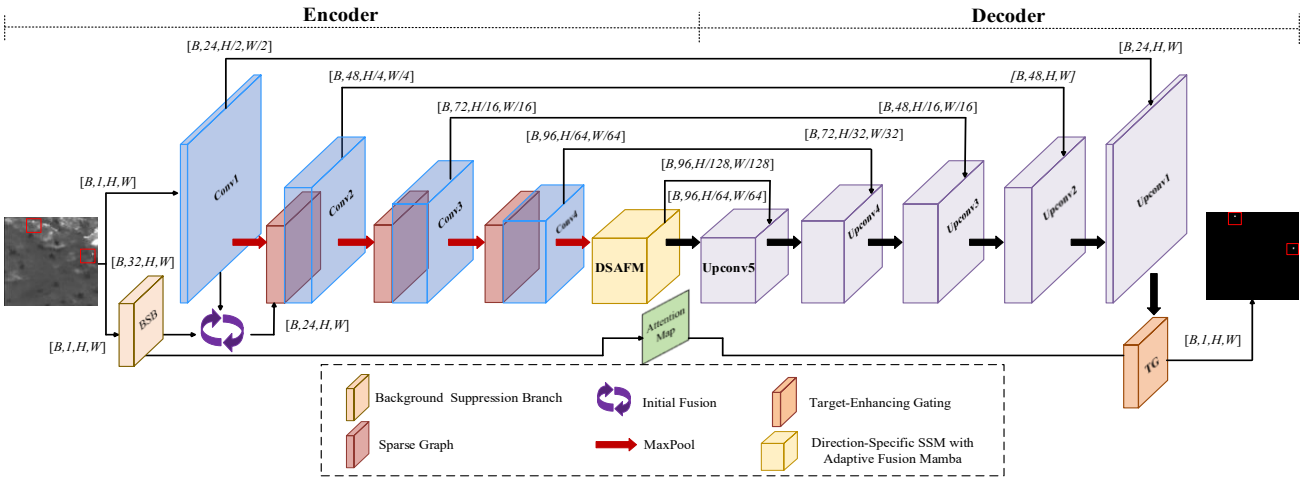


Figure 1. Overview of the BSG-Mamba model.

Algorithm 1 Inference for BSG-Mamba

Input: Infrared image $X_{\text{input}} \in R^{1\times H\times W}$, trained parameters Θ

Output: predicted target mask $\hat{Y} \in [0,1]^{1\times H\times W}$

$\{F_{bg}, A_{bg}, w_{fg}\} = \beta_\theta(X_{\text{input}})$

$F = \text{InitialFusion}(\text{X}_{\text{input}}, F_{bg})$

$e_1 = \text{RELU}(\text{BN}(\text{Co nv}_{3\times3}(F)))$; $e_1 = \text{MaxPool}_{2\times2}(e_1)$

$e_2 = \text{SGEncoder}(e_1)$; $e_2 = \text{MaxPool}_{2\times2}(e_2)$

$e_3 = \text{SGEncoder}(e_2)$; $e_3 = \text{MaxPool}_{2\times2}(e_3)$

$e_4 = \text{SGEncoder}(e_3)$; $e_4 = \text{MaxPool}_{2\times2}(e_4)$

$e_5 = \text{DSAFM}(e_4)$

$d_5 = \text{UpConv}(e_5)$; $d_5 = \text{ConvBlock}(\text{Concat}(d_5, e_5))$

$d_4 = \text{UpConv}(d_5)$; $d_4 = \text{ConvBlock}(\text{Concat}(d_4, e_4))$

$d_3 = \text{UpConv}(d_4)$; $d_3 = \text{ConvBlock}(\text{Concat}(d_3, e_3))$

$d_2 = \text{UpConv}(d_3)$; $d_2 = \text{ConvBlock}(\text{Concat}(d_2, e_2))$

$d_1 = \text{UpConv}(d_2)$; $d_1 = \text{ConvBlock}(\text{Concat}(d_1, e_1))$

$d_1 = d_1 \odot w_{fg}$

$A_{\text{fused}} = A_{\text{channel}} \odot A_{\text{spatial}}$; $A_{\text{enhanced}} = A_{\text{fused}}^{1.3}$

$A_s = \text{AvgPool2d}(A_{\text{enhanced}}, \text{kernel} = 3, \text{padding} = 1)$

$w_{\text{fusion}} = 0.6 \times A_{\text{final}} + 0.4 \times (1.0 - A_{\text{bg}})$

$d_1 = d_1 \odot (1.0 + 0.3 \times w_{\text{fusion}})$

$\hat{Y} = \sigma(\text{Conv}_{1\times1}(d_1))$

return $\hat{Y}$

Algorithm 1 defines the forward pass: background suppression, initial fusion, four-stage SG-encoder downsampling, Direction-specific SSM with adaptive fusion Mamba (DSAFM) processing, decoder upsampling with skip connections, target-enhancing gating, and final convolution to produce the probability map. Algorithm 2 defines training: random initialization, iterative mini-batch sampling, forward pass via Algorithm 1, composite binary cross-entropy and intersection over union (BCE-IoU) loss computation, gradient descent update, and convergence.

Algorithm 2 Training for BSG-Mamba

Input: dataset $D = \{(X_k, M_k)\}_{k=1}^{K}$, batch size B, epoch E, learning rate η, loss weight λ

Output: optimized parameters Θ

Initialize Θ randomly

for epoch = 1 to E do

for each mini-batch $\{(X_k, M_k)\}_{k=1}^{K} \subset D$ do

$$\hat{Y} = \text{BSG-Mamba}(X_i : \Theta)$$

$$L = \frac{1}{B}\sum_{i=1}^{B}[BCE(\hat{Y}, M_i) + \lambda \cdot IoU((\hat{Y}, M_i)]$$

$$\Theta = \Theta - \eta \nabla_{\Theta} L$$

end for
end for
return Θ

3.2. Background suppression

In infrared images, background noise typically dominates, while target signals are weak and small in size. Traditional methods that detect targets directly across the entire image are prone to background interference. This work models the background distribution to generate a background attention map, thereby suppressing background responses and enhancing the target region at the feature level. The background suppression module is a two-stage processing module that achieves dynamic suppression of the background and enhancement of the target in infrared images through forward background feature extraction and backward target saliency evaluation, as shown in Figure 2.

Let the infrared image be $X_{\text{input}} \in R^{B\times C\times H\times W}$, where B denotes the batch size, C is the input channel number, and H and W represent the height and width of the input feature map, respectively. All convolutional layers adopt same padding with a stride of 1. The max-pooling layer uses a stride of 2. Θ denotes element-wise multiplication, σ represents the sigmoid activation function, and the rectified linear unit (ReLU) is adopted as the default activation function for all intermediate layers.

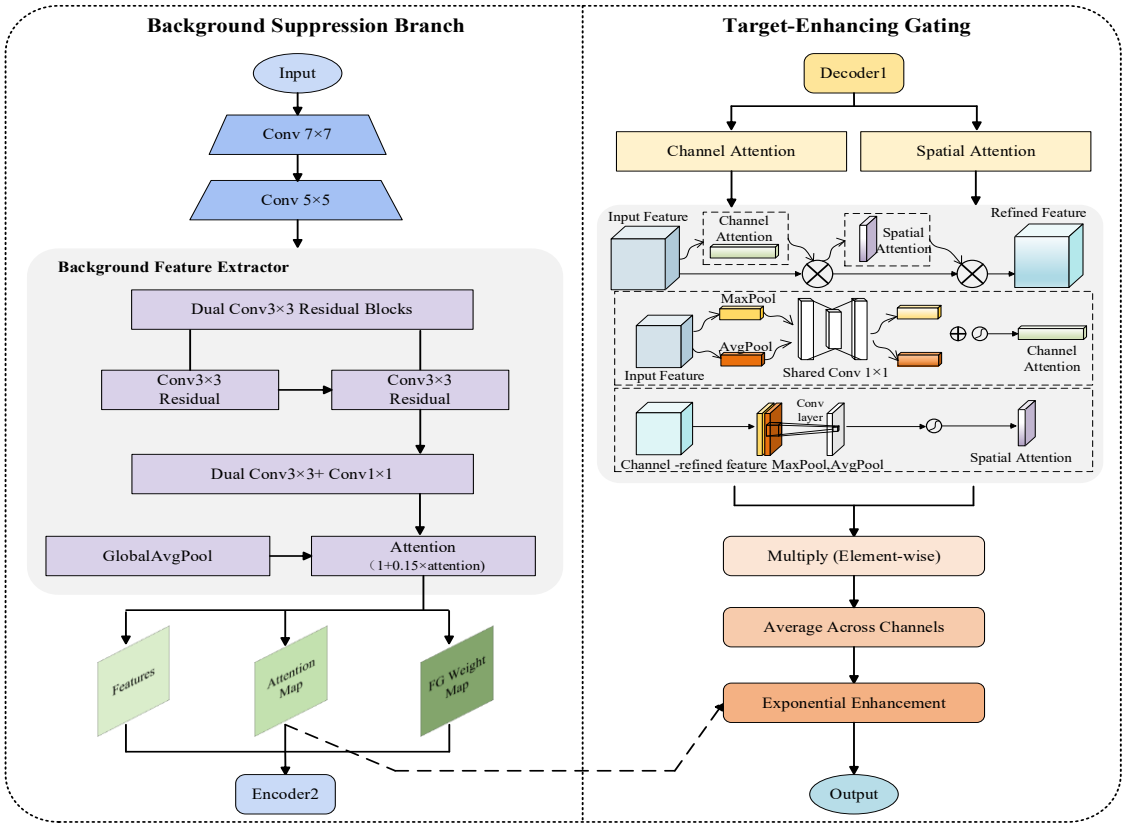


Figure 2. Background suppression and TEG mechanisms.

3.2.1. Background suppression branch (BSB)

First, the input image undergoes multi-scale background feature extraction via the background suppression branch (BSB). A three-layer convolutional network is used for progressive downsampling to capture background textures at different scales, ranging from coarse to fine features. A large 7 × 7 convolutional kernel is employed to capture the background over a wide area, and max-pooling is applied to downsample by a factor of 2, thereby reducing computational load while preserving the primary background structure. A 5 × 5 convolutional kernel is used to focus on medium-scale background textures, while a 3 × 3 kernel extracts local fine-scale features. The final step does not require downsampling, preserving more spatial information for subsequent feature reconstruction. As shown in Eqs (1)–(4),

$$F_i^{\text{conv}} = \text{Conv}_{k_i \times k_i}(F_{i-1}), \tag{1}$$

$$F_i^{\text{bn}} = \text{BatchNorm}(F_i^{conv}), \tag{2}$$

$$F_{\text{i}}^{\text{relu}} = \text{ReLU}(F_{\text{i}}^{\text{bn}}), \tag{3}$$

$$F_i = \text{MaxPool}_{2\times2}(F_i^{\text{relu}}), \tag{4}$$

where $k_1 = 7$, $k_2 = 5$, $k_3 = 3$, and $F_0 = X_{\text{input}}$ is the input image.

Add an additional convolutional layer to the third-level features and sum the result with the original features. Residual connections allow the network to learn incremental improvements to background features rather than completely altering the feature representations. As shown in Eqs (5) and (6),

$$F_3^{\text{ext}} = \text{Conv}_{3\times3}(F_3), \tag{5}$$

$$F_3^{\text{res}} = F_3 + \text{ReLU}(\text{BatchNorm}(F_3^{\text{ext}})). \tag{6}$$

Global average pooling (GAP) is utilized to aggregate global channel information, and two fully connected layers with a dropout layer are applied to generate channel attention weights α. As shown in Eqs (7) through (10),

$$g = \text{GAP}(F_3^{res}), \tag{7}$$

$$g_1 = \text{RELU}(\text{FC}_1(g)), \tag{8}$$

$$g_2 = \text{Dropout}(g_1), \tag{9}$$

$$\alpha = \sigma(FC_2(g_2)), \tag{10}$$

where GAP stands for global average pooling, and *FC* stands for fully connected layer. Attention weights are applied to enhance features as shown in Eq (11), where $\odot$ denotes element-wise multiplication:

$$F_{\text{bg}} = F_3^{\text{res}} \odot (1.0 + 0.15 \cdot \alpha). \tag{11}$$

The extracted background features are upsampled back to the original input size, and the spatial background attention map is obtained through two stacked convolutional layers: First, an intermediate feature map is generated, and then the final background attention map is produced via the second convolution and sigmoid activation. As shown in Eqs (12)–(14),

$$F_{\text{bg}} = \text{Upsample}(F_{\text{bg}}, \text{scale} = 4), \tag{12}$$

$$F_{\text{bg}}^{\text{mid}} = \text{RELU}(\text{Conv}_{3\times3}(F_{bg})), \tag{13}$$

$$A_{\text{bg}} = \sigma(\text{Conv}_{3\times3}(F_{bg}^{\text{mid}})). \tag{14}$$

The background feature map F_{bg} is upsampled to the original input resolution via bilinear interpolation with a scaling factor of 4. Then two stacked 3×3 convolutions and a sigmoid activation are used to generate the spatial background attention map A_{bg}.

The enhanced weights for the foreground regions are calculated based on the background attention map, as shown in Eq (15). This equation inverts the background attention map into a foreground weight map. In regions with high background probability, the weights approach 1.0, meaning neither enhancement nor suppression occurs; in regions with low background probability, the weights are increased to 1.15, thereby enhancing the target regions. The coefficient of 0.15 controls the intensity of the enhancement, balancing the need for background suppression and target enhancement.

$$w_{\text{fg}} = 1.0 + 0.15 \cdot (1.0 - A_{\text{bg}}). \tag{15}$$

The forward propagation of the entire background suppression branch can be expressed as shown in Eq (16),

$$\{F_{\text{bg}}, A_{\text{bg}}, w_{\text{fg}}\} = \beta_\theta(X_{\text{input}}). \tag{16}$$

The core of this background suppression branch models the background and foreground separately; by specifically learning background features, the network can accurately identify which information belongs to the distracting background, thereby reducing its impact in subsequent processing.

3.2.2. Target-enhancing gating (TEG)

At the end of decoder, feature maps contain both actual target information and residual background noise. Target-based gating evaluates the saliency of the target at each location through the collaborative action of channel-space attention mechanisms.

By fusing channel attention A_{channel} and spatial attention A_{spatial} through element-wise multiplication, high response values are obtained only when a location is significant in both the channel and spatial dimensions. A power transformation is then applied to the fused attention map to enhance contrast. As shown in Eqs (17) and (18),

$$A_{\text{fused}} = A_{\text{channel}} \odot A_{\text{spatial}}, \tag{17}$$

$$A_{\text{enhanced}} = A_{\text{fused}}^{1.3}. \tag{18}$$

We adopt the exponential function $f(x) = x^{1.3}$ to optimize the targetness map. This transformation suppresses low-value background noise while well preserving weak and strong target signals, as validated by quantitative results. It strikes a favorable balance between clutter suppression and target feature retention and further improves the signal-to-clutter ratio for infrared small target detection.

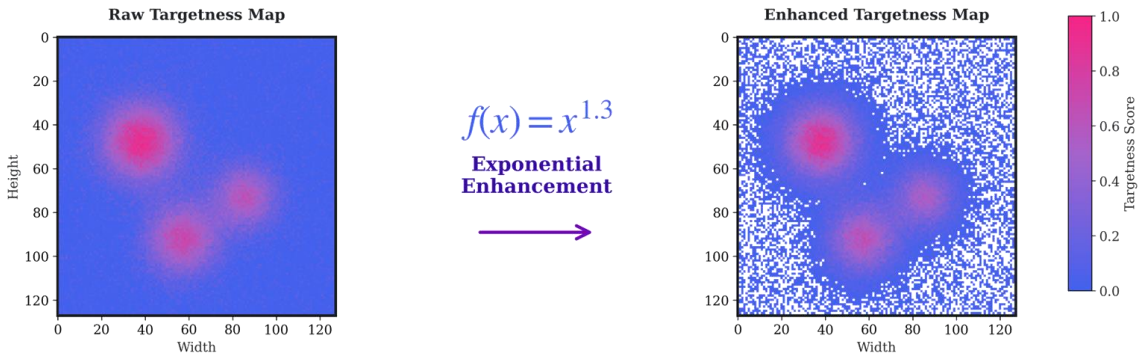
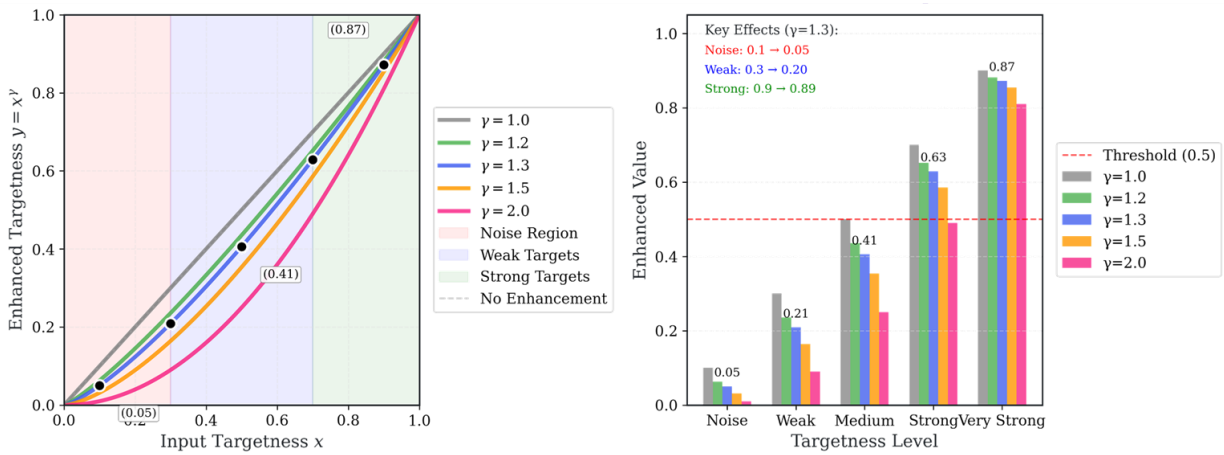


Figure 3. Targetness heatmap enhancement.



(a) (b)

Figure 4. Exponential enhancement for IRSTD. (a) Enhancement curves, (b) enhancement comparison.

Two-level average pooling eliminates single-point noise responses and enhances the continuity of the target region. As shown in Eqs (19) and (20),

$$A_s = \text{AvgPool2d}\left(A_{\text{enhanced}}, \text{kernel} = 3, \text{padding} = 1\right), \tag{19}$$

$$A_{\text{final}} = \text{AvgPool2d}\left(A_s, \text{kernel} = 3, \text{padding} = 1\right). \tag{20}$$

The background attention map obtained from the background suppression branch is weight-blended with the target saliency map provided by the target-specific gating, and the blended weights are applied to the decoder features to produce the final output features of the decoder. As shown in Eqs (21) and (22),

$$w_{\text{fusion}} = 0.6 \times A_{\text{final}} + 0.4 \times (1.0 - A_{\text{bg}}), \tag{21}$$

$$X_{\text{output1}} = X_{\text{output}} \odot (1.0 + 0.3 \times w_{\text{fusion}}). \tag{22}$$

To verify the rationality of fusion weights 0.6 and 0.4 in the TEG module, we conduct ablation experiments on different weight combinations. We restrict the sum of the two weights to 1 and keep all other network settings unchanged. All experiments are implemented on the NUAA-SIRST dataset, and the results are shown in Table 1.

Table 1. Performance of different fusion weights on NUAA-SIRST.

Weight Combination (λ_1, λ_2)	*IoU* (%)	*Pd* (%)	*Fa* (10^{-6})
(0.9,0.1)	76.70	97.41	15.21
(0.8,0.2)	76.82	97.56	14.33
(0.6,0.4)	**76.91**	**97.74**	**13.65**
(0.5,0.5)	76.85	97.62	14.02
(0.4,0.6)	76.61	97.28	16.15

In the fusion formula, λ_1 is the weight of target saliency map A_{final}, and λ_2 corresponds to the weight of reversed background attention map (1.0-A_{bg}). As observed from the table, the weight pair (0.6, 0.4) achieves the highest intersection-over-union (*IoU*) and probability of detection (*Pd*), along with the lowest false alarm ration (*Fa*). For infrared small target detection, accurate target enhancement is the primary objective. Assigning a larger weight 0.6 to target saliency features ensures the model focuses on highlighting weak small targets. The weight 0.4 serves as a moderate background constraint to suppress interference. When λ_1 is too large, the model ignores background noise and produces more false alarms. When λ_1 is smaller than 0.6, the ability to enhance target features declines. The equal-weight setting also fails to reach optimal performance. Therefore, (0.6, 0.4) is selected as the optimal fusion weight combination for the TEG module.

3.3. Sparse graph encoder (SG-Encoder)

Pixels in an image do not exist in isolation, rather, they form structures through spatial relationships and semantic associations. Traditional convolutional networks only consider pixel interactions within local neighborhoods, making it difficult to model non-local semantic associations. The SG-Encoder combines traditional CNNs with GNNs to model the non-local correlations of small objects in infrared images, as shown in Figure 5.

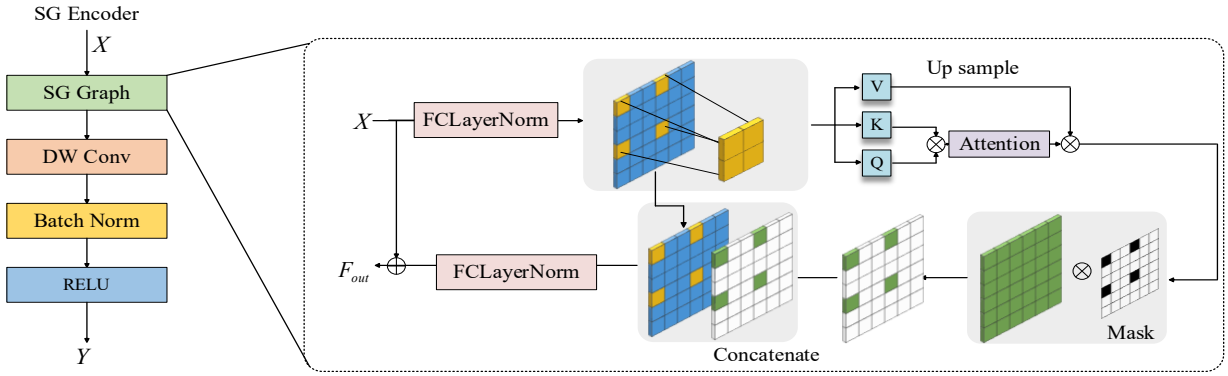


Figure 5. SG-encoder.

To balance computational efficiency and feature quality, we construct the map using sparse sampling, uniformly sampling the feature map at a stride of 4 to divide the image into (H/4) × (W/4) superpixel regions, with each superpixel representing a local area. As shown in Eq (23),

$$X_s[i,j,m,n] = X_{\text{input}}[i,j,4m,4n] \in R^{B\times C\times H_s\times W_s}. \tag{23}$$

In this structure, each superpixel node is connected to all other nodes, forming a complete graph. Rewrite the superpixel features as a sequence of nodes and compute self-attention. Here, $N = H_s \times W_s$ represents the number of nodes, and $d = \sqrt{C}$ is the scaling factor. Through the self-attention mechanism, each node updates its own features based on its similarity to all other nodes; features of similar nodes reinforce each other, thereby achieving global information aggregation. As shown in Eqs (24)–(27),

$$Q = \text{Reshape}(X_s) \in R^{B\times N\times C} \tag{24}$$

$$\text{Attn} = \text{Softmax}(\frac{QQ^T}{\sqrt{d}}) \in R^{B\times N\times N} \tag{25}$$

$$V = Q \tag{26}$$

$$X_{s1} = \text{Attn} \times V \in R^{B\times N\times C} \tag{27}$$

The node features are propagated back to the original grid and fused with the original features; sparse nodes are propagated to the dense network via upsampling and then fused with the original local features, thereby preserving both local details and global semantic information, as shown in Eqs (28)–(30),

$$X_{s1}^{2D} = \text{Reshape}(X_{s1}) \in R^{B\times C\times H_S\times W_s}, \tag{28}$$

$$X_1 = \text{Upsample}(X_{s1}^{2D}, \text{scale} = 4) \in R^{B\times C\times H\times W}, \tag{29}$$

$$F_{\text{out}} = \text{Conv}_{1\times1}(\text{Concat}(X, X_1)) \in R^{B\times C\times H\times W}. \tag{30}$$

We adopt a customized graph convolution module named MRConv4dSP for feature modeling. The full name of MRConv4dSP is multi-scale 4D sparse graph convolution. This module is specially designed for the sparse distribution characteristic of infrared small targets, which combines multi-scale convolution and sparse graph structure to capture non-local feature correlations while maintaining low computational complexity.

Different from traditional dense graph convolution, MRConv4dSP constructs a dynamic sparse adjacency matrix based on local sliding windows instead of KNN clustering, which greatly reduces computation overhead.

Given the input preprocessed feature $F_{\text{pre}} \in R^{B\times C\times H\times W}$, the calculation process of MRConv4dSP is formulated as

$$F_{\text{ms}}(F_{\text{pre}}) = g_{1\times1}(F_{\text{pre}}) \oplus g_{3\times3}(F_{\text{pre}}) \oplus g_{7\times7}(F_{\text{pre}}), \tag{31}$$

$$F_{\text{graph}} = A \odot F_{\text{ms}}(F_{\text{pre}}). \tag{32}$$

where $g_{k\times k}$ denotes standard $k \times k$ convolution operation; four branches with different kernel sizes extract multi-scale features of infrared targets. $\oplus$ represents element-wise addition for multi-scale feature fusion. $A \in R^{B\times C\times H\times W}$ is the dynamic sparse adjacency matrix, which is adaptively updated according to the similarity between feature nodes in each forward propagation. The dynamic sparse adjacency matrix A only retains strong correlation connections between feature nodes related to potential targets and abandons invalid connections in homogeneous background areas, which realizes sparse graph modeling.

Finally, graph convolutions are integrated into the grapher module: Pre-processing convolutions F_{pre} adjust the feature representation, graph convolutions F_{graph} capture non-local dependencies, post-processing convolutions F_{post} refine the features, and a residual connection ensures the complete transmission of feature information. As shown in Eqs (33)–(35),

$$F_{\text{pre}} = \text{BN}(\text{Conv}_{3\times3}(X)), \tag{33}$$

$$F_{\text{graph}} = \text{MRConv4dSP}(F_{\text{pre}}), \tag{34}$$

$$\text{Grapher}(X) = \text{Dropath}(F_{\text{post}}) + X. \tag{35}$$

3.4. Direction-specific SSM with adaptive fusion Mamba (DSAFM)

Traditional CNNs are limited by their local receptive fields, making it difficult to model long-range dependencies in images. In target detection tasks, targets can appear anywhere in an image, and their detection requires the integration of global contextual information. By using SSMs and scanning mechanisms to serialize images into directional scanning sequences, and by capturing long-range spatial dependencies through recursive state propagation, these models overcome the limitations of traditional convolution networks, as shown in Figure 6(b).

The given input feature map $x \in R^{B\times C\times H\times W}$ is processed into four 1D sequences, and the results of the four-directional scans are fused by averaging, as shown in Eqs (36)–(39),

$$s_0 = \text{Reshape}(x) + p_0 \in R^{B\times L\times C}, \tag{36}$$

$$s_1 = \text{Reshape}(\text{Flip}_W(x)) + p_1 \in R^{B\times L\times C}, \tag{37}$$

$$s_2 = \text{Reshape}(\text{Permute}_{H\leftrightarrow W}(x)) + p_2 \in R^{B\times L\times C}, \tag{38}$$

$$s_3 = \text{Reshape}(\text{Flip}_H(\text{Permute}_{H\leftrightarrow W}(x))) + p_3 \in R^{B\times L\times C}. \tag{39}$$

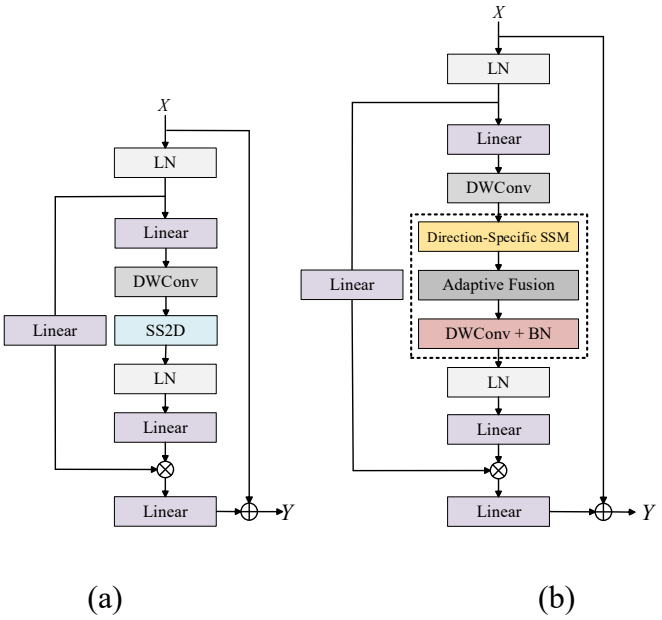


Figure 6. Mamba network, (a) Vmamba framework, (b) DSAFM.

Unidirectional scanning can only capture unidirectional dependencies, while four-directional scanning covers all directions in a 2D image. Each direction is equipped with independent position encoding Pd and independent SSM parameters.

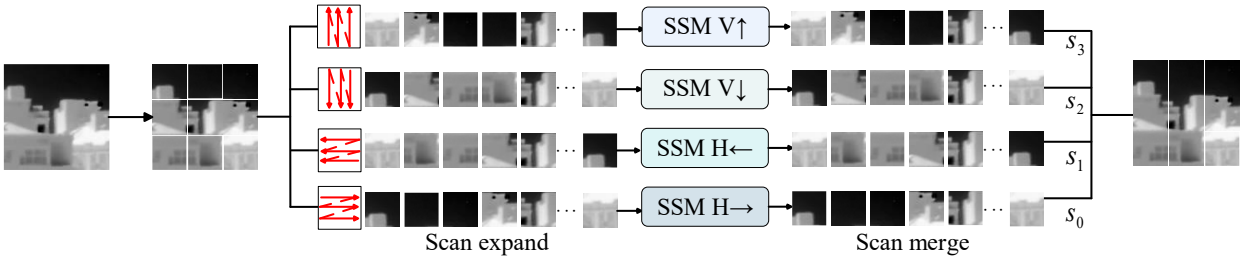


Figure 7. Directional SSM scan.

A SSM is a mathematical framework commonly used to model dynamic systems. At its core are the state equations, which describe how the hidden states *h(t)* change over time, as shown in Eq (40),

$$\frac{dh(t)}{dt} = A \cdot h(t) + B \cdot x(t), \tag{40}$$

$$y(t) = C \cdot h(t), \tag{41}$$

where $h(t) \in R^N$ is the hidden state, $x(t)$ is the input vector, A is the state transition matrix, which determines how the state naturally evolves over time, B is the input projection matrix, which determines how external inputs affect state changes, C is the output projection matrix, and $y(t)$ is the output.

Unlike traditional SSMs with fixed parameters, this model introduces input-dependent dynamic parameters,

$$\Delta_t = 0.1 + 0.9 \cdot \sigma(Linear_{\Delta}(x_t)), \tag{42}$$

$$B_t = Linear_B(x_t). \tag{43}$$

In this case, Δ_t is not a fixed value, but rather the x_t function of the input from the previous step t. $Linear_B(\cdot)$ is a linear transformation layer, and the B matrix for each time step is computed based on the current input.

$$A = -\exp(A_{\log}), \tag{44}$$

where A_{log} represents the trainable parameters; applying a negative exponent ensures that the eigenvalues of the matrix are negative, thereby guaranteeing the system's stability.

Using the first-order Euler discretization method, the continuous-time system is converted into a discrete-time system. For each time step t, the input projection is

$$x_t^{proj} = x_t \cdot W_B^T, \tag{45}$$

where W_B^T is the projection weight matrix.

Below are status updates and output calculations, as shown in Eqs (46) and (47),

$$h_t = h_{t-1} + \Delta_t \cdot [A \cdot h_{t-1} + B_t \odot x_t^{proj}], \tag{46}$$

$$y_t = C \cdot h_t. \tag{47}$$

This module employs direction-specific SSMs designed for scanning in different directions. The horizontal scanning SSM is used to capture left-right context, utilizing an asymmetric convolution kernel with kernel_size = 5 to enhance the horizontal receptive field. The A_{log} parameters are initialized to smaller values, enabling the model to better capture dependencies in the left-right domain. The vertically scanning SSM uses a standard 3 × 3 convolutional kernel to capture vertical context. It employs a larger A_{log} value and introduces a learnable vertical attenuation mechanism to adjust the strength of long-range dependencies in the vertical direction, thereby enhancing the model's ability to capture long-range dependencies in the vertical direction. Each direction is processed by an independent SSM, and each direction has its own independent spatial encoding to prevent confusion between spatial information from different directions:

$$Y_d = SSM_d(s_d), d \in \{0,1,2,3\}. \tag{48}$$

After restoring the image shape, perform weighted fusion using trainable weights $w = [w_0, w_1, w_2, w_3]$:

$$Y_{\text{fused}} = \sum_d \frac{e^{w_d}}{\sum_j e^{w_j}} \cdot Y_d. \tag{49}$$

SSM excels at handling long-range dependencies; by incorporating local convolutional enhancement modules, it achieves synergy between global and local features:

$$X_{\text{local}} = \text{Depthwise}Conv_{3\times3}(Y_{\text{fused}}), \tag{50}$$

$$Y_{\text{enhanced}} = Y_{\text{fused}} + 0.3 \cdot X_{\text{local}}. \tag{51}$$

The final output includes adaptive residual gating:

$$Y = Y_{\text{enhanced}} + \beta \cdot X_{\text{align}}, \beta = \text{res\_gate}, \tag{52}$$

where β is the learnable residual gate coefficient, and X_{align} is the input feature after channel alignment.

4. Experiments

4.1. Data sets and experimental details

In this work, we analyzed three publicly available datasets, including NUAA-SIRST [16], NUDT-SIRST [19], and IRSTD-1K [25]. The first two datasets are used for main experiments, and IRSTD-1K is adopted for supplementary quantitative evaluation. The NUAA-SIRST dataset contains 427 infrared images of varying sizes, while the NUDT-SIRST dataset contains 1327 infrared images of the

same size. All experiments were conducted on an NVIDIA GeForce RTX 3060 GPU platform under completely unified configurations to ensure fair comparison. During the training phase, each image in the dataset was normalized and randomly cropped to a size of 256 × 256 pixels.

The Adam optimizer was used for model training, with an initial learning rate set to 1e-4. For each dataset, the number of training epochs was set to 600, and the batch size was 16. The MultiStepLR scheduler was utilized to adjust the learning rate, with a decay factor of 0.5 at epochs 250, 350, 450, and 550. A dropout rate of 0.2 was adopted to prevent overfitting. We further performed hyperparameter sensitivity analysis with the control variable strategy. Multiple candidate values were tested for key hyperparameters, and the final settings were selected based on quantitative results on two standard datasets. This proves that our hyperparameter settings are reasonable and reliable.

The detailed results of hyperparameter sensitivity analysis are presented in Table 2. The bolded combinations achieved the highest *IoU* and *Pd* and the lowest *Fa* on both the NUAA-SIRST and NUDT-SIRST datasets. Excessively large or small hyperparameters would lead to underfitting, training oscillations, and an increase in the false alarm rate. Therefore, we adopted this combination in all experiments.

Table 2. The results of the hyperparameter sensitivity analysis for the BSG-Mamba model.

γ	lr	Batch size	Dropout	NUAA-SIRST			NUDT-SIRST		
				IoU (%)	*Pd* (%)	*Fa* (10^{-6})	*IoU* (%)	*Pd* (%)	*Fa* (10^{-6})
0.4	5e-5	8	0.1	73.01	96.54	24.08	89.88	97.56	6.78
0.5	1e-4	16	0.2	76.91	97.74	13.65	92.02	98.90	3.67
0.5	3e-4	16	0.2	76.12	96.89	17.46	90.03	97.90	5.83
0.6	5e-4	24	0.3	74.55	95.67	21.97	87.94	96.56	9.10
0.7	1e-3	32	0.4	73.42	95.46	31.23	85.45	92.34	10.23

To comprehensively evaluate algorithm performance, this paper selected 17 methods for comparison, including traditional methods such as Top-Hat [7], IPI [50], and NOLC [51] and deep learning algorithms such as U-Net [14], ACM [16], AGPCNet [17], UIU-Net [18], DNA-Net [19], U-Transformer [20], MTU-Net [21], APTNet [23], STASPPNet [24], IAA-Net [33], BPGA [40], GCLNet [41], MiM-ISTD [42], and MOU-Mamba [47]. This paper analyzes and trains the above algorithms to compare the performance of each network model under identical experimental conditions.

4.2. Assessment of indicators

This paper employs standard metrics to evaluate the model's performance in small object detection. The *IoU* and *nIoU* are used to measure prediction accuracy. The *Pd* and *Fa* indicate the proportion of correctly predicted objects. *Pd* measures the ratio of correctly predicted objects to the total number of objects. *Fa* measures the proportion of incorrectly predicted pixels out of all image pixels. The formulas for these metrics are as follows:

$$IoU = \frac{TP}{T + P - TP}, \tag{53}$$

$$nIoU = \frac{1}{N}\sum_{i}^{N}\frac{TP_i}{T_i + P_i - TP_i}, \tag{54}$$

$$Pd = \frac{T_{correct}}{T_{All}}, \tag{55}$$

$$Fa = \frac{P_{False}}{P_{All}}. \tag{56}$$

True positives (*TP*) represent the number of pixels correctly predicted as the positive class by the model. *T* stands for the number of pixels with the true label, *P* stands for the number of pixels with the predicted label, and *N* stands for the total number of pixels.

The *Fa* calculates the proportion of false target pixels to all image pixels. For intuitive comparison, we scale the original *Fa* results by 10^{-6} in all experimental tables. That is, the listed values stand for false alarm counts per one million pixels. This standard normalization rule is widely adopted in infrared small target detection tasks. Lower *Fa* corresponds to fewer false detections and stronger anti-interference ability.

The F1-Score (F1) mean of precision and recall provides an effective metric for evaluating performance, as shown in the following formula:

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R}. \tag{57}$$

The receiver operating characteristic (ROC) curve is used to describe the trend of *Pd* across different *Fa*. Plotting the false alarm rate on the x-axis and the detection rate on the y-axis yields a curve ROC for a given sequence. The more convex the curve ROC, the better the detection performance of the method for that sequence. Area under curve (*AUC*) is a metric for evaluating the curvature of the curve ROC, representing the algorithm's ability to distinguish between targets and background during detection.

In target detection networks, floating-point operations (FLOPs), the number of parameters (Params), and frames per second (FPS) are core metrics for measuring a model's computational efficiency, memory overhead, and inference speed. Among these, FLOPs are used to evaluate the model's forward computation complexity; a lower value indicates higher computational efficiency; Params reflect the model's scale; fewer parameters typically indicate a more lightweight model structure, which helps improve generalization ability and reduce the risk of overfitting; FPS directly characterizes the model's inference speed—a higher value indicates that the model requires less time to process a single image, making it easier to meet real-time requirements in practical applications.

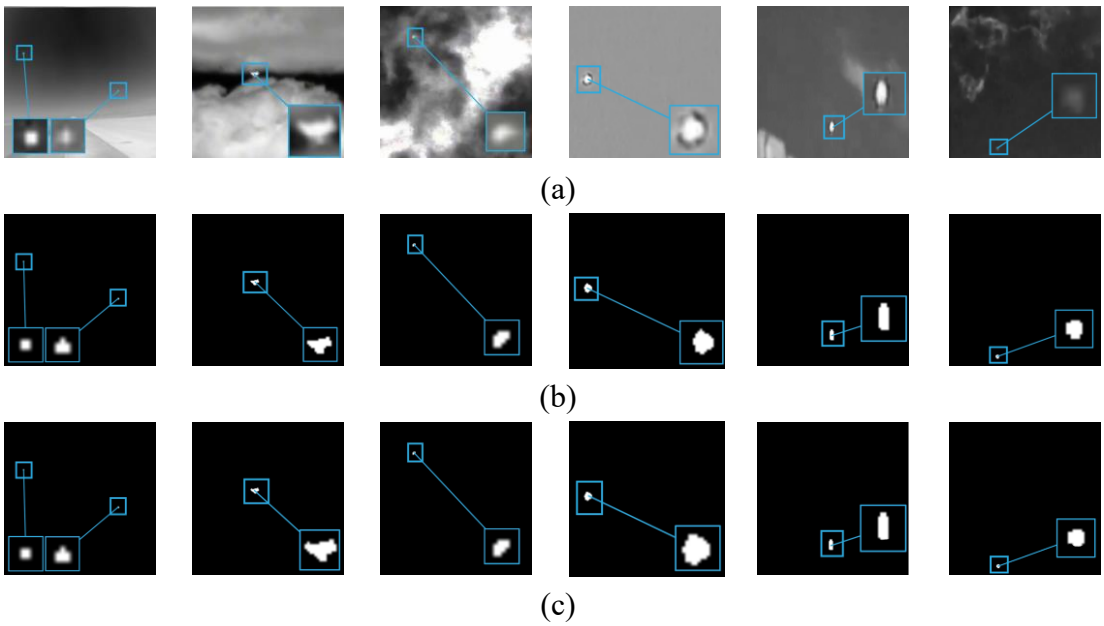
4.3. Qualitative analysis

This work conducted a qualitative evaluation on the NUAA-SIRST and NUDT-SIRST datasets. Infrared images depicting various complex scenes were selected from each dataset to reproduce and visually analyze the results of the BSG-Mamba model against representative methods from recent years, including AGPCNet, UIU-Net, DNA-Net, IAA-Net, MTU-Net, and GCLNet. As shown in Figures 8 and 9, the blue boxes indicate small target detections, while the

red boxes indicate false positives.

Figure 8 presents the detection results on the NUAA-SIRST dataset. For this experiment, images of the sky were selected, which were further categorized into simple cloud layers and high-brightness cloud layers. Figure 9 shows the detection results on the NUDT-SIRST dataset, where complex environments such as buildings, forests, fields, and water surfaces were selected for validation. The experiment in Figure 8 places greater emphasis on false alarm control, while the experiment in Figure 9 focuses more on target extraction capabilities in complex backgrounds. The experimental results indicate that small infrared targets occupy an extremely small pixel area in the image, with low contrast between the target and the background and blurred edges—this is the core challenge of this task. As shown by the ground truth (GT), the true targets vary in shape, ranging from point-like to irregular shapes, and multi-target scenarios are present.

As shown in Figure 8, with the exception of AGPCNet, UIU-Net, GCLNet, and DNA-Net—which exhibit obvious false positives in the seventh image—the remaining models can accurately identify targets. However, they differ in terms of background suppression and edge detection. In the second and fifth images, AGPCNet displays slight blurring of target edges, resulting in imprecise target shapes, and its false alarm rate control requires improvement; In the third and fourth images, UIU-Net exhibits overly smooth transitions along the target edges, resulting in the loss of detailed contour information; DNA-Net suffers from the same issue; and IAA-Net fails to fully fill the interior of larger targets. For example, in the fourth image, the target is not fully recognized and is prone to being misclassified as two separate targets. In contrast, BSG-Mamba performs consistently across the test samples, accurately locating the target position while maintaining clear edge contours.



continued on next page

Figure 8. The detection results of different network models on the NUAA-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.

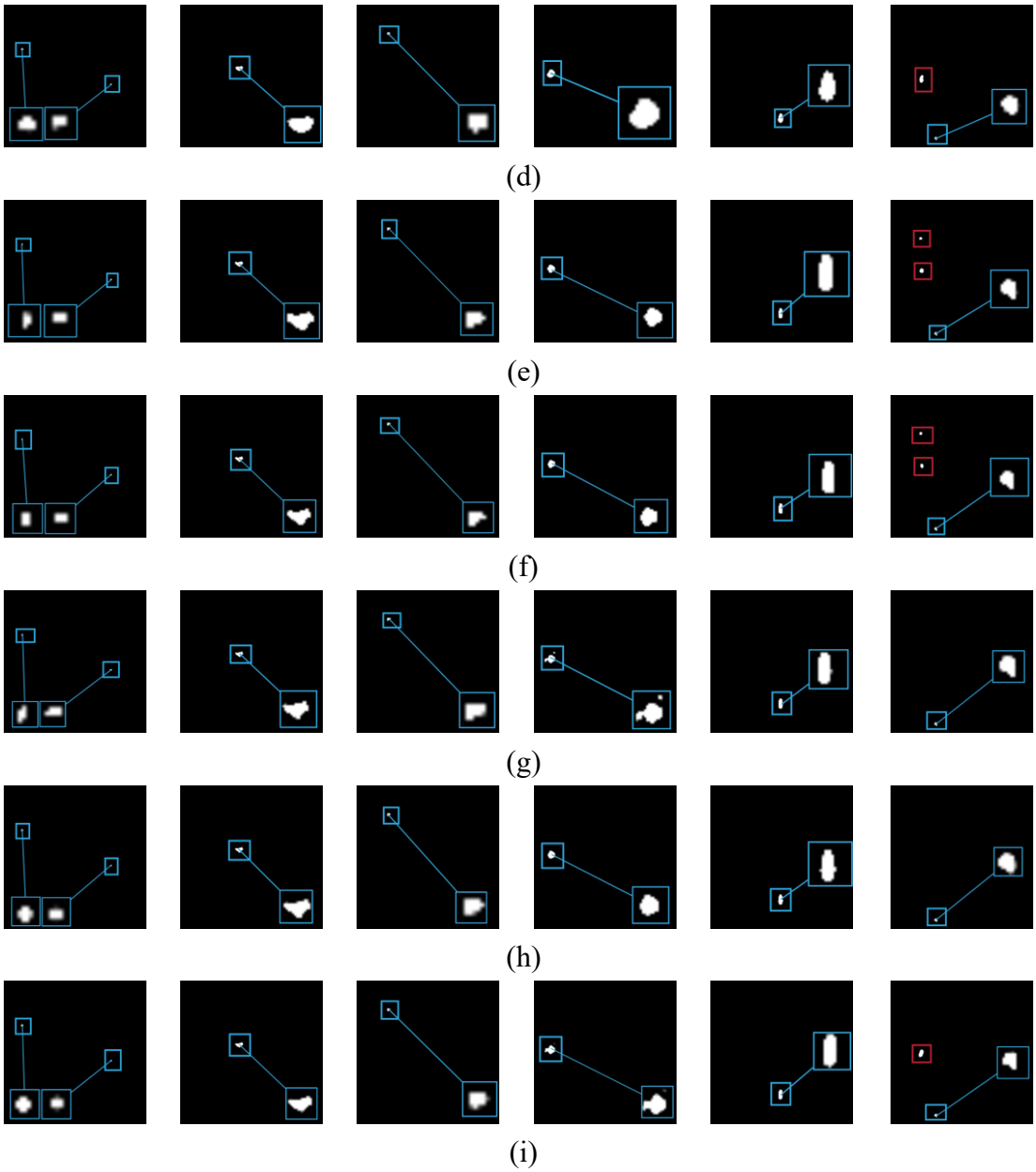
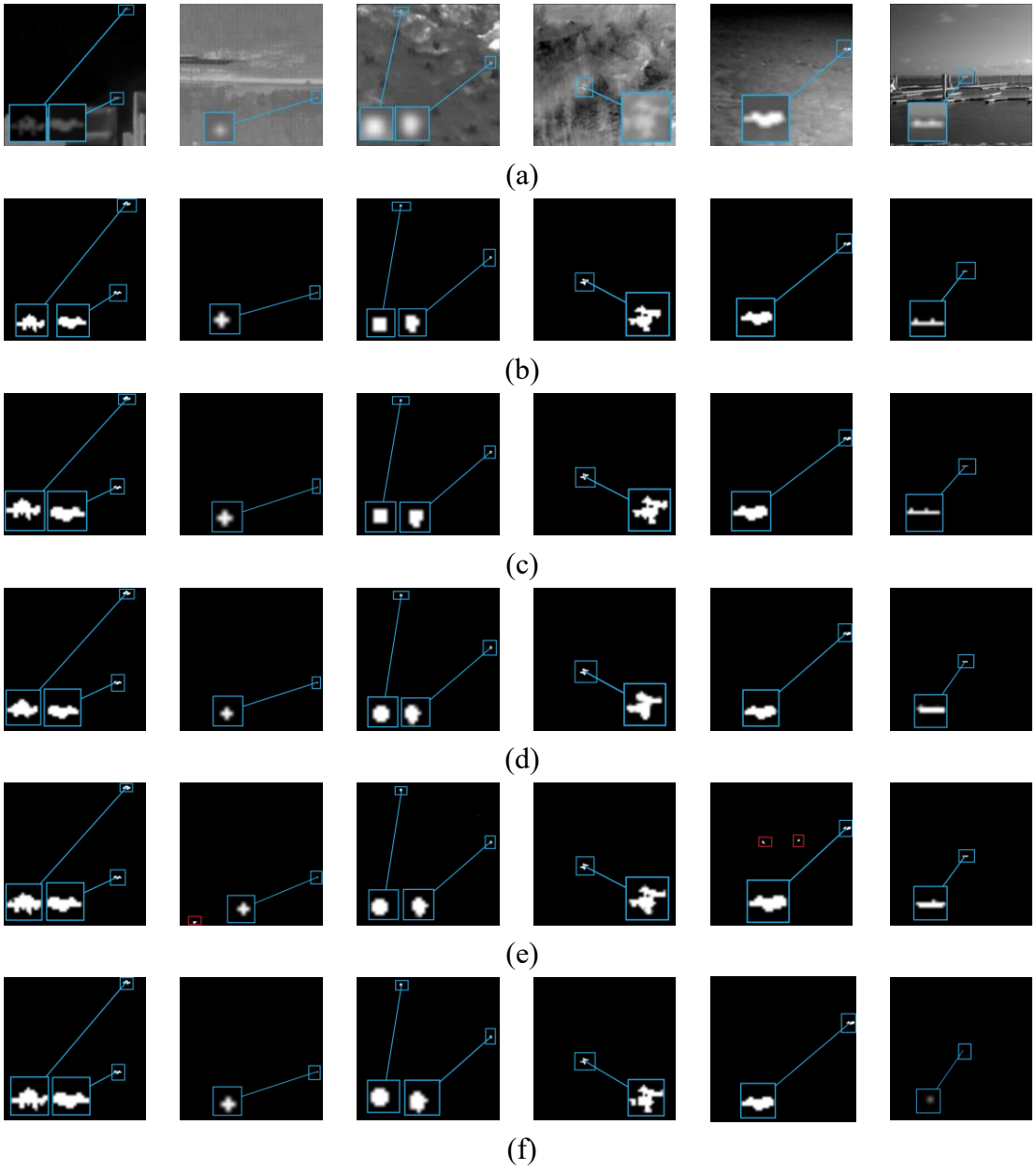


Figure 8. The detection results of different network models on the NUAA-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.

The image scenes selected from the NUDT-SIRST dataset are more complex, featuring a wider variety of object types and more challenging backgrounds. As shown in Figure 9, in the third image, both AGPCNet and UIU-Net exhibit morphological distortion, demonstrating insufficient segmentation capabilities; the object edges are blurred, and their ability to preserve object shapes is inferior to that of BSG-Mamba. In the drone detection task in the fourth image, DNA-Net exhibits fragmentation, with a single complete object being segmented into multiple parts; in the vessel recognition task in the seventh image, most models suffer from blurred edges and fragmentation of complete objects, while DNA-Net only identifies point-like targets and lacks precision in detecting area-like targets. Both MTU-Net and GCLNet exhibit fragmentation in

vessel recognition, whereas BSG-Mamba's shapes and edges are closer to the ground truth labels. BSG-Mamba outperforms most comparison models overall, particularly in multi-object detection and shape preservation, and demonstrates higher detection accuracy than most comparison models in high-brightness noise environments.



continued on next page

Figure 9. The detection results of different network models on the NUDT-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.

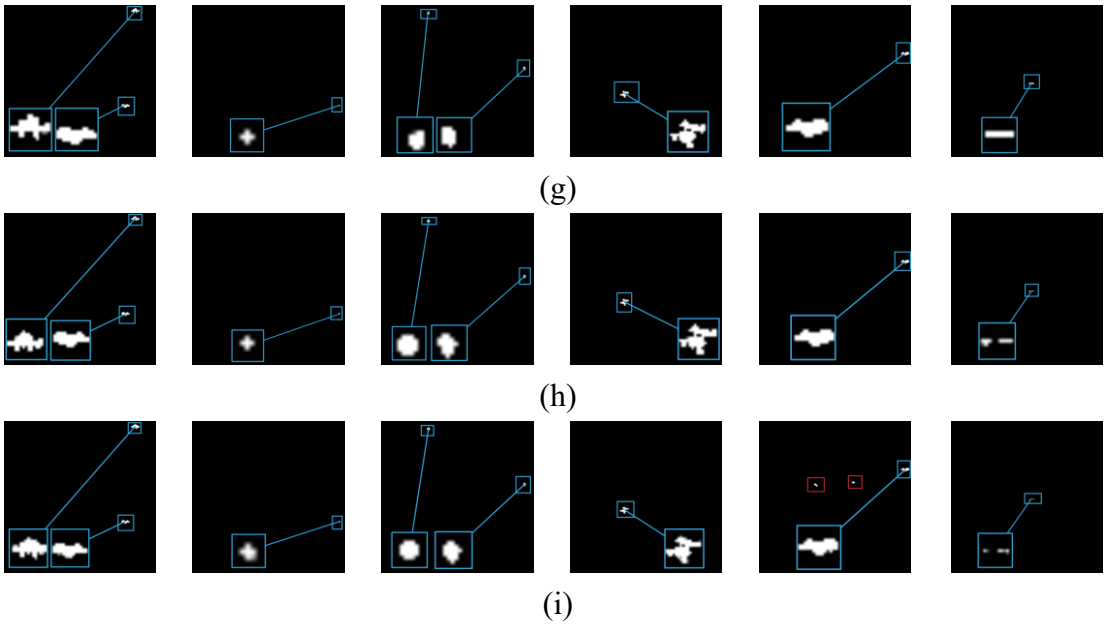
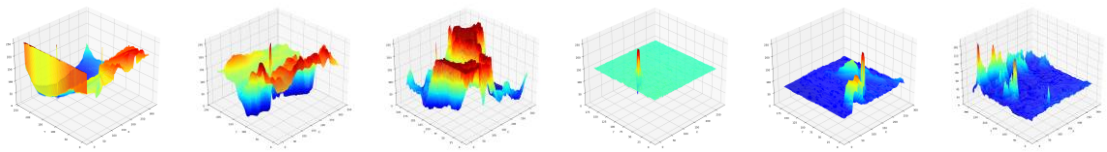


Figure 9. The detection results of different network models on the NUDT-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.

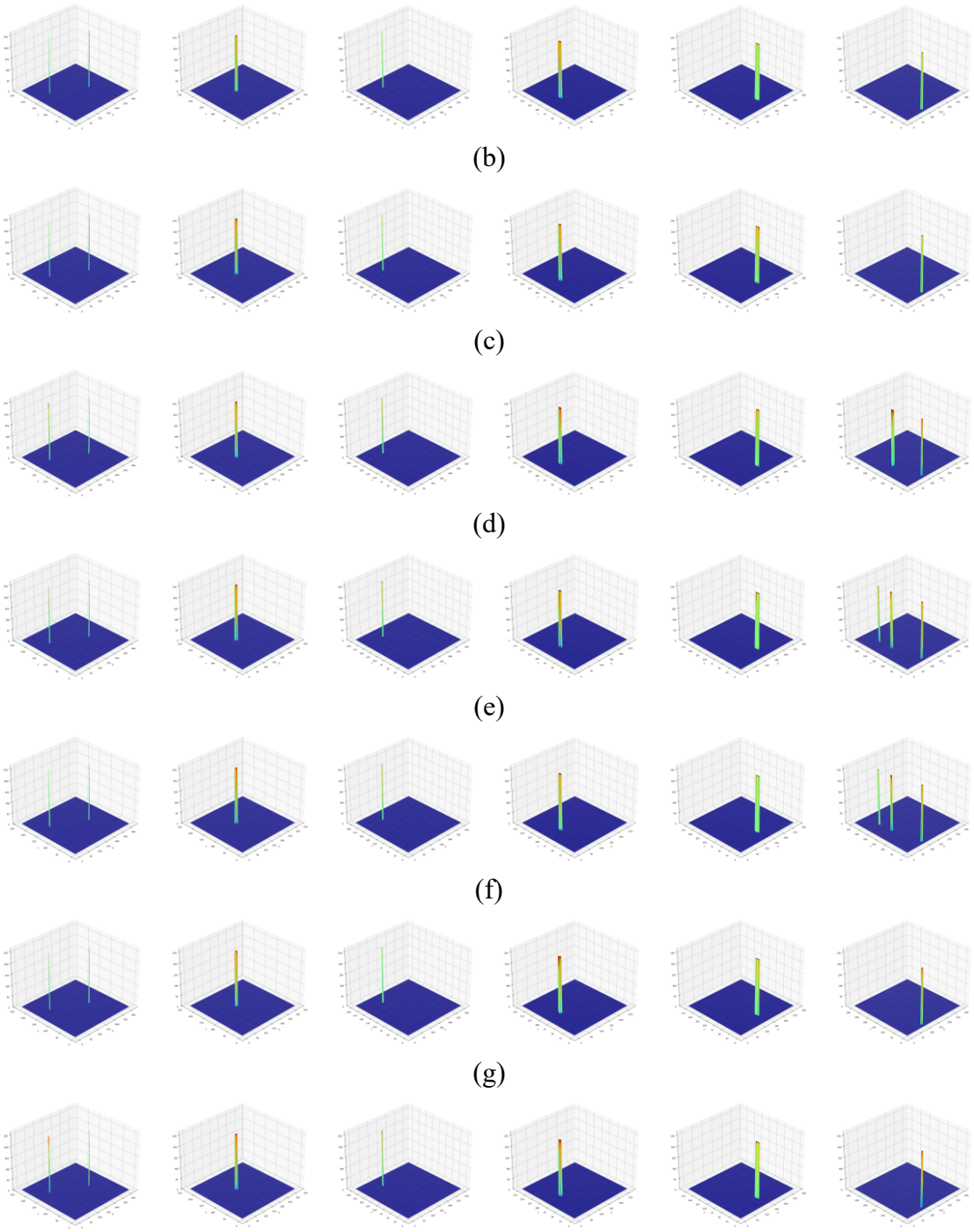
Figures 10 and 11 show 3D visualizations of the target distribution for the NUAA-SIRST and NUDT-SIRST datasets, effectively illustrating the detection of small targets in the images. The results indicate that the BSG-Mamba model exhibits a distinct peak response for targets, with the target region displaying a tall and narrow peak-like response that most closely matches the ground truth. The background remains relatively flat with no significant fluctuations in background noise, and the signal-to-noise ratio of the target response outperforms most comparison models. In contrast, as shown in Figure 10, the background of AGPCNet exhibits slight rippling fluctuations, indicating incomplete background suppression. Furthermore, in the sixth figure, a few models exhibit a bimodal distribution, suggesting the presence of false alarms. In the scenario shown in Figure 11, although most models can achieve target recognition, there are significant differences in recognition accuracy, and the overall accuracy is relatively low.



(a)

continued on next page

Figure 10. The 3D visualizations of different network models on the NUAA-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.



(b)

(c)

(d)

(e)

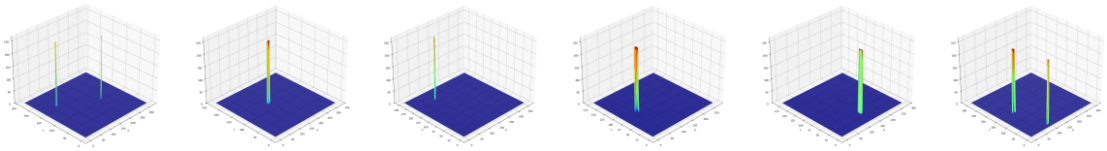
(f)

(g)

(h)

continued on next page

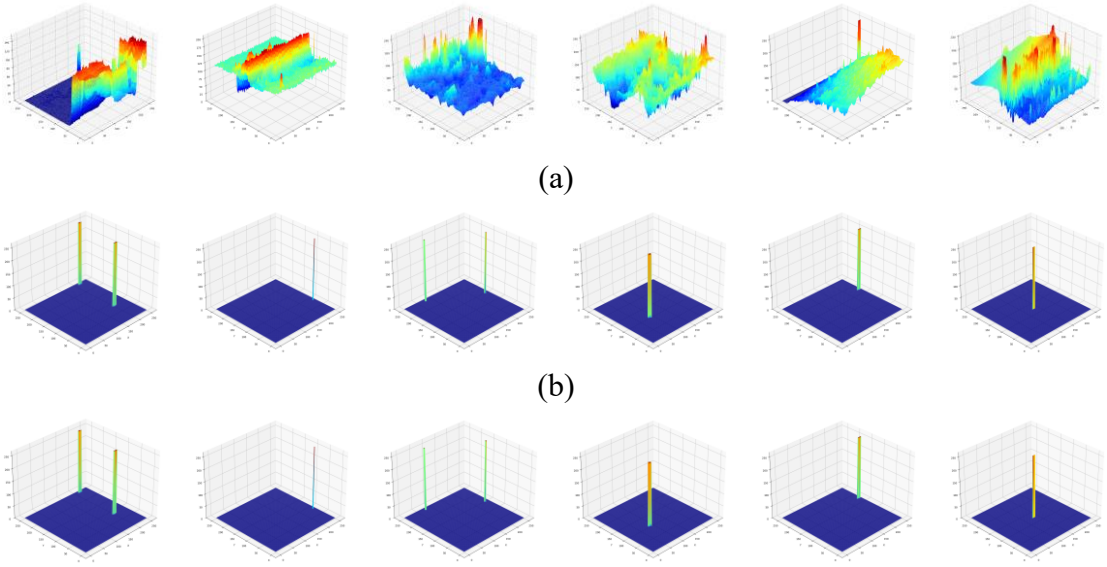
Figure 10. The 3D visualizations of different network models on the NUAA-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.



(i)

Figure 10. The 3D visualizations of different network models on the NUAA-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.

In comparison, the BSG-Mamba method proposed in this paper most closely aligns with the true distribution of 3D targets, elucidating the fundamental mechanism behind its excellent detection performance on the two datasets from the perspective of feature representation. This method demonstrates outstanding performance and detection accuracy across multiple robustness evaluations. Notably, it does not produce false alarms under high-brightness noise interference conditions and can achieve clear, stable identification of weak and small targets. This fully validates that BSG-Mamba possesses significant background suppression and target retention capabilities, further confirming its effectiveness and reliability in practical application scenarios.



(a)

(b)

(c)

continued on next page

Figure 11. The 3D visualizations of different network models on the NUDT-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.

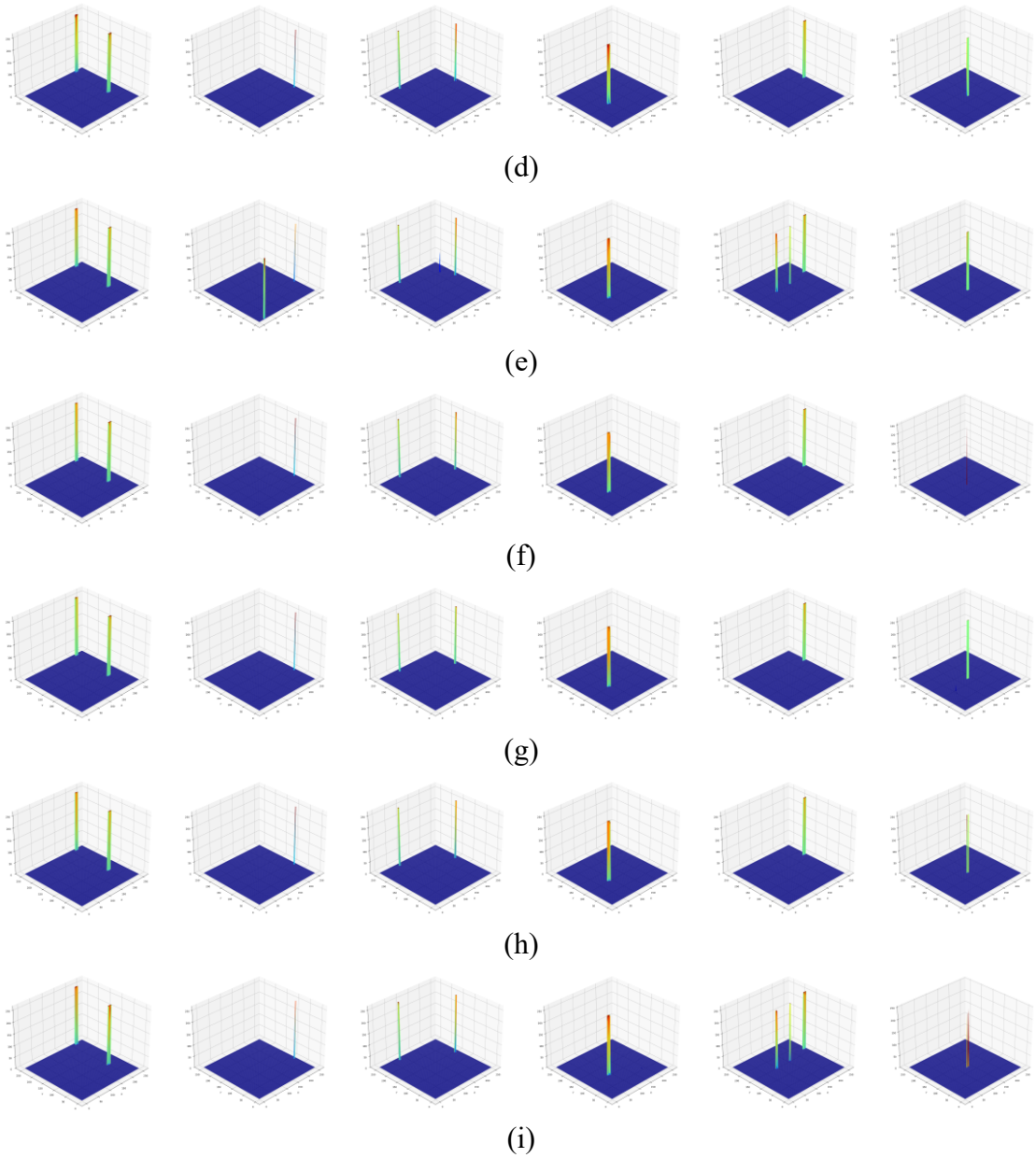


Figure 11. The 3D visualizations of different network models on the NUDT-SIRST dataset. (a) Input, (b) ground truth (GT), (c) BSG-Mamba, (d) AGPCNet, (e) UIU-Net, (f) DNA-Net, (g) IAA-Net, (h) MTU-Net, and (i) GCLNet.

4.4. Quantitative analysis

To comprehensively evaluate the detection performance of BSG-Mamba, this work conducted comparative quantitative experiments on two public datasets, NUAA-SIRST and NUDT-SIRST. To guarantee the statistical reliability of experimental results, all competing methods were implemented with five independent repeated trials, and all evaluation metrics were reported in the form of mean ± standard deviation to avoid random bias caused by parameter initialization and dataset random partitioning.

As shown in Tables 3 and 4, BSG-Mamba achieved superior performance compared to most of

the competing algorithms. The *IoU* metrics on both datasets reached the best results, particularly on the NUDT-SIRST dataset where target scale variations are more pronounced. Although the *nIoU* on NUAA-SIRST was slightly lower than that of UIU-Net, the *nIoU* on NUDT-SIRST improved by approximately 11.13 percentage points compared to UIU-Net, indicating that this method achieves more precise pixel-level localization of target boundary regions, while maintaining an optimal balance in overall performance. In terms of background suppression, the *Fa* on NUAA-SIRST is slightly higher than that of DNA-Net but still significantly outperforms other algorithms, indicating that BSG-Mamba possesses strong background discrimination capabilities in most scenarios. In the NUDT-SIRST dataset, the *Fa* is reduced by approximately 81.8% compared to IAA-Net and by 78.3% compared to U-Transformer. While some models exhibit low *Fa*, their *Pa* are insufficiently high, indicating that these models sacrifice recall for weak targets while reducing *Fa*. In contrast, BSG-Mamba achieves a balance between high *Pa* and low *Fa*, validating its comprehensive detection advantages in complex backgrounds. Compared with transformer-based APTNet and STASPPNet, our BSG-Mamba achieves the highest *IoU* and *Pd* with the lowest *Fa* on both NUAA-SIRST and NUDT-SIRST datasets, demonstrating superior detection and background suppression capabilities.

Table 3. Comparison of different methods for NUAA-SIRST dataset.

Model	NUAA-SIRST			
	IoU (%)	*nIoU* (%)	*Pd* (%)	*Fa* (10^{-6})
Top-Hat	7.85 ± 0.52	19.12 ± 1.13	78.90 ± 1.86	980.00 ± 32.75
IPI	34.11 ± 1.25	36.08 ± 1.47	89.42 ± 1.21	74.50 ± 4.12
NOLC	14.02 ± 0.76	18.21 ± 0.93	74.10 ± 2.03	38.90 ± 3.06
U-Net	62.40 ± 1.38	65.18 ± 1.42	76.08 ± 1.75	99.80 ± 5.37
ACM	67.89 ± 1.21	70.21 ± 1.35	96.24 ± 0.94	24.15 ± 2.16
AGPCNet	72.08 ± 1.05	74.15 ± 1.17	94.25 ± 1.08	25.60 ± 2.31
U-Transformer	76.55 ± 0.92	77.78 ± 1.04	96.32 ± 0.87	18.78 ± 1.85
APTNet	76.78 ± 0.86	78.13 ± 0.95	97.38 ± 0.76	11.98 ± 1.24
STASPPNet	74.89 ± 0.90	77.43 ± 0.98	97.34 ± 0.79	10.03 ± 1.07
IAA-Net	74.66 ± 0.93	75.45 ± 1.02	96.78 ± 0.82	23.48 ± 2.06
UIU-Net	76.88 ± 0.84	**79.85 ± 0.91**	92.50 ± 1.13	11.02 ± 1.19
DNA-Net	76.42 ± 0.88	79.75 ± 0.94	96.15 ± 0.85	**8.42 ± 0.96**
MTU-Net	74.67 ± 0.95	76.78 ±1.06	95.94 ± 0.89	22.13 ± 1.97
GCLNet	75.40 ± 0.91	78.55 ± 0.97	95.12 ± 0.93	23.10 ± 2.02
BPGA	76.74 ± 0.87	79.28 ± 0.92	96.99 ± 0.78	10.56 ± 1.12
MiM-ISTD	76.88 ± 0.85	79.18 ± 0.93	96.99 ± 0.77	13.62 ± 1.35
MOU-Mamba	75.80 ± 0.89	79.64 ± 0.95	97.36 ± 0.75	15.35 ± 1.48
BSG-Mamba	**76.91 ± 0.72**	79.40 ± 0.78	**97.74 ± 0.63**	13.65 ± 1.21

Table 4. Comparison of different methods for NUDT-SIRST dataset.

Model	NUDT-SIRST			
	IoU (%)	*nIoU* (%)	*Pd* (%)	*Fa* (10^{-6})
Top-Hat	21.05 ± 0.96	29.75 ± 1.24	77.83 ± 1.72	158.20 ± 7.36
IPI	33.25 ± 1.18	40.33 ± 1.35	84.60 ± 1.34	192.80 ± 8.72
NOLC	20.41 ± 0.89	31.24 ± 1.16	67.45 ± 2.15	57.33 ± 3.84
U-Net	65.72 ± 1.24	67.85 ± 1.31	91.68 ± 1.07	29.87 ± 2.43
ACM	76.58 ± 1.03	77.32 ± 1.12	96.88 ± 0.86	8.94 ± 1.02
AGPCNet	84.02 ± 0.91	84.57 ± 0.96	97.80 ± 0.79	6.89 ± 0.87
U-Transformer	87.78 ± 0.85	88.70 ± 0.90	96.79 ± 0.83	16.89 ± 1.54
APTNet	90.63 ± 0.78	91.89 ± 0.83	98.33 ± 0.71	6.17 ± 0.79
STASPPNet	91.45 ± 0.75	91.90 ± 0.81	97.96 ± 0.73	4.33 ± 0.65
IAA-Net	88.25 ± 0.83	90.44 ± 0.88	96.89 ± 0.82	20.11 ± 1.76
UIU-Net	81.40 ± 0.94	81.65 ± 0.99	96.20 ± 0.87	14.68 ± 1.41
DNA-Net	85.63 ± 0.88	86.88 ± 0.93	98.70 ± 0.72	5.64 ± 0.73
MTU-Net	78.80 ± 1.01	80.67 ± 1.05	96.78 ± 0.84	19.15 ± 1.69
GCLNet	85.25 ± 0.89	85.75 ± 0.94	98.35 ± 0.74	5.62 ± 0.72
BPGA	90.98 ± 0.76	92.60 ± 0.80	98.86 ± 0.70	5.66 ± 0.74
MiM-ISTD	91.89 ± 0.73	92.60 ± 0.79	98.24 ± 0.72	5.32 ± 0.70
MOU-Mamba	91.61 ± 0.74	92.45 ± 0.78	98.63 ± 0.71	4.01 ± 0.62
BSG-Mamba	**92.02 ± 0.68**	**92.78 ± 0.74**	**98.90 ± 0.65**	**3.67 ± 0.56**

To further verify the generalization ability and robustness of the proposed BSG-Mamba, we conduct experiments on the challenging IRSTD-1K dataset. This benchmark contains complex backgrounds, strong clutter interference, and low-contrast targets, making it more difficult than NUAA-SIRST and NUDT-SIRST and thus suitable for evaluating cross-scene generalization performance. As shown in Table 5, our BSG-Mamba achieves the best overall performance among all competing methods. Specifically, it obtains the highest *IoU* and *nIoU*, as well as the lowest false alarm rate. Although MiM-ISTD achieves a slightly higher *Pd* than our method, our approach maintains a significantly lower false alarm rate while achieving superior *IoU* and *nIoU*. This indicates that our method achieves a better balance between detection rate and false alarm suppression, which is critical for practical infrared small target detection tasks.

Table 6 and Figure 12 show that BSG-Mamba achieves an excellent balance between lightweight design, inference speed, and detection accuracy. Compared with DNA-Net, it improves the F1 score by 0.52%, reduces FLOPs by 6.2%, and increases FPS by 18.3%, confirming its architectural efficiency. Notably, BSG-Mamba maintains over 60 FPS with high accuracy, outperforming methods that sacrifice precision for speed.

Table 5. Comparison of different methods for IRSTD-1K dataset.

Model	IRSTD-1K			
	IoU (%)	*nIoU* (%)	*Pd* (%)	*Fa* (10^{-6})
AGPCNet	65.23 ± 1.36	66.56 ± 1.41	91.34 ± 1.12	27.12 ± 2.26
U-Transformer	66.56 ± 1.29	68.89 ± 1.35	92.25 ± 1.06	22.35 ± 2.01
APTNet	69.89 ± 1.15	69.28 ± 1.22	89.5 5± 1.23	29.04 ± 2.41
STASPPNet	68.17 ± 1.21	70.79 ± 1.26	91.92 ± 1.10	30.18 ± 2.47
IAA-Net	69.17 ± 1.17	71.36 ± 1.23	91.89 ± 1.11	29.90 ± 2.44
UIU-Net	67.35 ± 1.25	72.38 ± 1.28	93.38 ± 1.02	25.67 ± 2.19
DNA-Net	70.91 ± 1.08	72.89 ± 1.14	95.90 ± 0.91	24.45 ± 2.12
MTU-Net	71.21 ± 1.05	73.20 ± 1.11	94.31 ± 0.96	24.30 ± 2.10
GCLNet	71.29 ± 1.04	73.80 ± 1.09	91.57 ± 1.13	29.42 ± 2.39
BPGA	70.35 ± 1.11	71.62 ± 1.18	92.32 ± 1.05	30.41 ± 2.49
MiM-ISTD	70.36 ± 1.10	68.05 ± 1.25	**96.95 ± 0.88**	20.90 ± 1.94
MOU-Mamba	69.23 ± 1.16	65.65 ± 1.30	92.89 ± 1.03	23.74 ± 2.07
BSG-Mamba	**72.34 ± 0.92**	**74.21 ± 0.97**	95.78 ± 0.84	**18.70 ± 1.75**

Against transformer-based APTNet and STASPPNet, BSG-Mamba cuts parameters by 76.6% and 83.8%, reduces FLOPs by 12.1% and 71.6%, and achieves a higher FPS of 69.95, demonstrating clear advantages in lightweight design and real-time inference. Lower memory and hardware requirements also make it suitable for embedded devices.

Compared with the latest Mamba-based methods, our model further shows superior performance. Against MiM-ISTD, BSG-Mamba reduces FLOPs by 44.9%, decreases parameters by 72.0%, and improves F1 by 0.78% with similar FPS. Against MOU-Mamba, it boosts F1 by 3.72%, reduces parameters by 90.2%, and achieves over three times higher FPS.

These results confirm that BSG-Mamba effectively resolves the conflict between accuracy and efficiency, meeting the demands of real-time processing and resource-limited scenarios such as airborne monitoring, unmanned aerial vehicle (UAV) inspection, and portable field warning.

Table 6. Comparison of the model parameters (M), FLOPs (G), and FPS of different methods.

Model	Flops (G)↓	Params (M)↓	FPS↑	F1 (%)
AGPCNet	327.54	12.36	18.61	88.62
UIU-Net	54.43	50.54	52.22	91.13
DNA-Net	14.26	4.70	59.14	93.55
IAA-Net	18.13	14.05	55.75	86.94
MTU-Net	24.43	4.07	60.14	86.82
GCLNet	22.53	19.04	48.59	92.36
APTNet	15.21	5.60	68.12	93.32
STASPPNet	47.10	8.10	58.90	91.62
MiM-ISTD	24.37	4.67	69.39	93.29
MOU-Mamba	14.29	13.44	21	90.35
BSG-Mamba	**13.37**	**1.31**	**69.95**	**94.07**

As illustrated in Figure 13, the ROC curves of BSG-Mamba consistently outperform competing algorithms on both NUAA-SIRST and NUDT-SIRST datasets. The superior positioning of these curves indicates a higher true positive rate (TPR) across various false positive rate (FPR) thresholds. Notably, BSG-Mamba exhibits the most rapid TPR growth within the low FPR range, suggesting that the model effectively captures targets while strictly suppressing false alarms in practical applications. In summary, through precise feature modeling and a robust detection mechanism, BSG-Mamba surpasses existing mainstream methods, providing a promising technical framework for the reliable identification of faint targets in complex scenarios.

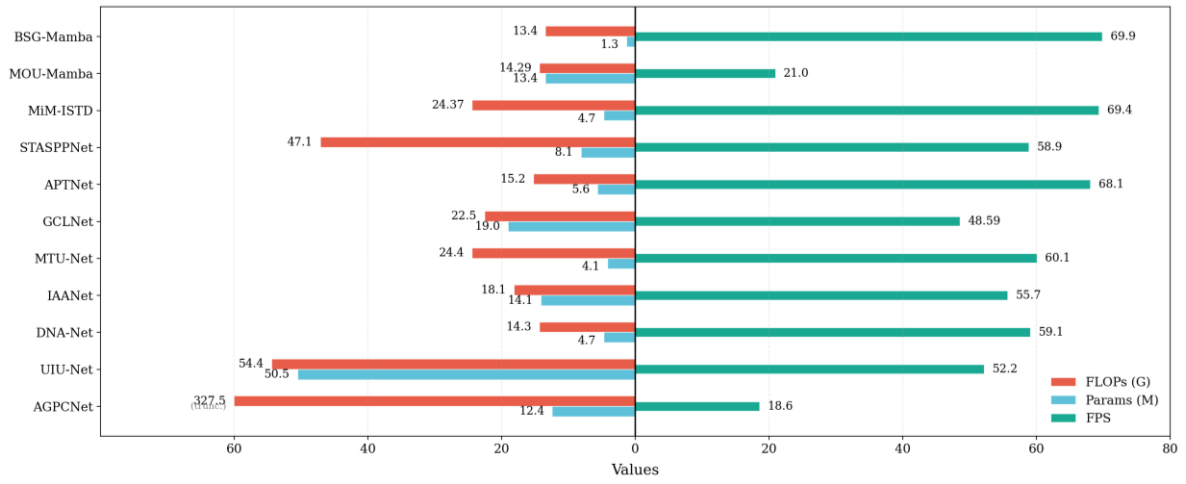
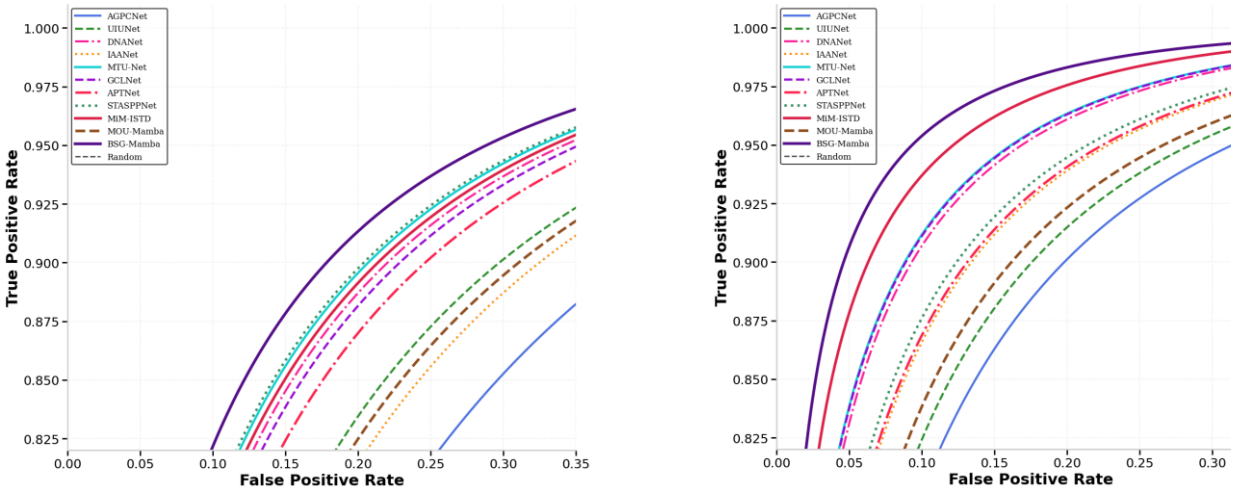


Figure 12. Model efficiency comparison: FLOPs (G) and FPS of different methods.



(a) (b)

Figure 13. The 2D ROC curves for comparison of different methods. (a) NUAA-SIRST dataset, (b) NUDT-SIRST dataset.

4.5. Ablation analysis

To comprehensively evaluate the contribution of each core component in the proposed framework, we conduct independent ablation experiments on the NUAA-SIRST dataset, with the vanilla U-Net as the baseline. Each module is validated separately under the same training and testing settings to clarify its unique function and performance improvement.

Table 7 analyzes the embedding position of the SG-Encoder module. The results show that adding the SG-Encoder at any encoder stage brings consistent improvements over the baseline. Embedding the module at deeper stages yields better performance than shallow-stage insertion, as higher-level features contain richer contextual information for capturing non-local structural correlations of small targets. The optimal configuration, inserting the SG-Encoder at conv2, conv3, and conv4 simultaneously, achieves 71.31% *IoU*, 72.61% *nIoU*, 93.16% *Pd*, and 34.99 × 10^{-6} *Fa*, confirming that multi-stage embedding enhances the discriminative power of sparse target features.

Table 7. Ablation experiments on the embedding positions of the SG-Encoder module.

Encoder				Metrics			
conv 1	conv 2	conv 3	conv 4	*IoU* (%)	*nIoU* (%)	*Pd* (%)	*Fa* (10^{-6})
√				65.44	66.89	90.23	58.10
	√			65.39	65.87	90.10	46.78
		√		67.21	68.92	91.34	38.19
			√	67.82	68.70	91.67	38.01
√	√			69.15	70.31	92.38	38.11
	√	√		70.21	71.13	92.32	39.16
		√	√	70.12	70.89.	92.78	37.21
√	√	√		70.85	71.93	92.90	37.28
	√	√	√	**71.31**	**72.61**	**93.16**	**34.99**
√	√	√	√	71.01	72.30	93.02	35.89

The superiority of the proposed DSAFM module over Vmamba is verified in Table 8. Compared with the baseline U-Net, conventional Vmamba already improves *IoU* to 72.85% by leveraging long-range state-space modeling. However, the proposed DSAFM further boosts performance across all metrics, reaching 75.10% *IoU*, 77.92% *nIoU*, 96.61% *Pd*, and 26.96 × 10^{-6} *Fa*. This consistent gain demonstrates that the direction-specific SSM and adaptive fusion design better fit the characteristics of infrared small targets, alleviating the over-smoothing problem in VMamba and enhancing both global context modeling and local detail preservation.

Table 8. Ablation comparison between the proposed DSAFM and Vmamba.

Baseline	Modules		Metrics			
U-Net	Vmamba	DSAFM	*IoU* (%)	*nIoU* (%)	*Pd* (%)	*Fa* (10^{-6})
√			62.40	65.18	76.08	99.80
√	√		72.85	75.60	95.12	32.45
√		√	**75.10**	**77.92**	**96.61**	**26.96**

To further validate the effectiveness of the directional modeling mechanism in DSAFM, we conduct an ablation study on three variants of the direction-aware SSM design: a shared-direction SSM that uses identical parameters across all scanning directions(S-SSM), a decoupled-direction SSM with independent parameters for each direction but no adaptive fusion weights(D-SSM), and our proposed direction-specific SSM with learnable adaptive fusion weights (DS-SSM).

As shown in Table 9, simply introducing decoupled parameters improves IoU by 0.59 percentage points and reduces Fa by 7.9%, confirming that independent modeling of different directions better captures the directional features of infrared small targets. Adding the adaptive fusion weights further boosts performance, demonstrating that the combination of direction-specific SSM parameters and learnable fusion is the key to the superior performance of the proposed mechanism.

Table 9. Ablation study on the directional-specific SSM and adaptive fusion mechanism in DSAFM.

Baseline	Modules			Metrics			
U-Net	SSSM	DSSM	DS-SSM	IoU (%)	$nIoU$ (%)	Pd (%)	Fa (10^{-6})
√	√			72.85	75.60	95.12	32.45
√		√		73.44	76.23	95.85	30.02
√			√	**75.10**	**77.92**	**96.61**	**26.96**

To verify the synergistic effects among the three components, we conduct progressive combination ablation experiments on both datasets. As shown in Tables 10 and 11, the performance improves consistently as modules are added, and the full combination yields the best results on all metrics.

Table 10. BSG-Mamba ablation experiments on the NUAA-SIRST dataset.

Baseline	Modules				Metrics			
U-Net	BSB	TEG	SG-Encoder	DSAFM	IoU (%)	$nIoU$ (%)	Pd (%)	Fa (10^{-6})
√					62.40	65.18	76.08	99.80
√	√				66.35	68.72	88.21	45.32
√		√			67.12	69.45	89.56	41.08
			√		71.31	72.61	93.16	34.99
√				√	75.10	77.92	96.61	26.96
√	√	√			68.04	70.39	91.63	36.70
√	√	√		√	76.19	78.75	96.96	18.66
√	√	√	√		76.85	79.31	97.34	14.26
√	√	√	√	√	**76.91**	**79.40**	**97.74**	**13.65**

On both NUAA-SIRST and NUDT-SIRST, pairing any two modules already outperforms using them alone, confirming their complementary functions. For the background suppression branch, standalone BSB or TEG each delivers moderate performance gains, while combining BSB and TEG yields further improvements by jointly filtering complex background clutter and amplifying weak target signals. Background suppression provides cleaner features, SG-Encoder enhances sparse target representation, and DSAFM captures global context. When BSB paired with TEG, SG-Encoder and

DSAFM are all integrated, the full model achieves the highest *IoU, Pd*, and the lowest false alarm rate on both datasets. These results demonstrate that each module addresses a distinct challenge, and their combination produces a more robust and accurate detection framework.

Table 11. BSG-Mamba ablation experiments on the NUDT-SIRST dataset.

Baseline	Modules				Metrics			
U-Net	BSB	TEG	SG-Encoder	DSAFM	IoU (%)	$nIoU$ (%)	Pd (%)	Fa (10^{-6})
√					65.72	67.85	91.68	29.87
√	√				73.26	74.37	93.86	28.79
√		√			75.79	77.93	93.54	16.71
			√		75.89	77.79	94.32	18.90
√				√	77.51	78.89	95.03	19.07
√	√	√			80.38	81.20	94.87	15.20
√	√	√		√	89.66	90.04	98.30	12.13
√	√	√	√		89.62	91.16	98.42	8.30
√	√	√	√	√	**92.02**	**92.78**	**98.90**	**3.67**

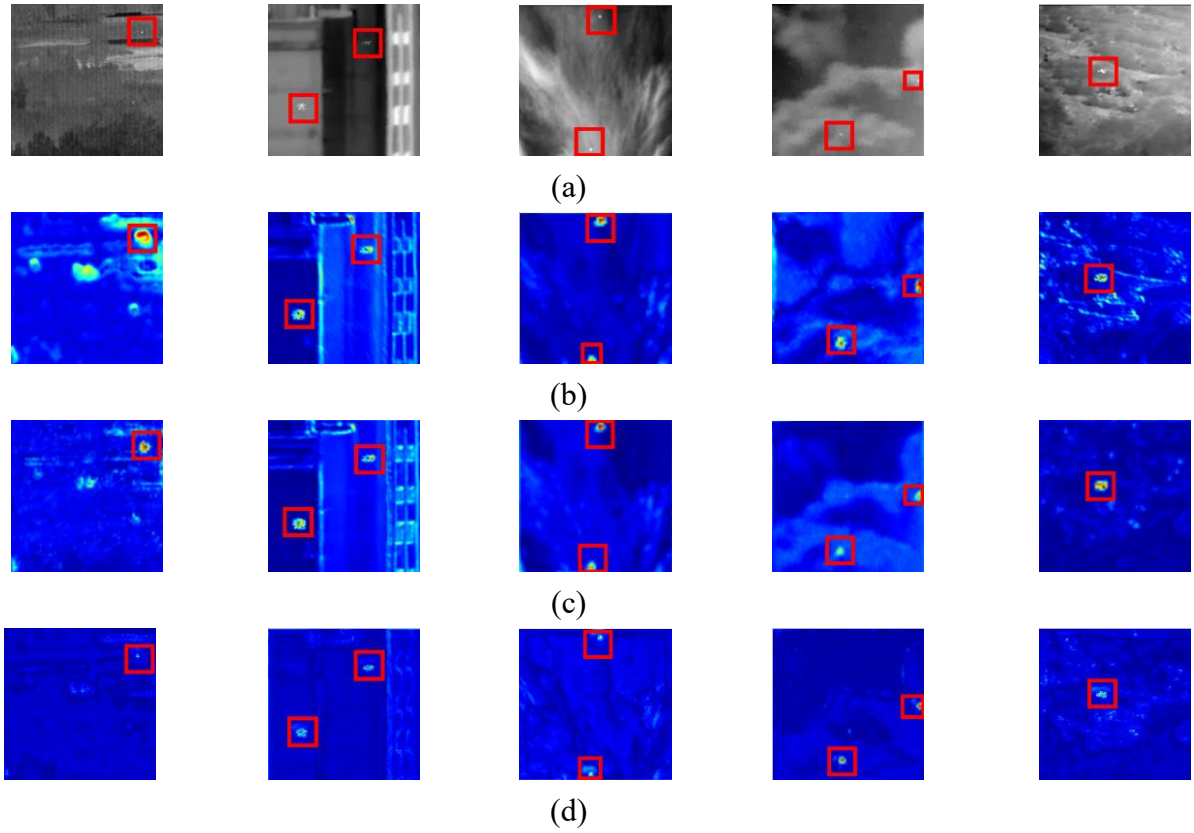
4.6. Validation based on encoder modules

We first visualize the intermediate feature maps of the background suppression branch to reveal its internal working mechanism. As shown in Figure 14, feature responses corresponding to complex background clutter are gradually restrained through layer-by-layer convolution processing, while the feature amplitude of tiny infrared targets is continuously strengthened. In the final layer feature map, most background interference is well suppressed, and the target region exhibits distinct high feature response. This visualization result demonstrates that the background suppression branch can effectively filter out background noise and highlight weak target features, which provides favorable feature representation for subsequent encoder feature extraction and accurate target detection.

Figure 14. Intermediate feature maps of the proposed background suppression branch. (a) Input image; (b) output of Conv 7 × 7; (c) output of Conv 5 × 5; (d) output of the background feature extractor; (e) final foreground weight map.

In this study, five grayscale images of small infrared targets were selected as input. Heatmaps were generated for the outputs of the encoder's core modules—BSB, SG-Encoder, and DSAFM— to evaluate the performance of feature extraction across these modules in terms of target response, background suppression, feature localization, and spatial context modeling. As the initial module, BSB effectively suppressed the bright background noise in the five images, causing the background to exhibit low-response characteristics. At the same time, it achieved basic separation between small targets and the background, completing preliminary target localization and providing low-noise feature inputs for subsequent modules.

Subsequently, the three-layer SG-Encoder from SG-Encoder2 to SG-Encoder4 shows a continuous decrease in background response in the heatmaps, a gradual increase in target response, and continuously enhanced feature focus, achieving a transition from preliminary target localization to mid-range refined feature representation. Specifically, SG-Encoder2 performs a second suppression of the background and an initial enhancement of target features; SG-Encoder3 achieves precise focusing of target features, serving as the key layer for feature transformation; and SG-Encoder4 highly condenses the target, providing high-purity feature inputs for deep modeling.



continued on next page

Figure 15. Validation of each module based on the Encoder. (a) Input, (b) input visualization, (c) BSB visualization, (d) SG-Encoder2 visualization, (e) SG-Encoder3 visualization, (f) SG-Encoder4 visualization, (g) DSAFM visualization, and (h) output.

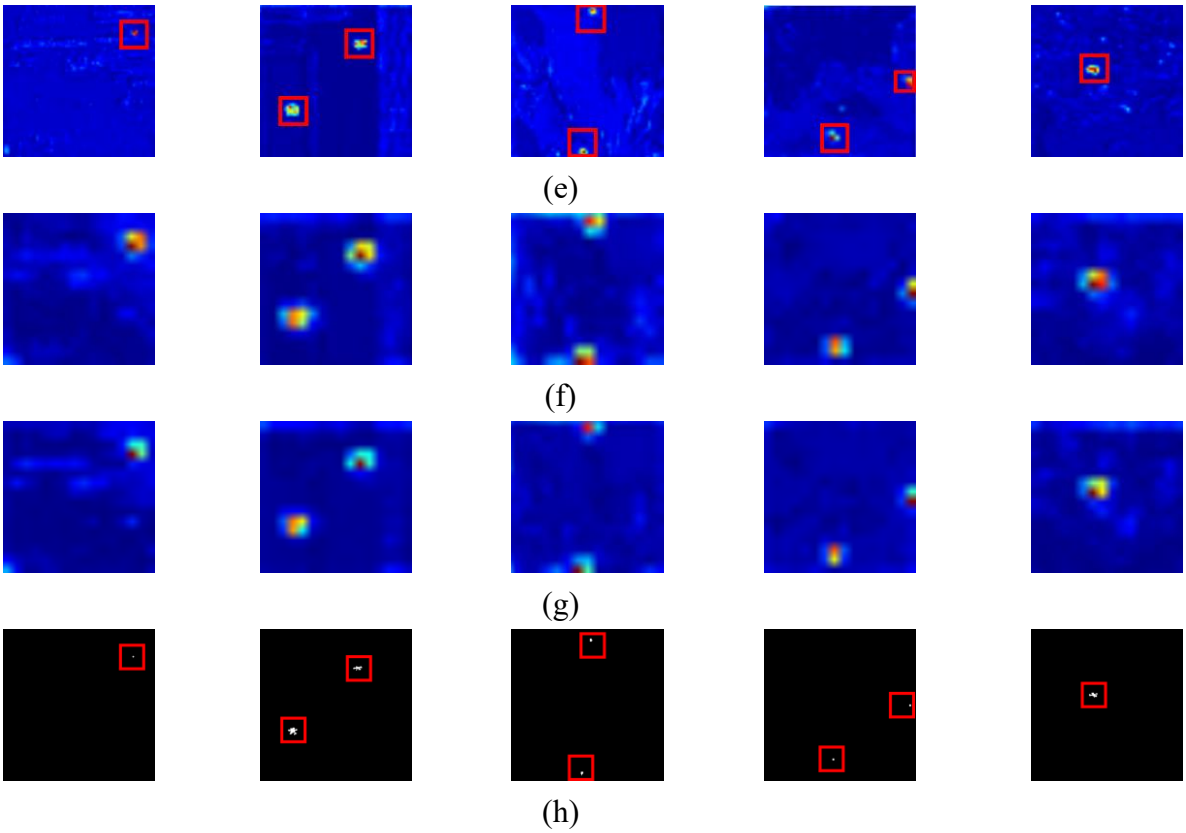


Figure 15. Validation of each module based on the Encoder. (a) Input, (b) input visualization, (c) BSB visualization, (d) SG-Encoder2 visualization, (e) SG-Encoder3 visualization, (f) SG-Encoder4 visualization, (g) DSAFM visualization, and (h) output.

As the deepest module of the encoder, DSAFM achieves superior long-range contextual modeling based on four-directional selective SSM. The small targets in its output exhibit the highest response in the encoder, with target-background contrast reaching its maximum. Simultaneously, four-directional modeling ensures precise association between target features and the surrounding spatial context, without any diffusion of target features or prominent background features.

These experimental results demonstrate that the hierarchical design of the encoder effectively addresses the challenges of small target detection, such as low signal-to-noise ratio, small scale, and complex backgrounds. The results show no feature discontinuities or loss of target features, and the high-purity, highly focused target representations provide an excellent foundation for subsequent decoder processing.

4.7. The robustness study of the BSG-Mamba in challenging scenarios

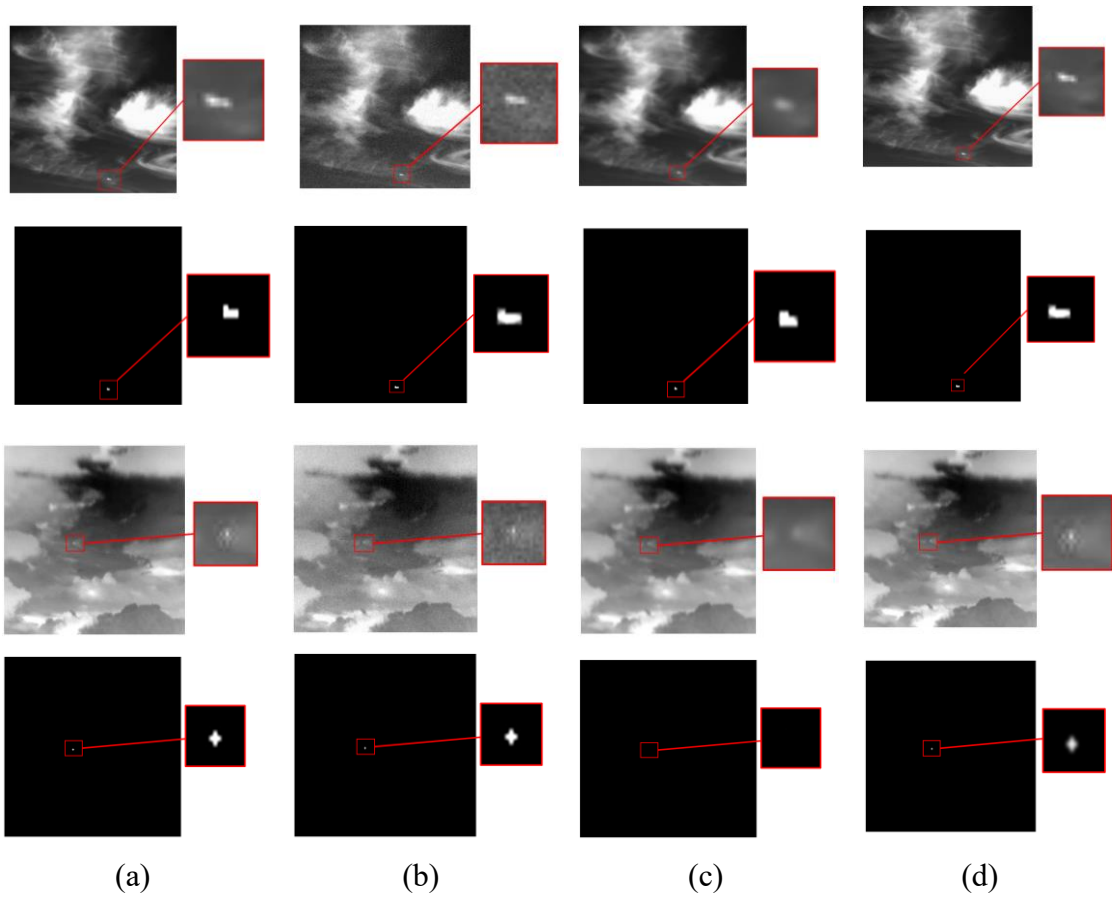


Figure 16. Detection results under four challenging conditions. (a) Original infrared scene; (b) scene with simulated sensor noise; (c) scene with atmospheric Gaussian blur; (d) scene with low signal-to-clutter ratio.

In practical airborne and UAV infrared monitoring systems, detection performance is frequently degraded by various adverse interference factors. To evaluate the anti-interference performance of the proposed BSG-Mamba, we select two typical infrared scenes with different background complexities and simulate three common degradation types: sensor noise, atmospheric Gaussian blur, and low signal-to-clutter ratio.

In the first group with moderate background clutter, although the segmented target edges show subtle differences under different interferences, our model can accurately locate small targets and extract their approximate contours in all scenarios. The method effectively eliminates noise interference, restores blurred target features, and enhances weak target signals, which verifies its favorable robustness in moderately complex environments.

For the second group with extremely complex cloud backgrounds, the model achieves accurate target detection under the original scene and sensor noise. However, obvious performance degradation appears in the atmospheric blur scenario, which is a typical failure case reflecting the limitation of our framework. Under the low signal-to-clutter ratio condition, the segmented target edge is relatively fuzzy, but the model can still identify the real target correctly and filter out massive background clutter interference.

In general, BSG-Mamba exhibits satisfactory robustness under most harsh conditions, while its feature extraction ability for blurred tiny targets in complex backgrounds still needs further improvement. We will optimize the model with multi-scale feature recovery strategies in future research.

5. Conclusions

The BSG-Mamba network proposed in this paper effectively addresses the key challenges in detecting small infrared targets against complex backgrounds, namely the difficulty of separating targets from the background, insufficient representation of non-local features, and the lack of directional specificity in global modeling. This is achieved through the collaborative design of four core modules: BSB and TEG, SG-Encoder, and DSAFM. Specifically, the dual-branch mechanism formed by BSB and TEG enables effective background suppression and target feature enhancement. The SG-Encoder enhances the representation of non-local structures of small targets by constructing graph-structured relationships in feature maps. The DSAFM, utilizing multi-directional independent SSMs and an adaptive fusion strategy, achieves precise modeling of both global and direction-specific contexts. Experiments demonstrate that the proposed model outperforms mainstream algorithms in both detection accuracy and false alarm control on two public datasets.

Despite the promising performance, the proposed framework still has certain limitations. The model is trained and validated on standard public datasets, and its performance degrades obviously under extremely complex backgrounds with severe atmospheric blur; meanwhile, detection accuracy may decline when facing ultra-low signal-to-noise ratio, strong sensor noise, and dense mixed cloud clutter. In addition, the current framework only processes single-frame static infrared images, without exploiting temporal information between consecutive frames. Although the model is lightweight and supports real-time inference, it has not been further optimized for extreme edge devices and long-term stable operation in practical deployment.

Aiming at the above limitations revealed in our robustness experiments, we outline several directions for future research. First, we will optimize the feature extraction module by introducing multi-scale feature recovery and detail enhancement strategies to remedy feature loss caused by atmospheric blur, strengthen the background suppression branch to adapt to dense clutter and ultra-low signal-to-noise ratio scenes, and further improve the model's robustness against image degradation and complex interference. Second, we will introduce motion features and inter-frame correlation to fuse spatiotemporal information, so as to improve detection reliability for infrared video sequences and reduce missed detection under blurred imaging conditions. Third, we will combine the detection network with target tracking algorithms to build an integrated detection and tracking system, which is more suitable for continuous airborne monitoring tasks, and promote its practical deployment on various embedded edge devices.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Natural Science Foundation of Fujian Province under Grant

No. 2024J01821, the High-Level Cultivation Project of Minnan Normal University under Grant No. MSGJB2024009, and the Principal Foundation of Minnan Normal University under Grant No. KJ18010.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. M. Teutsch, W. Krüger, Classification of small boats in infrared images for maritime surveillance, in *2010 International WaterSide Security Conference*, (2010), 1–7. https://doi.org/10.1109/WSSC.2010.5730289
2. C. Li, Z. Huang, X. Xie, W. Li, IST-TransNet: Infrared small target detection based on transformer network, *Infrared Phys. Technol.*, **132** (2023), 104723. https://doi.org/10.1016/j.infrared.2023.104723
3. H. Yi, C. Yang, R. Qie, J. Liao, F. Wu, T. Pu, et al., Spatial-temporal tensor ring norm regularization for infrared small target detection, *IEEE Geosci. Remote Sens. Lett.*, **20** (2023), 1–5. https://doi.org/10.1109/LGRS.2023.3236030
4. J. Yao, S. Xu, H. Feijiang, C. Su, Improved lightweight infrared road target detection method based on YOLOv8, *Infrared Phys. Technol.*, **141** (2024), 105497. https://doi.org/10.1016/j.infrared.2024.105497
5. S. Li, B. Hao, Adversarial camouflage for mobile targets against battlefield optical and infrared AI detection, *J. King Saud Univ. Comput. Inf. Sci.*, **38** (2026), 135. https://doi.org/10.1007/s44443-026-00542-8
6. N. Kumar, P. Singh, Small and dim target detection in infrared imagery: A review, current techniques and future directions, *Neurocomputing*, **630** (2025), 129640. https://doi.org/10.1016/j.neucom.2025.129640
7. X. Bai, F. Zhou, Analysis of new top-hat transformation and the application for infrared dim small target detection, *Pattern Recognit.*, **43** (2010), 2145–2156. https://doi.org/10.1016/j.patcog.2009.12.023
8. S. D. Deshpande, M. H. Er, R. Venkateswarlu, P. Chan, Max-mean and max-median filters for detection of small targets, in *Signal and Data Processing of Small Targets 1999*, **3809** (1999), 74–83. https://doi.org/10.1117/12.364049
9. Y. Li, Z. Li, C. Zhang, Z. Luo, Y. Zhu, Z. Ding, et al., Infrared maritime dim small target detection based on spatiotemporal cues and directional morphological filtering, *Infrared Phys. Technol.*, **115** (2021), 103657. https://doi.org/10.1016/j.infrared.2021.103657
10. C. L. P. Chen, H. Li, Y. Wei, T. Xia, Y. Y. Tang, A local contrast method for small infrared target detection, *IEEE Trans. Geosci. Remote Sens.*, **52** (2014), 574–581. https://doi.org/10.1109/TGRS.2013.2242477
11. Y. Wei, X. You, H. Li, Multiscale patch-based contrast measure for small infrared target detection, *Pattern Recognit.*, **58** (2016), 216–226. https://doi.org/10.1016/j.patcog.2016.04.002
12. X. Wang, G. Lv, L. Xu, Infrared dim target detection based on visual attention, *Infrared Phys. Technol.*, **55** (2012), 513–521. https://doi.org/10.1016/j.infrared.2012.08.004

13. H. Zhu, H. Ni, S. Liu, G. Xu, L. Deng, TNLRS: Target-aware non-local low-rank modeling with saliency filtering regularization for infrared small target detection, *IEEE Trans. Image Process.*, **29** (2020), 9546–9558. https://doi.org/10.1109/TIP.2020.3028457
14. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
15. M. Liu, H. Du, Y. Zhao, L. Dong, M. Hui, Image small target detection based on deep learning with SNR controlled sample generation, *Curr. Trends Comput. Sci. Mech. Autom.*, **1** (2022), 211–220. https://doi.org/10.1515/ 9783110584974-025
16. Y. Dai, Y. Wu, F. Zhou, K. Barnard, Asymmetric contextual modulation for infrared small target detection, preprint, arXiv:2009.14530. https://doi.org/10.48550/arXiv.2009.14530
17. T. Zhang, L. Li, S. Cao, T. Pu, Z. Peng, Attention-guided pyramid context networks for detecting infrared small target under complex background, *IEEE Trans. Aerosp. Electron. Syst.*, **59** (2023), 4250–4261. https://doi.org/10.1109/TAES.2023.3238703
18. X. Wu, D. Hong, J. Chanussot, UIU-Net: U-Net in U-Net for infrared small object detection, *IEEE Trans. Image Process.*, **32** (2023), 364–376. https://doi.org/10.1109/TIP.2022.3228497
19. B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, et al., Dense nested attention network for infrared small target detection, preprint, arXiv:2106.00487. https://doi.org/10.48550/arXiv.2106.00487
20. J. Lin, K. Zhang, X. Yang, X. Cheng, C. Li, Infrared dim and small target detection based on U-Transformer, *J. Vis. Commun. Image Represent.*, **89** (2022), 103684. https://doi.org/10.1016/j.jvcir.2022.103684
21. T. Wu, B. Li, Y. Luo, Y. Wang, C. Xiao, T. Liu, et al., MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–15. https://doi.org/10.1109/TGRS.2023.3235002
22. S. Yuan, H. Qin, X. Yan, N. Akhtar, A. Mian, SCTransNet: Spatial-channel cross transformer network for infrared small target detection, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–15. https://doi.org/10.1109/TGRS.2024.3383649
23. Y. Zhang, W. Bao, W. Wan, Q. Xiao, Y. Tang, X. Zou, et al., APTNet: Adaptive partial transformer network for infrared small target detection, *IEEE Sensors J.*, **25** (2025), 17960–17974. https://doi.org/10.1109/JSEN.2025.3559093
24. H. Wu, X. Huang, C. He, H. Xiao, S. Luo, Infrared small target detection with Swin transformer-based multiscale atrous spatial pyramid pooling network, *IEEE Trans. Instrum. Meas.*, **74** (2025), 1–14. https://doi.org/10.1109/TIM.2024.3522414
25. M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, J. Guo, ISNet: Shape matters for infrared small target detection, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 867–876. https://doi.org/10.1109/CVPR52688.2022.00095
26. T. Liu, Y. Liu, J. Yang, B. Li, Y. Wang, W. An, Graph Laplacian regularization for fast infrared small target detection, *Pattern Recognit.,* **158** (2025), 111077. https://doi.org/10.1016/j.patcog.2024.111077
27. A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, preprint, arXiv:2111.00396. https://doi.org/10.48550/arXiv.2111.00396
28. A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, preprint, arXiv:2312.00752. https://doi.org/10.48550/arXiv.2312.00752

29. J. Ma, F. Li, B. Wang, U-Mamba: Enhancing long-range dependency for biomedical image segmentation, preprint, arXiv:2401.04722. https://doi.org/10.48550/arXiv.2401.04722
30. J. Xu, HC-Mamba: Vision MAMBA with hybrid convolutional techniques for medical image segmentation, preprint, arXiv:2405.05007. https://doi.org/10.48550/arXiv.2405.05007
31. R. Zhao, S. H. Tang, J. Shen, E. E. B. Supeni, S. A. Rahim, Enhancing autonomous driving safety: A robust traffic sign detection and recognition model TSD-YOLO, *Signal Process.*, **225** (2024), 109619. https://doi.org/10.1016/j.sigpro.2024.109619
32. J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, L. Zhao, Infrared small target detection utilizing the multiscale relative local contrast measure, *IEEE Geosci. Remote Sens. Lett.*, **15** (2018), 612–616. https://doi.org/10.1109/LGRS.2018.2790909
33. K. Wang, S. Du, C. Liu, Z. Cao, Interior attention-aware network for infrared small target detection, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–13. https://doi.org/10.1109/TGRS.2022.3163410
34. X. Ling, L. Wu, C. Wu, S. Ji, Graph neural networks: Graph matching, in *Graph Neural Networks: Foundations, Frontiers, and Applications*, (2022), 277–295. https://doi.org/10.1007/978-981-16-6054-2_13
35. B. Jiang, Z. Zhang, D. Lin, J. Tang, B. Luo, Semi-supervised learning with graph learning-convolutional networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 11305–11312. https://doi.org/10.1109/CVPR.2019.01157
36. P. Veličković, A. Casanova, P. Liò, G. Cucurull, A. Romero, Y. Bengio, Graph attention networks, in *6th International Conference on Learning Representations (ICLR)*, 2018. https://doi.org/10.17863/CAM.48429
37. Y. Chen, M. Rohrbach, Z. Yan, S. Yan, J. Feng, Y. Kalantidis, Graph-based global reasoning networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 433–442. https://doi.org/10.1109/CVPR.2019.00052
38. K. Han, Y. Wang, J. Guo, Y. Tang, E. Wu, Vision GNN: An image is worth graph of nodes, preprint, arXiv:2206.00272. https://doi.org/10.48550/arXiv.2206.00272
39. M. Munir, W. Avery, R. Marculescu, MobileViG: Graph-based sparse attention for mobile vision applications, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2023), 2211–2219. https://doi.org/10.1109/CVPRW59228.2023.00215
40. J. Lin, Y. Ruan, S. Li, X. Yang, S. Niu, Z. Meng, BPGA: Enhancing infrared small and dim target detection with background prior guide and vision graph attention, *Opt. Laser Technol.*, **192** (2025), 113867. https://doi.org/10.1016/j.optlastec.2025.113867
41. Y. Shen, Q. Li, C. Xu, C. Chang, Q. Yin, Graph-based context learning network for infrared small target detection, *Neurocomputing*, **616** (2025), 128949. https://doi.org/10.1016/j.neucom.2024.128949
42. T. Chen, Z. Ye, Z. Tan, T. Gong, Y. Wu, Q. Chu, et al., MiM-ISTD: Mamba-in-Mamba for efficient infrared small-target detection, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–13. https://doi.org/10.1109/TGRS.2024.3485721
43. R. E. Kalman, A new approach to linear filtering and prediction problems, *J. Basic Eng.*, **82** (1960), 35–45. https://doi.org/10.1115/1.3662552
44. L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision Mamba: Efficient visual representation learning with bidirectional state space model, preprint, arXiv:2401.09417. https://doi.org/10.48550/arXiv.2401.09417

45. H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, S. T. Xia, MambaIR: A simple baseline for image restoration with state-space model, preprint, arXiv:2402.15648. https://doi.org/10.48550/arXiv.2402.15648
46. K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, Z. Shi, RSMamba: Remote sensing image classification with state space model, *IEEE Geosci. Remote Sens. Lett.*, **21** (2024), 1–5. https://doi.org/10.1109/LGRS.2024.3407111
47. W. Liu, X. Lu, J. Zhang, D. Li, X. Zhang, MOU-Mamba: Multi-order U-shape Mamba for infrared small target detection, *Opt. Laser Technol.*, **187** (2025), 112851. https://doi.org/10.1016/j.optlastec.2025.112851
48. D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, et al., SpectralGPT: Spectral remote sensing foundation model, *IEEE Trans. Pattern Anal. Mach. Intell.*, **46** (2024), 5227–5244. https://doi.org/10.1109/TPAMI.2024.3362475
49. C. Li, B. Zhang, D. Hong, J. Yao, J. Chanussot, LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–12. https://doi.org/10.1109/TGRS.2023.3279834
50. C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, A. G. Hauptmann, Infrared patch-image model for small target detection in a single image, *IEEE Trans. Image Process.*, **22** (2013), 4996–5009. https://doi.org/10.1109/TIP.2013.2281420
51. T. Zhang, H. Wu, Y. Liu, L. Peng, C. Yang, Z. Peng, Infrared small target detection based on non-convex optimization with Lp-norm constraint, *Remote Sens.*, **11** (2019), 559. https://doi.org/10.3390/rs11050559

AIMS Press