

Article

GIDNet: Infrared Small Target Detection Network Based on Gradient-Intensity Decoupled

Xianwei Gao ¹, Jingtao Wu ¹, Dafeng Cao ¹, Haotian Xu ¹, Yingjie Ma ¹, Lu Li ²  and Mingjing Zhao ^{1,*}

¹ Department of Electronic and Communication Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China; gaoxianwei@besti.edu.cn (X.G.); 20253908@mail.besti.edu.cn (J.W.); 20253711@mail.besti.edu.cn (D.C.); 20253815@mail.besti.edu.cn (H.X.); myj@besti.edu.cn (Y.M.)

² School of Automation, Beijing Information Science and Technology University, Beijing 100101, China; 20192380@bistu.edu.cn

* Correspondence: zmj@besti.edu.cn

Highlights

What are the main findings?

- A gradient-intensity decoupled network, GIDNet, is proposed to strengthen infrared small target representation by jointly modeling target thermal energy, local structural variation, and multi-scale contrast cues.
- Extensive experiments onIRSTD-1k, NUAA-SIRST, and NUDT-SIRST show that GIDNet maintains strong intersection over union, high probability of detection, and low false alarm rate in complex backgrounds.

What are the implications of the main findings?

- The proposed decoupling and shallow projection strategies provide an effective way to preserve weak target signatures while reducing the loss of fine spatial details caused by deep feature extraction.
- This framework offers a practical and efficient solution for infrared small target detection in cluttered and low-SNR scenes, with potential value for real-world remote sensing, surveillance, and early-warning applications.

Abstract

Infrared small target detection (IRSTD) plays a pivotal role in a comprehensive set of applications. Despite the extensive research alongside numerous algorithms proposed in recent years,IRSTD remains a formidable task, primarily stemming from the inherently low level of signal-to-noise ratios (SNR) as well as the presence of intricate background clutter. Current models remain constrained by three critical bottlenecks: the degradation of spectral coupling between intensity and gradient information in deep layers, limited scale adaptability of static filters, and the loss of spatial precision caused by iterative downsampling. We propose GIDNet, a gradient-intensity decoupled network that balances target energy preservation and noise suppression to address the aforementioned issues. Our GIDNet architecture incorporates three core components: a gradient-intensity synergistic convolution (GISC) designed to synergistically encode intensity and gradient information for robust target enhancement; a multi-scale difference contrast (MSDC) module for scale-adaptive detection via adaptive contrast modeling; and a shallow feature projection (SFP) strategy aimed at maintaining precise spatial localization by bridging the gap between deep semantics and shallow spatial details. Comprehensive evaluations, encompassing both quantitative metrics and qualitative visualizations, consistently demonstrate the preeminence of the developed GIDNet surpassing the performance of 16 counterparts.



Academic Editor: Tania Stathaki

Received: 11 March 2026

Revised: 29 April 2026

Accepted: 8 May 2026

Published: 12 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: infrared small target detection (IRSTD); gradient-intensity decoupling; spectral coupling; multi-scale difference contrast (MSDC); shallow feature projection (SFP)

1. Introduction

Infrared small target detection (IRSTD) constitutes a quintessential problem with diverse real-world utilities in surveillance and defense [1–3]. Particularly in the military domain, seizing the tactical initiative requires detecting targets at extremely long distances and across wide surveillance ranges. The targets typically exhibit small target characteristics, occupying a small number of pixels within the entire imaging area and are inherently lacking distinct structural, color, or texture details [4,5]. Furthermore, the extreme complexity of practical backgrounds (e.g., heavy sea clutter or clouds), the diversity of sensor noise, and the variations in target dimensions caused by changing sensor-to-target distances frequently cause the faint radiative signals to be completely submerged [6]. These compounded factors make discovering targets in the low level of signal-to-noise ratios (SNR) images exceptionally difficult, ensuring that the IRSTD task remains highly challenging.

Researchers have developed many methods over the past decades. Current methods are mainly divided into sequence-based methods and single-frame-based methods. Compared with the sequence, the single-frame methodologies offer several significant advantages. First and foremost, it does not rely on the prior information of target motion. This makes it more robust when the camera or the target moves quickly and unpredictably. Furthermore, single-frame methods have lower computational costs and less memory consumption. Therefore, they are more suitable for real-time applications on hardware platforms with limited power. Consequently, this paper focuses on the single. Filter-based methods like Top-Hat [7], Max-Median [8], and wavelet-domain filters [9] tried to enhance targets by removing smooth backgrounds. However, they failed when the background is complex or uneven. LCM (Local contrast-based methods) look for the brightness difference between the target and its neighbors [10–15]. These methods improve detection but reacted strongly to bright noise, which caused many false alarms. Low-Rank and sparse decomposition-based methods (LRSD) assume the background is low-rank and the target is sparse. Notable instances are IPI [16], RIPT [17], PSTNN [18], and NRAM [19]. However, they are slow to compute and fail when the background has strong edges [20].

Recently, deep learning-based (CNN) methods have improved IRSTD significantly [21]. These networks learn features from large datasets. MDvsFA [22] use GANs (Generative Adversarial Networks) to balance missed detections and false alarms. Other networks like ACMNet [23], ALCNet [24], DNANet [25], and UIU-Net [26] use attention or nested structures to find targets. Some researchers use transformers to catch long-distance connections in images [27–29]. However, transformers are slow for real-time use because they require too much computation [30]. To fix this, state space models (SSMs) like Mamba have become popular [31]. Mamba is very fast and can see the whole image efficiently. Vision Mamba (Vim) [32] and VMamba [33] use this for vision tasks by scanning images in different directions.

Despite the rapid development of deep learning, IRSTD remains a formidable challenge because it is difficult to capture subtle pixel variations and weak thermal signatures. Even current SOTA (state-of-the-art) models face three major limitations: primarily, standard convolutions often act as low-pass filters that mix intensity signals with high-frequency details, which caused small targets to disappear in deep layers [30,34]; moreover, most networks used fixed filters that could not effectively capture targets of varying sizes even with multi-scale structures [35–37]; equally importantly, the repetitive downsampling process in

deep architectures significantly reduced image resolution, which led to the degradation of precise spatial coordinates [38,39]. In response to these inherent limitations, we construct GIDNet, which focuses on both local brightness and sharp changes of small targets to achieve superior performance, and our main contributions are summarized into the following three aspects.

1. A gradient-intensity synergistic convolution (GISC) module is proposed to decouple the thermal energy and structural features of targets. In this component, an intensity path is designed to aggregate brightness information, while a gradient path is constructed through a difference mechanism to capture sharp edges. By combining these two paths, the robustness of the network against complex background noise is significantly enhanced.

2. A multi-scale difference contrast (MSDC) module is introduced to handle the variations in target dimensions. In this phase, multiple filter scales are utilized to scan for potential targets across different receptive fields. Through this multi-scale methodology, the functional effectiveness of our framework is effectively improved when dealing with targets of fluctuating sizes.

3. A shallow feature projection (SFP) strategy is developed to preserve critical spatial information. A direct connection is established between the initial layers and the ultimate prediction head, so that high-resolution details are transmitted directly to the output stage. Consequently, the precise spatial localization of small-scale targets is successfully achieved through this projection.

To verify the effectiveness of GIDNet, comprehensive evaluations are performed across three authoritative benchmark datasets, includingIRSTD-1k [26], NUAA-SIRST [23], and NUDT-SIRST [25]. Both quantitative and qualitative evaluations demonstrate that the GIDNet outperforms 16 representative algorithms.

2. Materials and Methods

The architecture and underlying mechanisms of GIDNet are delineated in this part. First, Section 2.1 provides a systematic breakdown of the overall network architecture. Then, the GISC module is introduced in Section 2.2 to handle feature decoupling. Subsequently, the MSDC module is detailed in Section 2.3, which aims to capture targets of various sizes. Furthermore, Section 2.4 explains the SFP strategy for precise spatial localization. Finally, the loss criteria employed for model training are formulated in the Section 2.5.

2.1. Overall Architecture

As schematized in Figure 1, the proposed GIDNet is designed to mitigate the loss of fine-grained details during deep feature extraction while effectively suppressing background clutter. The pipeline consists of three distinct stages: a hierarchical feature encoder, a semantic reconstruction decoder, and SFP.

The encoder is engineered to capture a hierarchical feature representation, evolving from fine-grained spatial cues to coarse-grained semantic abstractions, through four hierarchical stages. First, an initial convolution layer is applied to transform the source image I , introducing to the feature space. To amplify the discriminative cues of dim and small targets amidst intricate background clutter, the proposed MSDC module is utilized as the core building block, instead of using standard convolutional backbones. Within the encoder, the network distills hierarchical features by processing latent representations through successive MSDC blocks, while dimensionality reduction via max-pooling layers is employed to expand the receptive field. Let E_i stand for the feature map produced at the i -th encoder stage, where $i \in \{0, 1, 2, 3\}$. Consequently, the encoding stages which distill multi-scale representations are formulated as:

$$E_0 = F_{MSDC}(Conv_{init}(I)) \quad (1)$$

$$E_i = F_{MSDC}(P_{max}(E_{i-1})), \quad i = 1, 2, 3 \quad (2)$$

where $Conv_{init}(\cdot)$ represents the initial 1×1 convolution, $P_{max}(\cdot)$ denotes a 2×2 max-pooling layer with a subsampling factor of 2, and $F_{MSDC}(\cdot)$ represents the feature extraction via the MSDC module. $I \in R^{h \times w \times 1}$ denotes the input infrared grayscale image; the variables h and w correspond to the spatial height and width, with 1 indicating the channel depth. As the network deepens, the spatial resolution decreases while the channel dimension increases, capturing robust semantic context.

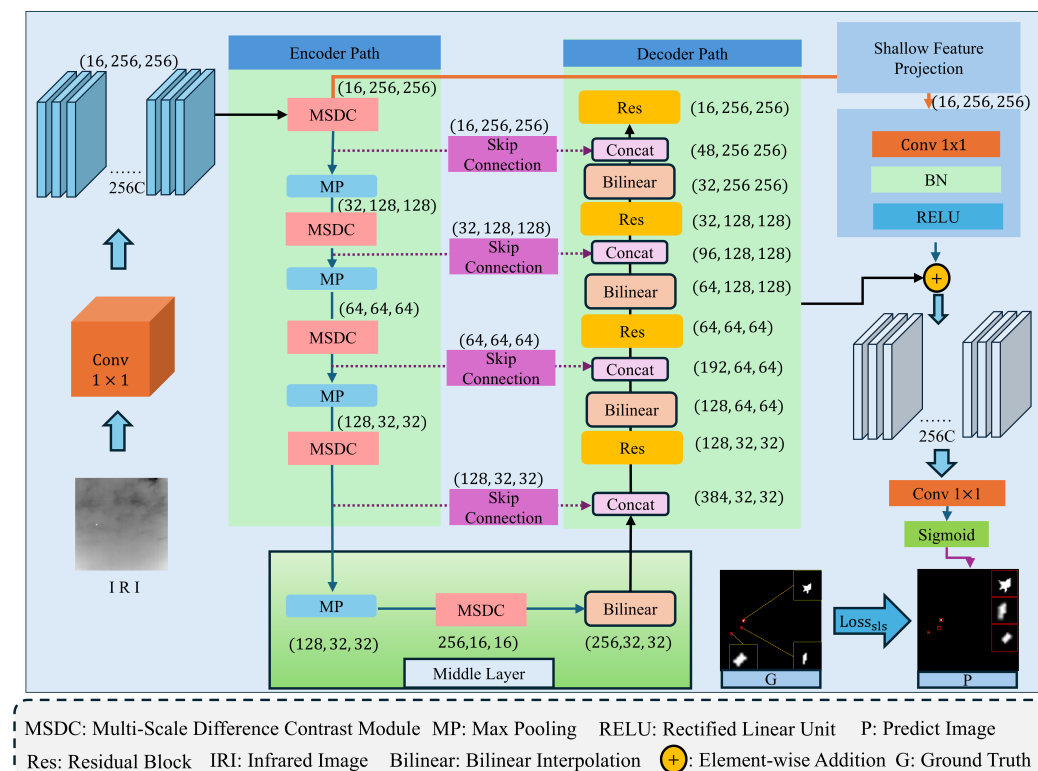


Figure 1. The holistic framework of the GIDNet.

Following the encoder, a middle layer acts as a bridge. This layer comprises a max pooling layer, an MSDC module, and a bilinear upsampling operation. It first consolidates the deepest semantic features and subsequently recovers their resolution to match the preceding encoder features, preparing for the decoding phase:

$$M = U(F_{MSDC}(P_{max}(E_3))) \quad (3)$$

where $U(\cdot)$ denotes the bilinear upsampling operation. M denotes a middle layer.

The decoder component is designed to incrementally restore the spatial dimensions of feature maps while finessing the intricate details of the targets. We employ residual blocks (Res) in the decoder to refine features effectively. Each decoder stage D_i receives features from the deeper level and fuses them with the corresponding encoder features E_i via skip connections [37].

The reconstruction process at each stage involves bilinear upsampling, concatenation, and residual processing:

$$D_3 = H_{Res}(Cat(M, E_3)) \quad (4)$$

$$D_i = H_{Res}(Cat(U_p(D_{i+1}), E_i)), \quad i = 0, 1, 2 \quad (5)$$

where $Cat(\cdot)$ represents the channel-wise concatenation operation, and $H_{Res}(\cdot)$ denotes a residual block augmented with channel and spatial attention, designed to dynamically accentuate features conducive to target detection.

A critical challenge in IRSTD is that repeated downsampling in the u-net structure often dilutes the signature of tiny targets in the deepest layers. To address this, we introduce the SFP strategy that directly bridges the shallowest, high-resolution details with the final reconstructed features.

Specifically, the initial encoder feature E_0 , which retains the richest spatial information, is projected and fused with the final decoder output D_0 . This is achieved through a dedicated projection module $\Psi(\cdot)$ comprising a 1×1 convolution, batch normalization (BN), and $RELU$ activation. The final feature map F_{final} is obtained via:

$$\Psi(E_0) = \sigma(BN(Conv_{1 \times 1}(E_0))) \quad (6)$$

$$F_{final} = D_0 + \Psi(E_0) \quad (7)$$

where $Conv_{1 \times 1}$ stands for a 1×1 convolution operation and σ is the $RELU$ activation function. This residual-style fusion ensures that high-frequency details lost during deep encoding are reintroduced before the final prediction.

Finally, F_{final} passes through a 1×1 convolution layer ($Conv_{out}$) which is then succeeded by:

$$P = Sigmoid(Conv_{out}(F_{final})) \quad (8)$$

where $P \in R^{h \times w \times k}$ is generated by passing the fused features, and k represents the number of output classes.

This architecture ensures that the network benefits from both the deep semantic context required for clutter suppression and the shallow spatial details necessary for precise localization of small targets.

2.2. Gradient-Intensity Synergistic Convolution (GISC)

As evidenced in Figure 2, we design the GISC component with a dual-branch topology. The goal is to balance semantic aggregation and fine-grained edge extraction. This module shares the same original weights W . It comprises three key stages: intensity perception, gradient embedding, and synergistic fusion.

As illustrated in the upper part of Figure 2, the intensity perception branch is designed to preserve both the energy information and low-frequency semantic context of the targets. To achieve this, a standard 2-dimensional (2D) convolution is employed to process the input feature map x . Within a local receptive field R , the branch aggregates local pixel intensities using the original weights W , representing the adaptive weights of the convolution operator. We can characterize this transformation as:

$$y_{int} = x \cdot W \quad (9)$$

where p_0 and p_n denote the central pixel coordinate and the neighbor offset, respectively. Furthermore, $y_{int}(p_0)$ signifies the resulting intensity feature. Although this process effectively extracts brightness patterns in a manner consistent with a traditional CNN, it tends to blur the edges of weak targets due to the smoothing nature of standard convolutions.

To compensate for the attenuation of high-frequency components, a gradient embedding branch is integrated into our framework, which is shown in the gray region of Figure 2. Instead of introducing extra edge operators, we reconstruct the original weights W to build a difference convolution kernel W_{diff} .

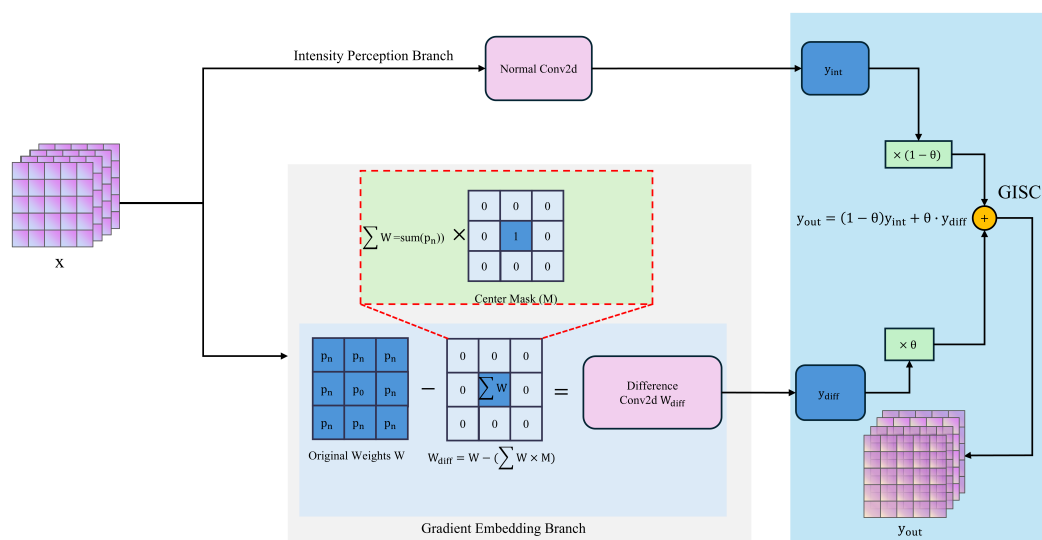


Figure 2. Schematic illustration of the proposed GISC component.

Subsequently, this total sum is deducted exclusively from the center position of the kernel by employing a center mask matrix M , where the center element is 1 and all others are 0. The corresponding mathematical expression is formulated as follows:

$$W_{diff} = W - \left(\sum_{p_n \in R} W(p_n) \right) \cdot M \quad (10)$$

Through this operation, a zero-sum constraint is strictly enforced on the weight W_{diff} , which ensures that $\sum W_{diff} = 0$. To provide a rigorous mathematical explanation for this gradient-capturing capability, the Taylor series expansion is introduced to demonstrate the interpretability of the module. Consequently, the convolution [40] is transformed into an aggregation of local pixel differences:

$$y_{diff} = x \cdot W_{diff} \quad (11)$$

By expanding the local intensity function, the relationship between the difference convolution and the image derivatives can be formally established. For any local position $p_0 + p_n$, the pixel value $x(p_0 + p_n)$ can be approximated via a first-order Taylor expansion at the center p_0 :

$$x(p_0 + p_n) \approx x(p_0) + (\nabla x(p_0))^T \cdot p_n + O(p_n^2) \quad (12)$$

where $x(p_0)$ is the zero-order term (constant intensity), $\nabla x(p_0)$ is the first-order derivative (gradient information), and $O(p_n^2)$ represents higher-order terms.

We substitute this approximation into the difference convolution formula. Since the sum of W_{diff} is zero ($\sum W_{diff}(p_n) = 0$), the zero-order term $x(p_0)$ is perfectly eliminated:

$$\begin{aligned} y_{diff}(p_0) &= \sum_{p_n \in R} W_{diff}(p_n) \cdot x(p_0 + p_n) \\ &\approx \underbrace{\left(\sum W_{diff}(p_n) \right)}_{=0} \cdot x(p_0) + \sum_{p_n \in R} W_{diff}(p_n) \cdot \left((\nabla x(p_0))^T \cdot p_n \right) \end{aligned} \quad (13)$$

This derivation proves that the zero-sum property of W_{diff} mathematically filters out the zero-order intensity term. It forces the kernel to focus exclusively on the first-order gradient term ∇x . Thus, this branch functions as a learnable high-pass filter. It keenly captures edge transitions of infrared small targets. As shown in the blue region of Figure 2,

we employ a linear weighted fusion strategy. We define a hyperparameter $\theta \in [0, 1]$ as the gradient focus factor. The final output y_{out} is formulated as:

$$y_{out} = (1 - \theta) \cdot y_{int} + \theta \cdot y_{diff} \quad (14)$$

To balance the contributions of thermal energy and structural details, we introduce the gradient focus factor θ . As illustrated in the preceding formula, θ is empirically set as a fixed hyperparameter, with its optimal value determined through sensitivity analysis (detailed in Section 3.5). Notably, unlike traditional attention mechanisms that rely on strict sigmoid constraints, our architecture purposely omits such restrictions on θ in the structural design. This design choice grants the network the theoretical potential to adopt negative weights, thereby facilitating active clutter suppression and noise cancellation in highly diverse and cluttered infrared backgrounds.

2.3. Multi-Scale Difference Contrast Module (MSDC)

IRSTD exhibits significant size variations. Single-scale convolution kernels fail to capture these diverse patterns simultaneously. Therefore, we design the MSDC module. This module employs a split-transform-merge strategy, which integrates the proposed GISC unit to extract multi-scale difference features. It should be noted that MSDC is different from common feature enhancement and attention-based modules. Many existing feature enhancement blocks mainly strengthen features by stacking convolution layers or by mixing channel information. Attention-based designs usually learn spatial or channel weights and then use these weights to highlight important regions. In contrast, MSDC does not only assign weights to the input feature map. It aims to build clear difference cues between small infrared targets and the nearby background under several receptive fields. By splitting the feature map into different branches and using GISC units with different dilation rates, MSDC can describe both local contrast and wider context. Thus, the proposed module is not a simple attention plug-in, but a multi-scale difference contrast unit designed for infrared small target detection.

The internal configuration of the MSDC module, designed for capturing multi-scale contrastive cues, is delineated in Figure 3. Initially, a 1×1 convolution is employed for the purpose of dimensionality reduction within the input feature space X_{in} . Subsequently, the projected feature map is partitioned along the spectral dimension into S feature subsets, denoted as $\{x_1, x_2, \dots, x_S\}$, which are then processed through individual branches to yield the corresponding outputs $\{y_1, y_2, \dots, y_S\}$. Specifically, the scale factor S is set to 4 in this implementation. Thereafter, these subsets are processed through parallel branches, each characterized by a distinct receptive field.

Inspired by the work in [41], a hierarchical cascade architecture is developed for these branches, which promotes efficient multi-scale information exchange.

Specifically, the first branch subset x_1 serves as a reference path, where the original information is directly preserved through an identity mapping:

$$y_1 = x_1 \quad (15)$$

To capture fine-grained local details, the second subset x_2 is processed by a GISC unit configured with a unit dilation rate ($d = 1$):

$$y_2 = F_{\text{GISC}}(x_2, \quad d = 1) \quad (16)$$

Furthermore, to effectively expand the receptive field, the third branch employs a progressive integration strategy. Specifically, the summation of x_3 and the preceding output y_2 is fed into a GISC unit with an increased dilation rate of $d = 2$:

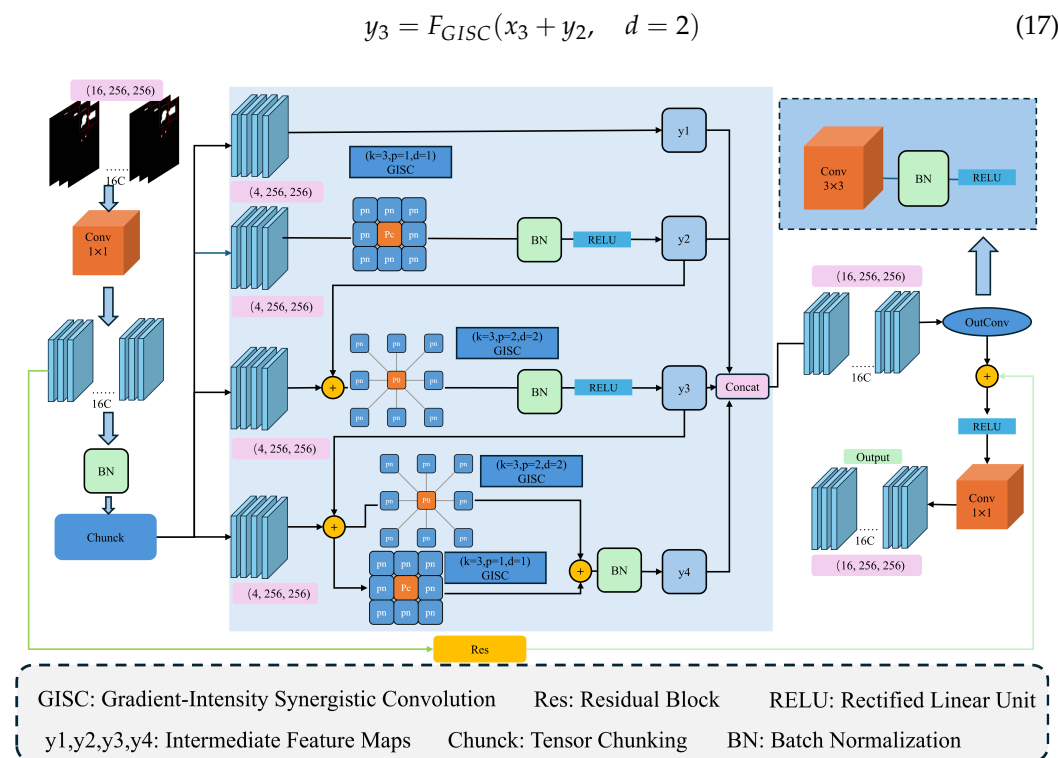


Figure 3. The architecture of the MSDC component.

To handle the most intricate patterns, the final branch receives the sum of x_4 and y_3 as input. A dual-path architecture is integrated within this branch to enhance feature richness. In this configuration, a GISC unit with $d = 1$ is applied in one path, while a GISC unit with $d = 2$ is utilized in the other. The output of the final branch is subsequently obtained by summing these two paths:

$$y_4 = F_{GISC}(temp, d = 1) + F_{GISC}(temp, \quad d = 2) \quad (18)$$

where $temp = x_4 + y_3$. This design captures both compact and extended target characteristics simultaneously.

Ultimately, all outputs $\{y_1, y_2, y_3, y_4\}$ are aggregated via channel concatenation, followed by a 3×3 convolution to integrate these multi-scale features. To alleviate the vanishing gradient problem and facilitate back-propagation, a residual component (Res) is further incorporated into the initial input. Consequently, the final output X_{out} is formulated as follows:

$$X_{out} = RELU(Conv_{3 \times 3}(Concat(y_1, y_2, y_3, y_4)) + X_{in}) \quad (19)$$

The role of each branch is also different from that in general multi-branch enhancement modules. In many multi-branch designs, different branches are only used to enlarge the feature space. In MSDC, the branches have a clear order and task. The first branch keeps the original feature as a stable reference. The second branch focuses on fine local contrast. The third branch receives the former output and further enlarges the receptive field. The last branch uses two GISC paths to cover compact and extended target patterns at the same time. This ordered design allows small-scale target cues to be passed to larger-scale branches step by step, which helps suppress background noise and keeps weak target responses.

By synergizing a split-transform-merge strategy with a hierarchical cascade, the MSDC module effectively addresses scale variations inIRSTD. This architecture captures multi-scale features via progressive summation, while incorporating 1×1 convolutions and

Res to alleviate parameter overhead and facilitate gradient flow. Despite these advantages, the structural complexity, which is primarily driven by the dual-path configuration, presents a trade-off regarding hardware implementation. Additionally, the efficacy of the MSDC module is contingent upon the precise tuning of the scale factor S and dilation rates to ensure robustness across varied infrared environments.

2.4. Shallow Feature Projection Strategy (SFP)

In the classic encoder-decoder architecture, the network expands the receptive field and extracts high-level semantic features through continuous downsampling operations (such as pooling or strided convolution). However, this operation is a double-edged sword forIRSTD. Although deep features can effectively distinguish targets from background clutter, for tiny targets that usually occupy only a few pixels, their precise geometric location information is extremely easily destroyed by quantization errors or completely lost during multiple downsampling processes [38]. Although existing U-net architectures transfer features through skip connections, these features typically need to undergo progressive fusion and processing through multiple decoder levels, and original spatial details are often diluted or smoothed by the time they reach the prediction generator.

Let the output feature map of the first stage of the encoder be $X_{e0} \in R^{h \times w \times k}$. Although the semantic information of this layer is weak, it retains the highest-resolution original edges, textures, and pixel coordinate information of the image, which are crucial for the localization of minute targets. Let the derived feature representation of the highest-resolution stage of the decoder be $X_{d0} \in R^{h \times w \times k}$. This feature has undergone the complete encoder-decoder process and contains rich semantic discrimination information, but spatial details may have become blurred.

When effectively fusing the shallow feature X_{e0} into the deep feature X_{d0} , direct addition may lead to a mismatch in feature distributions. Therefore, we design a projection mapping function $P(\cdot)$ to adapt and align the shallow features. The projection process includes a linear transformation and a non-linear activation, calculated as follows:

$$P(X_{e0}) = \text{RELU}(\text{BN}(\text{Conv}_{1 \times 1} \cdot X_{e0})) \quad (20)$$

where $\text{Conv}_{1 \times 1}$ represents a 1×1 convolution kernel. Crucially, this projection is restricted to channel-wise feature alignment and strictly keeps the spatial resolution unchanged. By avoiding any spatial downsampling (e.g., pooling or strided convolutions), it ensures that the fragile geometric structures and pixel coordinates of targets with weak signals are perfectly preserved without information loss during the alignment process.

After obtaining the aligned projection features, we adopt the idea of residual learning to inject them directly into the prediction generator of the decoder. The fused feature X_{d0}^{final} is derived as:

$$X_{d0}^{\text{final}} = X_{d0} + P(X_{e0}) \quad (21)$$

where $+$ represents element-wise addition. Unlike feature concatenation, which typically requires further convolutional fusion that could inadvertently blur or dilute weak signals, this residual-style addition directly superimposes the pristine shallow details onto the deep semantic map. This robust mechanism guarantees that the faint thermal energy of small targets is protected and highlighted, rather than being suppressed by dominant background features.

The SFP strategy is specifically designed to counteract the loss of precise geometric localization for tiny targets caused by continuous downsampling and quantization errors. By leveraging the first encoder feature map, the module preserves crucial high-resolution spatial details such as original edges and pixel coordinate information. Consequently, it

provides a robust structural guarantee to prevent information loss, particularly ensuring the reliable transmission of weak target signals to the final prediction head.

2.5. Loss Function

Due to the diminutive size of the targets, traditional loss functions (such as binary cross-entropy loss) tend to be dominated by a massive amount of background pixels, leading to difficulty in network convergence. Furthermore, for tiny targets, pixel-level localization deviation and scale fluctuation can have a huge impact on detection performance. To solve these challenges, we trained GIDNet using a scale and location sensitive loss (L_{SLs}) inspired by [42].

L_{SLs} consists of two parts: scale sensitive loss (L_S) and location sensitive loss (L_L). The total loss function is defined as follows:

$$L_{SLs} = L_S + L_L \quad (22)$$

L_S is designed to enhance robustness to minor target-scale changes. To this end, it optimizes the soft intersection over union (*SoftIoU* [6]) between the predicted target and the ground-truth target. In addition, a scale penalty term is incorporated.

The traditional loss is not directly amenable to gradient-based optimization; therefore, *SoftIoU* is employed as a continuous relaxation of the traditional loss. This formulation is particularly suitable forIRSTD, where the target region typically accounts for less than 0.01% of the entire image. Under such extreme class imbalance, conventional pixel-wise loss functions are often dominated by the extensive background and thus fail to provide sufficient supervision for tiny targets. By computing overlap directly from continuous probability maps and characterizing the target at the object level, *SoftIoU* effectively alleviates this limitation. As a result, the optimization landscape becomes smoother, which facilitates stable back-propagation. Meanwhile, the relative contributions of foreground and background are naturally balanced, regardless of the absolute target size. Guided by this consideration, the formulation of L_S starts from the basic *SoftIoU* defined between the predicted probability map P and the binary ground-truth map G :

$$SoftIoU = \frac{\sum(P \cdot G) + \epsilon}{\sum P + \sum G - \sum(P \cdot G) + \epsilon} \quad (23)$$

where ϵ is a smoothing term.

To further improve the responsiveness of the loss function to variations in target scale, an additional penalty coefficient α is introduced. This factor is formulated according to the discrepancy between P and G .

$$\alpha = \frac{\min(S_p, S_g) + D_{dis} + \epsilon}{\max(S_p, S_g) + D_{dis} + \epsilon} \quad (24)$$

where $S_p = \sum P$ and $S_g = \sum G$ represent the pixel sum of the P and G , respectively. D_{dis} measures the square term, defined as $D_{dis} = ((S_p - S_g)/2)^2$.

The L_S incorporates a weighting coefficient β (used to balance samples), defined as:

$$L_S = \sum(\beta \cdot (1 - \alpha \cdot SoftIoU)) \quad (25)$$

where β dynamically adjusts weights according to the target scale, ensuring that extremely small targets receive sufficient attention during training. Compared to the fixed parameter θ , the penalty factor α and the weight β in the loss function L_{SLs} rely entirely on data. The network calculates them during the training process. As shown in the above two

equations, the value of α is determined by the pixel difference D_{dis} between the predicted map P and the ground truth G at the current step. Because of this design, the loss function can automatically adjust itself to fit targets of different sizes. In this way, the model achieves a strong generalization ability across various infrared scenes without extra manual changes.

To solve the localization drift problem that easily occurs with small targets, L_L introduces geometric constraints based on the centroid. This loss function optimizes localization precision by minimizing the centroid distance and angle difference between the P and G .

For each prediction block, we calculate the centroid (x_p, y_p) of the P and the centroid (x_g, y_g) of G . The location loss includes length loss (L_{len}) and angle loss (L_{ang}):

$$L_L = \frac{1}{N} \sum_{i=1}^N (1 - L_{len}^{(i)} + L_{ang}^{(i)}) \quad (26)$$

where L_{len} measures the matching degree of centroid vector modulus lengths:

$$L_{len} = \frac{\min(L_p, L_g)}{\max(L_p, L_g) + \epsilon} \quad (27)$$

where L_p and L_g are the Euclidean distances from the P and G to the origin, respectively.

L_{ang} measures the consistency of centroid vector directions:

$$L_{ang} = \frac{4}{\pi^2} \left(\arctan\left(\frac{y_p}{x_p + \epsilon}\right) - \arctan\left(\frac{y_g}{x_g + \epsilon}\right) \right)^2 \quad (28)$$

By combining L_S and L_L , this loss function can accurately segment the shape of the target and precisely locate it in the correct position. Consequently, the proposed design markedly enhances the detection capability of GIDNet when operating in elaborate background environments.

3. Results

This section begins by outlining the experimental settings, covering the datasets, implementation details, and evaluation metrics employed. Subsequently, a comprehensive comparison, which encompasses both quantitative and qualitative analyses, is conducted between the proposed GIDNet and existing methods. Finally, ablation studies are performed to verify the effectiveness of each core module within the network.

3.1. Datasets and Experimental Settings

To thoroughly assess the proposed GIDNet, experiments are conducted on three widely recognized benchmark datasets forIRSTD:

IRSTD-1k [26] consists of 1001 real infrared images (IRIs) characterized by complex background clutter and targets with diverse spatial scales. The annotated targets vary in size from 1×1 to 40×40 pixels. Following the standard protocol, the dataset is partitioned into training and testing subsets with a ratio of 4:1.

NUAA-SIRST [23] includes 427 IRIs. This dataset is widely recognized for its severe background interference and extremely low SNR targets. Similar toIRSTD-1k, the samples are divided into training and testing sets using a 4:1 split.

NUDT-SIRST [25] contains 1327 images and covers a broader range of scenes with multiple types of interference sources. Owing to this diversity, this dataset is regarded as one of the most challenging benchmarks forIRSTD. In this dataset, the images are evenly divided into training and testing subsets with a 1:1 ratio.

The proposed GIDNet is implemented using the PyTorch 2.0.0 framework within a Python 3.8 environment on Ubuntu 20.04. Model training is performed on a high-

performance workstation equipped with a single NVIDIA GeForce RTX 4090 GPU. During training, all input images are uniformly resized to 256×256 pixels. To improve generalization ability and alleviate overfitting, several data augmentation techniques are employed, including random horizontal and vertical flips as well as stochastic rotations. The network parameters are optimized using the Adagrad optimizer with an initial learning rate of 5×10^{-2} . In addition, a cosine annealing schedule is adopted to gradually decrease the learning rate throughout 800 training epochs. To mitigate the issues of foreground as well as background imbalance and localization deviation that frequently arise in IRSTD tasks, the training objective incorporates the L_{SLC} loss function.

3.2. Evaluation Metrics

To quantitatively assess detection capability, three widely used criteria in IRSTD are adopted, namely the intersection over union (IoU), probability of detection (P_d), and false alarm rate (F_a) [4,5,21,25]. In the following table, $\uparrow$ indicates that higher values are better, while $\downarrow$ indicates that lower values are better.

Among them, IoU characterizes the algorithm's ability to preserve target morphology at the pixel level, with particular emphasis on the accuracy of boundary delineation [26]. Specifically, it is calculated as the ratio of the overlapping region between the P and G annotations to their combined area:

$$IoU = \frac{\sum_{i=1}^{N_{pixels}} TP_i}{\sum_{i=1}^{N_{pixels}} (TP_i + FP_i + FN_i)} \quad (29)$$

where TP , FP , and FN represent the numbers of true-positive, false-positive, and false-negative pixels. A larger IoU value indicates a greater overlap between the P and the G , reflecting more accurate contour localization and boundary characterization of the target.

P_d is a target-level indicator used to quantify the proportion of successfully detected targets among all annotated targets [23,24]. In this study, a prediction is regarded as correct when the Euclidean distance between the centroid of P and that of the G is smaller than a predefined threshold T . Following [22,25,43], T is fixed at 3 pixels in all experiments. A prediction is considered correct when the distance between its center and the ground truth is smaller than a threshold T . Since all input images are resized to 256×256 pixels before testing, this fixed threshold remains stable and fair for our evaluation. However, in practical applications, the sizes of images from cameras often change. A fixed threshold of 3 pixels may fail to apply to these changing sizes. To solve this problem, future studies could design a dynamic threshold that adjusts according to the input image size.

$$P_d = \frac{N_{correct}}{N_{total}} \quad (30)$$

where $N_{correct}$ denotes the volume of targets that satisfy the centroid-matching criterion, whereas N_{total} represents the total number of G in the dataset. A higher P_d value indicates superior detection sensitivity and demonstrates a stronger ability of the algorithm to identify tiny targets embedded in complex backgrounds.

F_a is defined as the proportion of F_a pixels to the total number of pixels in the dataset. This metric is commonly used to reflect the extent to which an algorithm is affected by background interference and noise contamination [5,34].

$$F_a = \frac{P_{false}}{N_{img} \times h \times w} \quad (31)$$

where P_{false} represents the total pixel area of predicted regions that do not match any ground truth target (i.e., false alarms), and $N_{img} \times h \times w$ represents the total pixel count of the dataset. Notably, a lower F_a is indicative of better performance, demonstrating the model's robustness in effectively suppressing background artifacts and minimizing the occurrence of erroneous detections.

Additionally, the receiver operating characteristic (ROC) curve is employed to assess detection behavior and model robustness under varying decision thresholds [44]. This curve depicts the relationship between the true positive rate (TPR) and the false positive rate (FPR), thereby providing a comprehensive view of the performance variation induced by threshold adjustment. By continuously changing the confidence threshold of the prediction results, a series of (FPR, TPR) points can be obtained to form the ROC curve:

$$TPR = \frac{TP}{TP + FN} \quad (32)$$

$$FPR = \frac{FP}{FP + TN} \quad (33)$$

In general, better detection performance is indicated by a curve located closer to the upper-left corner, which corresponds to a higher TPR achieved at a lower FPR.

3.3. Quantitative Comparison

We compared GIDNet with 16 representative methods, which include traditional methods, CNN-based methods, and hybrid CNN models. For the traditional methods, we selected eight well-established algorithms. These include filter-based methods such as Top-Hat [7], and local contrast-based methods like TLLCM [11] and WSLCM [15]. We also evaluated several patch-image-based models, including IPI [16], RIPT [17], NRAM [19], PSTNN [18], and MSLSTIPT [20]. To assess the performance of our proposed method, we conducted a comprehensive comparison with seven representative CNN architectures, focusing primarily on CNN-based paradigms. Specifically, we selected foundational models such as UIU-Net [26] and MSHNet [6], along with several recent high-performance networks, including MMLNet [45], HDNet [46], SDS-Net [43], and SCTransNet [27]. Additionally, for L2SKNet [47], both of its distinct architectural variants (L2SKNet-Unet and L2SKNet-FPN) were included in our evaluation. The methods are categorized as follows: Trad-F (filter-based traditional methods), Trad-C (local contrast-based traditional methods), Trad-L (low-rank-based traditional methods), CNN, and CNN-T (hybrid CNN and transformer-based methods).

Table 1 summarizes the quantitative evaluation on the IRSTD-1K dataset, where our GIDNet outperforms existing methods across all metrics. As observed from the data, such as Top-Hat, IPI, and the more recent MSLSTIPT generally exhibit suboptimal performance. These methods struggle to balance a higher P_d with a lower F_a , largely because hand-crafted features lack the representative power to distinguish dim targets from complex, cluttered backgrounds. For instance, while IPI achieves a relatively low F_a of 16.18%, its IoU remains stagnant at 27.92%, which is significantly lower than that of CNN competitors.

In contrast, CNN architectures demonstrate a substantial performance leap, underscoring the efficacy of deep feature representations. Among the recent SOTA models, SDS-Net and L2SKNet-Unet emerge as strong contenders. Specifically, SDS-Net achieves an impressive IoU of 66.67% and a P_d of 92.93%. However, our proposed GIDNet outperforms all compared methods across all three evaluation metrics. Notably, GIDNet achieves the highest IoU of 69.01% and a P_d of 93.54%, surpassing the second-best method (SDS-Net) by 2.34% and 0.61%. In addition, SCTransNet, a CNN-T network model, exhibited a remarkably low F_a of 11.84%.

Furthermore, GIDNet demonstrates superior noise suppression capabilities, yielding the lowest F_a of 10.32%. This represents a noteworthy improvement over L2SKNet-Unet (11.84%), which previously held the leading record for noise suppression among the listed CNN models.

Table 1. Performance benchmarks on the IRSTD-1K dataset for GIDNet and its competitors. The primary, secondary, and tertiary rankings are distinguished by bold, underlined, and italicized fonts.

	Type	Size	Year	IoU $\uparrow$	Pd $\uparrow$	Fa $\downarrow$
Top-Hat	Trad-F	256	2010	10.06	75.11	1432
IPI	Trad-L	256	2013	27.92	81.37	16.18
RIPT	Trad-L	256	2017	14.11	77.55	28.31
NRAM	Trad-L	256	2018	15.25	70.68	16.93
PSTNN	Trad-L	256	2019	24.57	71.99	35.26
MSLSTIPT	Trad-L	256	2021	11.43	79.03	1524
TLLCM	Trad-C	256	2020	3.31	77.39	6738
WSLCM	Trad-C	256	2021	3.45	72.44	6619
UIU-Net	CNN	256	2023	64.36	74.88	42.18
MSHNet	CNN	256	2024	63.37	92.86	27.25
SCTransNet	CNN-T	256	2024	66.18	91.92	14.58
L ² SKNet-Unet	CNN	256	2025	66.54	90.24	<u>11.84</u>
L ² SKNet-FPN	CNN	256	2025	63.93	91.58	20.01
MMLNet	CNN	256	2025	66.26	90.24	15.91
HDNet	CNN	256	2025	66.41	92.52	16.4
SDS-Net	CNN	256	2025	<u>66.67</u>	<u>92.93</u>	18.2
GIDNet(Ours)	CNN	256	2026	69.01	93.54	10.32

Table 2 summarizes the comparative results on the NUAA-SIRST, as indicated by the tabulated data, traditional methods such as Top-Hat and IPI, are observed to suffer from exceedingly high $F_a > 10,000$, whereby a significant limitation in suppressing complex background clutter is revealed. Although modest improvements in the IoU are achieved by methods like PSTNN and RIPT, their performance remains substantially inferior to the precision that is fundamentally required for reliable IRSTD.

In contrast, a decisive performance leap is demonstrated by CNN methods. Among the evaluated models, a remarkable IoU of 78.16% is achieved by the proposed GIDNet, by which the majority of recent SOTA models, including UIU-Net (76.41%), MMLNet (77.16%), and the recently proposed SDS-Net (78.01%), are outperformed. Notably, a perfect probability of detection ($P_d = 100\%$) is attained solely by GIDNet, which marginally surpasses the closely competing HDNet (99.89%). By this unparalleled detection capability, the exceptional robustness of our model in identifying infrared targets without omissions, even under severely challenging conditions, is explicitly highlighted. However, SCTransNet, which belongs to the CNN-T network type, did not perform well on the NUAA-SIRST dataset.

Furthermore, while a slight superiority in IoU is maintained by HDNet (78.65% compared to our 78.16%), the lowest false alarm rate ($F_a = 6.74\%$) among all comparable methods is exhibited by GIDNet. This marks a notable improvement over other high-performing architectures, such as HDNet (6.83%), L2SKNet-Unet (9.94%), and SDS-Net (12.7%). This superior balance, struck between high detection sensitivity and effective background suppression, strongly underscores the structural advantages of our proposed network. In summary, it is comprehensively validated by the quantitative results that an optimal trade-off between localization accuracy and target integrity is achieved by GIDNet, establishing it as a robust and efficient architecture for IRSTD.

Table 2. Performance benchmarks on the NUAA-SIRST dataset for GIDNet and its competitors. The primary, secondary, and tertiary rankings are distinguished by bold, underlined, and italicized fonts.

Method	Type	Size	Year	IoU $\uparrow$	Pd $\uparrow$	Fa $\downarrow$
Top-Hat	Trad-F	256	2010	1.51	79.74	16456
IPI	Trad-L	256	2013	1.09	87.05	30467
RIPT	Trad-L	256	2017	16.79	69.76	59.33
NRAM	Trad-L	256	2018	15.25	70.68	16.93
PSTNN	Trad-L	256	2019	30.3	72.8	48.99
MSLSTIPT	Trad-L	256	2021	1.08	0.052	8.18
TLLCM	Trad-C	256	2020	4.24	88.37	6243
WSLCM	Trad-C	256	2021	6.39	88.74	4462
UIU-Net	CNN	256	2023	76.41	87.92	79.54
MSHNet	CNN	256	2024	57.93	82.57	29.28
SCTransNet	CNN-T	256	2024	74.73	98.17	51.2
L2SKNet-Unet	CNN	256	2025	70.26	98.17	9.94
L2SKNet-FPN	CNN	256	2025	68.85	97.25	75.83
MMLNet	CNN	256	2025	77.16	96.33	38.05
HDNet	CNN	256	2025	78.65	<u>99.89</u>	<u>6.83</u>
SDS-Net	CNN	256	2025	<u>78.01</u>	<u>99.08</u>	12.7
GIDNet(ours)	CNN	256	2026	<u>78.16</u>	100	6.74

Table 3 presents a comprehensive quantitative evaluation on the NUDT-SIRST. As demonstrated by the experimental results, CNN architectures consistently and significantly surpass traditional mathematical models across all three evaluation metrics. Traditional methods, such as MSLSTIPT and TLLCM, struggle severely with elevated F_a and remarkably low IoU scores, reflecting their limited robustness against complex background clutter and varying target scales.

Notably, our proposed GIDNet excels in F_a suppression, achieving the optimal overall performance with an F_a of strictly 2.80. This marks a discernible improvement even over the most recent and competitive baselines, such as HDNet (2.82) and L2SKNet-Unet (5.72). Concurrently, GIDNet maintains a highly competitive P_d at 98.41%. While SDS-Net (98.73%) and HDNet (98.52%) report marginally higher detection probabilities, they inherently compromise by yielding higher F_a . This indicates that GIDNet strikes a superior balance, effectively isolating true infrared targets without aggressively misclassifying background artifacts.

With regard to pixel-level shape description, GIDNet registers an IoU of 83.51%. Although several SOTA networks, notably MMLNet (94.93%), SCTransNet (93.57%), and L2SKNet-Unet (93.48%), exhibit superior segmentation completeness, the proposed network remains highly viable. In practicalIRSTD scenarios, the operational priority is often heavily weighted toward absolute minimization of false alarms along with the reliable discovery of targets (P_d), both of which are critical dimensions where GIDNet demonstrates exceptional and leading capability.

Although MMLNet achieves a superior IoU of 94.93%, our GIDNet's lower score (83.51%) stems from a strategic emphasis on background clutter suppression rather than pixel-level morphological fidelity. As evidenced in Table 3, GIDNet achieves the most competitive false alarm rate ($F_a = 2.8$) and a high detection probability ($P_d = 98.41\%$). This suggests that while GIDNet slightly sacrifices boundary integrity due to aggressive feature filtering, it effectively minimizes false positives, making it more robust for practical infrared search and track systems where precision in detection outweighs the necessity for perfect target shape reconstruction.

Table 3. Performance benchmarks on the NUDT-SIRST dataset for GIDNet and its competitors. The primary, secondary, and tertiary rankings are distinguished by bold, underlined, and italicized fonts.

Method	Type	Size	Year	IoU $\uparrow$	Pd $\uparrow$	Fa $\downarrow$
Top-Hat	Trad-F	256	2010	20.72	78.41	166.7
IPI	Trad-L	256	2013	17.76	74.49	41.23
RIPT	Trad-L	256	2017	29.44	91.85	344.3
NRAM	Trad-L	256	2018	6.93	56.4	19.27
PSTNN	Trad-L	256	2019	14.85	66.13	44.17
MSLSTIPT	Trad-L	256	2021	8.34	47.4	888.1
TLLCM	Trad-C	256	2020	2.18	62.01	1608
WSLCM	Trad-C	256	2021	2.28	56.82	1309
UIU-Net	CNN	256	2023	82.98	89.81	47.14
MSHNet	CNN	256	2024	60.51	88.68	88.38
SCTransNet	CNN-T	256	2024	<u>93.57</u>	98.32	7.03
L2SKNet-Unet	CNN	256	2025	<u>93.48</u>	97.78	5.72
L2SKNet-FPN	CNN	256	2025	92.33	<u>98.41</u>	12.3
MMLNet	CNN	256	2025	94.93	98.2	16.36
HDNet	CNN	256	2025	85.58	<u>98.52</u>	<u>2.82</u>
SDS-Net	CNN	256	2025	92.91	98.73	6.75
GIDNet(ours)	CNN	256	2026	83.51	<u>98.41</u>	2.8

To comprehensively evaluate the robustness and detection capability of the proposed method against various background clutter, we plot the ROC curves across three datasets, as depicted in Figure 4. The ROC curves visually articulate the trade-off between the *TPR* and the *FPR*. It is worth noting that the x-axis (*FPR*) is strictly scaled down to the range of $[0, 10^{-4}]$ to emphasize model performance under highly stringent F_a constraints, which is critical for practical IRSTD.

As shown in the graphical comparisons, GIDNet consistently demonstrates superior performance across all evaluated benchmarks. In particular, on the IRSTD-1k dataset, the proposed method exhibits a competitive trend, where GIDNet quickly surpasses baseline models as the *FPR* increases. It ultimately reaches the highest *TPR*, highlighting its excellent ability to preserve target detection capability while minimizing background interference.

On the NUAA-SIRST, GIDNet maintains remarkable consistency. While models like SDS-Net and UIU-Net produce strong early responses, GIDNet follows closely behind their state-of-the-art performance, positioning itself in the top-performing cluster and outperforming older models, such as MSHNet and the L2SKNet variants.

Notably, GIDNet's excellence is particularly apparent on the NUDT-SIRST. At extremely low *FPR*, GIDNet achieves a rapid rise in *TPR*, surpassing all other networks in the early stages of the *FPR* threshold. This rapid approach to near-perfect *TPR* ensures highly sensitive responses to faint infrared targets, even in challenging environments.

In comparison, SCTransNet, known for its efficient CNN-T architecture, shows commendable performance as well, though it slightly trails behind GIDNet in terms of detection sensitivity at low *FPR*. SCTransNet performs well under certain conditions but does not consistently outperform GIDNet across all datasets.

In summary, the quantitative analysis of the ROC curves consistently demonstrates that GIDNet achieves an optimal balance between target localization accuracy and minimizing false alarms (*Fa*), thereby presenting a robust and effective framework for IRSTD tasks.

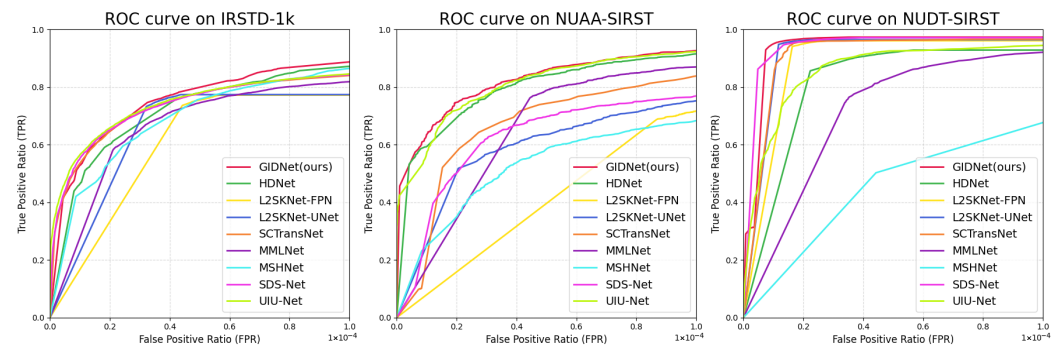


Figure 4. ROC curves of the proposed GIDNet and eight SOTA methods on the IRSTD-1k, NUAA-SIRST, and NUDT-SIRST datasets.

3.4. Qualitative Comparison

As illustrated in Figure 5, we conduct a qualitative analysis to evaluate the performance of our GIDNet against SOTA methods, including SDS-Net, HDNet, MMLNet, L2SKNet, SCTransNet, MSHNet, and UIU-Net. Due to page constraints, these visual examples are drawn from three datasets, which cover various challenging scenarios.

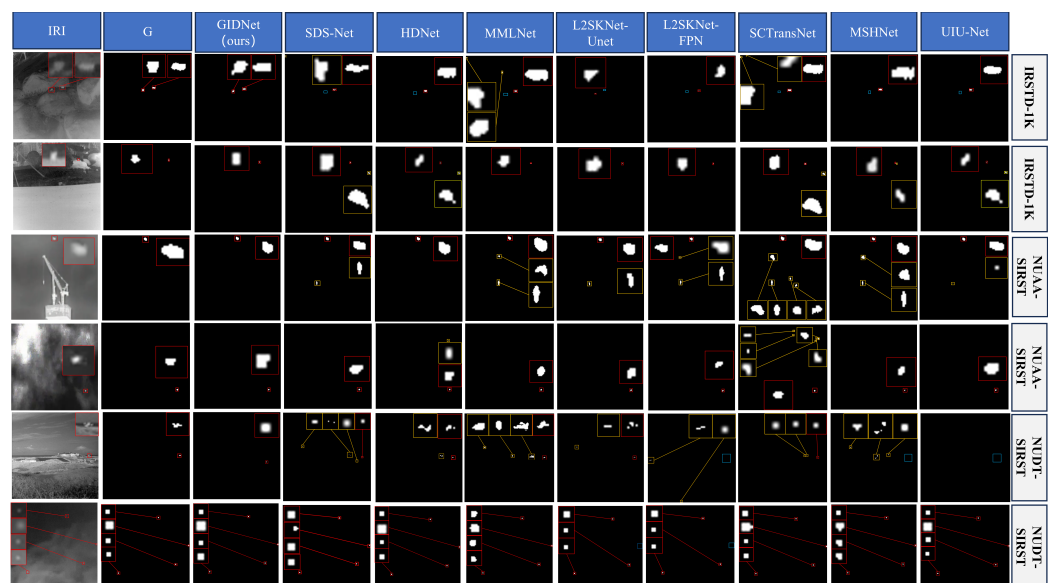


Figure 5. Comparison of the proposed GIDNet against eight SOTA methods. Red, yellow, and blue rectangles indicate correctly detected targets, false alarms, and missed targets. To facilitate clearer observation, magnified insets of the target areas are provided in the image corners.

Upon examining the original IRIs in the first column, two primary challenges for IRSTD are evident. First, the targets exhibit extremely small sizes and diverse morphologies. For example, in the first and second rows, the targets occupy only a few pixels and are buried in significant noise. In the fourth row, multiple targets of varying sizes appear simultaneously. Existing methods like UIU-Net, MSHNet, SDS-Net, and L2SKNet-FPN occasionally suffer from missed detections (marked by blue boxes) or incomplete shape segmentation when dealing with such multi-scale targets.

Second, the presence of complex background clutter and low SNR poses a considerable risk of false alarms. Observing the fifth and sixth rows, which feature building structures and heavy cloud interference, most comparative methods like SDS-Net, HDNet, MMLNet, L2SKNet, MSHNet, and UIU-Net struggle to differentiate real targets from high brightness background artifacts, resulting in numerous false alarms (marked by yellow boxes).

In contrast, our GIDNet demonstrates superior robustness and precision. GIDNet shows an exceptional ability to preserve the spatial integrity of diminutive targets while simultaneously mitigating the interference of pixel-level artifacts. As shown in the third and fifth rows, GIDNet achieves a much lower false alarm rate compared to HDNet and SDS-Net. Additionally, in multi-target scenarios in the fourth row, our model successfully identifies every target instance with precise boundaries, demonstrating its powerful spatial context modeling capabilities. In summary, GIDNet excels in both target highlighting and background clutter suppression, producing results that are most consistent with the G.

3.5. Ablation Experiments

The incremental performance gains resulting from the proposed components are summarized in Tables 4 and 5. To assess the effectiveness of the proposed architecture, we evaluate the individual contributions of the GISC and SFP modules.

Table 4. Quantitative evaluation of the individual contributions of the GISC and SFP modules within the proposed GIDNet on theIRSTD-1k benchmark. The optimal results have been highlighted in bold for easy reference.

No.	GISC	SFP	IoU ↑	Pd ↑	Fa ↓
1			60.78	85.8	44.45
2	✓		68.18	91.98	11.32
3		✓	63.73	89.74	17.38
4	✓	✓	69.01	93.54	10.32

Table 5. Quantitative evaluation of the individual contributions of the GISC and SFP modules within the proposed GIDNet on the NUAA-SIRST benchmark. The optimal results have been highlighted in bold for easy reference.

No.	GISC	SFP	IoU ↑	Pd ↑	Fa ↓
1			70.86	90.45	27.99
2	✓		78.05	99.07	8.94
3		✓	72.92	95.58	16.27
4	✓	✓	78.16	100	6.74

As reported in the Table 4, the baseline model exhibits suboptimal performance, yielding an IoU of 60.78% and a P_d of 85.8%. Upon integration of the GISC module, a substantial performance gain is observed, with IoU escalating to 68.18% and F_a being significantly suppressed to 11.32%. Similarly, the independent contribution of the SFP component is demonstrated, through which IoU is increased to 63.73%. Ultimately, the synergistic effect of both modules is demonstrated by the complete GIDNet architecture, which achieves the superior metrics across all categories, specifically reaching a peak IoU of 69.01% and a minimum F_a of 10.32%.

The generalizability of the proposed framework is further validated on the NUAA-SIRST. The baseline configuration achieves an IoU of 70.86% with a F_a of 27.99%. With the inclusion of the GISC component, the model's discriminative capability is notably strengthened, demonstrated by the IoU increasing to 78.05% and the P_d reaching 99.07%. The optimal performance is consistently delivered by GIDNet, which achieves a perfect P_d of 100% and a negligible F_a of 6.74%. These results reinforce the conclusion that the integration of GISC and SFP is vital for high-precision infrared small target detection.

To investigate the influence of the hyperparameter θ on model performance, we conduct a sensitivity analysis by varying its value. Table 6 presents quantitative results from the internal ablation experiments on the θ parameter within the GISC module of

GIDNet, conducted on the IRSTD-1k. The threshold parameter θ is varied from 1.0 to 0. To evaluate the performance, three metrics are employed: IoU , P_d , and F_a . Notably, as θ decreased from 1.0 to 0.7, a favorable trend in detection performance is observed, characterized by a steady increase in both IoU and P_d alongside a concurrent reduction in F_a .

However, further decreasing θ from 0.7 to 0 results in a performance degradation, characterized by a reduction in both IoU and P_d , accompanied by a deterioration in the F_a score. Consequently, $\theta = 0.7$ is identified as the optimal threshold, yielding superior results across all metrics. Specifically, at this setting, the model achieves an IoU of 69.01%, a P_d of 93.54%, and a minimized F_a of 10.32%. Based on these empirical findings, we select 0.7 as the final value for our proposed framework.

Table 6. Sensitivity analysis of the hyperparameter θ within the proposed GISC module on the IRSTD-1k benchmark. Boldface entries denote the optimal performance.

No.	θ	$IoU \uparrow$	$P_d \uparrow$	$F_a \downarrow$
1	1.0	60.12	84.32	45.33
2	0.9	60.78	85.8	44.45
3	0.8	67.18	91.98	11.32
4	0.7	69.01	93.54	10.32
5	0.6	67.01	91.89	14.05
6	0.5	66.98	91.74	16.72
7	0.4	66.64	91.07	17.42
8	0.3	63.57	89.05	18.77
9	0.2	63.01	87.45	23.09
10	0.1	61.98	86.74	27.89
11	0	60.23	84.92	32.13

3.6. Computational Efficiency

To evaluate model complexity and computational efficiency, several metrics are employed, including network parameters ($Params$), floating-point operations ($FLOPs$), and frames per second (FPS). The quantitative results are summarized in Table 7, where the proposed GIDNet has been benchmarked with SOTA models developed in recent years. As evidenced by the data, GIDNet demonstrates superior efficiency; notably, a minimal parameter footprint is maintained while achieving the lowest $FLOPs$ among all competitors. Compared to conventional CNN architectures, SCTransNet (CNN-T architecture) not only incurs a higher number of $FLOPs$ but also exhibit slower FPS . Simultaneously, an impressive inference speed is attained in terms of FPS . These results collectively validate that the proposed approach facilitates high-speed processing capability characterized by significantly reduced computational complexity.

Table 7. Analysis of computational overhead in terms of $Params$, $FLOPs$ and FPS , between our GIDNet and the eight most recent advanced models

Method	Year	$Params(M) \downarrow$	$FLOPs (G) \downarrow$	$FPS (f/s) \uparrow$
UIU-Net	2023	50.54	54.43	9.87
MSHNet	2024	4.07	6.11	68.96
SCTransNet	2024	11.19	20.24	34.36
L2SKNet-FPN	2025	1.07	6.008	162.67
L2SKNet-Unet	2025	0.899	6.89	110.68
MMLNet	2025	3.58	20.41	38.17
HDNet	2025	3.84	5.96	78.81
SDS-Net	2025	2.701	6.823	41.5
GIDNet(ours)	2026	3.65	5.703	101.13

4. Discussion

In this study, we propose GIDNet to solve the problem of finding small targets in complex backgrounds. The results show that separating gradient and intensity can help the network keep target energy and remove noise. Compared with previous methods, which usually mix features and lose small details during deep learning processes, our method keeps the exact position of targets. This proves our working hypothesis that protecting shallow features and using difference convolution without strict limits can actively suppress background noise.

The main strength of our study is the good balance between finding targets and making few mistakes. In practical applications, keeping a low false alarm rate is very important. Our model achieves the lowest false alarm rate on the test datasets while keeping a high detection rate. This means our network can be a useful tool for actual warning systems and remote sensing tasks.

However, our study has some limitations, and there are sources of uncertainty in our results. First, we use a fixed distance of three pixels to judge if a target is found. If the input picture size changes from different cameras, this fixed number might cause measurement errors. Second, we lack a comparison with the newest Mamba models. We tried to test these models, but we failed to run their code successfully, so we do not know how our network performs against them. Third, we only use three basic metrics to test our network. We do not include other important metrics, such as precision-recall curves, F1-score, or mAP. This makes our evaluation not complete enough to show the full ability of the model.

In the future, we will improve our research in several directions. We will add more metrics like F1-score and precision-recall curves to test our model from different sides. We will also continue to study the code of Mamba models to finish the comparison. Finally, we plan to use time information from continuous video frames to help the network find moving targets more easily.

5. Conclusions

In this paper, a novel GIDNet was developed to mitigate SNR and background interference inIRSTD by addressing “spectral coupling”, a limitation in standard CNNs where simultaneous processing often resulted in the loss of subtle target signatures. To circumvent this, the GISC module was designed to decouple target energy from edge details via a Taylor-series-derived zero-sum convolution, while multi-scale features were adaptively captured by the MSDC module through a hierarchical split-transform-merge strategy. High-resolution spatial localization was further preserved by the SFP strategy, which established a direct information highway to the output. Rigorous quantitative and qualitative evaluations were carried out on theIRSTD-1k, NUAA-SIRST, and NUDT-SIRST datasets, which demonstrated that GIDNet consistently outperformed 15 methods in terms of P_d , F_a , and IoU . Ultimately, a robust balance between background suppression and target enhancement was achieved, providing an effective solution for real-world detection that could be further enhanced by integrating temporal information in future research.

Author Contributions: Conceptualization, X.G. and M.Z.; methodology, X.G. and J.W.; software, D.C. and H.X.; validation, X.G., J.W. and D.C.; formal analysis, X.G. and Y.M.; investigation, X.G., J.W. and L.L.; resources, M.Z.; data curation, X.G. and H.X.; writing—original draft preparation, J.W.; writing—review and editing, X.G., J.W. and M.Z.; visualization, Y.M. and L.L.; supervision, M.Z.; project administration, M.Z.; funding acquisition, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62501025, 62471049), the Fundamental Research Funds for the Central Universities (3282025011), and the National Natural Science Foundation of China (62476013).

Data Availability Statement: The source code is available at <https://github.com/besti-irstd/GIDNet> (accessed on 27 April 2026).

Acknowledgments: The authors thank the editor and anonymous reviewers for their valuable comments and suggestions to this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Teutsch, M.; Krüger, W. Classification of small boats in infrared images for maritime surveillance. In Proceedings of the International WaterSide Security Conference, Carrara, Italy, 3–5 November 2010; pp. 1–7.
2. Zhang, J.; Tao, D. Empowering Things with Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. *IEEE Internet Things J.* **2021**, *8*, 7789–7817. [\[CrossRef\]](#)
3. Ying, X.; Wang, Y.; Wang, L.; Sheng, W.; Liu, L.; Lin, Z.; Zhou, S. Local Motion and Contrast Priors Driven Deep Network for Infrared Small Target Superresolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5480–5495. [\[CrossRef\]](#)
4. Zhao, M.; Li, W.; Li, L.; Hu, J.; Ma, P.; Tao, R. Single-Frame Infrared Small-Target Detection: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 87–119. [\[CrossRef\]](#)
5. Cheng, Y.; Lai, X.; Xia, Y.; Zhou, J. Infrared Dim Small Target Detection Networks: A Review. *Sensors* **2024**, *24*, 3885. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Liu, Q.; Liu, R.; Zheng, B.; Wang, H. Review on recent development in infrared small target detection algorithms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2145–2156.
7. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2012**, *43*, 2145–2156. [\[CrossRef\]](#)
8. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. Max-mean and max-median filters for detection of small targets. *Signal Data Process. Small Targets* **1999**, *3809*, 74–83.
9. Wang, Z.; Tian, J.; Liu, J.; Zheng, S. Small infrared target fusion detection based on support vector machines in the wavelet domain. *Opt. Eng.* **2006**, *45*, 076401. [\[CrossRef\]](#)
10. Chen, C.P.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A Local Contrast Method for Small Infrared Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 574–581. [\[CrossRef\]](#)
11. Han, J.; Moradi, S.; Faramarzi, I.; Liu, C.; Zhang, H.; Zhao, Q. A Local Contrast Method for Infrared Small-Target Detection Utilizing a Tri-Layer Window. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1822–1826. [\[CrossRef\]](#)
12. Deng, H.; Sun, X.; Liu, M.; Ye, C.; Zhou, X. Small Infrared Target Detection Based on Weighted Local Difference Measure. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4204–4214. [\[CrossRef\]](#)
13. Qin, Y.; Li, B. Effective Infrared Small Target Detection Utilizing a Novel Local Contrast Method. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1890–1894. [\[CrossRef\]](#)
14. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [\[CrossRef\]](#)
15. Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; Zhao, L. Infrared Small Target Detection Utilizing the Multiscale Relative Local Contrast Measure. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 612–616. [\[CrossRef\]](#)
16. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared Patch-Image Model for Small Target Detection in a Single Image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [\[CrossRef\]](#)
17. Dai, Y.; Wu, Y. Reweighted Infrared Patch-Tensor Model With Both Nonlocal and Local Priors for Single-Frame Small Target Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3752–3767. [\[CrossRef\]](#)
18. Zhang, L.; Peng, Z. Infrared Small Target Detection Based on Partial Sum of the Tensor Nuclear Norm. *Remote Sens.* **2019**, *11*, 382. [\[CrossRef\]](#)
19. Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; Peng, Z. Infrared Small Target Detection via Non-Convex Rank Approximation Minimization Joint l2,1 Norm. *Remote Sens.* **2018**, *10*, 1821. [\[CrossRef\]](#)
20. Zhang, X.; Ding, Q.; Luo, H.; Hui, B.; Chang, Z.; Zhang, J. Infrared small target detection based on an image-patch tensor model. *Infrared Phys. Technol.* **2019**, *99*, 55–63. [\[CrossRef\]](#)

21. Kou, R.; Wang, C.; Peng, Z.; Zhao, Z.; Chen, Y.; Han, J.; Huang, F.; Yu, Y.; Fu, Q. Infrared small target segmentation networks: A survey. *Pattern Recognit.* **2023**, *143*, 109788. [[CrossRef](#)]
22. Wang, H.; Zhou, L.; Wang, L. Miss Detection vs. False Alarm: Adversarial Learning for Small Object Segmentation in Infrared Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 8509–8518.
23. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 950–959.
24. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional Local Contrast Networks for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [[CrossRef](#)]
25. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense Nested Attention Network for Infrared Small Target Detection. *IEEE Trans. Image Process.* **2023**, *32*, 1745–1758. [[CrossRef](#)]
26. Wu, X.; Hong, D.; Chanussot, J. UIU-Net: U-Net in U-Net for Infrared Small Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 364–376. [[CrossRef](#)] [[PubMed](#)]
27. Yuan, S.; Qin, H.; Yan, X.; Akhtar, N.; Mian, A. SCTransNet: Spatial-Channel Cross Transformer Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5002615. [[CrossRef](#)]
28. Yang, H.; Mu, T.; Dong, Z.; Zhang, Z.; Wang, B.; Ke, W.; Yang, Q.; He, Z. PBT: Progressive Background-Aware Transformer for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5004513. [[CrossRef](#)]
29. Zhang, M.; Bai, H.; Zhang, J.; Zhang, R.; Wang, C.; Guo, J.; Gao, X. RKformer: Runge-Kutta Transformer with Random-Connection Attention for Infrared Small Target Detection. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 1730–1738.
30. Hu, C.; Huang, Y.; Li, K.; Zhang, L.; Long, C.; Zhu, Y.; Pu, T.; Peng, Z. DATransNet: Dynamic Attention Transformer Network for Infrared Small Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 7001005. [[CrossRef](#)]
31. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
32. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. [[CrossRef](#)]
33. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
34. Lin, F.; Bao, K.; Li, Y.; Zeng, D.; Ge, S. Learning Contrast-Enhanced Shape-Biased Representations for Infrared Small Target Detection. *IEEE Trans. Image Process.* **2024**, *33*, 3047–3058. [[CrossRef](#)]
35. Sun, H.; Bai, J.; Yang, F.; Bai, X. Receptive-Field and Direction Induced Attention Network for Infrared Dim Small Target Detection With a Large-Scale Dataset IRDST. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5000513. [[CrossRef](#)]
36. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
38. Wu, T.; Li, B.; Luo, Y.; Wang, Y.; Xiao, C.; Liu, T.; Yang, J.; An, W.; Guo, Y. MTU-Net: Multilevel TransUNet for Space-Based Infrared Tiny Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5601015. [[CrossRef](#)]
39. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 7000805. [[CrossRef](#)]
40. Yu, Z.; Zhao, C.; Wang, Z.; Qin, Y.; Su, Z.; Li, X.; Zhou, F.; Zhao, G. Searching Central Difference Convolutional Networks for Face Anti-Spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5295–5305.
41. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [[CrossRef](#)]
42. Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; Fu, Y. Infrared Small Target Detection with Scale and Location Sensitivity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 17490–17499.
43. Yue, T.; Lu, X.; Cai, J.; Chen, Y.; Chu, S. SDS-Net: Shallow–Deep Synergism-Detection Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 3001113. [[CrossRef](#)]
44. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [[CrossRef](#)]
45. Li, Q.; Zhang, W.; Lu, W.; Wang, Q. Multibranch Mutual-Guiding Learning for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5605710. [[CrossRef](#)]

46. Xu, M.; Yu, C.; Li, Z.; Tang, H.; Hu, Y.; Nie, L. HDNet: A Hybrid Domain Network With Multiscale High-Frequency Information Enhancement for Infrared Small-Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5004115. [[CrossRef](#)]
47. Wu, F.; Liu, A.; Zhang, T.; Zhang, L.; Luo, J.; Peng, Z. Saliency at the Helm: Steering Infrared Small Target Detection with Learnable Kernels. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5000514. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.