MDIGCNet: Multidirectional Information-Guided Contextual Network for Infrared Small Target Detection

Luping Zhang, *Student Member, IEEE*, Junhai Luo, *Member, IEEE*, Yian Huang, *Student Member, IEEE*, Fengyi Wu, *Graduate Student Member, IEEE*, Xingye Cui, and Zhenming Peng, *Member, IEEE*

***Abstract*—Infrared small target detection (ISTD) technology has extensive applications in the military field. Due to the quality of imaging equipment and environmental interference, infrared small target images lack texture and structural information. Deep learning-based algorithms have achieved superior accuracy in this field compared to traditional algorithms; however, these methods are often not designed with domain knowledge integration. In this article, we propose a multidirectional information-guided contextual network (MDIGCNet) for ISTD. The primary structure of this network adopts the U-Net architecture. To address the issue of lacking texture and structural information in the target images, we employ an integrated differential convolution (IDConv) module to extract richer image features during both the encoding and decoding stages. Skip connections in the network utilize a multidirectional gradient information extraction block (MGIEB) to obtain gradient features of infrared small targets. Our domain-inspired multidirectional Gaussian differential convolution (MGDC) module is employed to extract features of Gaussian-distributed small targets, enhancing the distinction between targets and backgrounds. Additionally, we designed a local-global feature fusion (LGFF) module incorporating an attention mechanism to merge shallow and deep features, thereby improving the efficiency of feature utilization within the model. Furthermore, since both IDConv and MGDC are parallel multiconvolutional kernel structures, reparameterization techniques are used to avoid excessive parameters and computational load. Experimental results on public datasets NUDT-SIRST, IRSTD-1k, and SIRST-Aug demonstrate that our algorithm outperforms other state-of-the-art methods in detection performance.**

***Index Terms*—Difference convolution, infrared small target detection (ISTD), multidirectional gradient information extraction, reparameterization.**

I. Introduction

INFARED small target detection (ISTD) research is crucial in civilian and military applications. The stealth and convenience of this technology enable ISTD to operate both day and night, and it is widely used in areas such as reconnaissance, maritime surveillance, and precision guidance [1].

Unlike conventional target detection, ISTD has the following characteristics. 1) Dim: Infrared small targets typically have low image resolution and a high level of clutter, resulting in low contrast and a low signal-to-noise ratio. 2) Small: Due to the long imaging distance, the size of infrared small targets is generally between 2×2 and 9×9 pixels [2]. 3) Varying shape: The shape and size of the targets vary due to different target types and imaging scenes. These characteristics pose significant challenges for ISTD in complex scenarios. Particularly, under interference from clouds and buildings, it is often difficult to distinguish targets from background clutter. Therefore, detecting small targets in infrared images remains a topic worthy of research.

Current ISTD algorithms can be broadly categorized into model-driven methods and deep learning-based methods. Model-driven methods include those based on background consistency assumptions, optimization techniques, and human visual salience (HVS) assumptions. However, these methods have limitations. Methods based on background consistency assumptions [3], [4] can only detect targets in uniform backgrounds and are not suitable for complex detection scenarios. Optimization-based methods [5], [6], [7] can detect targets in low-contrast situations and perform well among model-driven methods, but they are prone to false detections in cluttered scenes and often lack real-time performance, limiting practical applications. HVS-based methods are mostly designed based on local contrast [8], [9], [10], but the shallow structure of manually designed feature extractors cannot adapt to complex detection scenes, resulting in a high false alarm rate. These traditional methods rely on domain knowledge for modeling, making them sensitive to parameter changes and limiting their generalization across different scenarios. In recent years, deep learning has achieved significant success in image processing, and some researchers [11], [12], [13] have started to introduce deep learning into ISTD. For instance, some algorithms use generative adversarial networks (GANs) to balance detection and false alarm rates [12], and others design networks by integrating multiple feature fusion strategies [13].

However, two major issues remain in the methods based on deep learning. 1) There is a lack of effective utilization of existing information in images. Due to the small size of

Received 14 August 2024; revised 21 October 2024; accepted 15 November 2024. Date of publication 28 November 2024; date of current version 18 December 2024. This work was supported by the Natural Science Foundation of Sichuan Province of China under Grant 2022NSFSC40574 and Grant 2023NSFSC0508. *(Corresponding author: Junhai Luo.)*

The authors are with the School of Information and Communication Engineering, Chengdu 611731, China, and also with the Laboratory of Imaging Detection and Intelligent Perception, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: anguing@foxmail.com; junhai_luo@uestc.edu.cn; huangyian@std.uestc.edu.cn; wufengyi98@163.com; cxy011211@163.com; zmpeng@uestc.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3508255

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

infrared small targets, deep network designs often lose target response information. Most existing studies employ shallow network designs [14], [15], yet extracting multilevel, multiscale information from infrared small targets is challenging due to the lack of texture and structural information. Some researchers have designed networks based on dense nesting strategies [16], [17], [18], achieving higher detection accuracy. Additionally, transformer-based methods [19], [20] search for useful information by exploring long-distance dependencies in images. However, these methods are significantly more computationally expensive than others, hindering their practical and widespread application [21]. Therefore, it is necessary to maximize the use of information within the network while minimizing computational overhead. 2) The effectiveness of ordinary convolutions is relatively low. HVS-based methods [8], [22], [23] have demonstrated that well-designed priors aid in the detection of infrared small targets. Most current deep learning-based methods [20], [24], [25], [26] use standard convolutions for feature extraction without leveraging prior information. Without constraints, standard convolutions have a vast solution space, potentially limiting their expressive power. Some approaches combine HVS methods with deep learning designs [27], [28], [29], [30], but these methods typically involve simple concatenations and leave room for optimization. Thus, an ideal solution would involve designing convolutional kernels using prior information to enhance the network's ability to learn from infrared small targets.

To address the aforementioned issues, a multidirectional gradient information extraction block (MGIEB) has been designed, comprising integrated differential convolution (IDConv) and multidirectional Gaussian differential convolution (MGDC). The IDConv consists of four 3×3 differential convolutions and one 3×3 standard convolution, deployed in parallel for feature extraction. Specifically, central difference convolution (CDC) [31], angle-based difference convolution (ADC) [32], horizontal difference convolution (HDC), and vertical difference convolution (VDC) are employed to enhance the extraction of information from infrared small targets. By utilizing differential convolutions, a computational strategy between specific pixel pairs is devised, integrating traditional local descriptors into the convolutional neural network (CNN). The MGDC includes four 5×5 differential convolutions deployed in parallel to extract gradient features in four directions. Designed with prior knowledge, the MGDC enhances the model's ability to distinguish small targets from clutter. The MGIEB is incorporated into the skip connections of the network to extract multilevel, multiscale information from infrared small targets.

Furthermore, the IDConv is implemented in the encoder and decoder of U-Net [33] as the backbone network. Specifically, the standard 3×3 convolutions in U-Net are replaced with IDConv to further enhance the network's ability to extract detailed information.

To avoid excessive computational overhead, the IDConv and MGDC are reparameterized to reduce the number of parameters. Since both are parallel convolution structures, multiple parallel convolutions can be simplified into a single standard convolution. Thus, the IDConv and MGDC can improve model performance while maintaining the same number of parameters and computational load as standard convolutions.

Finally, the local-global feature fusion (LGFF) module merges shallow detail features and deep semantic features, which is beneficial for highlighting and preserving the features of small targets.

The proposed algorithm is named MDIGCNet: multidirectional information-guided contextual network (MDIGCNet) for ISTD. Subsequent experiments have demonstrated that this algorithm achieves superior detection results on multiple datasets compared to other state-of-the-art (SOTA) algorithms.

In summary, the main contributions of our work are as follows.

1) To address the lack of texture and detail information in infrared small target images, IDConv is designed to extract rich details. This module is integrated into both the encoder and decoder of the network, effectively enhancing performance in pixel-level tasks.
2) Incorporating domain knowledge, we designed the MGDC module, which integrates Gaussian gradient operators into the CNN to extract multidirectional gradient information of small targets. By utilizing the MGIEB, a module that combines IDConv and MGDC, the network's ability to extract multilevel and multiscale information from infrared small targets is enhanced. Additionally, to avoid excessive computational overhead, multibranch parallel convolution kernels are reparameterized after training.
3) To improve the efficiency of feature utilization in the model, we designed the LGFF module. Leveraging an attention mechanism, this module fully utilizes both global and local information, integrating deep and shallow features. This approach enhances pixel-level accuracy while ensuring effective target detection.

The rest of this article is organized as follows. Section II reviews related work in recent years; Section III outlines the structure design of MDIGCNet; Section IV analyzes the performance of the proposed network structure through ablation and comparative experiments; and finally, Section V concludes this article.

II. Related Work

A. Infrared Small Target Detection

1) Model-Driven Methods: Model-driven approaches encompass methods based on background consistency assumptions, optimization, and human visual saliency hypotheses. Methods based on background consistency assume that the background of a small target image is usually similar, while the small target disrupts this correlation. These methods [3], [4], [34] typically estimated background information, subtracted the background from the original image, and then applied adaptive filtering to obtain the target image. However, in practical applications, the backgrounds of small target images are often complex, making these detection methods susceptible to noise and clutter, resulting in a higher false alarm rate. Optimization-based methods model an infrared image as a linear combination of background, small target, and random noise images.

Due to the consistency of the background, the data structure of the background image is low-rank. In contrast, the small target image is sparse, occupying only a few pixels of the entire image. Low-rank sparse decomposition can separate the background and target images from the original image. These methods include matrix-based approaches [5], [35], [36] and tensor-based approaches [6], [7], [37], [38], [39]. While these optimization-based methods achieve high detection accuracy, their real-time performance is often poor, limiting their practical applications. Methods based on human visual saliency assumption hypothesize that regions containing small targets in infrared images are visually salient. Algorithms can be designed to compute saliency maps of the image, followed by adaptive threshold segmentation to detect small targets. The local contrast method (LCM) [8] was first proposed to describe the difference between a location and its neighborhood. Many researchers have proposed improved algorithms to define and calculate the contrast between the target and its surrounding background. For instance, Han et al. [40] introduced the relative local contrast method (RLCM) to handle variations in small target size, Qin et al. [41] developed a novel local contrast method (NLCM) using Gaussian bandpass filters, and Wei et al. [9] proposed a multiscale patch-based contrast measure (MPCM) to address scale variations of small targets. Further improvements include a weighted enhanced local contrast measure [42], a multiscale local contrast measure with three-layer windows [43], and a double neighborhood gradient measure [10], all aimed at enhancing computational performance. However, the operators designed by these methods extract shallow features, leaving room for significant improvement in detection performance.

2) Deep Learning Methods: These methods involve designing various networks to generate saliency maps, extract features from infrared small target images, and then perform adaptive segmentation to achieve accurate detection results. Some studies [12], [44], [45] have employed generative adversarial strategies for ISTD. Other approaches [14], [24], [25], [46] used asymmetric context modulation modules to fuse shallow and deep features. To address the challenge of varying target sizes, several networks [24], [46] employ multiscale feature fusion, utilizing dilated convolutions and adaptive pooling to construct multiscale feature maps. To compensate for the lack of information in infrared images, some methods [47], [48], [49] have attempted to enhance the feature extraction and fusion capabilities of the network. For example, some approaches use dual-modality feature fusion [47] and dual-domain feature extraction [48] to improve detection performance. To fully leverage the contextual information of small targets, methods based on dense nested strategies [17], [18], [50] and those incorporating self-attention mechanisms [19], [20], [51] have been proposed. However, most of these approaches utilize common network design strategies without incorporating domain knowledge into the network design.

To address these issues, some methods [27], [28], [52], [53] have attempted to integrate model-driven designs into deep learning networks. For example, one study [27] modularized the traditional MPCM algorithm into skip connections within the network, replaced the concept of blocks with dilation rates, and used a cyclic shift strategy to accelerate computation. Hou et al. [29] used handcrafted fixed-weight convolution kernels for feature extraction at the network front end, where each kernel's output represents the difference between the mean values inside and outside the kernel. Some researchers have integrated target shape reconstruction for ISTD, collecting and enhancing comprehensive edge information at different levels to improve target-background contrast [54]. Sun et al. [55] introduced a receptive field and direction-induced attention network (RDIAN), which uses convolutional layers with various receptive fields to capture target features in different local regions and a multidirectional guided attention mechanism to enhance target features in low-level feature maps. However, these methods only use some traditional techniques to assist CNNs in detection and do not effectively network HVS-based operators within the CNN. Thus, there is still room for optimization in these approaches.

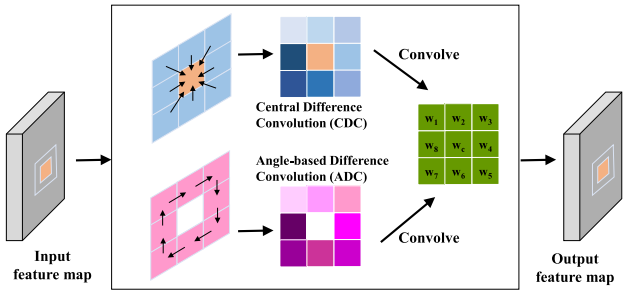


Fig. 1. Structure of CDC and ADC.

B. Difference Convolution

Modern CNNs employ multilayer image features with rich hierarchical information, guided by deep supervision for end-to-end detection, yielding excellent results. However, the optimization of CNN convolution kernels starts from random initialization and lacks explicit encoding, making it difficult for the network to focus on specific representations. Traditional image processing convolution kernels incorporate expert knowledge, enabling rapid image processing, but their structure is shallow, and their performance is subpar.

To address these issues, Yu et al. [31] proposed the central difference convolution (CDC) operator, which encodes 3×3 convolution kernels by computing the difference between the central pixel and its neighboring pixels, thus extracting more detailed image information. The structure is shown in Fig. 1. Building on this, Yu et al. [56] decoupled CDC into two symmetric suboperators—horizontal–vertical and diagonal-introducing the cross-central difference convolution (Cross-CDC). Su et al. [32] proposed pixel difference convolution (PDC), which employs three-pixel difference methods: central difference with neighboring features, pairwise differences in a clockwise direction, and differences between the outer and inner rings of a 5×5 neighborhood. Furthermore, Yu et al. [57] introduced spatiotemporal difference convolution (3D-CDC) to efficiently extract spatiotemporal difference features, directly replacing Vanilla 3-D convolution in any 3-D CNN without additional parameter overhead.

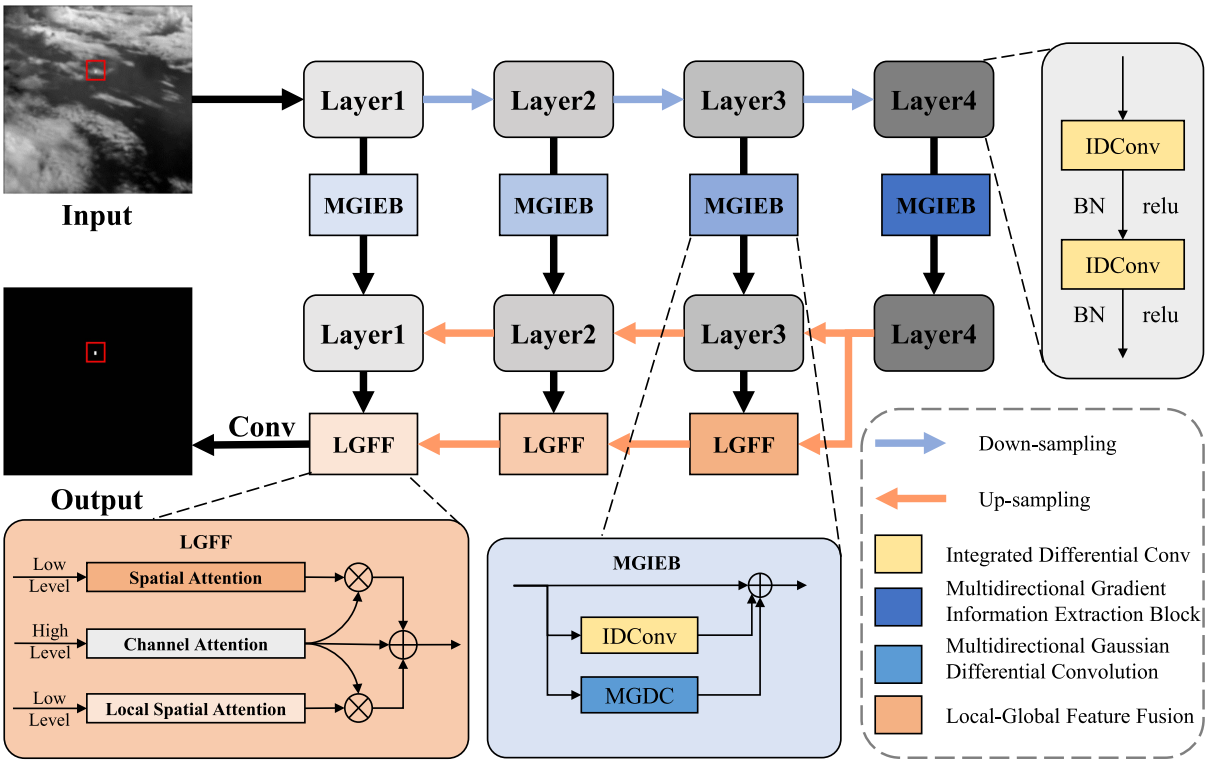


Fig. 2. Overview of the proposed MDIGCNet, which has a U-Net structure with MGIEB and LGFF module.

In the field of ISTD, Wu et al. [21] proposed a simple yet efficient ISTD network (RepISD-Net), merging the multibranch topology incorporating CDC into a single branch composed solely of cascaded 3 × 3 convolutions for rapid inference. Ying et al. [58] introduced a deep network driven by local motion and contrast priors for infrared small target super-resolution, integrating a central difference residual group that embeds CDC into the feature extraction backbone, achieving gradient-aware feature extraction centered on the target and further enhancing target contrast. However, these methods were not designed with the characteristics of infrared small targets in mind. In particular, simply using CDC may result in performance degradation. Designing appropriate differential convolutions based on the principles of HVS is a feasible approach.

C. Reparameter

In CNNs, various methods have been employed to address the issue of target scale variation by using multibranch structures to alter the network's receptive field and enhance performance [59], [60]. However, the multibranch convolutional structures significantly increase the computational and parameter overhead. Consequently, some studies [61], [62] have proposed reparameterization techniques to reduce computational costs. Specifically, networks with multibranch structures are trained first, and then merged into a single branch with a standard 3 × 3 convolution before inference. For instance, Ding et al. [62] transformed 3 × 3 convolutions, 1 × 1 convolutions, and residual structures into a single 3 × 3 convolution kernel, achieving improved results. Similarly, Chen et al. [63] converted four 3 × 3 differential convolutions and one standard convolution into a single standard convolution, enhancing performance and generalization ability.

In the field of ISTD, Wu et al. [21] introduced a simple yet efficient ISTD network (RepISD-Net), which merges multibranch topologies containing CDC into a single branch composed solely of cascaded 3 × 3 convolutions. Peng et al. [64] proposed a dynamic reparameterization network (DRPN), which employs multiple branches with different convolution kernel sizes and a dynamic convolution strategy. After training, the multibranch structures are further transformed into a single branch using reparameterization techniques.

III. Methodology

A. Overall Architecture

The MDIGCNet consists of three primary components: the backbone network, the MGIEB, and the LGFF module. The network's structure is depicted in Fig. 2.

Our backbone utilizes a U-Net [33] architecture. It employs skip connections between the downsampling and upsampling paths to merge deep and shallow features, facilitating the extraction of weak and small targets. The 3 × 3 convolution layers in the encoder and decoder are replaced with the designed IDConv module. This replacement enables the extraction of richer image features. Through reparameterization techniques, it transforms multiple convolution kernels into a single standard convolution, enhancing model performance without increasing parameters or computational complexity.

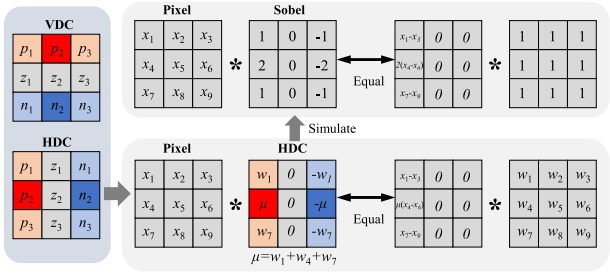


Fig. 3. Derivation process of the VDC and HDC modules.

Our MGIEB is integrated into skip connections, combining IDConv and MGDC. The MGDC, based on the assumption that the grayscale distribution of infrared small targets follows a Gaussian distribution, calculates Gaussian gradient information in four directions. It then reparameterizes this information into standard convolutions, enhancing the extraction of prior information with minimal additional parameters and computational overhead.

Finally, the LGFF module merges shallow and deep features. This module fully leverages global and local image information by utilizing attention mechanisms, allowing the network to focus on target regions while preserving detailed information. This approach enhances the model's feature utilization, improving both accuracy and efficiency in target detection.

B. IDConv Module

General convolutional layers explore a vast solution space during training, often resulting in disappointing outcomes for specific tasks due to insufficient constraints [56], [57]. In ISTD, gradient information is crucial in distinguishing small target regions. For instance, some ISTD networks utilize image edge information to assist in target detection [29], [54]. However, these networks employ fixed-weight convolutional kernels to aid the main network in learning detection without internalizing gradient convolution operators into the CNN.

Inspired by existing works [27], [58], [61], we introduce the IDConv module, which integrates a general convolution and four difference convolutions deployed in parallel to extract finer target features. Apart from the common CDC and angle-based difference convolution (ADC), which is shown in Fig. 1, we employ two difference convolutions in different directions to mimic the Sobel operator. The derivation process is illustrated in Fig. 3, where x_i represents the pixels in the current patch, p_i, z_i, n_i denote the weights of a 3×3 convolutional kernel, where $p_2 = \sum_{i=1}^{3} p_i, z_i = 0, n_i = -p_i$.

During the training phase, five convolutional layers are used for model training, encoding features such as horizontal and vertical gradients into the convolutional layers to enhance the extraction of image details. After completing the training, the learned convolutional kernel weights are rearranged and reparameterized into a single convolutional kernel, as formulated as follows:

$$F_{\text{out}} = \sum_{j=1}^{5} F_{\text{in}} * U^j = F_{\text{in}} * \sum_{j=1}^{5} U^j = F_{\text{in}} * U_{\text{sum}} \quad (1)$$

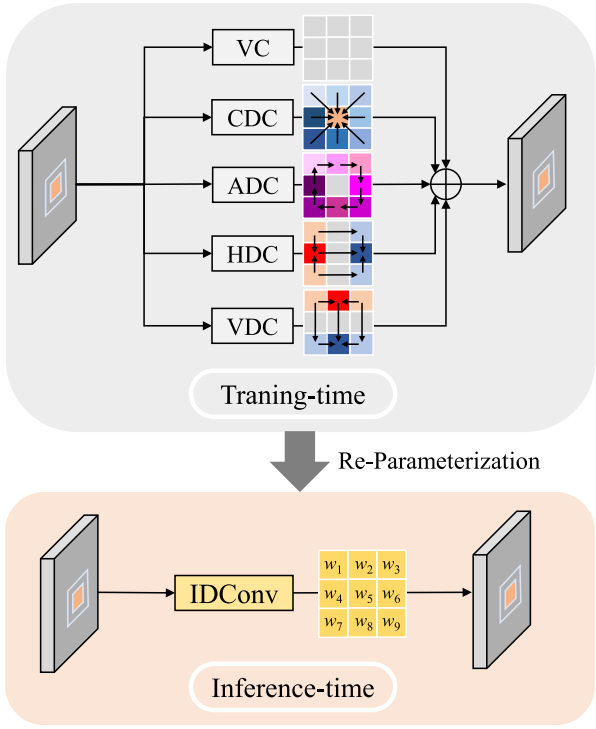


Fig. 4. Architecture diagram of the IDConv module.

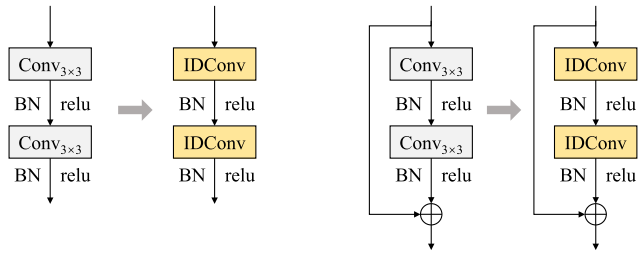


Fig. 5. Architecture diagram of the ablation experiment of the backbone network.

U_{sum} represents the convolution kernel of IDConv, whose weight is

$$w_i = \sum_{j=1}^{5} u_i^j. \quad (2)$$

The weights of each convolutional kernel before reparameterization are denoted as $u_i^j, i = 1 \sim 9$. This design offers the advantage of enhancing the model's fitting capacity without introducing additional computational overhead. During inference, the weights of the transformed convolutional kernels remain fixed and no longer undergo updates. The parameter count and computational load are equivalent to a standard 3×3 convolutional kernel yet capable of extracting richer image information, as illustrated in Fig. 4.

The 3×3 convolution kernel in the U-Net network is replaced with the designed IDConv, as shown in Fig. 5. The residual structure and IDConv module will be analyzed in the experimental part using ablation experiments.

C. Multidirectional Gradient Information Extraction Block

In this section, we delineate the design rationale and specifics of the MGIEB, aimed at better extracting fine details of small targets. Due to the distant imaging range of infrared and the attenuation during atmospheric transmission, infrared small target imaging resembles a Gaussian blob [65]. Its mathematical model can be approximated as a 2-D Gaussian function

$$f(x,y) = A * \exp\left(-\frac{1}{2}\left(\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2\right)\right) \tag{3}$$

where A denotes the peak intensity of the target, (x, y) represents the pixel of a small target, σ_x and σ_y represent the horizontal and vertical spread functions, respectively. Several algorithms [8], [9], [66] model infrared small targets based on human visual saliency. However, these methods are sensitive to salient edges and high-brightness regions, making it difficult to distinguish targets from texture clutter. Therefore, some researchers [22], [23] employ the facet kernel model [67] to approximate the 2-D Gaussian function for polynomial approximation. By calculating the gradient information in four directions, the fixed convolution kernel weights of the convolution template are determined, thereby computing multidirectional gradient information that conforms to the Gaussian distribution. However, these manually designed detection operators have shallow structures, which may lead to performance degradation when dealing with target scale variations and complex backgrounds.

To address the limitations of these algorithms, the concept of differential convolution is introduced to incorporate multidirectional Gaussian filtering templates into a deep learning network, leading to the design of the MGDC module to enhance detection performance. Unlike traditional methods with fixed convolution kernel weights, this convolution module is designed as a paradigm capable of backpropagation within CNNs, allowing it to be integrated into the designed network for joint training. Specifically, the weights of the 5×5 convolution kernel are constrained within the differential convolution paradigm to simulate the computation of Gaussian gradient information in different directions. Inspired by CDC [31] and PDC [32], four convolutional kernels are designed, each primarily distinguished by the encoded directions. To achieve the effect of differencing the pixel values in the red region from those in the blue region, we derived the expression based on the convolutional kernel in Fig. 6 as follows.

For general convolution, its expression can be written as

$$y = f(x) = \sum_{i=1}^{k\times k} u_i \cdot x_i. \tag{4}$$

For the convolutional kernel in Fig. 6, the expression is

$$\begin{aligned} y &= u_1(x_1 - x_{25}) + \cdots + \lambda \cdot u_7(x_7 - x_{19}) + \cdots \\ &\quad + \theta \cdot u_{19}(x_{19} - x_7) + \cdots + u_{25}(x_{25} - x_1) \\ &= (u_1 - u_{25})x_1 + \cdots + (\lambda \cdot u_7 - \theta \cdot u_{19})x_7 + \cdots \\ &\quad + (\theta \cdot u_{19} - \lambda \cdot u_7)x_{19} + \cdots + (u_{25} - u_1)x_{25} \end{aligned}$$

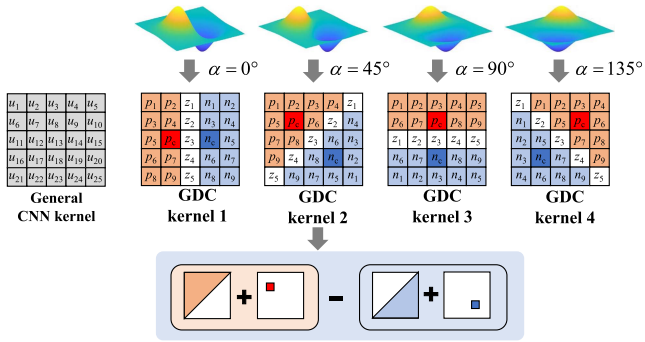


Fig. 6. Derivation process of the MGDC module.

$$= \sum \hat{u}_i x_i = \hat{u} * x \tag{5}$$

where x_i represents the pixels in the current patch and u_i denotes the weights in the convolutional kernel. λ and θ are the gain weights on both sides of the convolution kernel, used to enhance the contrast of small targets. Assuming parameter $\lambda = \theta$, $\lambda \cdot u_7 - \theta \cdot u_{19}$ can be simplified to $\lambda(u_7 - u_{19})$. The parameter λ is referred to as the gain factor, with its determination inspired by the Sobel and LoG operators, which weight information from the neighborhood. It is simultaneously observed that $\hat{u}_i = -\hat{u}_{25-i}$ holds within this expression. To better describe and simplify the equation, $\hat{u}_i$ is replaced with p_i, n_i, and z_i, as depicted in Fig. 6.

Thus, the weights of this convolutional kernel can be described by the following:

$$\begin{aligned} p_c &= \sum_{i=1}^{8} p_i \\ z_i &= 0 \\ n_i &= -p_i \end{aligned} \tag{6}$$

where p_c represents the central position on one side of the GDC kernel segmentation line, aligning with significantly higher central significance in the Gaussian distribution than the neighborhood, effectively aggregating local information. The values of p_i and n_i are opposite to each other, and through careful design, rich gradient information can be effectively extracted by encoding four convolutional kernels in four directions.

To avoid the significant computational overhead introduced by four convolution kernels, the weights of the four kernels are rearranged and reparameterized into a single convolution kernel after training, as derived in the following:

$$F_{\text{out}} = \sum_{j=1}^{4} F_{\text{in}} * U^j = F_{\text{in}} * \sum_{j=1}^{4} U^j = F_{\text{in}} * U_{\text{sum}} \tag{7}$$

where U_{sum} represents the reparameterized convolution kernel, whose weight is shown in (8). The weights of each convolutional kernel before reparameterization are denoted as $u_i^j, i = 1 \sim 4$

$$w_i = \sum_{j=1}^{4} u_i^j. \tag{8}$$

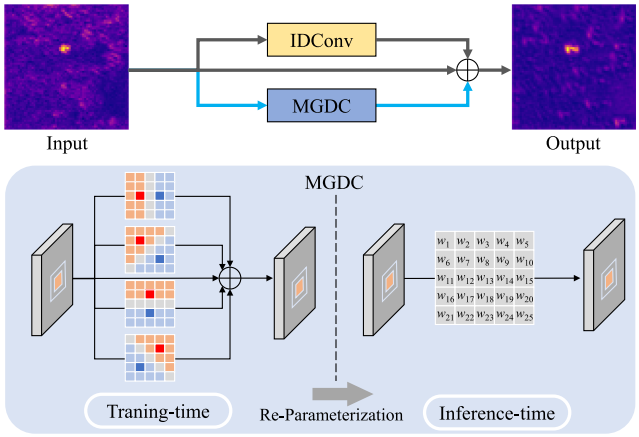


Fig. 7. Architecture diagram of the MGIEB.

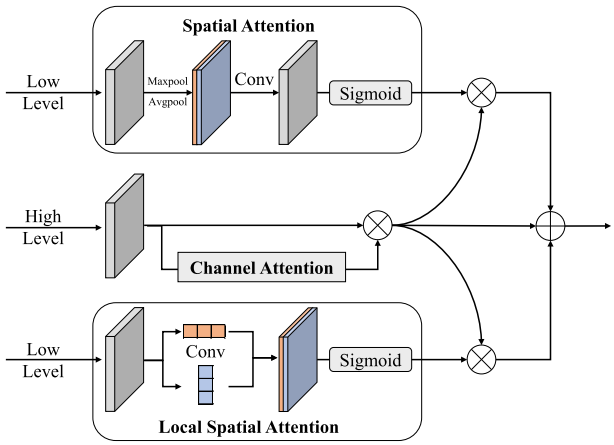


Fig. 8. Architecture diagram of the LGFF module.

Due to the reparameterization technique employed during the inference process, the parameters and computational cost of this module are comparable to that of a single convolution kernel, as illustrated in Fig. 7. The MGDC, IDConv, and residual results are concatenated to form the MGIEB, as shown in (9). Since MGDC is only used in the skip connections of the network, employing a 5×5 convolution kernel does not introduce significant computational overhead. Conversely, the aforementioned IDConv is used throughout the entire encoding and decoding stages of the network, making the 3×3 kernel a more optimal choice

$$\mathrm{MGIEB(X)} = X + \mathrm{IDConv}(X) + \mathrm{MGDC}(X). \quad (9)$$

D. LGFF Module

The design of the LGFF module is described in this section. In CNNs, shallow features often contain rich detail information but lack higher-level semantic information, in contrast to deep features. Detailed information is usually lost during downsampling in small infrared target detection networks due to the small size of the targets. To address this, an LGFF module is proposed to fuse deep and shallow features for fine-grained detection of small targets, as illustrated in Fig. 8.

Spatial attention mechanism (SAM) [68] is utilized to compute global spatial information on shallow features, providing pixel-level global attention guidance on the spatial level. Attention mechanisms are employed to calculate attention information in the row and column directions, refining detail features in space. For deep features, CAM is used to provide semantic information and channel guidance for low-level features, which are then fused with shallow features. This process is illustrated in the formula below:

$$\begin{aligned}
&\mathrm{SA}(X) = \sigma(\mathrm{f}^{7\times7}([\mathrm{AvgPool}(X) : \mathrm{MaxPool}(X)])) \\
&\mathrm{CA}(Y) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(Y)) + \mathrm{MLP}(\mathrm{MaxPool}(Y)))) \\
&\mathrm{LSA}(X) = \sigma(\mathrm{f}^{3\times3}([\mathrm{f}^{1\times3}(X) : \mathrm{f}^{3\times1}(X)])) \\
&\mathrm{LGFF}(X, Y) = \mathrm{CA}(Y) \otimes Y \otimes (1 + \mathrm{SA}(X) + \mathrm{LSA}(X))
\end{aligned} \quad (10)$$

where $\sigma(\cdot)$ represents the sigmoid function operation, $f^{3\times3}(\cdot)$ denotes the 3×3 convolution operation, $\otimes$ indicates element-wise multiplication, and X, Y represent the shallow and deep features. SA stands for spatial attention, CA stands for channel attention, and LSA stands for local spatial attention.

When computing $\mathrm{LSA}(X)$, we utilize 1×3 and 3×1 convolutions to aggregate the details of shallow features in the row and column directions. We then used a 3×3 convolution to aggregate the two features, fully utilizing the available information to extract local shape information from the low-level features. Simultaneously, for calculating $\mathrm{SA}(X)$, MaxPool and AvgPool are employed to obtain global spatial information. Finally, these are modulated with deep features, enabling the integration of rich detail features into high-level semantic information, thus enhancing the network's pixel-level accuracy.

E. Loss Function

Due to the small size of ISTD, which occupies very few pixels in the image, a severe imbalance exists in the samples. Therefore, we adopt the softIoU loss [50], expressed as follows:

$$\mathrm{Loss}_{\mathrm{softIoU}}(p, y) = \frac{\sum_{i,j}(\sigma(p_{i,j}) \cdot y_{i,j}) + \sigma}{\sum_{i,j}(\sigma(p_{i,j}) + y_{i,j} - \sigma(p_{i,j}) \cdot y_{i,j}) + \sigma} \quad (11)$$

where $p_{i,j}$ and $y_{i,j}$ represent the predicted value and the ground truth mask value at point (i, j), respectively. $\sigma(\cdot)$ denotes the sigmoid activation function operator, and the smoothing factor c is set to 1.

IV. Results

A. Evaluation Metrics

We utilize probability of detection (P_d) and false alarm rate (F_a) to assess the algorithm's capacity in target detection and its ability to preserve target shape, employing pixel-level metrics such as intersection over union (IoU) and F-measure (F_1). These metrics are computed using a fixed threshold of 0.5. The receiver operating characteristic (ROC) curve evaluates the algorithm's detection performance across various thresholds.

P_d reflects the ability to correctly detect targets, defined as the ratio of correctly detected targets to the total number of targets detected. Its formula is expressed as

$$P_d = \frac{T_{\text{correct}}}{T_{\text{act}}} \tag{12}$$

F_a reflects the accuracy of target detection, calculated as the ratio of the sum of incorrectly predicted pixels to the total number of pixels in the entire image. Its definition formula is as follows:

$$F_a = \frac{P_{\text{false}}}{P_{\text{all}}} \tag{13}$$

F_1 is a classic semantic segmentation metric that balances Precision and Recall, where Precision represents the ratio of correctly classified pixels to all labeled target and predicted target pixels. Recall represents the ratio of correctly classified pixels to all labeled target pixels. Its expression is as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ Fmeasure &= \frac{2 * Precsion * Recall}{Precision + Recall} \end{aligned} \tag{14}$$

IoU reflects the similarity of predicted targets to actual targets, calculated as the ratio of their intersection to their union, expressed as

$$\text{IoU} = \frac{\#\text{Area of Overlap}}{\#\text{Area of Union}}. \tag{15}$$

The ROC curve represents the classifier's classification performance at different thresholds, characterizing the dynamic relationship between false positive rate (FPR) and true positive rate (TPR). The expression is shown in (16). The closer the curve approaches the coordinate (0,1), the better the algorithm's performance

$$FPR = \frac{\sum FP}{\sum FP + TN}, TPR = \frac{\sum TP}{\sum TP + FN}. \tag{16}$$

B. Implementation Details

Ablation experiments are conducted on three publicly available mainstream datasets (NUDT-SIRST [17], IRSTD-1K [54], and SIRST-Aug [24]) for the proposed network, and comparisons are made with other algorithms. The NUDT-SIRST dataset comprises 1327 images, with 663 images for training and 664 for testing. This dataset covers urban, rural, highlight, ocean, and cloud scenes, with images sized at 256 × 256 pixels. The IRSTD-1K dataset consists of 1001 images, with 800 for training and 201 for testing. It includes targets of various sizes and shapes with rich, cluttered backgrounds, and the image size is 512 × 512 pixels. The SIRST-Aug dataset contains 9070 images, with 8525 for training and 545 for testing. This dataset augments the SIRST dataset, with images sized at 256 × 256 pixels.

During network training, the batch size is set to 8, the initial learning rate is 0.0005, and the number of epochs is set to 400. The network is optimized using the Adam optimizer [69] and the multistepLR scheduler. The software platform for this model is based on Python 3.8 and PyTorch 1.12.1. All models in the experimental results are implemented on a machine equipped with an Intel(R) Xeon(R) Gold 6133 central processing unit (CPU) and an Nvidia GeForce 4090 graphics processing unit (GPU).

C. Comparison to SOTA Methods

We selected various algorithms for comparison, including MPCM [9], IPI [5], PSTNN [37], ALCNet [27], AGPCNet [24], RDIAN [55], ISTDU [70], DNANet [17] and UIUNet [16], MSHNet [71].

1) Quantitative Comparison: The quantitative comparison of these algorithms is presented in Table I. Overall, deep learning-based algorithms exhibit superior detection performance compared to traditional algorithms. The MPCM algorithm is a classical algorithm for ISTD. It shows satisfactory performance on the target-level metric P_d, even approaching the detection rate P_d of deep learning algorithms on the SIRST-Aug dataset. However, it performs the worst on the pixel-level metrics IoU and F_1, indicating insufficient detail preservation. The data structure-based IPI and PSTNN algorithms outperform MPCM, showing significant improvements in the pixel-level metrics IoU and F_1. Nevertheless, the performance of these algorithms is related to the sparsity of the targets. Their detection performance may decline when multiple targets are present in a single image, as evidenced by the lower P_d values on the SIRST-Aug dataset. ALCNet and RDIAN demonstrate significant improvements over traditional methods, but their performance is generally average among deep learning algorithms. AGPCNet performs well on SIRST-Aug but is average on the first two datasets. DNANet and UIUNet show overall good performance but have the most significant model parameter count and computational cost among all methods.

The proposed MDIGCNet achieves competitive performance with relatively low computational overhead. Compared to the parameter-efficient ALCNet and RDIAN, our MDIGCNet performs better across all metrics on the three datasets. On NUDT-SIRST, our method achieves an IoU that is 7.87% higher than ALCNet. Compared to ISTDU-Net, which has slightly higher parameters and computational costs, our method improves IoU, P_d, and F_a by approximately 2%. Compared to DNANet, AGPCNet, and UIUNet, which have higher parameters and computational costs, our MDIGCNet achieves comparable or even better detection performance. Our method has three times fewer parameters and two times less computational cost than DNANet, and an order of magnitude lower than AGPCNet and UIUNet. This indicates that our method balances performance and computational overhead, offering greater practicality.

We plotted the ROC curves of different methods on NUDT-SIRST, IRSTD-1k, and SIRST-Aug. As shown in Fig. 9, our MDIGCNet achieves competitive performance across all datasets. It performs excellently on NUDT-SIRST and is comparable to UIUNet and DNANet on IRSTD-1k and SIRST-Aug. Traditional methods (MPCM, IPI, PSTNN) exhibit

TABLE I
COMPARISON OF DETECTION PERFORMANCE [IoU (%),F_1(%), P_d (%), AND F_a ($\times 10^{-5}$)] AND MODEL EFFICIENCY (THE NUMBER OF PARAMETERS (M) AND THEORETICAL FLOPS (G)) OF DIFFERENT METHODS ON THE NUDT-SIRST, IRSTD-1K, AND SIRST-AUG DATASETS

Methods	Params(M)	FLOPs(G)	Time(s)	NUDT-SIRST				IRSTD-1K				SIRST-Aug			
				IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a
MPCM	-	-	-	8.18	15.12	66.14	444.8	13.52	23.83	68.69	50.16	21.25	35.05	90.78	35.05
IPI	-	-	-	33.94	50.68	85.93	88.31	19.61	32.80	72.73	75.82	21.79	35.78	67.13	54.65
PSTNN	-	-	-	24.99	39.98	71.64	101.8	15.07	26.20	61.62	81.86	13.07	23.12	43.05	65.63
ALCNet	0.378	3.74	0.0139	85.70	92.26	97.35	8.36	64.15	78.16	88.22	18.69	70.08	82.41	97.39	56.36
RDIAN	0.217	3.72	0.0053	89.71	94.58	97.88	10.36	63.12	77.42	89.89	24.16	69.97	82.33	97.38	110.9
ISTDU-Net	2.752	7.94	0.0245	92.32	96.00	97.88	4.18	64.93	78.73	89.56	28.96	70.81	82.79	96.97	45.58
MSHNet	4.065	6.11	0.0206	83.21	72.06	93.65	30.84	64.61	68.73	87.21	8.31	69.74	74.83	96.01	80.74
AGPCNet	12.36	43.18	0.1114	85.50	92.18	97.04	7.28	64.71	78.56	90.23	10.04	73.97	85.04	99.04	**20.07**
DNANet	4.697	14.26	0.0305	93.41	96.59	**98.31**	5.38	65.98	79.53	90.91	**9.51**	71.81	83.46	97.11	70.41
UIUNet	50.54	54.42	0.0308	92.39	96.03	97.78	6.55	65.20	78.93	89.89	18.67	71.24	83.19	92.71	195.4
Ours	1.505	6.557	0.0146	**93.57**	**96.67**	97.99	**3.72**	**66.99**	**80.06**	**91.25**	24.03	**74.09**	**85.12**	**99.04**	48.63

The best results are in bold, and the second best results are underlined.

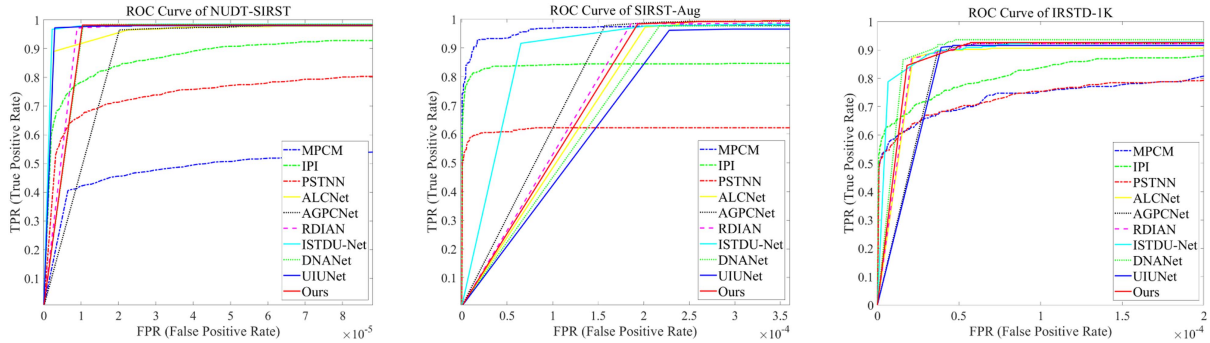


Fig. 9. ROC curves of different algorithms. (a) ROC curves of different algorithms with the NUDT-SIRST dataset. (b) ROC curves of different algorithms with the IRSTD-1K dataset. (c) ROC curves of different algorithms with the SIRST-Aug dataset.

unstable performance across different datasets, indicating poor robustness.

2) Visual Comparison: The visualization results of different methods are shown in Fig. 10. Two images from various scenes are selected from NUDT-SIRST, IRSTD-1k, and SIRST-Aug, respectively. In each figure, yellow circles indicate false positives, and blue circles represent missed detections. The detected target areas are highlighted in red and enlarged, placed in the corners of the detection image. Among traditional algorithms, MPCM demonstrates poor detection performance, particularly in scenes with strong edges, resulting in high false alarm rates, as shown in images NUDT-SIRST 00063 and SIRST-Aug 008808. The IPI algorithm has a high detection rate but exhibits high false alarm and missed detection rates in multitarget scenes. The PSTNN algorithm achieves good detection and false alarm rates but fails to accurately restore target shape information. For example, in image SIRST-Aug 008876, the detected target shape significantly differs from the ground truth, sometimes even mistaking a single target for multiple targets. Overall, these traditional algorithms heavily rely on prior information, lack robustness, and exhibit inferior detection performance compared to deep learning-based algorithms.

Compared to other deep learning algorithms, the proposed MDIGCNet accurately detects targets with the lowest miss rate. Our method exhibits a lower false alarm rate and is more effective at detecting targets in multitarget and high-brightness scenes compared to ALCNet. In comparison to DNANet, our method maintains a higher detection rate. The introduction of IDConv and MGIEB enables MDIGCNet to extract richer detail information and edge information of targetlike objects, enhancing the detection capability and improving the ability to restore the shape of small targets. The use of LGFF for multiscale information fusion enhances feature utilization efficiency, thereby reducing the occurrence of missed targets.

D. Ablation Study

To validate the effectiveness of each module in the proposed algorithm, ablation experiments are conducted on the aforementioned datasets. Specifically, in the first part, network performance is compared on the backbone network with and without residual structures and IDConv. This experiment is also conducted on the complete network to ensure optimal performance. In addition, we compared the performance of MSHNet, DNANet, and UIUNet with and without the use of IDConv. In the second part, to validate the effectiveness of MGIEB, comparisons are made with CDC and standard convolution. To demonstrate that IDConv and MGDC in MGIEB effectively extract detailed information, heatmap examples of four images are generated. Finally, in the third part, network performance is compared with and without the MGIEB and LGFF modules.

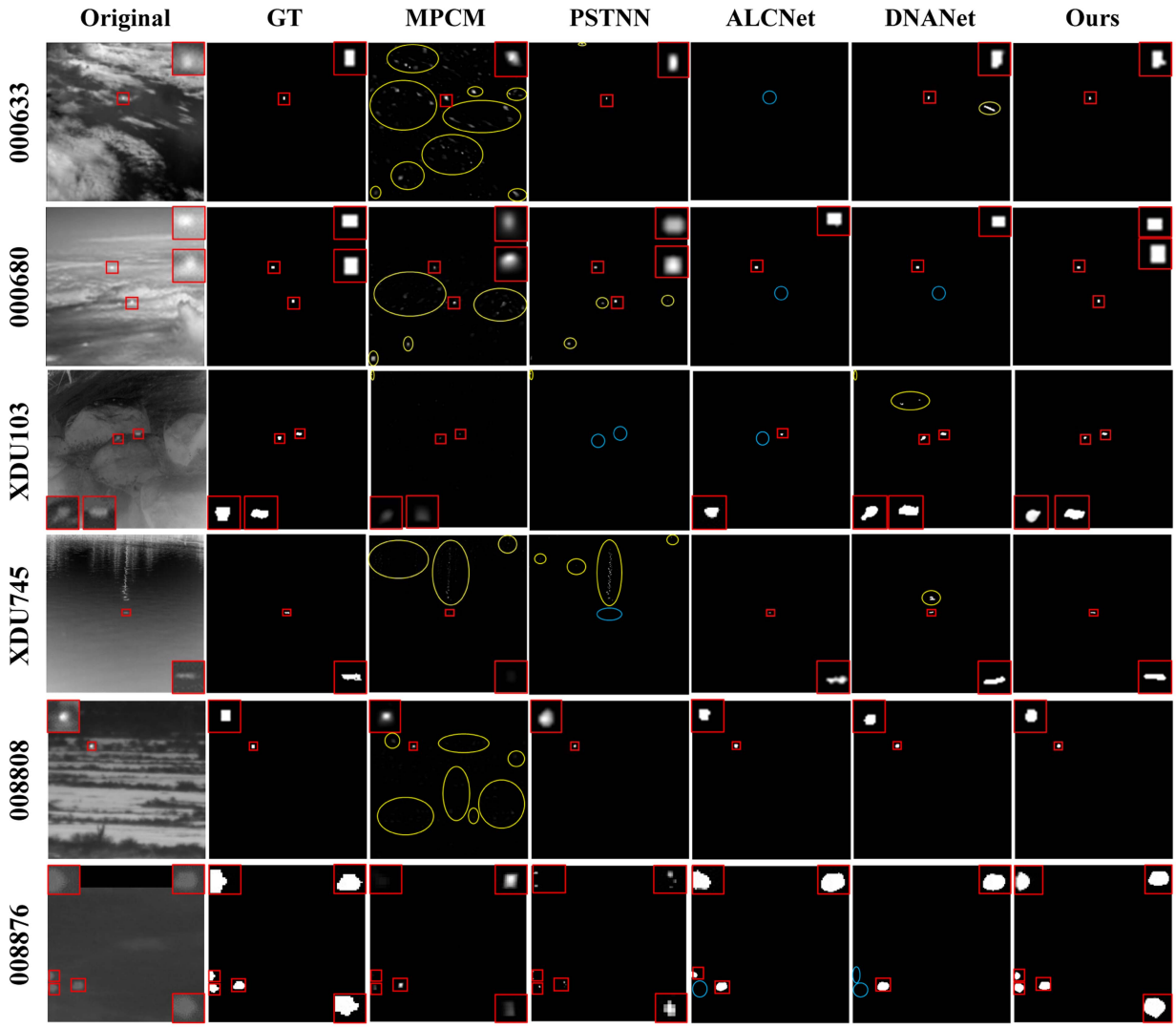


Fig. 10. Visual examples of some representative methods. In each figure, yellow circles indicate false positives, and blue circles represent missed detections. The detected target areas are highlighted in red and enlarged, placed in the corners of the detection image.

TABLE II
IoU (%),F_1(%), P_d (%), AND F_a ($\times 10^{-5}$) VALUES ACHIEVED IN THE NUDT-SIRST, IRSTD-1K, AND SIRST-AUG DATASETS ON ABLATION EXPERIMENTS ABOUT IDCONV AND RESIDUAL STRUCTURES

Methods	IDConv	Res	NUDT-SIRST				IRSTD-1K				SIRST-Aug			
			IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a
backbone	×	×	90.77	95.12	96.72	9.28	64.38	78.27	**87.54**	14.63	72.83	84.25	97.25	**60.48**
	×	✔	88.45	93.83	94.50	**9.10**	64.22	78.10	86.87	14.92	73.56	84.76	97.11	84.76
	✔	×	**90.84**	**95.16**	**96.83**	9.79	**64.52**	**78.41**	86.87	**11.54**	**73.73**	**84.88**	**99.17**	83.52
	✔	✔	90.40	94.95	96.61	12.25	60.03	74.99	87.21	17.25	73.27	84.57	98.07	83.29
backbone + MGIEB&LGFF	×	×	93.41	96.57	97.78	3.84	66.21	79.70	90.91	16.23	70.65	82.79	96.15	41.16
	×	✔	93.05	96.40	**98.20**	7.15	66.25	79.73	90.57	**12.66**	73.62	84.80	98.21	**37.13**
	✔	×	**93.57**	**96.67**	97.99	3.72	**66.99**	**80.06**	**91.25**	24.03	74.09	85.12	**99.04**	48.63
	✔	✔	93.01	96.38	97.35	**3.49**	65.54	79.19	88.89	14.39	**75.10**	**85.78**	98.90	47.96

The bold values represent the optimal values, while underlined values indicate the second-best values.

Throughout the experiments, the structure of other parts of the network remained unchanged.

1) Effect of IDConv: We compared the detection performance of using residual networks and not using residual networks in the backbone network, as well as the performance with and without IDConv. Table II presents the comparison of their detection performance in IoU(%), F_1 (%), P_d (%), and F_a (10^{-6}). It can be observed that when only residual structures are used, the network may experience performance degradation. Remarkably, there is a significant decline in performance, especially in pixel-level metrics like IoU, indicating that residual structures may not be suitable for our network. On the other

TABLE III
COMPARISON OF QUANTITATIVE METRICS [IoU (%),F_1(%), P_d (%), AND F_a ($\times 10^{-5}$)] FOR THE APPLICATION OF IDCONV ON OTHER NETWORKS(MSHNET, DNANET, AND UIUNET)

Methods	IDConv	Params(M)	FLOPs(G)	NUDT-SIRST				IRSTD-1K				SIRST-Aug			
				IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a
MSHNet	×	4.065	6.11	83.21	72.06	93.65	30.84	64.61	**68.73**	**87.21**	**8.31**	69.74	**74.83**	96.01	80.74
	✔	4.065	6.11	**89.14**	**78.92**	**95.03**	**24.70**	**66.05**	65.20	81.81	15.32	**72.07**	73.04	**97.52**	**27.05**
DNANet	×	4.697	14.26	93.41	96.59	98.31	**5.38**	65.98	79.53	90.91	**9.51**	71.81	83.46	97.11	70.41
	✔	4.697	14.26	**94.82**	**97.34**	**98.41**	6.41	**66.57**	**79.95**	**91.58**	24.35	**73.40**	**84.66**	**97.94**	**68.73**
UIUNet	×	50.54	54.42	92.39	96.03	97.78	6.55	65.20	78.93	89.89	**18.67**	71.24	83.19	92.71	195.4
	✔	50.54	54.42	**94.65**	**97.26**	**98.52**	**4.18**	**67.57**	**80.67**	**92.26**	22.32	**72.77**	**84.23**	**95.87**	**46.08**

The bold values represent the optimal values.

TABLE IV
COMPARISON OF QUANTITATIVE METRICS [IOU (%),F_1(%), P_d (%), AND F_a ($\times 10^{-5}$)] FOR DIFFERENT CONVOLUTIONAL BLOCKS

Methods	Re-Para	Params(M)	FLOPs(G)	Time(s)	NUDT-SIRST				IRSTD-1K				SIRST-Aug			
					IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a
General	-	1.486	6.203	0.0145	93.03	96.37	97.14	5.93	64.25	78.23	87.21	**12.26**	66.69	80.01	95.87	134.8
CDC	-	1.486	6.203	0.0319	92.03	95.83	96.93	9.54	61.07	75.78	87.88	20.17	71.84	83.61	97.52	**65.46**
MGIEB	×	3.901	13.65	0.0922	**93.51**	**96.63**	**97.99**	**2.96**	**64.61**	**78.36**	**89.90**	25.30	**72.61**	**84.13**	**97.52**	85.59
MGIEB	✔	**1.486**	**6.203**	0.0146	**93.51**	**96.63**	**97.99**	**2.96**	**64.61**	**78.36**	**89.90**	25.30	**72.61**	**84.13**	**97.52**	85.59

The bold values represent the optimal values.

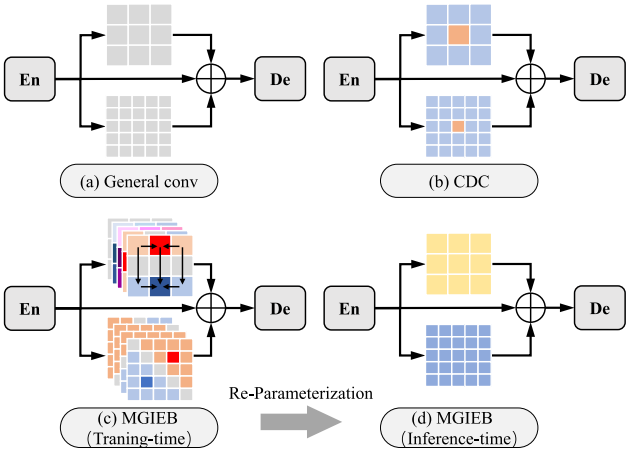


Fig. 11. Skip connection structures with different convolutional blocks. (a) General conv. (b) CDC. (c) MGIEB (Training-time). (d) MGIEB (Inference-time).

hand, the network performs better when only IDConv is used, especially with IoU and F_1 scores reaching optimal values on all three datasets, suggesting that our designed IDConv structure can effectively assist the backbone network in extracting texture details of small targets. When both structures are used simultaneously, the network performance is moderate.

We also conducted this experiment on the complete network to ensure optimal performance. As shown in the second row of Table II, it can be observed that the network performs better when only IDConv is used. Although the network using both structures performs better on the SIRST-Aug dataset, its robustness is inferior to the former. Therefore, we will adopt the backbone network in subsequent experiments using only IDConv.

To demonstrate the robustness and effectiveness of IDConv, its performance is compared across MSHNet, DNANet, and UIUNet. As shown in Table III, the use of IDConv resulted in an improvement in most metrics on both NUDT-SIRST and SIRST-Aug datasets. Specifically, MSHNet achieved an approximate 5% increase in IoU and F_1 scores on NUDT-SIRST, despite underperforming on the IRSTD-1K dataset. DNANet and UIUNet exhibited consistent improvement across most metrics.

2) Effect of MGIEB: The reparameterized MGIEB is compared with standard convolution and CDC, as shown in Fig. 11 and Table IV. Standard convolution, which does not model prior knowledge, exhibited poor performance, particularly on the SIRST-Aug dataset. Although CDC encodes differential information, it is not designed based on the characteristics of small targets and performed worse than standard convolution on the NUDT-SIRST and IRSTD-1K datasets. Our MGIEB, which integrates multidirectional Gaussian gradient information, achieved the best results across all three datasets. Through reparameterization, the module's parameters and FLOPs are reduced by 62% and 55%, respectively, reaching the same level as standard convolution while maintaining performance.

Additionally, to demonstrate that IDConv and MGDC within MGIEB effectively extract detailed information, images from four scenarios are utilized. As illustrated in Fig. 12, the columns from left to right represent the original image, ground truth, feature map input to MGIEB, heatmap output from IDConv, heatmap output from MGDC, and heatmap output from the MGIEB module. Target areas are marked with red boxes, and color intensity indicates energy levels from low to high. It is evident that the input features contain significant background noise, hindering precise detection of small targets. IDConv processing extracts detailed information from the images, facilitating the recovery of small target shapes. Following MGDC processing, clutter information outside the target area is reduced, aiding the network in focusing on targetlike regions. The concentrated energy on the output heatmaps confirms the effectiveness of the module's design.

3) Effect of LGFF: The impact of LGFF on performance improvement is assessed, as shown in Table V. Compared to the baseline network, the addition of the LGFF module resulted in significant increases in all metrics on the NUDT-SIRST

TABLE V
COMPARISON OF QUANTITATIVE METRICS [IOU (%),F_1(%), P_d (%), AND F_a ($\times 10^{-5}$)] FOR ABLATION EXPERIMENTS IN MGIEB AND LGFF

MGIEB	LGFF	Params(M)	FLOPs(G)	NUDT-SIRST				IRSTD-1K				SIRST-Aug			
				IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a	IoU	F_1	P_d	F_a
×	×	0.549	3.293	90.84	95.16	96.83	9.79	64.52	78.41	86.87	**11.54**	73.73	84.88	**99.17**	83.52
×	✔	0.568	3.649	91.50	95.54	97.67	8.62	64.34	78.14	89.56	24.67	73.07	84.44	98.35	**43.76**
✔	×	1.486	6.203	93.51	96.63	97.99	**2.96**	64.61	78.36	89.90	25.30	72.61	84.13	97.52	85.59
✔	✔	1.505	6.558	**93.57**	**96.67**	**97.99**	3.72	**66.99**	**80.06**	**91.25**	24.03	**74.09**	**85.12**	99.04	48.63

The bold values represent the optimal values, while underlined values indicate the second-best values.

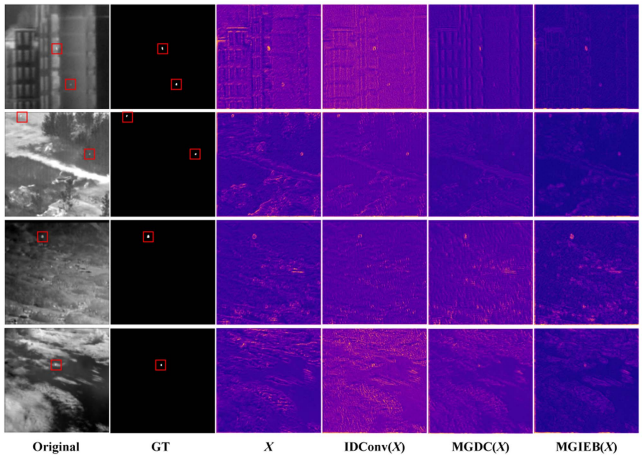


Fig. 12. Illustration of heatmap. The columns from left to right represent the original image, ground truth, feature map input to MGIEB, heatmap output from IDConv, heatmap output from MGDC, and heatmap output from MGIEB.

dataset. Performance on the other two datasets is moderate. When compared to the network with only MGIEB added, the inclusion of the LGFF module led to optimal results in nearly all metrics across the three datasets. Specifically, IoU increased by 0.06%, 2.38%, and 1.48% on the NUDT-SIRST, IRSTD-1K, and SIRST-Aug datasets, respectively, while F_1 increased by 0.04%, 2.24%, and 1.48%. P_d showed improvements of 1.35% and 1.52% on the latter two datasets. Although F_a performance declined on NUDT-SIRST, it improved on the other two datasets. Overall, the computational overhead introduced by LGFF is minimal, enhancing model performance across all datasets.

V. CONCLUSION

In this article, we propose a novel network for infrared target detection—MDIGCNet. To address the issues of low image information utilization and the lack of prior information in existing methods, we design IDConv, MGIEB, and LGFF. We use IDConv as both encoder and decoder within the U-Net architecture, enabling the extraction of rich image features. The skip connections employ the MGIEB module to capture gradient features of infrared small targets, enhancing network performance by fusing gradient features from multiple directions. With reparameterization techniques, performance improvement is achieved with low computational costs. Moreover, the LGFF module integrates shallow and deep features, enhancing the model's utilization of features. Extensive experimental results on various datasets scientifically demonstrate that our proposed network significantly improves detection accuracy, particularly in complex ISTD tasks, validating the soundness of our approach. Notably, the model achieves a lower false alarm rate and higher precision compared to existing methods.

While the MDIGCNet shows promising results, future research will focus on optimizing the network's computational efficiency and reducing false positives in more complex real-world scenarios. For example, in the MGIEB module, although we employ a 5 × 5 convolution kernel to model gradient features and use reparameterization techniques to limit parameter growth, further reduction in computational complexity is possible. Future work will also explore refining the differential convolution-based modules and applying differentiated designs at various network layers to reduce redundancy. Moreover, we plan to introduce transformer-based modules to further mitigate missed and false detections in challenging environments.

ACKNOWLEDGMENT

The authors would like to thank all those who provided invaluable comments on this article.

REFERENCES

[1] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 87–119, Jun. 2022.
[2] Y. Cheng, X. Lai, Y. Xia, and J. Zhou, "Infrared dim small target detection networks: A review," *Sensors*, vol. 24, no. 12, 2024, Art. no. 3885.
[3] V. T. Tom, T. Peli, M. Leung, and J. E. Bondaryk, "Morphology-based algorithm for point target detection in infrared backgrounds," *Proc. SPIE*, vol. 1954, pp. 2–11, 1993.
[4] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," *Proc. SPIE*, vol. 3809, pp. 74–83, 1999.
[5] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.
[6] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.
[7] F. Wu, H. Yu, A. Liu, J. Luo, and Z. Peng, "Infrared small target detection using spatiotemporal 4-D tensor train and ring unfolding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5002922.
[8] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.
[9] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, 2016.
[10] L. Wu, Y. Ma, F. Fan, M. Wu, and J. Huang, "A double-neighborhood gradient method for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1476–1480, Aug. 2021.
[11] X. Wang et al., "Multiscale feature extraction U-Net for infrared dim- and small-target detection," *Remote Sens.*, vol. 16, no. 4, 2024, Art. no. 643.

[12] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8509–8518.

[13] S. Liu, P. Chen, and M. Woźniak, "Image enhancement-based detection with small infrared targets," *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3232.

[14] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 950–959.

[15] X. Zhang, X. Zhang, S.-Y. Cao, B. Yu, C. Zhang, and H.-L. Shen, "MRF3Net: An infrared small target detection network using multireceptive field perception and effective feature fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5629414.

[16] X. Wu, D. Hong, and J. Chanussot, "UIU-NET: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.

[17] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023.

[18] C. Li, Y. Zhang, Z. Shi, Y. Zhang, and Y. Zhang, "Moderately dense adaptive feature fusion network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5616712.

[19] G. Chen, W. Wang, and S. Tan, "Irstformer: A hierarchical vision transformer for infrared small target detection," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3258.

[20] X. Tong et al., "ST-Trans: Spatial-temporal transformer for infrared small target detection in sequential images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5001819.

[21] S. Wu, C. Xiao, L. Wang, Y. Wang, J. Yang, and W. An, "RepISD-Net: Learning efficient infrared small-target detection network via structural re-parameterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5622712.

[22] R. Lu, X. Yang, W. Li, J. Fan, D. Li, and X. Jing, "Robust infrared small target detection via multidirectional derivative-based weighted contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7000105.

[23] X. Bai and Y. Bi, "Derivative entropy-based contrast measure for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2452–2466, Apr. 2018.

[24] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 4, pp. 4250–4261, Aug. 2023.

[25] T. Guo, B. Zhou, F. Luo, L. Zhang, and X. Gao, "DMFNet: Dual-encoder multistage feature fusion network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5614214.

[26] R. Zhang et al., "Part-aware correlation networks for few-shot learning," *IEEE Trans. Multimedia*, vol. 26, pp. 9527–9538, 2024.

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

[28] C. Yu et al., "Infrared small target detection based on multiscale local contrast learning networks," *Infrared Phys. Technol.*, vol. 123, 2022, Art. no. 104107.

[29] Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, and W. Zhang, "RISTDnet: Robust infrared small target detection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7000805.

[30] R. Zhang et al., "Cognition-driven structural prior for instance-dependent label transition matrix estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3347633.

[31] Z. Yu et al., "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5295–5305.

[32] Z. Su et al., "Pixel difference networks for efficient edge detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5117–5127.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, Munich, Germany, 2015, pp. 234–241.

[34] C. Xia, S. Chen, X. Zhang, Z. Chen, and Z. Pan, "Infrared small target detection via dynamic image structure evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5003318.

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

[36] T. Zhang, Z. Peng, H. Wu, Y. He, C. Li, and C. Yang, "Infrared small target detection via self-regularized weighted sparse model," *Neurocomputing*, vol. 420, pp. 124–148, 2021.

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

[38] Y. Luo, X. Li, S. Chen, and C. Xia, "4DST-BTMD: An infrared small target detection method based on 4-d data-sphered space," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5000520.

[39] C. Xia, S. Chen, R. Huang, J. Hu, and Z. Chen, "Separable spatial–temporal patch-tensor pair completion for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5001620.

[40] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, "Infrared small target detection utilizing the multiscale relative local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 612–616, Apr. 2018.

[41] Y. Qin and B. Li, "Effective infrared small target detection utilizing a novel local contrast method," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1890–1894, Dec. 2016.

[42] J. Han et al., "Infrared small target detection based on the weighted strengthened local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1670–1674, Sep. 2021.

[43] J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, and Q. Zhao, "A local contrast method for infrared small-target detection utilizing a tri-layer window," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1822–1826, Oct. 2020.

[44] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Xie, and Z. Li, "PixelGame: Infrared small target segmentation as a nash equilibrium," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8010–8024, 2022.

[45] B. Zhao, C. Wang, Q. Fu, and Z. Han, "A novel pattern for infrared small target detection with generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4481–4492, May 2021.

[46] Z. Zuo et al., "Affpn: Attention fusion feature pyramid network for small infrared target detection," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3412.

[47] R. Zhang et al., "Differential feature awareness network within antagonistic learning for infrared-visible object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6735–6748, Aug. 2024.

[48] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, "Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation," *IEEE Trans. Multimedia*, vol. 24, pp. 1735–1749, 2022.

[49] R. Kou et al., "Multi-scale small target detection techniques in single-frame infrared images: A review," *J. Image Graph.*, vol. 29, pp. 0193–0217, Sep. 2024.

[50] X. He, Q. Ling, Y. Zhang, Z. Lin, and S. Zhou, "Detecting dim small target in infrared images via subpixel sampling cuneate network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6513005.

[51] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small and dim target detection with transformer under complex backgrounds," *IEEE Trans. Image Process.*, vol. 32, pp. 5921–5932, 2023.

[52] F. Wu, T. Zhang, L. Li, Y. Huang, and Z. Peng, "RPCANet: Deep unfolding RPCA based infrared small target detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 4809–4818.

[53] R. Kou, C. Wang, Y. Yu, Z. Peng, F. Huang, and Q. Fu, "Infrared small target tracking algorithm via segmentation network and multistrategy fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612912.

[54] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 877–886.

[55] H. Sun, J. Bai, F. Yang, and X. Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5000513.

[56] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, "Dual-cross central difference network for face anti-spoofing," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Z. Zhi-Hua, Ed., Aug. 2021, pp. 1281–1287, doi: 10.24963/ijcai.2021/177.

[57] Z. Yu et al., "Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 5626–5640, 2021.

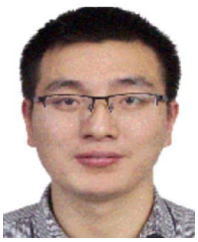
[58] X. Ying et al., "Local motion and contrast priors driven deep network for infrared small target superresolution," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5480–5495, 2022.

[59] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[60] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6053–6062.
[61] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1911–1920.
[62] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style convnets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13733–13742.
[63] Z. Chen, Z. He, and Z. Lu, "DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention," *IEEE Trans. Image Process.*, vol. 33, pp. 1002–1015, 2024.
[64] J. Peng, H. Zhao, Z. Hu, K. Zhao, and Z. Wang, "DRPN: Making CNN dynamically handle scale variation," *Digit. Signal Prog.*, vol. 133, 2023, Art. no. 103844.
[65] R. Kou et al., "Infrared small target segmentation networks: A survey," *Pattern Recognit.*, vol. 143, 2023, Art. no. 109788.
[66] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 33–37, Jan. 2018.
[67] S. Qi, G. Xu, Z. Mou, D. Huang, and X. Zheng, "A fast-saliency method for real-time infrared small target detection," *Infrared Phys. Technol.*, vol. 77, pp. 440–450, 2016.
[68] S. Woo, J. Park, J. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[69] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Comput. Sci.*, 2014.
[70] Y. Xi et al., "Infrared moving small target detection based on spatial-temporal local contrast under slow-moving cloud background," *Infrared Phys. Technol.*, vol. 134, 2023, Art. no. 104877.
[71] Q. Liu, R. Liu, B. Zheng, H. Wang, and Y. Fu, "Infrared small target detection with scale and location sensitivity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17490–17499.

Luping Zhang (Student Member, IEEE) received the B.E. degree in electronic information engineering in 2022 from the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, where he is currently working toward the M.E. degree in communication engineering.

His research interests include image processing, computer vision, and infrared small target detection.



Junhai Luo (Member, IEEE) received the B.S. degree in computer science and appliance from the University of Electronic Science and Technology of China, Chengdu, China, in 2003, the M.S. degree in computer appliance technology from the Chengdu University of Technology, Chengdu, in 2006, and the Ph.D. degree in information and communication engineering from the University of Electronic Science and Technology of China.

He was a Visiting Scholar with McGill University, Montreal, Canada, and the University of Tennessee, Knoxville, TN, USA. He was promoted to Associate Professor in 2011. His research interests include target detection and information fusion.

Yian Huang (Student Member, IEEE) received the B.S. degree in communication engineering from Sun Yat-sen University (SYSU), Guangzhou, China, in 2022. He is currently working toward the Ph.D. degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China.

His research interests include image processing, computer vision, and infrared target recognition.

Fengyi Wu (Graduate Student Member, IEEE) received the joint B.E. degree in electric and electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, and the University of Glasgow (UoG), Glasgow, U.K., in 2021. He is currently working toward the Ph.D. degree with the School of Information and Communication Engineering, UESTC.

His research interests include image processing, computer vision, and small target detection.

Xingye Cui was born in Nanjing, China, in 2001. She received the B.E. degree in communication engineering from the School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, China, in 2023. She is currently working toward the M.E. degree in electronic information with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

Her research interests include image processing and infrared small target detection.

Zhenming Peng (Member, IEEE) received the Ph.D. degree in geodetection and information technology from the Chengdu University of Technology, Chengdu, China, in 2001.

From 2001 to 2003, he was a Postdoctoral Researcher with the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu. He is currently a Professor with the University of Electronic Science and Technology of China, Chengdu. His research interests include image processing, machine learning, objects detection, and remote sensing applications.

Dr. Peng is a Member of many academic organizations, such as the Institute of Electrical and Electronics Engineers (IEEE), Optical Society of America (OSA), China Optical Engineering Society (COES), Chinese Association of Automation (CAA), Chinese Society of Astronautics (CSA), Chinese Institute of Electronics (CIE), and China Society of Image and Graphics (CSIG), etc.