Single-Frame Infrared Small Target Detection Network Based on Multibranch Feature Aggregation

Ziqiang Hao, Zheng Jiang, Xiaoyu Xu, Zhuohao Wang, and Zhicheng Sun

***Abstract*—Single-frame infrared small target detection is critical in fields, such as remote sensing, aerospace, and ecological monitoring. Enhancing both the accuracy and speed of this detection process can substantially improve the overall performance of infrared target detection and tracking. While deep learning-based methods have shown promising results in general detection tasks, increasing network depth vertically to improve feature extraction often results in the loss of small targets. To address this challenge, we propose a network framework based on multibranch feature aggregation, which expands the network depth horizontally. The parallel auxiliary branches are carefully designed to provide the main branch with semantic information at varying depths and scales. Furthermore, we introduce a differential correction module that corrects erroneous target features through differential methods, significantly boosting detection accuracy. Lastly, we develop a joint attention module that combines channel and spatial attention mechanisms, enabling the network to accurately localize and reconstruct small targets. Extensive experiments on the NUDT-SIRST, SIRST, and NUST-SIRST datasets demonstrate the clear superiority of our approach over other state-of-the-art infrared small target detection methods.**

***Index Terms*—Channel spatial attention, deep learning, feature interaction, infrared small target detection (ISTD), information aggregation.**

I. Introduction

INFRARED small target detection (ISTD), by offering high-precision detection of small targets on the Earth's surface and in the atmosphere, has greatly advanced research in the fields of geoscience and remote sensing. Unlike traditional visible light detection methods, infrared detection offers distinct advantages in complex backgrounds, low-light conditions, and completely dark environments [1], [2]. Infrared small target detection leverages infrared imaging to identify and locate faint targets that exhibit temperature differences from their surroundings, making it possible to detect targets that are often missed by visible light imaging [3], [4]. This technique has widespread applications in military, aviation, aerospace, and security fields, including missile warning, airborne target tracking, and night vision surveillance [5]. However, infrared small target detection presents unique challenges compared to general target detection tasks. 1) Small targets typically have a low signal-to-noise ratio, making them prone to noise interference. 2) The shape, size, and location of small targets are often unclear, appearing as faint signals within the image. 3) The backgrounds are frequently complex and dynamic, complicating the task of distinguishing the target from its background.

Single-frame infrared small target detection remains a persistent challenge in the field of infrared image detection. The faint characteristics of small targets against complex backgrounds significantly increase the difficulty of detection. In addition, unlike multiframe sequences, single-frame images lack motion information and temporal context, forcing the detection process to rely solely on features like grayscale, gradients, and contrast [6]. As a result, there is still a need for more robust methods to address the limitations of single-frame infrared small target detection.

Traditional single-frame infrared small target detection methods, which rely on manual feature engineering and fixed hyperparameter settings, struggle to handle the complexities of these tasks [7]. In recent years, the rapid development of artificial intelligence has led to the proposal of large models capable of quickly segmenting remote sensing images, such as Alibaba's AIE-SEG and Meta AI's "Segment Anything Model" [8]. Although these large models can achieve "zero-shot generalization," neural network-based methods still hold advantages in scenarios where computational resources are limited or specific tasks need to be addressed. Therefore, methods based on convolutional neural networks (CNNs), with their learnable parameters and trainable models, have remained a key focus and challenge in the field of infrared target detection for remote sensing image processing.

We found that many CNN-based ISTD networks, in an effort to enhance feature extraction capabilities, tend to design deeper networks. This often leads to the loss of small target pixels in the deeper layers of the network. On the other hand, shallower networks lack the necessary feature extraction capacity. To address this issue, we explored and designed a new network structure that introduces parallel auxiliary branches to balance this tradeoff.

As shown in Fig. 1, we compared the visualization results of the same infrared image input into two different network structures [see Fig. 1(a) and (b)]. It can be observed that the feature map ($(F_1$ and $F_2)$ output at the same network depth shows more prominent target pixels in structure Fig. 1(b) as seen in F_2. In contrast, for structure Fig. 1(a), the deeper output

Received 2 April 2024; revised 10 May 2024, 17 September 2024, and 25 October 2024; accepted 16 December 2024. Date of publication 23 December 2024; date of current version 17 January 2025. This work was supported by the Chongqing Natural Science Foundation through project "Research on Target Recognition Technology in Multisource Image Fusion under Complex Backgrounds" under Grant CSTB2022NSCQ-MSX1071. The project is scheduled for the period 2022-2024. *(Corresponding author: Zheng Jiang.)*

The authors are with the College of Electronic Information Engineering, Changchun University of Science and Technology, Changchun 130000, China (e-mail: haoziqiang@cust.edu.cn; 2022100918@mails.cust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3521057

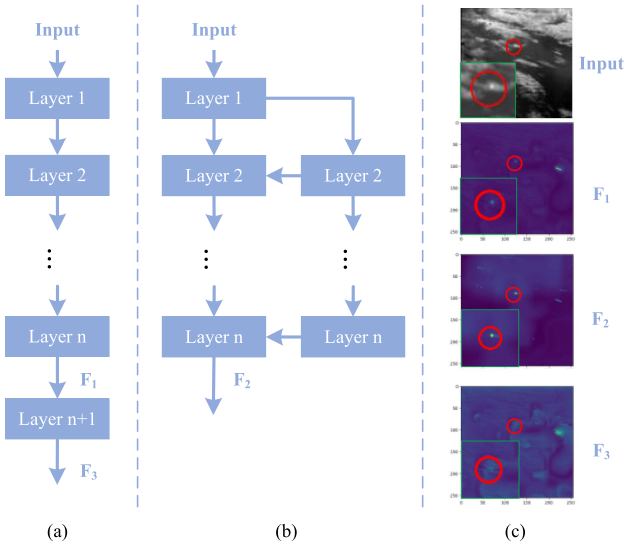

Fig. 1. Comparison of small target retention effects under two network design structures. (a) Increasing vertical depth. (b) Adding parallel auxiliary branch. (c) Input image and visualized feature maps.

feature map (F_3) shows that small target pixels are almost lost. Therefore, we designed a CNN-based ISTD network that uses a main-branch and subbranch multipath structure with integrated attention mechanisms.

The main contributions of this article are as follows.

1) We propose a main-secondary multibranch structure, which expands the network depth by adding auxiliary branches, preventing small targets from being lost in deeper network layers.
2) We propose a joint attention module (JAM) that extracts weights from both the spatial and channel dimensions, enhancing the network's ability to localize infrared small targets.
3) We designed a differential correction module (DCM) that utilizes the primary feature maps containing target information to correct the deeper feature maps that have lost target information through differential methods, enhancing the sensitivity to concealed targets.

II. Related Work

In this section, we will review the relevant literature on ISTD networks, specifically focusing on approaches related to the method we propose.

A. Single-frame ISTD

Currently, in the field of single-frame ISTD, there are many traditional methods that can be classified into three categories: ISTD methods based on background features, methods based on target features and methods based on the imaging characteristics of infrared images [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. Although the aforementioned traditional methods have achieved significant success, they still face challenges, such as low universality and long detection times.

Differing from traditional methods, deep learning approaches address this problem through a data-driven manner, overcoming the drawbacks of model-driven approaches in traditional methods. Methods based on CNNs, such as MDvsFA [21] and the Faster R-CNN [22], have been adopted. Given the characteristics of infrared small targets, including few pixels and class imbalance, some researchers have proposed targeted strategies. The most common approach is to design feature fusion strategies, such as [23], [24], [25], and [26]. Alternatively, some researchers have leveraged the concept of patch classification, dividing the input image into blocks and feeding them into the network to enhance attention to small targets, as in LPNet [27]. Meanwhile, other scholars have regarded small targets in infrared images as "noise" and proposed an ISTD model based on a denoising autoencoder network [28]. Some researchers have also been dedicated to enhancing the lightweight nature of detection networks by 1) reducing the depth of the backbone, the number of channels, the number of convolutional layers, and the complexity of integration methods. 2) replacing the convolutional layer with group or depthwise-separable convolution. 3) removing the fully connected layer [29], among other approaches. These efforts aim to improve the network's real-time performance and feasibility for embedded deployment, such as [30].

While the aforementioned detection methods achieve good results, they still struggle to preserve small target features in the deep layers of the network while enhancing the network's feature extraction capabilities. The proposed method not only enhances the network's feature extraction capabilities but also horizontally extends the network depth, aiming to retain small target features as much as possible. Therefore, it represents a robust detection approach.

B. ISTD Based on Attention Mechanism

Attention mechanisms play a crucial role in the design of ISTD networks. They assist the network in focusing on target regions in the image while suppressing background interference, thereby enhancing target discernibility and contrast. In the field of single-frame ISTD, the most commonly designed attention mechanisms include the following.

1) Channel Attention: The core idea is to learn the importance of each feature channel, thereby assigning weights to the channels. Several channel attention modules have been proposed based on this concept, such as the SE module [31], ECA module [32], and FcaNet [33].

2) Spatial Attention: This involves learning the importance of spatial positions within the image. Examples include PSANet [34]. In recent years, several spatial attention mechanisms based on self-attention have also been proposed, such as nonlocal [35], SASA [36], and ViT [37].

3) Channel-Spatial Mixed Attention: This approach simultaneously applies both channel attention and spatial attention mechanisms, combining the two in a complementary manner. Examples include CBAM [38] and scSE [39].

Based on the core ideas of these attention mechanisms, many single-frame ISTD algorithms have designed various attention modules and integrated them into their networks.

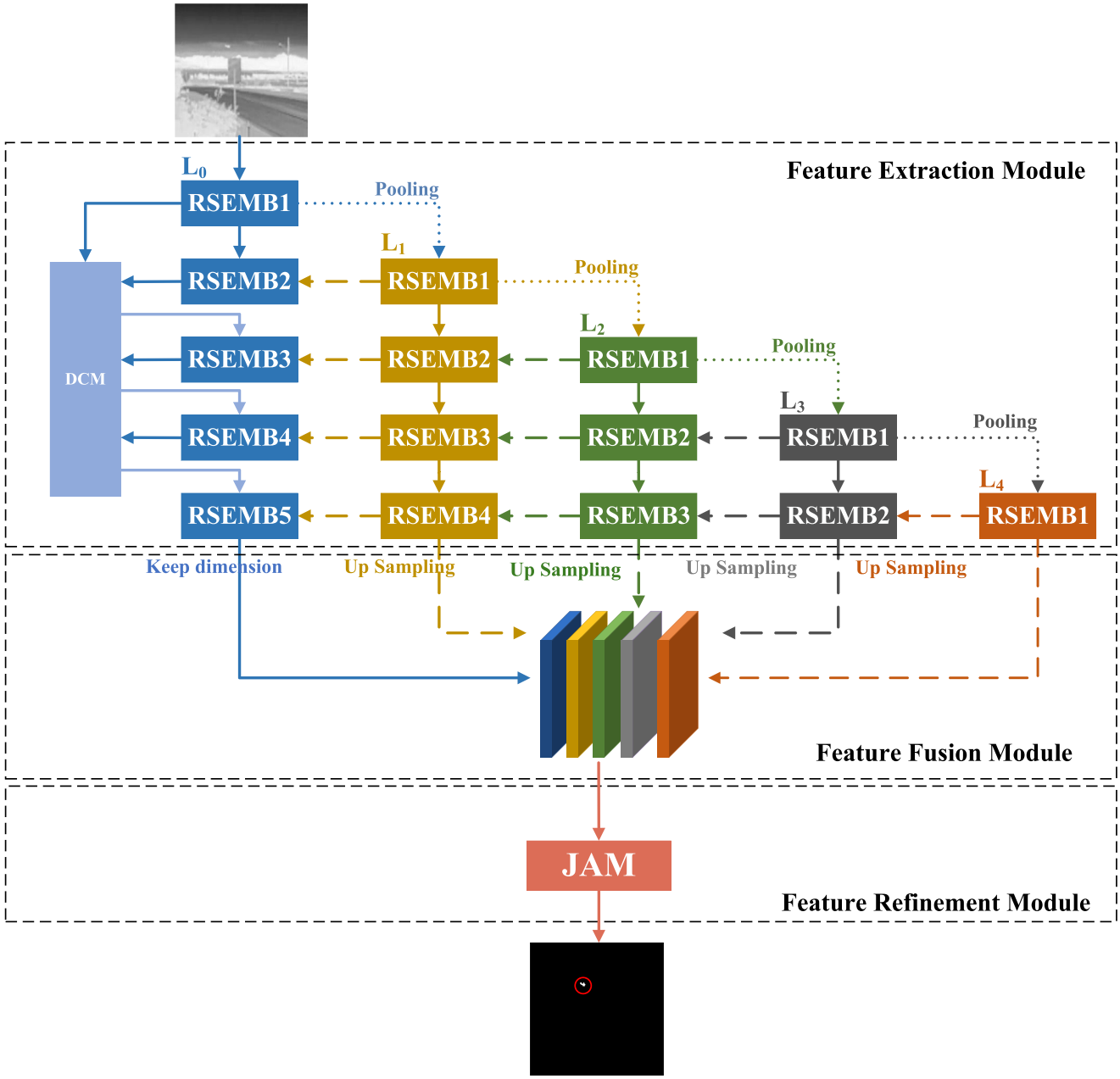

Fig. 2. Overall network architecture.

Examples include ACM [40], AGPCNet [41], MDFENet [42], SCTransNet [43], APAFNet [44], MCAF [45], and interactive-cross attention Module [46]. These all reflect the principles of different attention mechanisms.

In summary, enhancing the representation of small target features efficiently can be achieved by designing and incorporating attention modules into the network, proving to be an effective means of improving detection performance. The proposed method introduces an effective channel-spatial attention module. This module, tailored to the characteristics of small targets, employs dilated convolutions with various dilation rates to capture small targets, simultaneously enlarging the receptive field to suppress background information.

III. Proposed Method

In this section, we introduce the proposed multibranch feature aggregation network (MBFANet) for ISTD. The network's overall architecture is shown in Fig. 2. It processes a single-frame infrared image through three stages: the feature extraction module, feature fusion module, and feature refinement module, ultimately producing the detection results. The feature extraction module consists of a main branch and auxiliary branches. The main branch extracts multilevel features, while the auxiliary branches capture semantic information at different depths. This information is progressively integrated back into the main branch, enhancing its feature extraction capabilities.

The module utilizes the RSEMB (Res-SE-Module-Block), an improved version of the block used in ResNet18 [47] designed to enhance both feature extraction and interbranch interaction. In addition, the DCM refines feature extraction during training, improving the detection of weak small targets. A detailed explanation of the feature extraction module, RSEMB, and DCM is provided in Section III-A. The feature fusion module merges spatial information from the main branch with multiscale semantic information from the auxiliary branches, producing a more comprehensive global feature map. The core principles of this module are discussed in Section III-B. Finally, the feature refinement module, incorporating a JAM, refines the prediction map to improve detection accuracy. A detailed description of its components is provided in Section III-C.

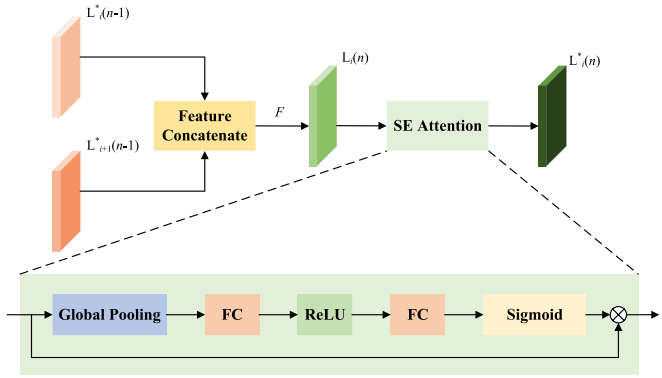

Fig. 3. Illustration of the RSEMB module.

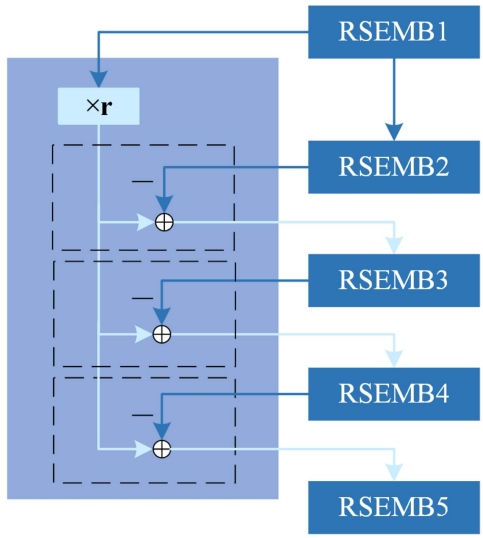

Fig. 4. Specific composition of DCM.

A. Feature Extraction Module

1) Network Architecture: The existing feature extraction backbones of ISTD networks often have a single depth, making it easy for small targets to disappear in the deep layers of the network, resulting in poor detection performance. Therefore, instead of extending the feature extraction network vertically to a great depth, we expand the network horizontally. Our feature extraction module consists of one main branch (L_0) and four auxiliary branches (L_1, L_2, L_3, and L_4). In this module, all branches are composed of different numbers of RSEMB connected in series. The inputs to the auxiliary branches are obtained by pooling the output of the first RSEMB of the preceding branch. The main branch L_0 is composed of five connected RSEMB to extract features. The outputs of the first four RSEMB are connected to the last one, aggregating all useful feature information at once. This connection method, as compared to dense connections, avoids receiving redundant information, making it more efficient. The auxiliary branch L_1 consists of four connected RSEMB and its sinput is pooled from the output of the first RSEMB block in the main branch L_0. After passing through each RSEMB, the extracted feature information is transmitted to the main branch L_0 as supplementary semantic information at different depths. Similarly, auxiliary branches L_2, L_3 and L_4, successively act as the auxiliary branches of the preceding branch, continually expanding the network depth, extracting richer semantic information and eventually converging into the main branch L_0. Meanwhile, each auxiliary branch also outputs its respective feature maps (L_{1_out}, L_{2_out}, L_{3_out}, L_{4_out}), which will be fused with the output (L_{0_out}) of the main branch in the next module, further utilizing semantic information at different scales.

2) RSEMB—The key feature extraction block: Our main component block for the feature extraction module, RSEMB, is an improved version of the ResNet18 block. In each block, an SE Attention Module is added before the residual connection. After the multilevel features from different branches are fused, the RSEMB is used for adaptive feature enhancement, significantly improving the feature extraction capability of the backbone. As shown in Fig. 3, $L_i^*(n)$ denotes the output of the nth RSEMB in the ith branch, $L_i^*(n-1)$ denotes the output of the (n-1)th RSEMB in the ith branch, $L_{i+1}^*(n-1)$ denotes the output of the (n-1)th RSEMB in the (i+1)th branch. After feature concatenation, the process continues through the convolutional layers in the ResNet18 block (denoted by F in this process), resulting in $L_i(n)$. Then, it goes through the SE attention module, the process of which involves. First, performing global average pooling on the feature map, where each channel is represented by a single value, essentially possessing the global receptive field for that channel. Subsequently, through two fully connected layers, the network generates the required weight information, obtaining feature relevance and assigning different weights to channels. The entire process can be represented by the following equation:

$$L_i^*(n) = F_{\text{SE}}\left[F\left[\text{Concat}(L_i^*(n-1), L_{i+1}^*(n-1))\right]\right] \quad (1)$$

where Concat denotes splicing in the channel dimension, F_{SE} denotes the SE attention module, F denotes the convolutional layers in the ResNet18 block.

3) DCM: We creatively introduce a DCM aimed at reducing the model's sensitivity to noise and interference in images. The specific structure of the DCM is shown in Fig. 4. We use the output feature map of the first RSEMB block in the main branch

as the reference image, denoted as image A, for differencing. This choice is motivated by the rich background and target information typically contained in the feature maps at this stage. Specifically, we scale the pixel values of image A by a factor of r, referred to as r× A. The output feature maps B1, B2, and B3 from the second, third, and fourth RSEMB blocks in the main branch serve as the difference images. We subtract the difference images B1, B2, and B3 from the scaled reference image r× A to generate the differentially processed images.

The fundamental cause of false alarms (Fa) and missed detections in ISTD lies in the fact that in some images, the pixel values of target objects are nearly indistinguishable from certain noise or clutter pixels in the background. This similarity makes it challenging for neural networks to differentiate between background and targets. As a result, when these images are fed into the designed neural network, the convolution operations often erroneously increase the pixel values of background clutter (or noise) while decreasing the pixel values of target objects. However, by subtracting the feature maps where the erroneously enhanced background clutter (or noise) pixel values and weakened target pixel values are present from the feature maps containing both intact target pixel values and background clutter (or noise) pixel values, we can reduce the heightened background clutter (or noise) pixel values and enhance the pixel values of targets. The differentially processed feature maps are then fed into the next convolutional block in the main branch to extract features anew. This constitutes the correction process.

It is important to note that when the neural network correctly focuses on the targets and suppresses background clutter, this differential correction operation does not improve performance; instead, it may degrade it. Therefore, we first multiply the pixel values of the difference images by a factor of r. In this way, even if the feature maps successfully extract target features, the differential process will not overly affect the pixel values of the correct targets. Simultaneously, with the assistance of other auxiliary branches, this correction operation does not affect the detection accuracy of images where targets and backgrounds are easily distinguishable by the network. In summary, this differential correction operation yields significant improvements in detecting images where targets and backgrounds are prone to confusion. Fig. 5 illustrates the role of the DCM module in different scenarios. It can be observed that the DCM corrects deep-layer feature maps prone to Fa and missed detections by utilizing the information contained in shallow-layer feature maps regarding targets and backgrounds.

We define the pixel value at the target position in the differentially corrected image A as x, and the pixel value at the noise position as y. The differential image B is the output feature map of the subsequent convolutional block in the main branch, with the pixel value at the target position denoted as x', and the pixel value at the noise position denoted as y'. Thus, the corresponding pixel values of the target and noise in the differentially corrected image F are $2x - x'$ and $2y - y'$, respectively. For the correction module to be effective, the pixel value at the target position in the differentially corrected image F must be greater than x, and the pixel value at the noise position must be less than y, enhancing the target and weakening the noise. This can be represented as

$$\begin{cases} 2x - x' \geq x \\ 2y - y' \leq y \end{cases} \Rightarrow \begin{cases} x \geq x' \\ y \leq y'. \end{cases} \tag{2}$$

This indicates that if, after convolutional feature extraction, the target is erroneously weakened while noise is erroneously emphasized, the DCM will effectively rectify this error. It will then pass the corrected feature map to the next convolutional block for re-extraction of features.

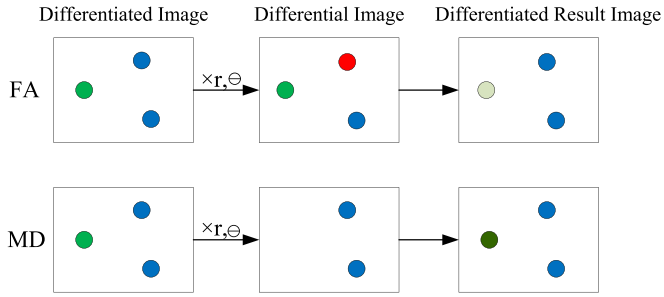

Fig. 5. Operational principle of DCM in handling Fa and missed detections. In this figure, green circles represent correctly identified targets, red circles represent noise falsely recognized as targets, and blue circles represent interfering noise. Light green and dark green circles, respectively, denote weakened and enhanced target pixels.

B. Feature Fusion Module

To effectively utilize the information extracted from each branch, after the feature extraction module, we designed a feature fusion module to merge the outputs of the main branch L_0 with the four auxiliary branches L_1, L_2, L_3, and L_4, thus integrating features from different levels. The outputs [L_{1_Output}, L_{2_Output}, L_{3_Output}, L_{4_Output}] of the auxiliary branches L_1, L_2, L_3, and L_4 are first upsampled to match the size of the output of the main branch L_0, denoted as L_{0_Output}, in both length and width. Due to the relatively shallow depth of the main branch, which extracts spatial contour information and the deeper depth of each auxiliary branch, which captures semantic information, we concatenate the outputs of all branches to obtain a more robust feature map F_C that integrates rich features from each source

$$F_C = \text{Concat}[L_{0\_\text{Output}}, L_{1\_\text{Output}}, L_{2\_\text{Output}}, L_{3\_\text{Output}}, L_{4\_\text{Output}}]. \tag{3}$$

C. Feature Refinement Module

To improve the precision of small target localization and enhance the network's ability to detect and reconstruct their shapes, we designed a JAM to refine the feature map F_C. As shown in Fig. 6, the JAM integrates two types of attention mechanisms: channel attention and spatial attention. The channel attention mechanism starts with average pooling of the input feature map along the spatial dimensions, followed by processing through a fully connected layer to obtain channel-wise correlation weights. After applying the Sigmoid activation function for normalization, these weights are multiplied by the input feature map to

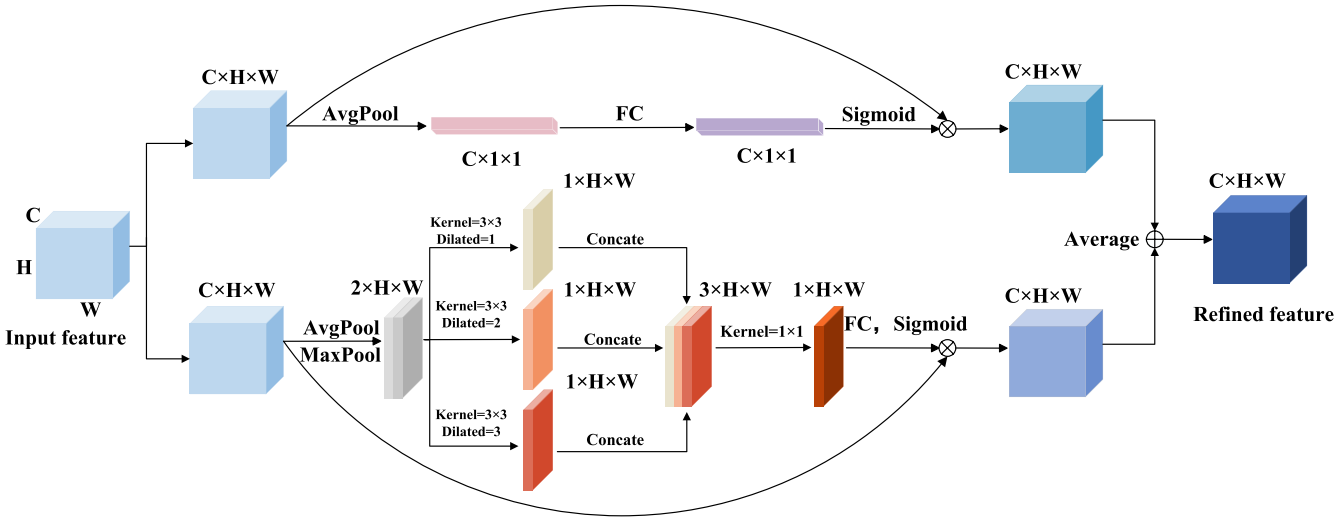

Fig. 6. Structure of the JAM.

produce F_{C_SA}

$$F_{C\_\mathrm{SA}} = F_C \otimes \mathrm{Sigmoid}\left[F_{\mathrm{FC}}\left[F_{AP}\left[F_C\right]\right]\right] \tag{4}$$

where $\otimes$ denotes pixel-wise multiplication, F_{AP} denotes the global average pooling operation, and F_{FC} denotes the fully connected operation.

The spatial attention mechanism applies both average pooling and max pooling along the channel dimension of the input feature map. The pooled results are concatenated, then passed through dilated convolutions with 3×3 kernels and dilation rates of 1, 2, and 3 to capture spatial correlation information. After concatenating the convolution outputs, a 1×1 convolution reduces the channel dimension. The result is processed through fully connected layers and a Sigmoid activation function to generate the spatial weight information. Finally, this is multiplied with the input feature map to obtain F_{C_CA}

$$F' = \mathrm{Concat}\left[F_{AP}\left[F_C\right], F_{\mathrm{MP}}\left[F\right]\right] \tag{5}$$

$$F'' = F_{1\times1}\left[\mathrm{Concat}\left[F_1\left[F'\right], F_2\left[F'\right], F_3\left[F'\right]\right]\right] \tag{6}$$

$$F_{C\_\mathrm{CA}} = F_C \otimes \mathrm{Sigmoid}\left[F_{\mathrm{FC}}\left[F''\right]\right] \tag{7}$$

where F_{MP} denotes global max pooling operation.

The outputs of F_{C_SA} and F_{C_CA} are combined by taking their average, resulting in the final output of the feature refinement module, denoted as F_{Final}

$$F_{\mathrm{Final}} = \mathrm{Mean}\left[F_{C\_\mathrm{CA}} \oplus F_{C\_\mathrm{SA}}\right] \tag{8}$$

where $\oplus$ denotes pixel-wise addition, and mean denotes the operation of averaging.

IV. Experiment

In this section, we will first introduce our evaluation metrics and implementation details. Then, we will compare our proposed network with several state-of-the-art SIRST detection methods. Finally, we present ablation experiments to study the effectiveness of our network.

A. Evaluation Metrics

The evaluation metrics commonly used in the field of ISTD with deep learning focus on pixel-level metrics from the segmentation domain, such as intersection over union (IoU), accuracy and recall. These metrics emphasize the evaluation of target shapes. However, infrared small targets often lack clear shapes and textures and the crucial aspect of ISTD is the ability to accurately locate the targets. Therefore, we choose IoU to evaluate shape description ability and use probability of detection (Pd) and Fa rate to assess the localization capability.

1) IoU: IoU is a pixel-level evaluation metric that assesses the contour description ability of an algorithm. It is calculated as the ratio of the area of the intersection between the predicted and ground truth regions to the area of their union

$$\mathrm{IoU} = \frac{A_{\mathrm{Intersection}}}{A_{\mathrm{Union}}}. \tag{9}$$

2) Pd: Pd is an object-level evaluation metric. It characterizes the ratio of correctly predicted objects to the total number of objects

$$P_d = \frac{T_{\mathrm{Correct}}}{T_{\mathrm{All}}}. \tag{10}$$

If the centroid deviation of a target is less than the predefined deviation threshold Dthresh, we consider these targets to be correctly predicted. In this article, we set the predefined deviation threshold to 3.

3) Fa: Fa is also a target-level evaluation metric. It is used to measure the ratio of incorrectly predicted pixels to all pixels in the image

$$F_a = \frac{P_{\mathrm{False}}}{P_{\mathrm{All}}}. \tag{11}$$

If the centroid deviation of a target is greater than the predefined deviation threshold, we consider those pixels to be incorrectly predicted. In this article, we set the predefined deviation threshold to 3.

TABLE I
QUANTITATIVE COMPARISON RESULTS (ON NUDT-SIRST AND SIRST), WHERE THE BEST VALUES ARE MARKED IN RED AND THE SECOND-BEST VALUES IN BLUE (ALL TABLES FOLLOW THIS RULE)

Method	NUDT-SIRST (Train and Test)			SIRST (Verify)		
	IoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	Fa ($\times 10^{-6}$)	IoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	Fa ($\times 10^{-6}$)
Top-Hat	19.87	72.55	187.9	7.882	70.25	1125
ALCNet	68.11	83.20	9.16	39.24	64.55	31.55
DNANet(Res18)	77.30	90.40	8.06	61.69	83.79	7.98
AMFU-Net	75.12	93.11	6.47	63.78	82.79	8.64
UIU-Net	78.25	91.22	5.12	63.22	84.99	7.16
SCTransNet	80.13	96.55	4.91	61.43	82.32	6.01
Ours	84.94	98.20	3.86	64.87	86.12	7.12

4) Receiver operation characteristics (ROC): ROC is used to describe the change trend of Pd at different Fa.

B. Implementation Details

We selected the NUDT-SIRST [26], SIRST [40], and NUST-SIRST [21] dataset. The dataset was split into 50% for training and 50% for testing, specifically using the NUDT-SIRST dataset for training and validating on the other two datasets. t should be noted that the SIRST dataset is a collection of real infrared small target images, including 427 images of different wavelengths of infrared, including infrared images at 950 nm wavelength. The NUDT-SIRST dataset contains 1327 images synthesized using real infrared targets, similar in style to the SIRST dataset. The NUST-SIRST dataset is also a synthetic dataset, but the style of small targets in it differs significantly from the other two datasets. Therefore, using it directly for testing is not rigorous. Hence, we trained and tested separately on the NUST-SIRST dataset. It is worth mentioning that preprocessing of input images is required before network training, including normalization, random flipping and cropping. The final resolution of input images is 256×256.

In this study, we trained our network using the Soft-IoU loss function and optimized it using the Adagrad method with a CosineAnnealingLR scheduler. We initialized the model's weights and biases using the Xavier method. The learning rate, batch size, and epoch size were set to 0.05, 16, and 1350, respectively. Pytorch was the deep learning framework employed, and the computational setup consisted of a 12th generation Intel (R) Xeon (R) Platinum 8255C CPU @ 2.50GHz and an Nvidia GeForce 3080 GPU.

C. Comparison to the State-of-the-Art Methods

To compare with other single-frame ISTD methods, we selected six existing state-of-the-art methods for comparison. The compared methods include one traditional algorithm, Top-Hat [48] and five deep learning-based methods, ALCNet [49], DNANet [26], AMFU-Net [50], UIU-Net [46], and SCTransNet [43]. It is worth noting that, due to the presence of significant noise in the test results of traditional methods, we employed a threshold segmentation method to denoise and obtain the position results of small targets as accurately as possible. The deep learning-based methods were trained and tested on the NUDT-SIRST dataset using the provided original code and default parameters and validation was conducted on the SIRST dataset.

TABLE II
QUANTITATIVE COMPARISON RESULTS (ON NUST-SIRST)

Method	NUST-SIRST (Train and Test)		
	IoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	Fa ($\times 10^{-6}$)
Top-Hat	16.36	70.21	165.23
ALCNet	61.45	80.17	10.74
DNANet(Res18)	78.12	93.12	6.89
AMFU-Net	78.61	91.46	7.50
UIU-Net	79.44	91.78	6.54
SCTransNet	80.56	94.36	7.21
Ours	82.22	95.32	5.32

1) Quantitative Results: The quantitative comparison between the proposed method and other approaches is presented in Table I. The results indicate that our method surpasses traditional techniques across various metrics. In comparison to current state-of-the-art deep learning methods, our approach achieves superior performance in all metrics on the NUDT-SIRST dataset. Furthermore, the strong validation results on the SIRST dataset highlight the robustness and generalization capability of the proposed model.

To further demonstrate the robustness and generalization of our proposed method, we conducted training and testing on the NUST-SIRST dataset (see Table II). The proposed method still achieves the best values across all metrics.

2) Qualitative Results: The qualitative results are shown in Figs. 7–10. It can be observed that traditional algorithms have some effect on the localization of small targets but suffer from high Fa rates and poor shape reconstruction.

We selected weak small target images from the NUST-SIRST and NUDT-SIRST datasets for testing. From the test results, it can be seen that the method proposed in this article not only accurately locates small targets but also achieves good shape reconstruction for them. This indicates that our model is capable of adapting to various complex backgrounds and tasks involving targets of different sizes and types. In contrast, other deep learning algorithms exhibited many missed detections and Fa under the influence of background clutter.

The validation results on the SIRST dataset also demonstrate acceptable performance of the proposed method, further confirming its outstanding generalization ability.

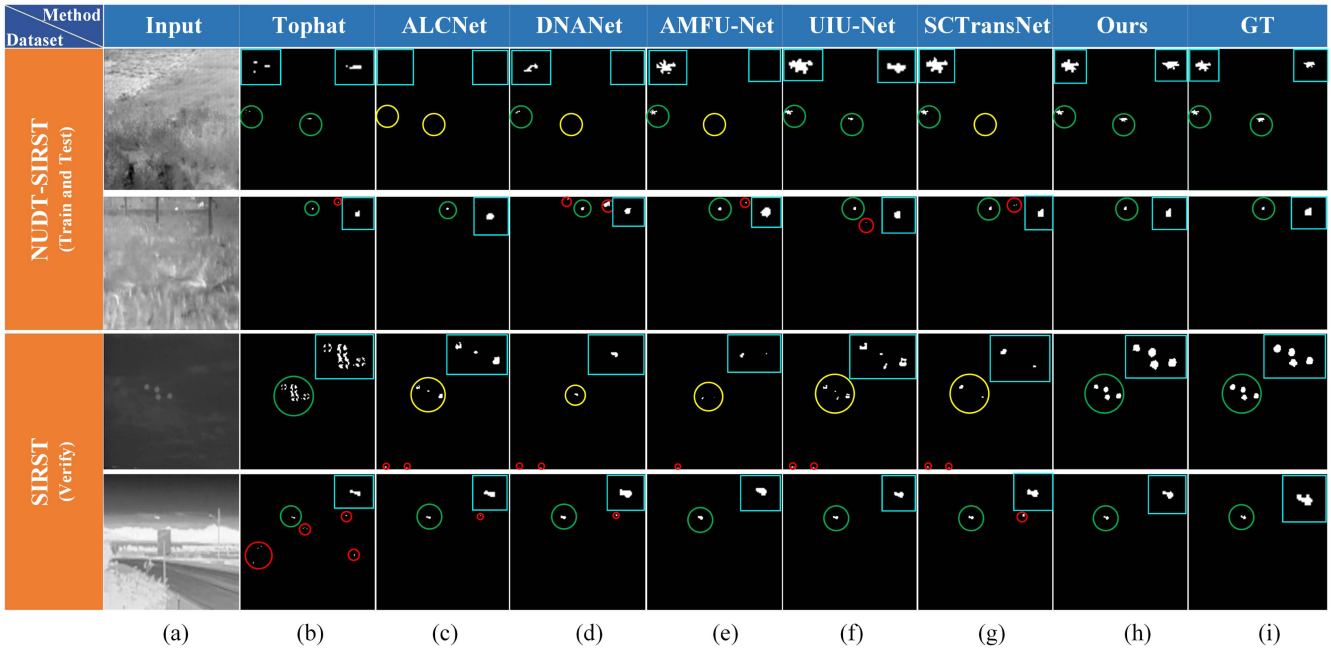

Fig. 7. Qualitative results of different detection methods trained and tested on the NUDT-SIRST dataset and validated on the SIRST dataset. For better visualization, enlarged views of the target areas are displayed on the side. Correctly detected targets, Fa and missed detection regions are highlighted with green, red, and yellow circles, respectively.

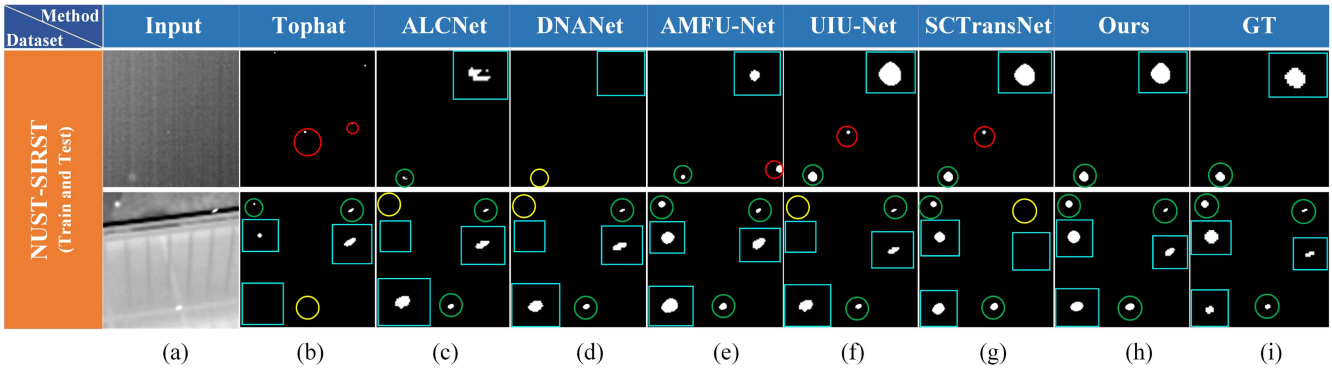

Fig. 8. Qualitative results obtained from training and testing different detection methods on the NUST-SIRST dataset. For enhanced visualization, enlarged regions of interest are displayed alongside. Correctly detected targets, Fa and missed detection regions are highlighted with green, red, and yellow bounding boxes, respectively.

In addition, we plotted ROC curves (see Fig. 11) to compare with five other deep learning-based single-frame ISTD algorithms. The larger the area under the ROC curve, the better the model's performance at different thresholds. It is evident from the graph that the curve corresponding to the proposed method has a larger area, proving the superior performance of our model.

To comprehensively demonstrate the network's performance, we also tested its inference speed and the overall size of the network model parameters, and compared the relevant metrics of partial models, as shown in Table III.

Further, to gain a clearer and more intuitive understanding of the specific training process of the network, we also present the process of loss reduction and mIOU improvement during network training. As shown in Fig. 12.

TABLE III
MODEL COMPARISON: PARAMETERS (M), FLOPS (G), AND INFERENCE TIME (S) PER IMAGE

Module	*Params (M)*	*Flops (G)*	*Inference Time (s)*
ALCNet	1.48	9.74	0.021
DNANet	4.7	14.26	0.18
AMFU-Net	2.17	5.42	0.12
UIU-Net	50.54	54.42	0.98
SCTransNet	11.19	20.24	0.52
MBFANet(Ours)	4.3	11.95	0.22

D. Ablation Study

To ensure the effectiveness and rationality of each module in the proposed method, we specifically designed four ablation

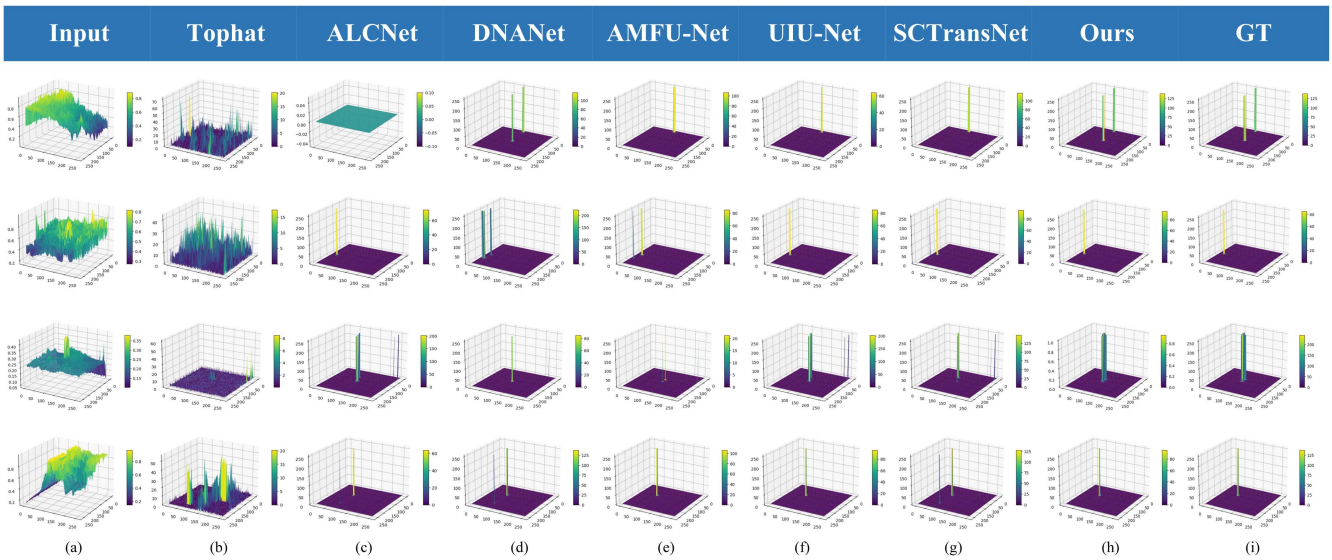

Fig. 9. 3-D visualization results for different methods on four test images [(1) and (2) are from the NUDT-SIRST dataset and (3) and (4) are from the SIRST dataset].

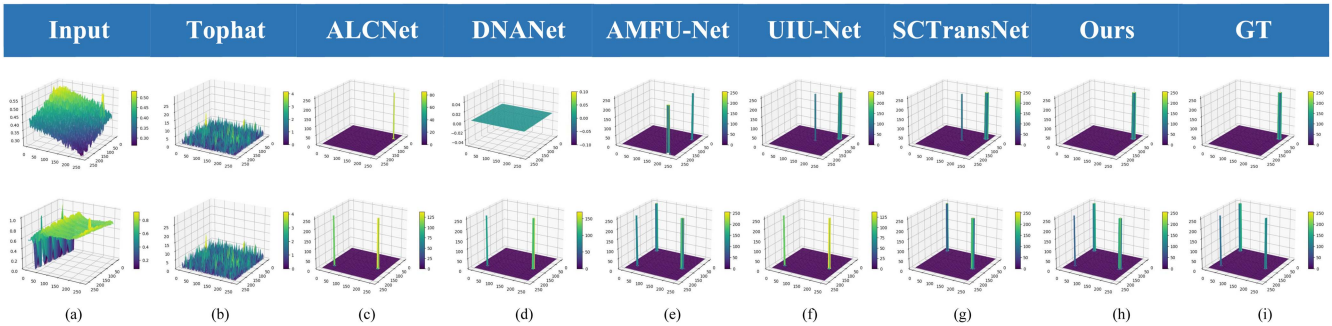

Fig. 10. 3-D visualization results for different methods on 2 test images [(1) and (2) are from the NUST-SIRST dataset].

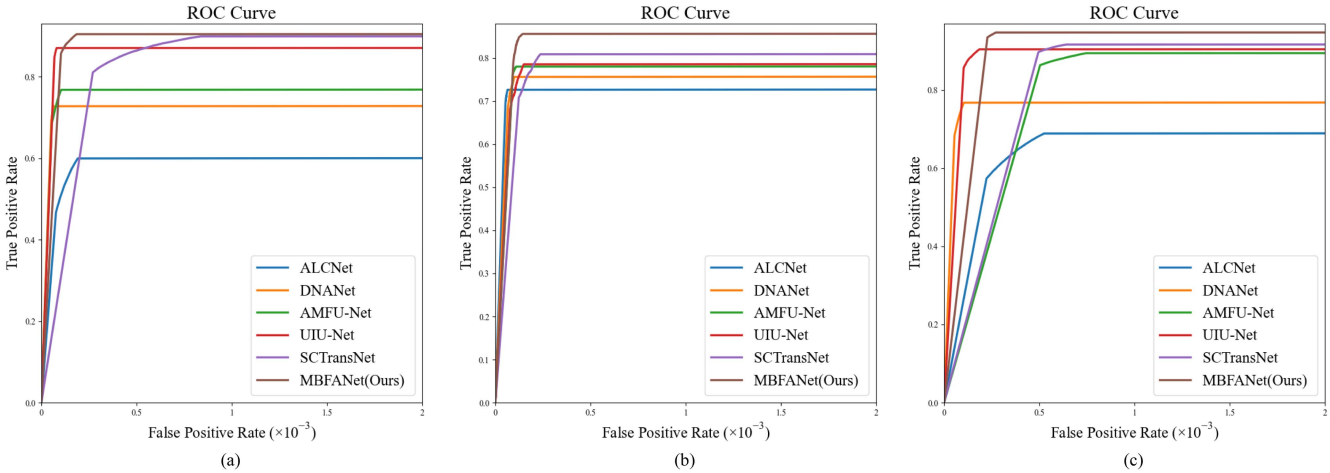

Fig. 11. ROC curves of various deep learning-based single-frame ISTD methods, where (a) results trained and tested on the NUDT-SIRST dataset, (b) results directly validated on the SIRST dataset, and (c) results trained and tested on the NUDT-SIRST dataset.

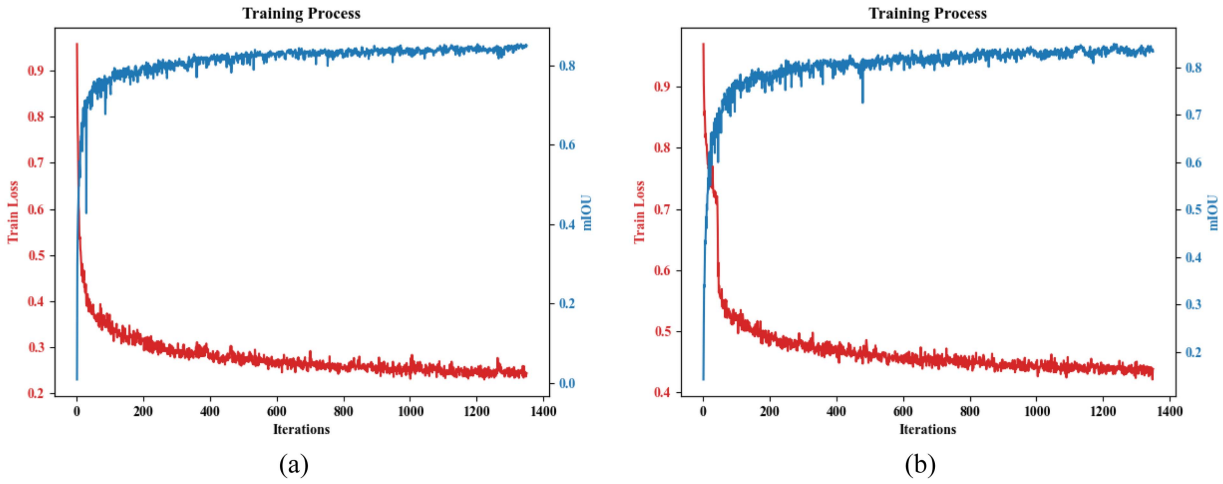

Fig. 12. Line charts depicting the changes in loss value and mIOU during the training process. (a) Results trained on the NUDT-SIRST dataset, while (b) results trained on the NUST-SIRST dataset.

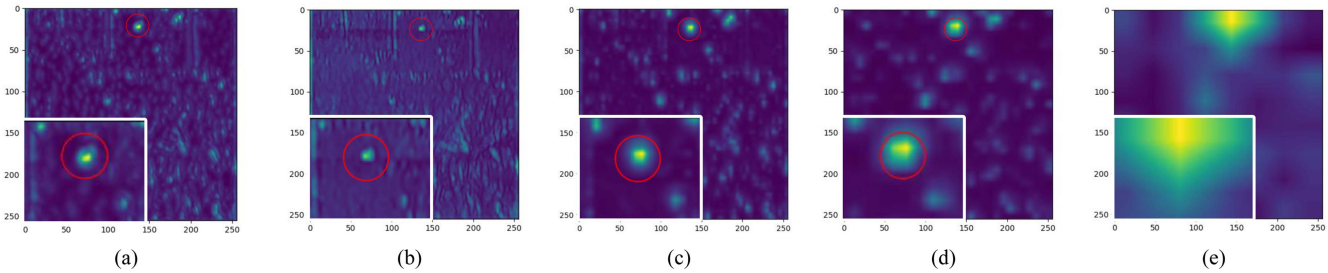

Fig. 13. Input front feature maps of each auxiliary branch L_1, L_2, L_3, and L_4 correspond to (a), (b), (c), and (d). The feature map after another pooling operation corresponds to (e).

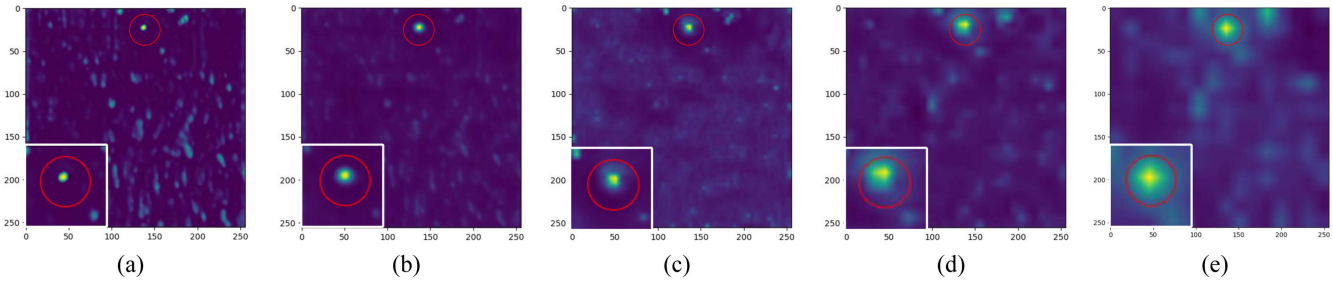

Fig. 14. Output feature maps of each branch L_0, L_1, L_2, L_3, and L_4 correspond to (a), (b), (c), (d), and (e).

experiments for verification. These experiments aim to replace different modules within the method to validate the proposed approach's rationality and effectiveness.

1) Research on the Feature Extraction Structure: Since the inputs of the auxiliary branches are pooled from the output of the first block of the previous branch and we know that pooling (downsampling) is usually not conducive to small object detection tasks, we visualized the feature maps before the pooling operation for the four auxiliary branches (as shown in Fig. 13). It can be observed that despite the pooling operation, the feature maps before pooling still retain the position information and partial shape information of small objects, as well as suppressed background information. These feature maps, after passing through each auxiliary branch, can provide the main branch with semantic features at different scales. To illustrate the rationality of the auxiliary branch design, we also observed the feature maps after pooling again [see Fig. 13(e)]. It can be seen that if pooled again, the feature map almost completely loses its shape and position information, rendering it useless.

Furthermore, to illustrate the contributions of the main branch and the four auxiliary branches, we observed their feature maps (as shown in Fig. 14). It can be seen that the main branch accurately indicates the position and shape information of small objects, while the auxiliary branches help suppress a large

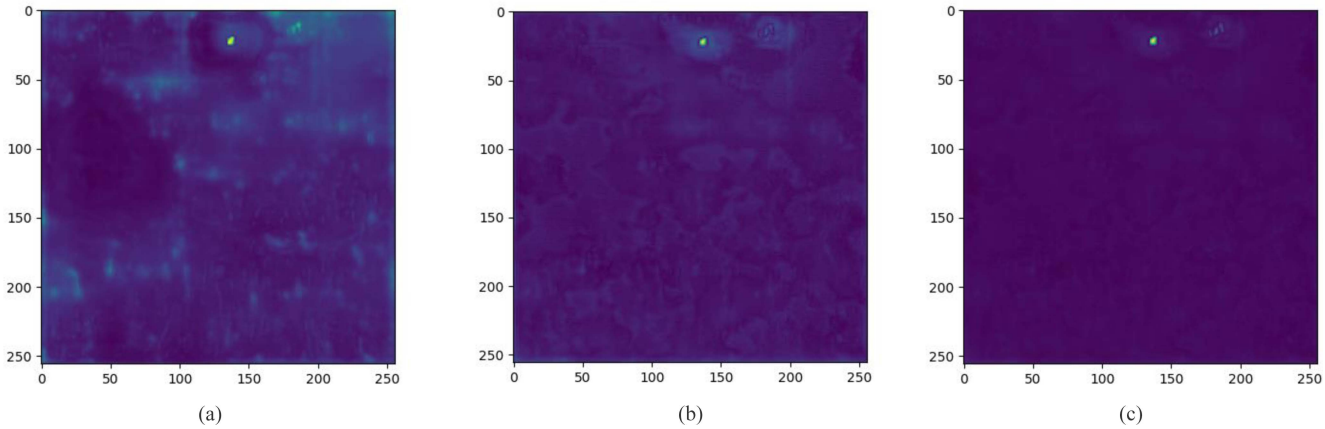

Fig. 15. (a) Feature map before the attention module. (b) Feature map after the CBAM attention module. (c) Feature map after the JAM attention module.

TABLE IV
QUANTITATIVE RESULTS FOR THREE ATTENTION MODULES: NONE, CBAM, AND JAM (OURS)

Attention Module	*IoU* ($\times 10^{-2}$)	*Pd* ($\times 10^{-2}$)	*Fa* ($\times 10^{-6}$)
None	78.12	97.65	7.12
CBAM	80.89	95.88	5.12
JAM (Ours)	84.94	98.20	3.10

amount of background redundant information and provide a small amount of position and shape information for small objects.

2) Research on the JAM: To demonstrate the effectiveness of this module, we designed two sets of variants: one without the JAM module and another with the JAM module replaced by the channel and spatial attention module (CBAM), which is another classic attention module. Visualizations of the feature maps before the attention module are presented in Fig. 15(a), showing some background noise. After applying the CBAM attention module [see Fig. 15(b)], the background noise is partially suppressed, and the enhancement of small targets is not significant. However, with the JAM attention module [see Fig. 15(c)], the background noise is noticeably weakened and the small targets are significantly enhanced.

To further illustrate, we conducted a comparison of three network variants (without attention module, with CBAM attention module and with JAM attention module) and the quantitative results are presented in Table IV. It is evident from the results that our JAM attention module significantly enhances the performance of object detection.

3) Research on the DCM: DCM promptly corrects feature maps that incorrectly extract small targets, facilitating successful extraction of small targets in subsequent feature extraction processes. To demonstrate the effectiveness of this module, we compared the detection results of MBFA-Net without DCM and MBFA-Net with DCM. The comparative results are shown in Fig 16. In comparison, the network with DCM can more

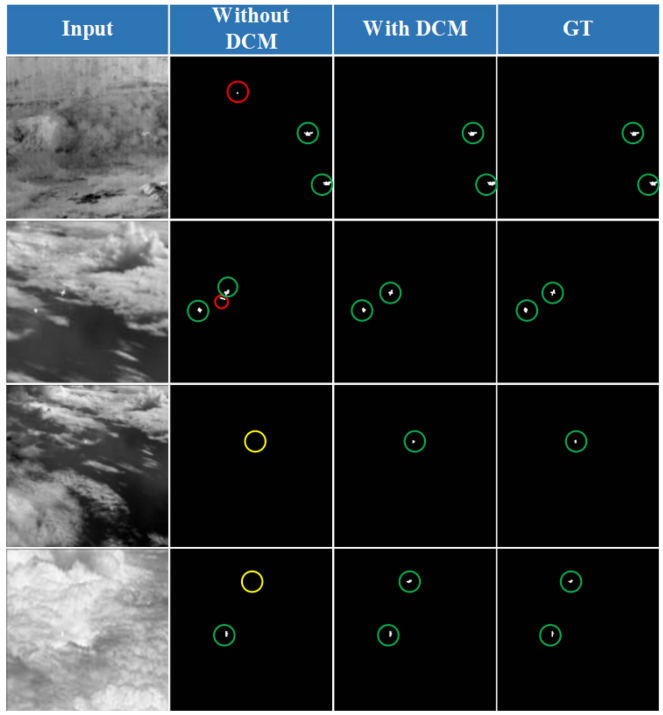

Fig. 16. Comparison of the effects with and without DCM.

TABLE V
NETWORK PERFORMANCE COMPARISON: MODELS WITH AND WITHOUT DCM

	IoU ($\times 10^{-2}$)	*Pd* ($\times 10^{-2}$)	*Fa* ($\times 10^{-6}$)
MBFANet w/o DCM	83.62	98.31	3.86
MBFANet w/ DCM	84.94	98.20	3.10

sensitively detect concealed small targets and suppress interference from certain bright noise, thereby effectively reducing Fa rates and missed detection rates. In addition, we compared the quantitative results under both conditions, as shown in Table V. It can be observed that the network with DCM exhibits superior performance in terms of metrics.

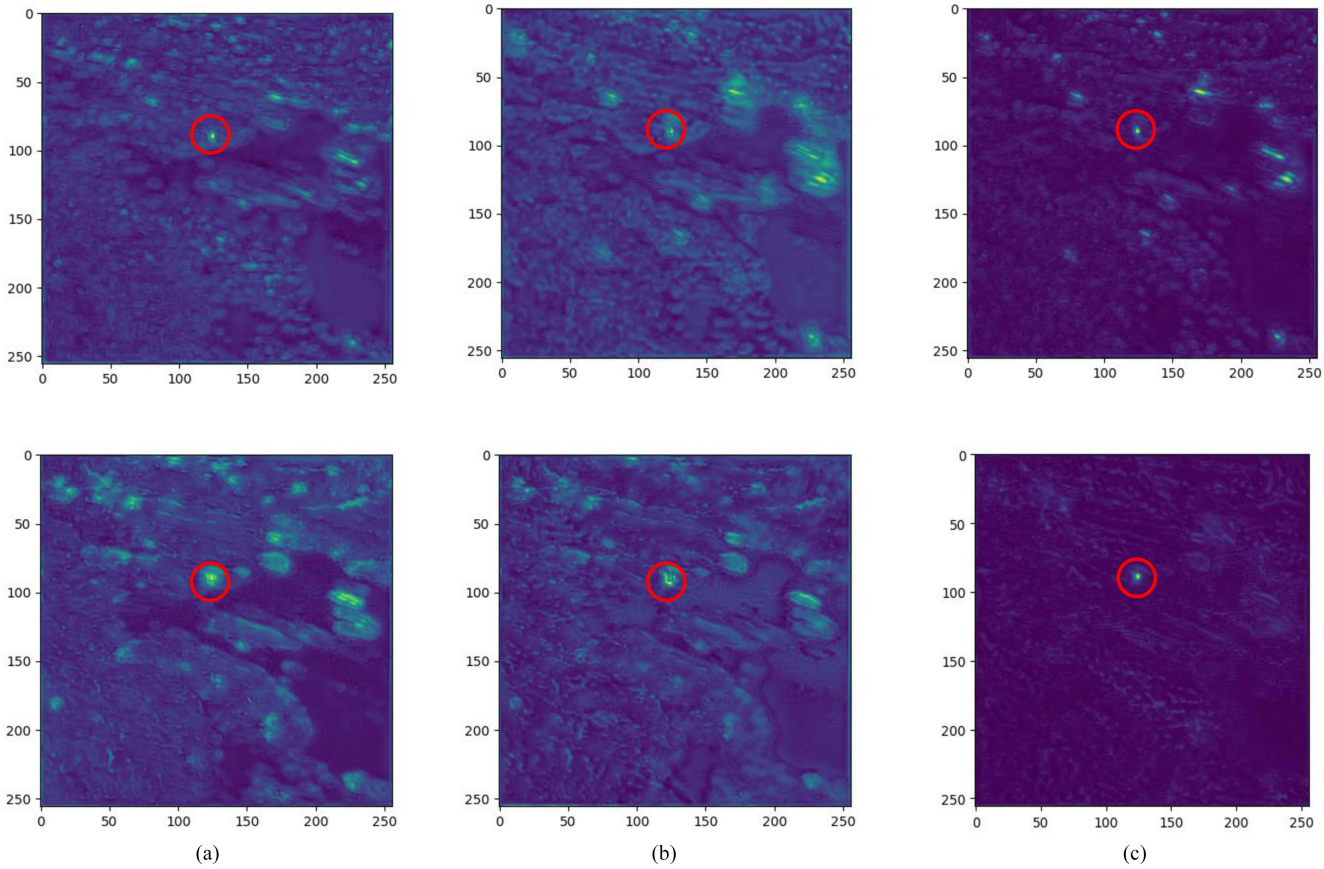

Fig. 17. Comparison of the output feature maps of (a) third, (b) fourth, and (c) fifth RSEMB in the main branch (the first row represents the results without DCM, and the second row represents the results with DCM).

To further illustrate the effectiveness of DCM, we visualized the output feature maps of the third, fourth, and fifth RSEMB in the main branch and compared them with the output feature maps of these three RSEMB without DCM (as shown in Figs. 17 and 18). In Fig. 17, for instance, the target in the output feature map of the third RSEMB is significantly enhanced, although some background clutter is also amplified, the target is more prominently highlighted. Consequently, after subsequent convolutional blocks' feature extraction (especially the last one), the concealed small targets are completely extracted, whereas the results without DCM indicate that some background is erroneously amplified, leading to the failure of small target reconstruction. Similarly, upon observing the results in Fig. 18, although in both cases, the two small targets in the test image are extracted, there are bright interference points near the small targets, which would cause Fa in the detection results of the network without DCM. Conversely, the network with DCM can successfully distinguish between targets and background, accurately locating the targets.

4) Research on the Feature Extraction Blocks (RSEMB): To demonstrate the effectiveness of this design, We designed two sets of variants: one where the SE module is removed (reverting to the original Res18 block) and another where the SE module is replaced with the SAM [38]. The SAM is a lightweight spatial attention mechanism that differs from channel attention, aiming to compare the impact of processing in the channel and spatial dimensions on small target detection. The quantitative results comparison in Table VI shows that the network metrics of the improved block with the added SE attention module are higher than the other two variants. This indicates that the improvement in small target detection is superior to both the unimproved block and the improved block with the added SAM attention module. This also suggests a perspective: in the feature extraction process of ISTD tasks, emphasizing processing in the channel dimension may lead to better detection results.

TABLE VI
QUANTITATIVE RESULTS FOR THREE NETWORK VARIANTS: WITHOUT IMPROVED BLOCK, WITH SAM, AND WITH SE

Block	IoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	Fa ($\times 10^{-6}$)
Res18	74.65	94.89	6.11
Res18_SAM	83.21	96.64	4.16
Res18_SE(Ours)	84.94	98.20	3.10

V. DISCUSSION

To address the issue of small targets being easily lost during the feature extraction process, this article creatively proposes a

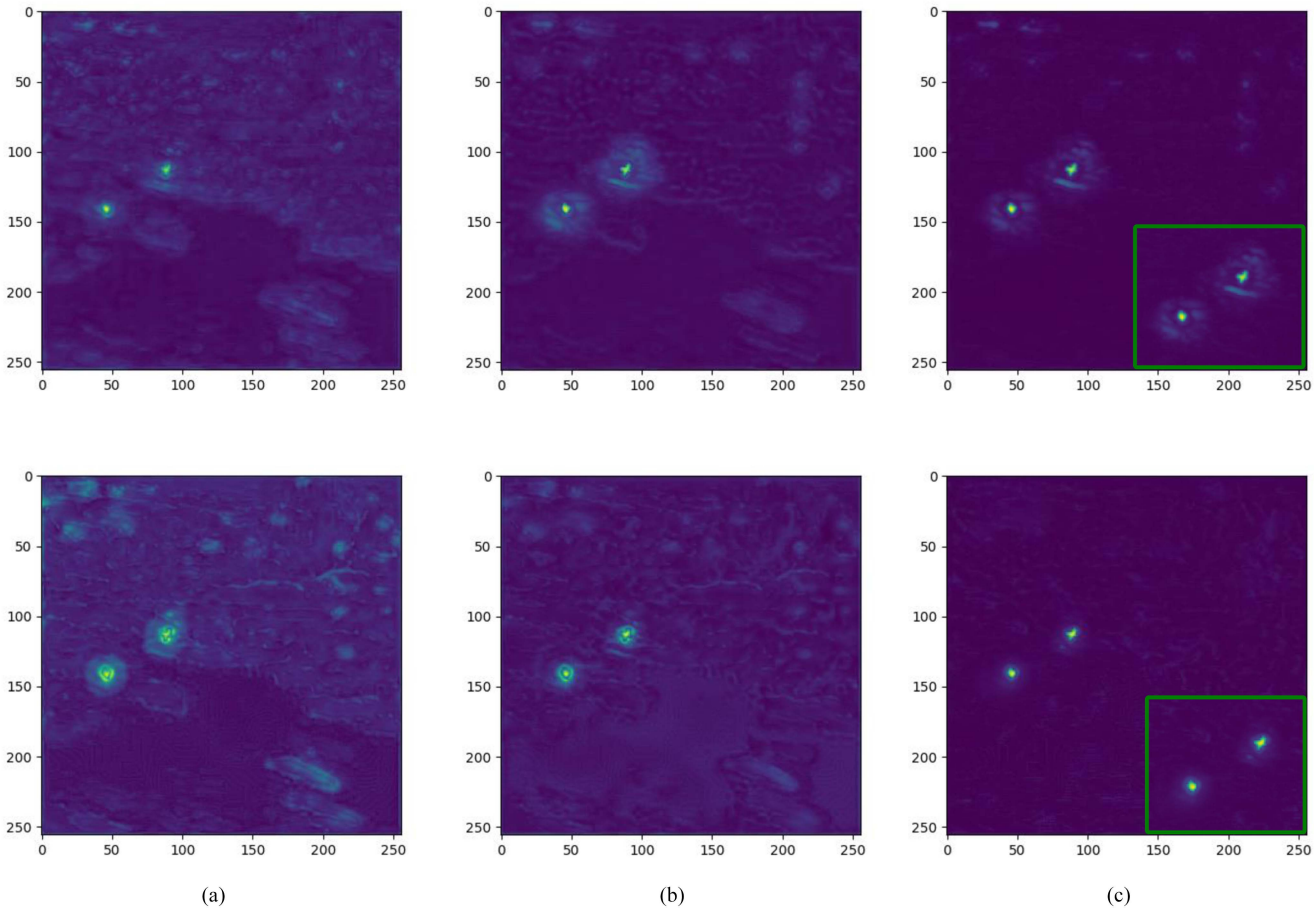

Fig. 18. Comparison of the output feature maps of (a) third, (b) fourth, and (c) fifth RSEMB in the main branch (the first row represents the results without DCM, and the second row represents the results with DCM).

differential correction approach. This will inspire researchers in this field, potentially leading to the emergence of new methods based on this approach in the future. In subsequent studies, we will attempt to apply the proposed differential correction approach to other target detection networks to validate its effectiveness and generalization. In addition, we will focus on improving the network inference speed, as it is of practical significance. Techniques, such as model distillation or other lightweight methods, can be explored for this purpose.

VI. Conclusion

In this article, we propose an MBFANet for ISTD. Unlike the common approach in deep learning methods of increasing network depth vertically to enhance feature extraction capability, we design a main-secondary multibranch structure to horizontally broaden network depth. This progressive supplementation of small target information from feature maps of different scales and depths into the main branch effectively improves detection performance while preserving small targets. To enhance the network's sensitivity to concealed small targets, we introduce a DCM to effectively reduce Fa rates and missed detection rates, offering a new approach to improving ISTD performance. In addition, we design a JAM to enhance the network's ability to localize small targets and reconstruct their shapes. Experimental results on three public datasets demonstrate that our method outperforms state-of-the-art methods.

References

[1] Y. Sun, J. Yang, and W. An, "Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3737–3752, May 2021.

[2] D. Wu, L. Cao, P. Zhou, N. Li, Y. Li, and D. Wang, "Infrared small-target detection based on radiation characteristics with a multimodal feature fusion network," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3570.

[3] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 91–124, Mar. 2022.

[4] 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.

[5] T.-W. Bae, F. Zhang, and I.-S. Kweon, "Edge directional 2D LMS filter for infrared small target detection," *Infrared Phys. Technol.*, vol. 55, no. 1, pp. 137–145, 2012.

[6] T. Ma, Z. Yang, J. Wang, S. Sun, X. Ren, and U. Ahmad, "Infrared small target detection network with generate label and feature mapping," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6505405.

[7] T. R. Goodall, A. C. Bovik, and N. G. Paulter, "Tasking on natural statistics of infrared images," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 65–79, Jan. 2016, doi: 10.1109/TIP.2015.2496289.

[8] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges and perspectives," *Innovation*, vol. 5, no. 5, 2024, Art. no. 100691.

[9] X. Shao, H. Fan, G. Lu, and J. Xu, "An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system," *Infrared Phys. Technol.*, vol. 55, no. 5, pp. 403–408, 2012.
[10] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4204–4214, Jul. 2016.
[11] J.-F. Rivest and R. Fortin, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, pp. 1886–1893, 1996.
[12] 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. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.
[13] C. 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.
[14] 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.
[15] S. Kim and J. Lee, "Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track," *Pattern Recognit.*, vol. 45, no. 1, pp. 393–406, 2012.
[16] X. Wang, G. Lv, and L. Xu, "Infrared dim target detection based on visual attention," *Infrared Phys. Technol.*, vol. 55, no. 6, pp. 513–521, 2012.
[17] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, "Infrared small target detection via non-convex rank approximation minimization joint l 2, 1 norm," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1821.
[18] 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.
[19] H. Zhu, S. Liu, L. Deng, Y. Li, and F. Xiao, "Infrared small target detection via low-rank tensor completion with top-hat regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1004–1016, Feb. 2020.
[20] Y. Dai, Y. Wu, Y. Song, and J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infrared Phys. Technol.*, vol. 81, pp. 182–194, 2017.
[21] 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.
[22] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[23] 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.
[24] C. Yu et al., "Pay attention to local contrast learning networks for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2022, Art. no. 3512705, doi: 10.1109/LGRS.2022.3178984.
[25] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, "EAAU-Net: Enhanced asymmetric attention U-Net for infrared small target detection," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3200.
[26] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, Aug. 2023, doi: 10.1109/TIP.2022.3199107.
[27] F. Chen et al., "Local patch network with global attention for infrared small target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 5, pp. 3979–3991, Oct. 2022.
[28] M. Shi and H. Wang, "Infrared dim and small target detection based on denoising autoencoder network," *Mobile Netw. Appl.*, vol. 25, pp. 1469–1483, 2020.
[29] R. Kou et al., "Infrared small target segmentation networks: A survey," *Pattern Recognit.*, vol. 143, 2023, Art. no. 109788.
[30] R. Kou et al., "LW-IRSTNet: Lightweight infrared small target segmentation network and application deployment," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 5621313, doi: 10.1109/TGRS.2023.3314586.
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[32] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 11534–11542.
[33] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 783–792.
[34] H. Zhao et al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283.
[35] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[36] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, *arXiv:1906.05909*.
[37] D. Zhou et al., "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.
[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[39] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Intervention–MICCAI 2018, 21st Int. Conf.*, Granada, Spain, Sep. 16–20, 2018, pp. 421–429.
[40] 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.
[41] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," in *IEEE Trans. Aerosp. Elect. Syst.*, vol. 59, no. 4, pp. 4250–4261, Aug. 2023, doi: 10.1109/TAES.2023.3238703.
[42] Y. Zhang, B. Nian, Y. Zhang, Y. Zhang, and F. Ling, "Lightweight multimechanism deep feature enhancement network for infrared small-target detection," *Remote Sens.*, vol. 14, no. 24, 2022, Art. no. 6278.
[43] S. Yuan, H. Qin, X. Yan, N. Akhtar, and A. Mian, "SCTransNet: Spatial-channel cross transformer network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Apr. 2024, Art. no. 5002615, doi: 10.1109/TGRS.2024.3383649.
[44] Z. Wang, J. Yang, Z. Pan, Y. Liu, B. Lei, and Y. Hu, "APAFNet: Single-frame infrared small target detection by asymmetric patch attention fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Dec. 2022, Art. no. 7000405, doi: 10.1109/LGRS.2022.3230415.
[45] B. Nian, B. Jiang, H. Shi, and Y. Zhang, "Local contrast attention guide network for detecting infrared small targets," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5607513, doi: 10.1109/TGRS.2023.3266447.
[46] 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, Dec. 2022, doi: 10.1109/TIP.2022.3228497.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[48] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognit.*, vol. 43, no. 6, pp. 2145–2156, 2010.
[49] 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.
[50] W. Y. Chung, I. H. Lee, and C. G. Park, "Lightweight infrared small target detection network using full-scale skip connection U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, May 2023, Art. no. 7000705, doi: 10.1109/LGRS.2023.3276326.

Ziqiang Hao received the Ph.D. degree in optical engineering from the Changchun University of Science and Technology (CUST), Changchun, China.

He is currently an Associate Professor with the National Demonstration Center for Experimental Electrical, School of Electronic and Information Engineering, CUST. His research interests include medical image processing, infrared target detection, and tracking.

Zheng Jiang received the B.S. degree in electronic information engineering from Yanbian University, Yanbian, China, in 2022. He is currently working toward the M.S. degree in control engineering with Changchun University of Science and Technology, Changchun, China. His research interests include infrared target detection and semantic segmentation with a focus on infrared small target detection.

Xiaoyu Xu received the B.S. degree in electronic information engineering in 2019 from the Changchun University of Science and Technology, Changchun, China, where he is currently working toward the Ph.D. degree in information and communication engineering.

His research interests include infrared small target detection, infrared pedestrian detection, and target data enhancement.

Zhuohao Wang received the B.S. degree in optoelectronic information science and engineering in 2022 from the Changchun University of Science and Technology, Changchun, China, where he is currently working toward the M.S. degree in control science and engineering.

His research interests include infrared target detection and semantic segmentation with a focus on infrared small target detection.

Zhicheng Sun received the B.S. degree in electronic information engineering in 2022 from the Changchun University of Science and Technology, Changchun, China, in 2022, where he is currently working toward the M.S. degree in control engineering.

His research interests include target detection and pedestrian recognition.