Infrared Small Target Detection Based on Prior Guided Dense Nested Network

Chang Liu, Xuedong Song, Dianyu Yu, Linwei Qiu, Fengying Xie, *Member, IEEE*,
Yue Zi, and Zhenwei Shi, *Senior Member, IEEE*

***Abstract*— Infrared small target detection (IRSTD) has been widely applied and developed in military and civilian fields, playing a vital role. Despite the extensive research foundation of traditional manual feature-based methods, they are still constrained by the inherent problem of infrared small targets lacking prior features. In recent years, the advancement of deep learning methods has enriched the research landscape in this field, yet they are still constrained by the imbalance of positive and negative samples between the target and the background. To address these issues, we propose a novel prior guided dense nested network (PGDN-Net), which ingeniously integrates traditional manual features with a deep learning network model. First, three prior features are extracted, including the high-order Riesz transform feature, the compactness and heterogeneity feature (CH), and the corner feature of the structure tensor (ST). Then, these features are input into a dense nested network for guidance, supported by a two-orientation attention aggregation module and a channel and spatial attention module. Different features play their respective guiding roles in different depths of the network. Through multiple attention mechanisms and feature fusion operations on the interested target area, the extraction and preservation of target features can be improved, while easily removing irrelevant backgrounds. Experiments on public datasets demonstrate the effectiveness and progressiveness of our PGDN-Net. Compared with other state-of-the-art methods, it achieves better performance in background suppression, target enhancement, probability of detection, and false alarm rate. In addition, the PGDN-Net model can effectively maintain and restore the original shape of the target while performing robust detection, which is beneficial for subsequent fine-grained recognition tasks.**

Received 14 September 2024; revised 18 January 2025 and 10 February 2025; accepted 11 February 2025. Date of publication 20 February 2025; date of current version 27 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62475006 and Grant 62125102 and in part by the National Key Research and Development Program of China under Grant 2022ZD0160401. *(Corresponding author: Fengying Xie.)*

Chang Liu is with the Tianmushan Laboratory, Hangzhou 311115, China, also with the Department of Aerospace Information Engineering, School of Astronautics, Beihang University, Beijing 100191, China, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: by1915010@buaa.edu.cn).

Xuedong Song is with Shanghai Institute of Satellite Engineering, Shanghai 201109, China (e-mail: owensong96@163.com).

Dianyu Yu, Linwei Qiu, and Fengying Xie are with the Tianmushan Laboratory, Hangzhou 311115, China, and also with the Department of Aerospace Information Engineering, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: tcdd20010521@buaa.edu.cn; qiulinwei@buaa.edu.cn; xfy_73@buaa.edu.cn).

Yue Zi is with the School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: ziyue91@csust.edu.cn).

Zhenwei Shi is with the Department of Aerospace Information Engineering, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3542232

***Index Terms*— Deep learning, dense nested network, high-order Riesz transform, infrared small target detection (IRSTD), traditional manual feature.**

I. Introduction

INFRARED small target detection (IRSTD) is currently widely used in military early warning [1], search and tracking [2], guidance and anti-missile [3], remote sensing, surveillance, and reconnaissance [4], among others. It also plays an important role in the civilian field [5], such as search and rescue of maritime personnel, warning of forest fires, and so on. Due to the long distance of detection, small targets in infrared images usually occupy a few pixels and exhibit minimal shape or texture characteristics [6]. Furthermore, various factors like hardware system errors and disturbances from atmospheric turbulence lead to a reduced signal-to-noise ratio (SNR) in the infrared image, potentially causing the target to be obscured by background clutter. In the early years, the Society of Photo-Optical Instrumentation Engineers (SPIE) defined small targets as those with image sizes smaller than 81 pixels (9 × 9) [7]. This type of target typically exhibits strong central infrared characteristics and radiates attenuation in all directions [8].

The current challenges faced by IRSTD can be summarized into the following three points.

1) *Lack of prior features of the target:* The long transmission distances result in smaller target sizes, with a lack of texture and shape features, making traditional target detection methods unsuitable for detecting such point targets [9].
2) *High background complexity:* The background in infrared images is complex and varied. In strong noise or high brightness backgrounds, targets are easily submerged or disturbed, and sometimes even become invisible, which poses a great challenge to the detection performance and multiscene adaptability of the detectors [10].
3) *Imbalance of positive and negative sample information:* Background information occupies the vast majority of pixels in infrared images, and the proportion of pixels for infrared small targets is extremely low. Therefore, there is an extreme imbalance of information between the positive and negative samples of the target and background, which can easily lead to false alarms in detection [11].

The field of IRSTD has evolved over several decades, with experts and scholars consistently proposing and refining solutions to address the aforementioned challenges, aiming to improve detection performance. The IRSTD methods can be divided into two categories [12]: sequence-frame detection method and single-frame detection method. The sequence-frame detection method can obtain satisfactory detection results by extracting inter-frame feature, while existing methods have richer and more diverse extraction of inter-frame features. The triple-domain strategy (Tridos) method [13] extracts infrared target feature learning comprehensively in spatiotemporal-frequency domains, while a sliced spatio-temporal network (SSTNet) method [14] explores the cross-slice spatio-temporal motion modeling for infrared dim-small targets. However, due to the advantages of low data demand, high computational efficiency, and strong algorithm flexibility in single-frame detection technologies, the single-frame detection method has received much attention in recent years. By using nonlinear filtering [15], morphological operations [16], and other techniques to focus on background information, the infrared background can be predicted, and small targets can be detected and extracted through difference calculation. There are also many methods from the perspective of the human visual system (HVS) that enhance the local salient features of the target based on measures such as contrast measurement [17], saliency enhancement [18], and image information entropy [19], while suppressing the interference from clutter and noise in the background to achieve robust IRSTD. In addition, some methods utilize the structural information of images and transform the detection task into an optimization problem of matrix decomposition [20], [21]. On the one hand, it can restore the low-rank background matrix, and on the other hand, it can also restore the sparse target matrix, ultimately achieving the separation of both matrices and facilitating target detection. These methods are all traditional model-driven methods, and although they have lower computational complexity, most of them rely heavily on manual feature settings. In recent years, data-driven deep learning methods have been widely applied in the field of image processing owing to the powerful feature extraction ability of convolutional neural networks. Several studies on IRSTD, based on Nash equilibrium [22], [23], task transformation [24], [25], attention and semantic features [26], [27], multiscale information [28], [29], dense nested structures [30], [31], and multisupervision mechanisms [32], [33], have achieved satisfactory results. However, the problem of imbalanced positive and negative sample information in infrared images remains a significant challenge for deep learning methods. The loss of target features in deep networks and the false detection of interfering elements in the background significantly degrade detection performance.

To address the current challenges in IRSTD and alleviate the shortcomings of traditional and deep learning methods, we propose a prior guided dense nested network (PGDN-Net) model, which combines traditional prior feature information with a deep learning model to achieve robust and accurate IRSTD. Fig. 1 shows the comparison of the different typical methods and the proposed method. On the one hand, traditional prior features can guide the focus of the network, effectively preserving target feature information within deep networks. On the other hand, dense nested network structures can achieve feature extraction of targets at multiple scales, enabling progressive feature interaction and enhancement. The proposed PGDN-Net model can combine the advantages of traditional and deep learning methods to improve detection performance. Specifically, we introduce the high-order Riesz transform features, compactness and heterogeneity features (CHs), and corner features of the structure tensor (ST) of the target into a dense nested network, and combine two-orientation attention aggregation module and channel and spatial hybrid attention module to complete feature extraction and segmentation detection of the target.

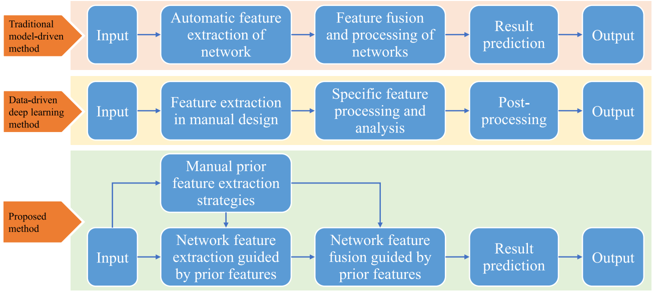
Traditional model-driven method
Input
Automatic feature extraction of network
Feature fusion and processing of networks
Result prediction
Output
Data-driven deep learning method
Input
Feature extraction in manual design
Specific feature processing and analysis
Post-processing
Output
Proposed method
Manual prior feature extraction strategies
Input
Network feature extraction guided by prior features
Network feature fusion guided by prior features
Result prediction
Output

Fig. 1. Scheme comparison diagram of the traditional model-driven method, the data-driven deep learning method, and the proposed method for IRSTD.

In this article, we propose a novel PGDN-Net to achieve IRSTD. The main contributions of the proposed method can be summarized as follows.

1) To explore the potential features of infrared small targets, we propose a method for extracting high-order Riesz features (HRs). This method is based on the Riesz transform theory and effectively extracts the target's local feature information using corresponding convolutional templates.
2) To address the challenges of limited prior features, high background complexity, and imbalanced positive and negative samples, we propose the PGDN-Net, which combines traditional and deep learning methods and guides the dense nested network using prior features. This method has demonstrated superior performance in various indicators.
3) To effectively integrate manually extracted features with the deep learning network, multiple attention mechanisms are introduced to preserve and focus on the features of the target area at each layer of the PGDN-Net. This approach ensures that the output not only detects the target but also accurately reflects its original shape.

The remainder of this article is organized as follows. Section II gives a brief review of related work on IRSTD methods. In Section III, we introduce the architecture of our PGDN-Net model. In Section IV, we conduct experiments on comparison with state-of-the-art methods, as well as experiments on the ablation of our method. At last, the conclusion of this article is presented in Section V.

II. Related Work

Based on the feature information used in IRSTD, the detection method can be classified into four categories: 1) background feature-based method; 2) target feature-based method; 3) low-rank and sparse matrix decomposition-based method; and 4) deep learning-based methods. This section provides a concise overview of the relevant research on these four types of detection methods.

A. Background Feature-Based Method

Background feature-based methods primarily utilize the continuity and similarity features of the background as the main basis for the design of the detection method. These methods typically begin by filtering or smoothing out target information unrelated to the background to predict the background information. Then, they use the difference operation between the original image and the predicted background image to achieve rough detection of infrared small targets. Subsequently, threshold suppression is used to suppress the background of the infrared image, achieving the purpose of detecting and extracting targets. Representative techniques include Max Mean and Max Median nonlinear filtering [15], 2-D least-mean-square (TDLMS) filtering [34], Top Hat morphological filtering [16], bilateral filtering [35], etc. While these methods perform well in flat backgrounds, they exhibit poor robustness in complex backgrounds and are highly susceptible to strong edges and noise interference, resulting in false alarms and missed detections.

B. Target Feature-Based Method

Target feature-based methods primarily utilize the local saliency features of the target as the prior information for detection. Although this prior information is relatively scarce, potential feature information can still be mined based on this prior knowledge to enhance the target's saliency while suppressing background clutter and noise. After the classic local contrast measurement (LCM) method [17] was first proposed, a large number of optimized and variant versions have been derived, such as improved LCM (ILCM) [7], relative LCM (RLCM) [36], tri-layer LCM (TLLCM) [37], and local energy factor (LEF) [38]. The saliency filtering enhancement strategy uses the pixel value distribution characteristics of infrared small targets with high brightness at the center and attenuated radiation in all directions. By considering pixel intensity and gradient within the target area, local window filters have been designed to enhance the target and suppress the background, including methods like fast saliency [18], novel local contrast descriptor (NLCD) [39], fast adaptive masking and scaling with iterative segmentation (FAMSIS) [40], and variance difference (VARD) [41]. The grayscale difference between small target areas and their surrounding background areas in infrared images makes image entropy an effective information for IRSTD, such as multiscale local gray dynamic range (MLGDR) [42], derivative entropy-based contrast measure (DECM) [43], and multidirectional local difference measure weighted by entropy (MDLDE) [19]. In this type of method, effective detectors can be designed based on the local salient features of the target, leading to good generalization ability. However, due to limited prior feature information of the target, some false alarms may be detected in the background.

C. Low-Rank and Sparse Matrix Decomposition-Based Method

The method based on low-rank and sparse matrix decomposition transforms the detection task into an optimization problem involving the decomposition of a low-rank background matrix and a sparse target matrix. Based on constraints on background components, various effective strategies have emerged, such as infrared patch-image (IPI) model [20], non-negative infrared patch-image model based on partial sum minimization of singular values (NIPPS) [44], non-convex rank approximation minimization (NRAM) [45], and the edge and corner awareness-based spatial–temporal tensor (ECA-STT) [46]. Similarly, constraints on the target component have also been introduced into some innovative methods, such as low-rank and sparse representation (LRSR) [47], reweighted infrared patch-tensor model (RIPT) [21], tensor creation, and Tucker decomposition (TCTD) [48]. While this method effectively decomposes the background and target autonomously, it sometimes inaccurately categorizes background clutter and corners as sparse target components. In addition, the construction of tensor blocks and the iterative optimization of matrix decomposition decreases computational efficiency, leading to longer execution periods and hindering the achievement of real-time demands.

D. Deep Learning-Based Method

Data-driven deep learning methods place significant demands on the quantity, quality, and richness of the datasets. However, IRSTD datasets are limited, with the amount and variety of publicly available samples far inferior to those of other image types. Despite these challenges, deep learning methods have achieved remarkable progress in the field of IRSTD, leveraging their formidable capabilities in automatic feature extraction and representation. To tackle the Nash equilibrium problem between false alarms and missed alarms in detection tasks, two models have emerged: miss detection versus false alarm (MDFA) model [22] based on conditional GAN and the PixelGame model [23] based on fully dilated convolution network (FDCN) [49]. Both models have devised distinct adversarial strategies to optimize the equilibrium issue. Given the extremely low pixel ratio of infrared small targets and the potential issue of target loss in the network model after multiple layers of convolution, some methods have approached the problem by treating the target as noise, thereby converting the detection task into a denoising task, such as denoising autoencoder network (DAE) [24], dilated residual U-Net (DRUNet) [25], and mask-aware dynamic filtering (MADF) [50]. Some network models, such as asymmetric contextual modulation (ACM) [51], attentional local contrast network (ALCNet) [28], and enhanced asymmetric attention U-Net (EAAU-Net) [29], further emphasize the attention to the target area by combining

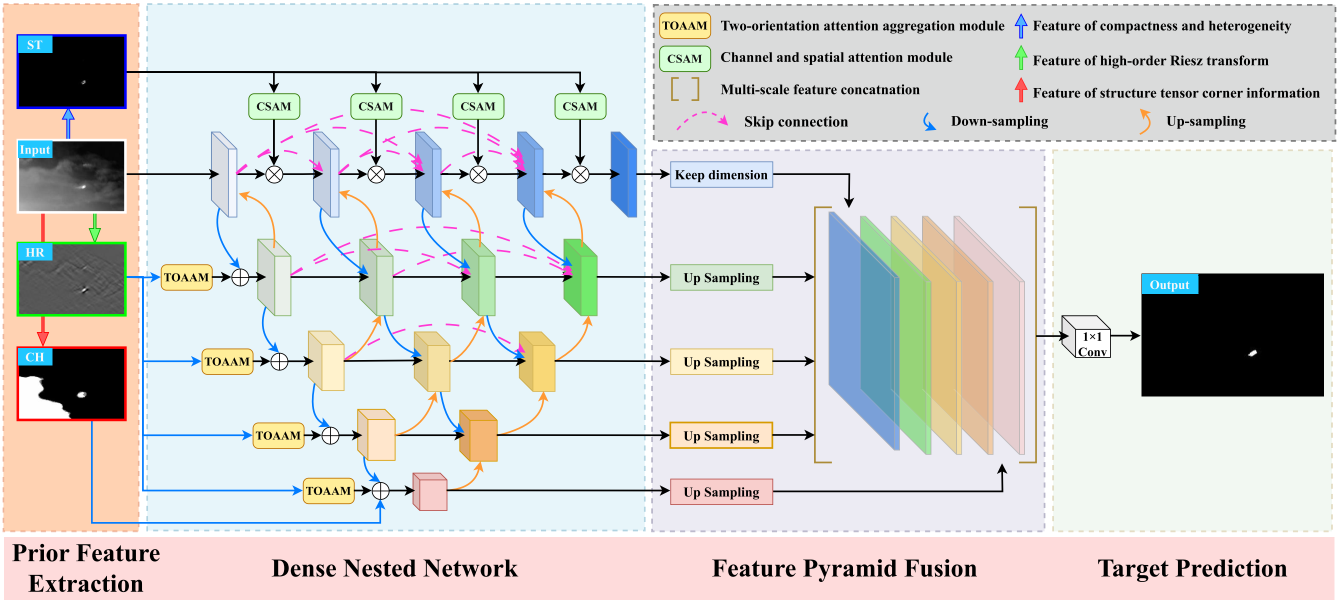
TOAAM Two-orientation attention aggregation module
CSAM Channel and spatial attention module
[ ] Multi-scale feature concatnation
Skip connection
Down-sampling
Up-sampling
Feature of compactness and heterogeneity
Feature of high-order Riesz transform
Feature of structure tensor corner information
ST
Input
HR
CH
CSAM
CSAM
CSAM
CSAM
TOAAM
TOAAM
TOAAM
TOAAM
Keep dimension
Up Sampling
Up Sampling
Up Sampling
Up Sampling
1×1 Conv
Output
Prior Feature Extraction
Dense Nested Network
Feature Pyramid Fusion
Target Prediction

Fig. 2. Overview of the proposed method. The blue, green, red, and white boxes represent the corner feature of the ST, the HR, the CH, and the input image, respectively. The blue, green, and red arrows, respectively, represent the extraction process of traditional manual features about ST, HR, and CH.

semantic features with contextual information. Furthermore, to address uncertainty in target sizes, multiscale feature analysis and fusion methods have been introduced. Models like infrared small-target detection U-Net (ISTDU-Net) [32], multiscale local contrast learning network (MLCL-Net) [11], and one-stage cascade refinement network (OSCAR) [33] have all employed multiscale feature extraction and propagation to improve the detection robustness. Dense nested networks [30], [31], [52] are similar to multiscale feature methods, which utilize a progressive feature interaction and fusion strategy to effectively integrate low-level detailed features with high-level semantic features. This approach is one of the effective methods for IRSTD. Nowadays, with the deepening exploration of deep learning, remote sensing image captioning [53], [54] is constantly developing. In the field of target detection, incorporating object counts into remote sensing image captioning [55] is beneficial for improving the intuitiveness of detection results and is one of the future development directions. Deep learning methods heavily depend on the quantity and diversity of training datasets, which remains a significant limitation in the field. Additionally, the lack of shape features and prominent textures in infrared small targets poses difficulties in feature learning and extraction, which may sometimes hinder their practical application in engineering.

III. Prior Guided Dense Nested Network

A. Overall Architecture

The overall architecture of the PGDN-Net model is shown in Fig. 2. It consists of a prior feature extraction module, a dense nested network module, a feature pyramid fusion module, and a target prediction module.

1) The prior feature extraction module extracts three different traditional handcrafted features, including the HR, the CH, and the corner feature of the ST.
2) The dense nested network module is composed of multiple U-shape sub-networks stacked on top of each other. Multiple nodes are applied on the path between the sub-networks, which consist of encoders and decoders, to receive feature information from the current layer and the adjacent upper and lower layers.
3) The feature pyramid fusion module is used to aggregate multilayer features at different depths within the network. It upsamples the multilayer features to a unified size, and then fuses the shallow-layer features rich in spatial and contour information with the deep-layer features rich in semantic information to obtain a hybrid feature map.
4) The target prediction module inputs mixed feature maps into a 1 × 1 convolutional kernel module to achieve feature extraction and information exchange across different channels, ultimately achieving the detection of infrared small targets.

B. Prior Feature Extraction

The dense nested network structure was initially designed to accommodate targets of different scales. The feature fusion process between the upper, lower, and current layers is primarily implemented to avoid the loss of small infrared targets in the deeper layers. However, due to the small size and limited features of the targets, there is an extreme imbalance in the data volume compared to the background information. The target in the network is easily affected by background clutter, which makes feature extraction difficult and leads to bias in the learning of the network model. Introducing prior features aims to guide the dense nested network in learning, extracting, and retaining feature information of the target area more accurately. In this method, three main types of target prior features are extracted, such as the HR, the CH of the target, and the corner feature of the ST.

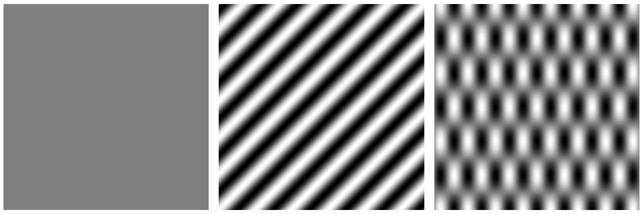

Fig. 3. From left to right: constant signal with intrinsic dimension 0, linear signal with intrinsic dimension 1, and other signal with intrinsic dimension 2.

1) High-Order Riesz Feature: Two-dimensional image signals can be divided according to different intrinsic dimensions to express the number of degrees of freedom required to describe the local structure [56]. The intrinsic dimension of a constant signal that remains constant is 0, the intrinsic dimension of lines and edges is 1, and the intrinsic dimension of other patterns is 2, as shown in Fig. 3 [56]. The features of the local signal with intrinsic dimension 2 have geometric and structural information, which determines the quality of the interpretation of the image and the analysis of the processing.

The Riesz transform is a complex transformation commonly used to analyze the phase and amplitude of the 2-D signal. It is a natural multidimensional extension of the Hilbert transform [57] and can obtain orthogonal pairs R_1 and R_2 of non-directional images, which is 90° phase-shifted with respect to the predominant orientation at each point in the original image [58]. Under 2-D condition, the frequency-domain representation of the Riesz transform is as follows:

$$(R_1(\mathbf{w}), R_2(\mathbf{w}))^T = \left(-j\frac{w_1}{\|\mathbf{w}\|}, -j\frac{w_2}{\|\mathbf{w}\|}\right)^T I_F \tag{1}$$

where $\mathbf{w} = [w_1, w_2]$ is a 2-D vector, I_F is the Fourier transform of image I, and R_1, R_2 represent the two components of the Riesz transform, respectively. The product in the frequency domain corresponds to the convolution operation in the spatial domain, so the 2-D Riesz transform in the spatial domain can be expressed as

$$(R_1(\mathbf{x}), R_2(\mathbf{x}))^T = \left(-j\frac{x}{\|(\mathbf{x})\|}, -j\frac{y}{\|(\mathbf{x})\|}\right)^T * I \tag{2}$$

where $\mathbf{x} = [x, y]$, and "$*$" represents the convolution operation.

The first-order Riesz transform is used to extract the local linear features of the image [59], but it cannot fully characterize its features. Therefore, the second-order Riesz transform is utilized to extract the intrinsic 2-D structure of local regions in the image to provide a more detailed expression of the image features in this method. The second-order Riesz transform is obtained by performing the Riesz transform again based on the results of the first-order Riesz transform

$$\begin{cases} R_{11} = R_1\{R_1(I)\} \\ R_{12} = R_2\{R_1(I)\} \\ R_{22} = R_2\{R_2(I)\}. \end{cases} \tag{3}$$

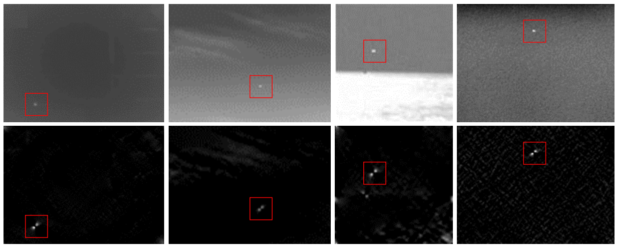

Fig. 4. R_{12} result of high-order Riesz transform for infrared small target. The target regions are marked with red boxes.

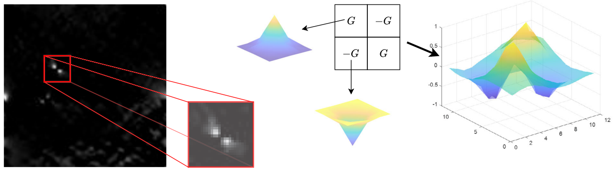

Fig. 5. Schematic of target region and feature extraction operator for high-order Riesz transform.

The second-order Riesz transform further emphasizes the features of the edge and the line through R_{11} and R_{22}, while using R_{12} to characterize the feature information of the intersection area of edges and lines. This is because R_{12} of the second-order Riesz transform is used for feature extraction on both the x- and y-axes, which can effectively perceive prominent point signals in multiple directions in the image, perfectly matching the features of infrared small target areas. Therefore, the target area in R_{12} exhibits significant features of local symmetry, as shown in Fig. 4.

A simple detection operator F_{HR} is designed to extract locally symmetric salient target regions in R_{12}, and generate prior information about the high-order Riesz transform features HR. The grid-shaped symmetric detection operator F_{HR} is constructed by a 2-D Gaussian function. Its shape structure and intensity distribution are similar to the target area in R_{12}, both symmetrical about the diagonal axis, as shown in Fig. 5. The detection operator F_{HR} is used to enhance the saliency of the target area, providing effective prior feature information HR. Subsequently, the two-orientation attention aggregation module is used to characterize the shape information of the target, guiding the subsequent training and learning of the network model to achieve precise target detection.

2) Compactness and Heterogeneity Feature: The compactness feature is reflected in the similarity and high intensity of the pixel values within the target area, while the heterogeneity feature is manifested in the significant difference in pixel intensity between the target and its surrounding background [39]. In this method, these two types of potential features are introduced into the network to guide the model to learn the features of the potential target regions while preserving the target features in deep layers.

Specifically, we use a three-layer detection window designed in our previous IRSTD research work [60], consisting of an inner layer L_i, a middle layer L_m, and an outer layer L_o. These three layers correspond to the center area of the target,

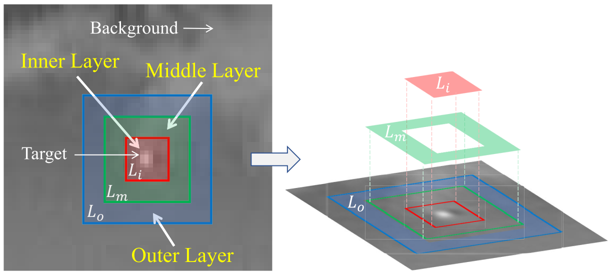

Fig. 6. Schematic of the three-layer local detection window.

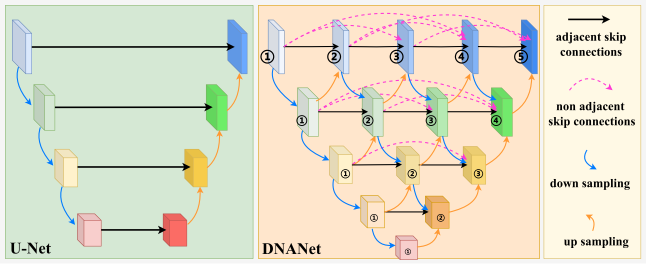

Fig. 7. Comparison diagram of vanilla U-Net and dense nested network structure.

the transition area between the target and the surrounding background, and the surrounding background area of the target, as shown in Fig. 6.

Here, the average pixel value ($\bar{P}_i$) of the inner layer L_i and the maximum pixel value ($\hat{P}_o$) of the outer layer L_o are calculated separately. Based on the prior characteristics of compactness and heterogeneity, the binary CHs can be obtained, which is the potential region of the target

$$\mathrm{CH}(x, y) = \begin{cases} 1, & \bar{P}_i(x, y) \geq \hat{P}_o(x, y) \\ 0, & \bar{P}_i(x, y) < \hat{P}_o(x, y) \end{cases} \tag{4}$$

where (x, y) represents the pixel coordinates of the center point of the three-layer local detection window.

The characteristics of compactness and heterogeneity provide important reference information to guide the model in subsequent target extraction.

3) Corner Feature of the ST: The ST of the image can effectively estimate the local structure information [21], providing a rational basis for distinguishing the flat, edge, and corner areas. Small targets in the infrared image exhibit significant geometric properties in the corners. The ST is defined as follows:

$$\begin{aligned} \mathbf{J}_\alpha(\nabla \mu_\tau) &= G_\alpha * (\nabla \mu_\tau \otimes \nabla \mu_\tau) \\ &= \begin{pmatrix} G_\alpha * {I_x}^2 & G_\alpha * I_x I_y \\ G_\alpha * I_x I_y & G_\alpha * {I_y}^2 \end{pmatrix} \end{aligned} \tag{5}$$

where μ_σ represents Gaussian smoothed image μ with variance $\tau > 0$, G_α indicates a Gaussian function with standard deviation α, $\otimes$ represents the Kronecker product, ∇ represents the gradient, I_x and I_y represent the gradient of μ_σ along the x- and y-axes, respectively. $\mathbf{J}_\alpha$ is a symmetric positive semi-definite matrix, therefore there are two orthogonal eigenvectors and corresponding eigenvalues λ_1 and λ_2. We define λ_1 as the larger of the two eigenvalues. Corresponding to the image, each pixel has two feature values λ_1 and λ_2. By comparing their relationships, the local structure information of the image can be reflected as follows.

1) *The flat area:* $\lambda_1 \approx \lambda_2 \approx 0$.
2) *The edge area:* $\lambda_1 \gg \lambda_2 \approx 0$.
3) *The corner area:* $\lambda_1 \gg \lambda_2 \gg 0$.

In this method, the corner feature of the ST ST in infrared image is derived based on the corner strength function [61], [62]

$$\mathrm{ST} = \frac{\lambda_1 \cdot \lambda_2}{\lambda_1 + \lambda_2} \cdot \lambda_1. \tag{6}$$

The corner feature prior information ST is input to various shallow nodes in the network through the channel and spatial hybrid attention module to emphasize the importance of the target area and suppress the simple background.

C. Dense Nested Network Structure

The dense nested network forms the main structure of the PGDN-Net model, which is composed of multiple subnets of U-Net stacked together. By adding multiple nodes along the path composed of the encoder and decoder, a dense nested structure is constructed, as illustrated in Fig. 7. The progressive feature interaction strategy in the dense nested structure allows each node in the intermediate layers of the network to receive feature maps from both the current depths and adjacent depths, effectively overcoming the limitations of vanilla U-Net where feature information is transmitted only within the same network depth. This enables the network to adapt to targets of different sizes. Moreover, feature information can also be transmitted between non-neighboring nodes of the same network depth to preserve the target's feature information as much as possible and prevent the loss of target features.

Specifically, the number of layers in a dense nested network is defined as i $(i = 0, 1, 2, \ldots)$, and the number of nodes in each layer is defined as j $(j = 0, 1, 2, \ldots)$. $L_{i,j}$ represents the output feature map of the jth node in the ith layer. The input of each initial node (i.e., $j = 0$) is only the downsampling feature map of the previous initial node, where $L_{i,0}$ is

$$L_{i,0} = P_{\max}\left[C\left(L_{i-1,0}\right)\right], \quad i > 0 \tag{7}$$

where $P_{\max}[\cdot]$ represents the max pooling operation and $C[\cdot]$ represents the cascaded convolution operation.

The input to the nodes in the intermediate layers of the dense nested network consists of feature maps from three different depth levels. Considering the existence of non-neighboring skip connections at the same depth, the calculation formula for intermediate layer node $L_{i,j}$ is as follows:

$$L_{i,j} = \left\{C\left[\left(L_{i,k}\right)\right]_{k=0}^{j-1}, P_{\max}\left[C\left(L_{i-1,j}\right)\right], U\left[C\left(L_{i+1,j-1}\right)\right]\right\} \tag{8}$$

where $i > 1$, $j > 0$, $U[\cdot]$ represents the upsampling operation, and $C[(L_{i,k})]_{k=0}^{j-1}$ represents the convolution result of connecting the feature maps of all nodes before the node in the current layer.

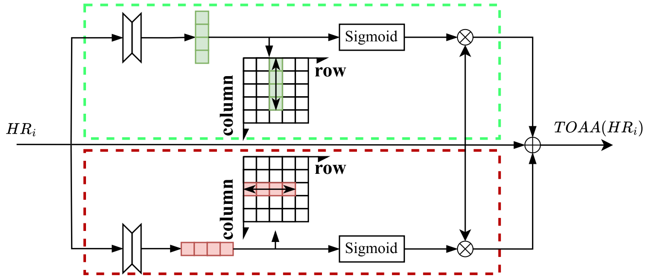

Fig. 8. Schematic of two-orientation attention aggregation module.

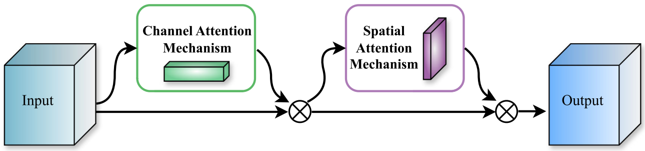

Fig. 9. Schematic of channel and spatial attention module.

D. Attention Mechanism

1) Two-Orientation Attention Aggregation Module: The two-orientation attention aggregation module is used to refine the high-order Riesz transform features HR at various scales. Here, the high-order Riesz transform feature HR corresponding to the size of the ith layer is defined as HR_i, and the output result of the two-orientation attention aggregation module is TOAA_i. As shown in Fig. 8, the two-orientation attention aggregation module consists of two parallel attention modules, each of which generates attention feature maps along one direction (column or row) to extract information about the target in two directions in the high-order Riesz transform features. Finally, summarize the attention feature maps into the output of the entire module.

The output of the two-orientation attention aggregation module can be expressed as

$$\mathrm{TOAA}(\mathrm{HR}_i) = F_s\{F_r[F_b(\mathrm{HR}_i)]\}\mathrm{HR}_i + F_s\{F_c[F_b(\mathrm{HR}_i)]\}\mathrm{HR}_i + \mathrm{HR}_i \qquad (9)$$

where F_s denotes the sigmoid function, F_b is a bottleneck architecture to restrict high-frequency noise in images, F_r and F_c represent the deformable convolution in the row and column directions, respectively.

The two-orientation attention aggregation module promotes the PGDN-Net model to extract shape information of significant regions of high-order Riesz prior features of the target from two directions. It inserts the integrated output into the initial nodes of each layer to guide the network in preserving target features and extracting target shapes.

2) Channel and Spatial Attention Module: As shown in Fig. 9, the channel and spatial attention module is composed of two cascaded units of channel attention and spatial attention, which are connected in series to adaptively enhance the region of interest in the image.

a) Channel attention module: Channel attention focuses more on "what useful information is" in the image. Therefore, each channel of the feature map is used as a feature detector.

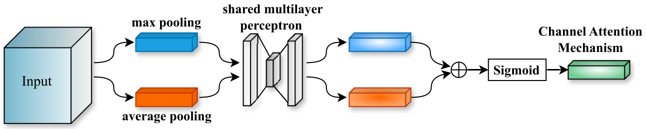

Fig. 10. Schematic of channel attention module.

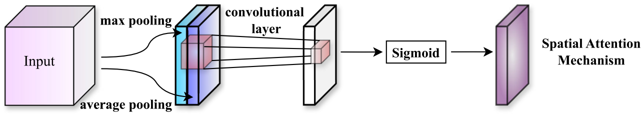

Fig. 11. Schematic of spatial attention module.

As shown in Fig. 10, the module first compresses the spatial dimension of the input feature map, aggregates spatial information through max pooling and average pooling, and then uses a shared network to learn the feature information on the channel dimension, assigning different weight information to each channel. Finally, the two feature vectors produced by the shared network are merged, and the weight matrix of the channel dimension, namely, the channel attention information, is obtained through activation function mapping processing.

The channel attention operation can be represented by the following formula:

$$M_c(F) = F_S\{\mathrm{MLP}[P_{\max}(F)] + \mathrm{MLP}[P_{\mathrm{avg}}(F)]\} \qquad (10)$$

where M_c denotes the weight matrix of channel attention, F is the input feature information to the channel attention module, MLP represents the shared multilayer perceptron, $P_{\max}$ and P_{avg} represent the max-pooling and average-pooling operation, respectively.

b) Spatial attention module: Spatial attention focuses on "where the effective information is" in the image, and it works in conjunction with channel attention to achieve functional complementarity. As shown in Fig. 11, the module first compresses the number of channels, performs max-pooling and average-pooling operations in the channel dimension, and then connects to obtain a feature map with two channels. Following convolution, the spatial attention feature is derived by applying an activation function.

The spatial attention operation can be represented by the following formula:

$$M_S(F) = F_S\{\mathrm{Conv}[P_{\max}(F); P_{\mathrm{avg}}(F)]\} \qquad (11)$$

where M_S denotes the weight matrix of spatial attention and Conv is the convolution operation.

E. Prior Feature Guidance

In the proposed PGDN-Net model, three traditional manual features are extracted: high-order Riesz transform features, CHs, and corner features derived from ST of the target. As shown in Fig. 12, these three types of prior targets act at different depths of the network, thus exerting different guiding functions.

First, the corner features of the ST are input into the bottom layer of the network for simple background filtering and

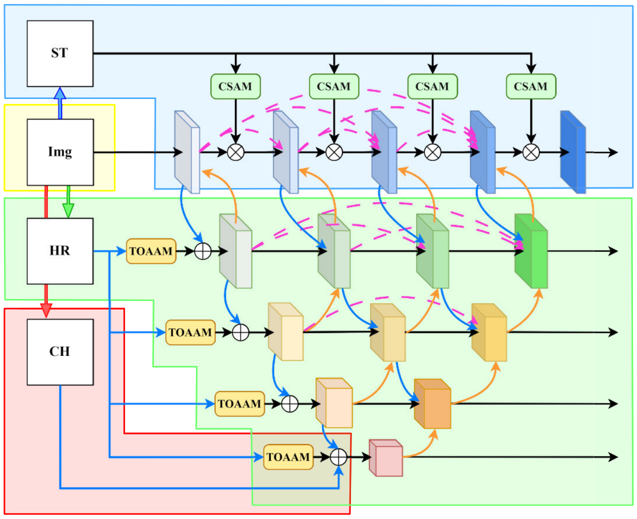

Fig. 12. Schematic of three prior feature guidance strategies. The blue area represents the guidance section of the corner feature of the ST, the green area represents the guidance section of the HR, and the red area represents the guidance section of the compact and heterogeneity feature CH.

elimination, in order to focus on the strong corner regions in the image and avoid information in the continuous flat irrelevant regions. Second, the HR is inputted in the initial nodes of each layer in the network to enhance the feature strength of the target in each layer. Under the action of the two-orientation attention aggregation module, the extraction of target shape information can be achieved. Finally, the compact and heterogeneity features CH are inputted into the top level of the network to preserve the deep target features. Due to the fact that compactness and heterogeneity are the most fundamental features of the target, any possible target area can be effectively preserved, thereby avoiding the occurrence of missed alarms.

F. Loss Function

In the proposed PGDN-Net model, we use the binary cross-entropy loss (BCELoss) function for network parameter optimization. BCELoss is commonly used for loss calculation in binary classification problems. In this method, the introduction of BCELoss can effectively evaluate the performance of IRSTD models in predicting targets. The calculation formula for BCELoss is as follows:

$$\mathrm{Loss}_{\mathrm{BCE}}(\mathrm{Pre},\mathrm{GT}) = -\big[\mathrm{Pre}\cdot\log(\mathrm{GT}) + (1-\mathrm{Pre})\cdot\log(1-\mathrm{GT})\big] \quad (12)$$

where Pre and GT represent the predicted result of the network and the ground truth of the training sample, respectively, and "·" indicates the multiplication.

IV. Experimental Results and Discussion

To analyze the detection performance of the PGDN-Net model, we utilize some commonly objective evaluation metrics and compare them with other effective methods through rich experiments. Initially, we present the utilized public datasets, followed by an in-depth explanation of the definitions and computational methods for the objective evaluation metrics employed. Then, the implementation details of the experiments are listed, and qualitative and quantitative experiments with other state-of-the-art methods are conducted. Finally, the ablation experiment is conducted on the PGDN-Net model to verify the effectiveness and functionality of each module.

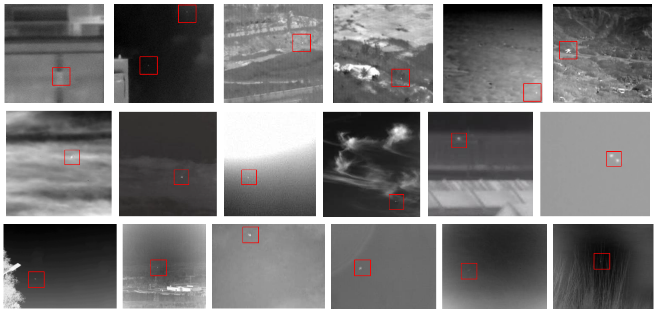

Fig. 13. Examples of dataset image. The first line is the images from NUDT-SIRST for training. The second and third lines are the images from the single-frame and sequence-frame datasets for testing. The targets are marked with a red box.

A. Dataset

In this article, we train various deep learning models on the NUDT-SIRST dataset [31], which recognizes the background through a scene-aware model and adds appropriate and reasonable targets. The background includes clouds, cities, oceans, fields, and strong light. The target sizes and clutter types are diverse, and the targets have rich situations and high-precision ground truth. The test data consists of an infrared single-frame dataset [51] and six groups of sequence-frame datasets [63], [64], [65]. Fig. 13 shows some examples of training data and test data, with targets marked in red boxes.

B. Evaluation Metrics

Despite the multitude of objective metrics for evaluating IRSTD methods, they can generally be categorized into several aspects, including but not limited to: the effect of background suppression, the effect of target enhancement, and detection performance (such as detection rate and false alarm rate).

In this article, we use background suppression factor (BSF) and contrast gain (CG) to evaluate the ability of background suppression and target enhancement, respectively. We also introduce the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), and the intersection over union (IoU) to analyze the overall detection performance of the IRSTD methods.

BSF is calculated in the following manner [46]:

$$\mathrm{BSF} = \frac{\sigma_{\mathrm{in}}}{\sigma_{\mathrm{out}}} \quad (13)$$

where σ represents the gray standard deviations of the whole background in the image, and "in" and "out" indicate the input image and the predicted result, respectively. The background suppression ability is directly proportional to the value of BSF.

There are two calculation approaches for CG [39], [66], which are used to reflect the enhancement effect of detection methods on the overall target and the center point of the target

$$\mathrm{CG}_o = \frac{|\bar{t} - \bar{b}|_{\text{out}}}{|\bar{t} - \bar{b}|_{\text{in}}}$$

$$\mathrm{CG}_c = \frac{|\hat{t} - \bar{b}|_{\text{out}}}{|\hat{t} - \bar{b}|_{\text{in}}} \tag{14}$$

where $\bar{t}$ and $\hat{t}$ indicate the average and maximum pixel value of the target region, respectively. $\bar{b}$ is the average pixel value of the background area surrounding the target. $|\cdot|$ is the symbol for solving absolute values. An increased CG value leads to an improved target enhancement effect.

The ROC curve is one of the effective statistical analysis tools used in IRSTD to evaluate the performance of binary classification models between the target and the background. The ROC curve presents the relationship between the detection probability P_d and the false alarm rate F_a in an intuitive graphical form, reflecting the performance of the IRSTD model. They are defined as follows [67]:

$$P_d = n_d / n_{\text{all}} \tag{15}$$

$$F_a = p_f / p_{\text{all}} \tag{16}$$

where n_d and n_{all} represent the number of detected true targets and all true targets, respectively. Similarly, p_f and p_{all} represent the number of detected false pixels and all pixels in the image, respectively.

AUC is defined as the area under the ROC curve, typically ranging between 0.5 and 1. The AUC index further reports the classifier with better performance based on the ROC curve. The classifier with a larger AUC value performs better. The method of calculating AUC is to sum the areas of each part under the ROC curve. And IoU is used to calculate the intersection-to-union ratio between the predicted results and the ground truth labels, enabling a more nuanced assessment of a model's capability to differentiate targets from backgrounds and to detect targets of varying sizes.

C. Implementation Details

All the deep learning models are implemented in Python 3.9.7 and PyTorch 2.0.1 on a computer with an Intel Xeon 6252 @ 2.10 GHz CPU and an Nvidia A100 GPU. The other traditional methods are implemented in MATLAB 2019b on an Intel Core i5-9400 at 2.90 GHz CPU. The number of dense nested downsampling layers in the PGDN-Net model is set to 4, and ResNet's U-Net paradigm is used as the segmentation backbone. The network model is trained using the Adagrad optimizer with a learning rate of 0.05, a batch size of 10, and 500 iterations. This method uses the BCE loss function for training.

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

To verify the superiority of our PGDN-Net model, we conduct quantitative and qualitative experiments with several typical state-of-the-art methods on single-frame and sequential-frame datasets. The compared methods comprehensively cover the four main types of IRSTD algorithms currently available, including background feature-based methods like TopHat [16], target feature-based methods like NLCD [39] and LEF [38], low-rank and sparse matrix decomposition-based methods like PSTNN [62] and NTFRA [63], as well as deep learning-based methods like MDFA [22], ISTDU [32], DNANet [31], HCFNet [68], and SCTNet [69]. All of the above methods use the public code provided and the reproduced code, and carry out relevant experiments according to the default parameter settings in the original paper.

1) Qualitative Analysis: Fig. 14 shows some example results achieved by different methods based on a single-frame dataset. These example images contain various complex detection scenarios, such as strong light interference, multitarget coexistence, complex background clutter, and target tailing.

In Image1, the strong background interference greatly reduces the saliency of the target, leading to the failure of LEF, ISTDU, and HCFnet. Meanwhile, the rapidly changing background signal causes confusion in NTFRA, resulting in a large number of background false alarms in the result. In Image2, there are two targets with similar sizes but different intensities. NLCD, LEF, ISTDU, and HCFnet cannot fully detect all targets, resulting in poor detection robustness. In addition, it also indicates that these methods have poor enhancement effects on weak targets and can only enhance the more significant targets. There are many easily confused pseudo target points in Image3, leading to a large number of false alarms in TopHat, LEF, and MDFA. There is also a problem of inconsistent enhancement effects on targets of different intensities. In Image4, there is obvious trailing interference on the right side of the target, which causes some methods to fail to recognize the target or mistakenly detect the trailing area as the target. The proposed PGDN-Net model can effectively cope with the complex detection scenarios mentioned above, with strong resistance to strong light interference, and can achieve high detection rates while preserving and restoring the true shape of the target as much as possible, which is beneficial for practical detection tasks.

Fig. 15, respectively, shows the results achieved by different methods based on sequential-frame datasets. It indicates that the proposed PGDN-Net model can accurately capture the target, and with the guidance of prior feature information and the role of the attention mechanism, the extracted target is more in line with the true form and feature distribution of the target. For example, in Seq.1, the target is an aircraft flying from left to right with a narrow tail on the left side. The output of our PGDN-Net displays an approximately trapezoidal target with a narrow left side and a wider right side, which is similar to the actual shape feature of the target. Similarly, in Seq.3 and Seq.4, the targets have certain shape features rather than a simple dot. Our PGDN-Net model can output the results that fit the actual shape and size of the target. TopHat and NLCD have poor enhancement effects on weak targets, and the target detection results are incomplete. In addition, NLCD is prone to the influence of high-brightness noise points in the background, leading to an increase in false alarms. The background suppression ability of LEF is poor, especially in low SNR scenarios such as Seq.4 to Seq.6, and this method has a significant bias in determining target size. PSTNN has a good inhibitory effect on the background. NTFRA is easily affected

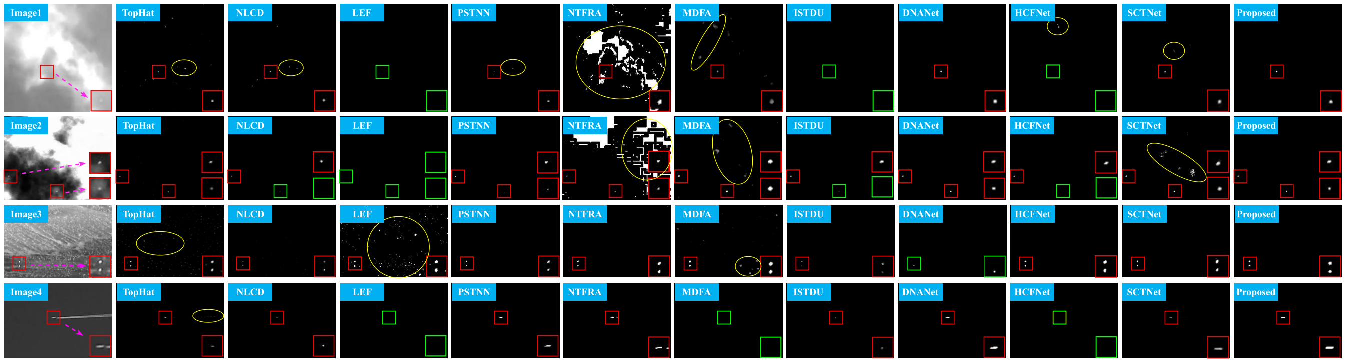

Fig. 14. Examples of the original image and the corresponding results by each method for the single-frame dataset. The red box indicates that the real target is detected, the green box indicates that the real target is lost, and the yellow circle marks the clutter and false target points present in the background. The bottom-left corner of each subplot shows a close-up display of the target area.

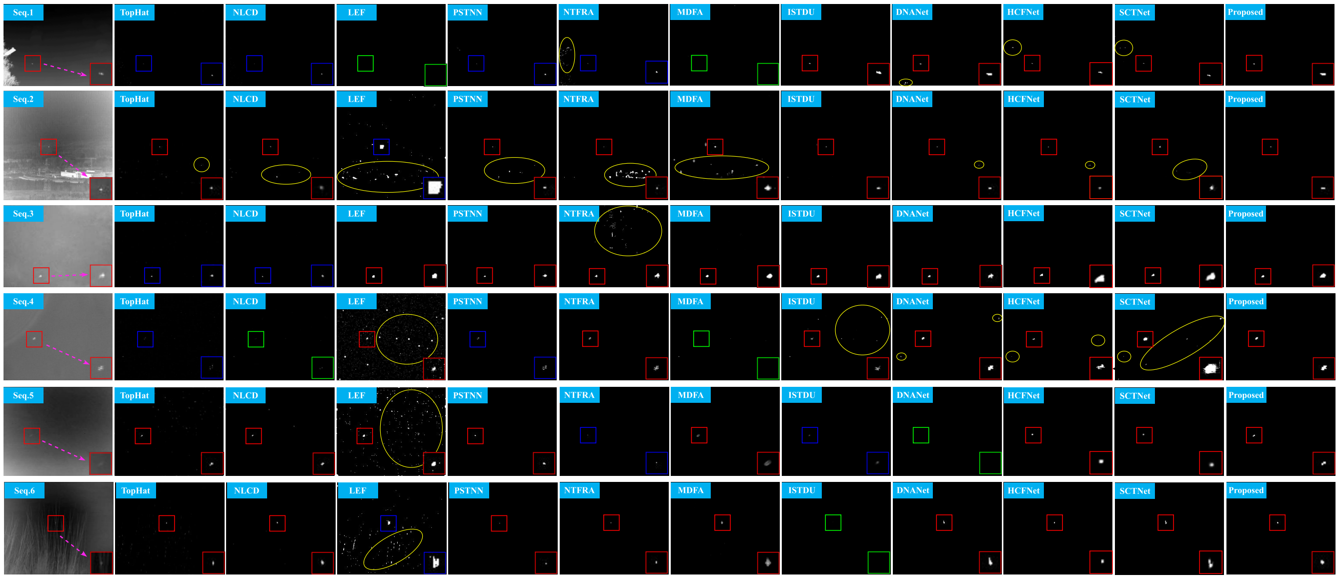

Fig. 15. Examples of the original image and the corresponding results by each method for six groups sequence-frame datasets. The red box indicates that the real target is detected, the green box indicates that the real target is lost, the blue box indicates that the detected target is incomplete, and the yellow circle marks the clutter and false target points present in the background. The bottom-left corner of each subplot shows a close-up display of the target area.

by strong edge signal and clutter in the background, resulting in a large number of false alarms, such as Seq.2 and Seq.3. MDFA and ISTDU have excellent performance on salient targets, but their enhancement effects on weak targets are poor, sometimes even lost them, resulting in a decrease in detection rate. DNANet has good performance, but its perception ability for target edge detail is poor, resulting in jagged target edges or misidentification of the background, such as Seq.3 and Seq.6. Since the model lacks prior information on features, it struggles to precisely distinguish between the target edge and the adjacent background region. Although HCFNet and SCTNet can capture the target, their inhibitory effect on the background is average. In contrast, our PGDN-Net model has an outstanding detection performance, good generalization in various scenarios, and higher detection robustness.

2) Quantitative Analysis: We quantitatively evaluate the detection performance of our PGDN-Net model and other compared methods on multiple datasets using the metrics mentioned earlier.

First, we calculated the performance metrics of various methods on the single-frame dataset. Table I lists the average BSF, CG, AUC, and IoU of the methods on the single-frame dataset, and Fig. 16 shows the ROC curves of the methods on the single-frame dataset. As can be seen, the PGDN-Net model has outstanding comprehensive detection performance in the single-frame dataset, with strong generalization ability. Thanks to the guidance of prior features of the target, the model can more effectively extract target features and enhance them. Concurrently, under the action of the attention module, the network can more accurately focus on the interested target area, suppress irrelevant background signals, and ultimately achieve outstanding comprehensive indicators.

Next, we quantitatively evaluate the detection performance of each method on six groups of sequence-frame datasets. Table I lists the average BSF, CG, AUC, and IoU of the methods on the sequence-frame dataset, and Fig. 16 shows the results of the ROC curve of the methods on the sequence-frame dataset.

TABLE I

AVERAGE BSF, CG, AUC, AND IoU OBTAINED ON THE SINGLE-FRAME DATASET AND SEQUENTIAL-FRAMES THROUGH DIFFERENT METHODS. THE FIRST TWO BEST RESULTS ARE REPRESENTED IN BOLD AND UNDERLINED, RESPECTIVELY

Dataset	Metrics	TopHat	NLCD	LEF	PSTNN	NTFRA	MDFA	ISTDU	DNANet	HCFNet	SCTNet	Proposed
Single	BSF	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
	CG_o	2.9657	2.0595	3.0820	2.3320	2.6940	4.4438	4.6387	4.7660	4.4545	4.1043	**4.8489**
	CG_c	3.5703	3.4234	2.2739	2.9669	2.8439	3.0695	3.6206	4.0531	3.6705	3.3098	**4.0773**
	AUC	0.8552	0.7027	0.7952	0.6812	0.7288	0.9551	**0.9949**	0.9750	0.9417	0.9806	0.9927
	IoU	0.4248	0.2694	0.2671	0.3697	0.2744	0.5331	**0.7266**	0.7006	0.6646	0.6151	0.7059
Seq.1	BSF	66.4624	**109.939**	——	55.455	2.8577	41.5505	40.5359	50.4214	51.0605	48.5757	37.0280
	CG_o	0.8687	0.8055	——	1.9210	2.0718	0.0009	2.0385	2.1735	**2.1797**	2.1119	2.0521
	CG_c	1.3827	1.4760	——	1.8686	1.8703	0.0007	1.6810	**1.9769**	1.9690	1.8092	1.8807
	AUC	0.6985	0.6234	——	0.7807	0.8632	0.5029	**1.0000**	0.9999	0.9970	**1.0000**	**1.0000**
	IoU	0.0113	0.0210	——	0.2657	0.0072	0.0000	**0.7412**	0.4379	0.5333	0.5369	0.5447
Seq.2	BSF	22.8628	102.1924	2.4220	9.3379	2.3198	6.0345	29.5026	**Inf**	3537.7742	37.6209	**Inf**
	CG_o	1.3842	1.7539	**5.7978**	1.3517	1.2680	4.8347	2.1905	2.2066	1.8763	3.6400	2.4051
	CG_c	1.5539	1.4642	1.5248	**1.5547**	1.5543	1.5546	1.5516	1.5512	1.4384	1.5215	1.5518
	AUC	0.9837	0.9140	0.9981	0.8007	0.7827	0.9999	0.9949	0.9689	0.9950	**1.0000**	**1.0000**
	IoU	0.5496	0.3643	0.0326	0.1863	0.0040	0.2165	0.4001	0.7953	0.5540	0.5722	**0.8440**
Seq.3	BSF	20.2974	**Inf**	**Inf**	**Inf**	1.2720	94.5509	417.3249	271.4807	978.4830	**Inf**	264.0487
	CG_o	1.2944	1.4915	3.4086	2.7304	3.1571	3.0898	3.3648	3.2945	3.1786	3.2367	**3.4111**
	CG_c	2.3009	2.4523	2.5206	2.5225	2.5065	2.3883	2.4913	**2.6086**	2.5917	2.4837	2.6024
	AUC	0.6921	0.6192	0.9538	0.7373	0.8888	0.9718	0.9915	0.9694	0.9663	**0.9925**	0.9896
	IoU	0.0831	0.0753	0.7326	0.3964	0.0837	0.6309	0.8348	0.7222	0.6728	0.6930	**0.8390**
Seq.4	BSF	1.8916	**6.9962**	0.1709	3.9527	2.4130	2.9762	1.0888	2.1456	2.1284	0.4849	2.0054
	CG_o	0.3675	0.1790	**4.7120**	1.0364	2.9545	0.2183	3.3971	4.5003	3.3872	0.9927	4.1343
	CG_c	0.3323	0.5774	1.8563	2.0740	2.0555	0.2765	1.8658	2.0438	**2.2172**	0.4876	1.9978
	AUC	0.5421	0.5625	0.8850	0.6640	0.7947	0.5591	0.9549	0.9850	0.8836	0.9940	**0.9989**
	IoU	0.0134	0.0255	0.0545	0.2711	0.5645	0.0030	0.4410	0.6426	0.3699	0.3278	**0.7055**
Seq.5	BSF	28.9230	74.5219	4.3014	**Inf**	**Inf**	1218.8020	926.1254	**Inf**	**Inf**	**Inf**	**Inf**
	CG_o	7.7758	6.2637	**13.6871**	7.2673	4.6038	9.9574	11.7116	11.9295	10.5487	5.5790	11.9654
	CG_c	9.3278	9.3286	9.2640	9.3288	9.3264	7.7938	8.9515	9.4816	**10.5487**	5.5699	9.7214
	AUC	0.9717	0.8324	0.9922	0.7214	0.6390	0.9999	**1.0000**	0.9806	0.9534	0.9523	**1.0000**
	IoU	0.4016	0.4313	0.1382	0.4994	0.2701	0.4502	0.5895	0.6605	0.5818	0.5098	**0.7214**
Seq.6	BSF	27.6528	7409.4240	**Inf**	**Inf**	**Inf**	45.6079	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
	CG_o	1.7673	1.8860	**4.7348**	0.7184	0.7077	4.4590	1.0445	2.4431	2.6445	3.4452	3.6552
	CG_c	2.2324	2.2359	2.2505	2.0995	2.0548	2.2063	0.9784	2.4575	2.3469	2.1039	**2.5113**
	AUC	0.9302	0.8946	0.9894	0.5884	0.5867	0.9972	0.8859	0.9806	0.9381	0.9989	**1.0000**
	IoU	0.2580	0.2944	0.5410	0.1920	0.1938	0.7050	0.1475	0.5907	0.4373	0.6380	**0.7064**

Due to the failure of LEF in Seq.1, its output result is a black background, so no calculation is performed on the objective indicators of Seq.1. As shown in Table I, the PGDN-Net model can almost achieve the optimal AUC and IoU on all groups of sequence-frame datasets, indicating that PGDN-Net obtains the best detection performance and can detect more accurately. Although the target enhancement of NLCD is poor, its background suppression ability is prominent, due to the random walk algorithm in this method, which can better distinguish background information. LEF focuses more on enhancing the overall target, therefore its CG_1 value is relatively high. In contrast, PSTNN has a better enhancement effect on the target center with higher CG_2. Although the detection performance of MDFA, ISTDU, and DNANet is good, their detection performance on different datasets varied widely. Compared to SCTNet, HCFNet has a better target enhancement effect, but SCTNet has better detection ability. Since these deep learning methods are based solely on network parameters acquired during training and do not incorporate objective prior features of the input image, their ability to generalize is inadequate.

Fig. 16 shows the ROC curves of the methods on the six groups of sequence-frame datasets. The ROC curves further demonstrate the superiority of PGDN-Net. Specifically, at the same detection rate, PGDN-Net achieves the lowest false alarm rate, and at the same false alarm rate, it can obtain the highest detection rate.

The comparative experimental result in this section indicates that the proposed PGDN-Net model has better detection performance than traditional methods, stronger generalization ability than other deep learning methods, and combines the advantages of traditional and deep learning methods.

E. Ablation Study

In this section, we conduct the ablation experiment on three prior feature modules in our PGDN-Net model, such as the

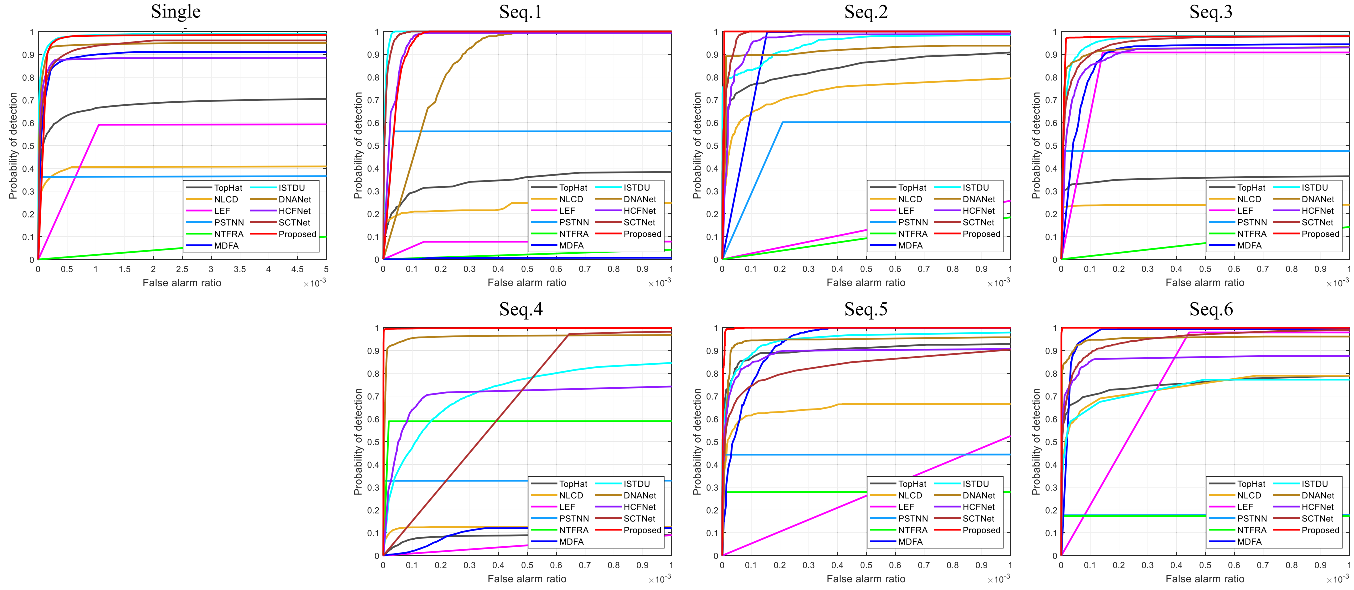

Fig. 16. ROC curves achieved by different methods on single-frame dataset and sequence-frame dataset.

TABLE II

DETAILED SETUP OF ABLATION EXPERIMENT

Method	HR	CH	ST	AUC
DNA-Net	✗	✗	✗	0.9750
PGDN-Net1	✗	✓	✓	0.9743
PGDN-Net2	✓	✗	✓	0.9728
PGDN-Net3	✓	✓	✗	0.9768
PGDN-Net4	✓	✓	✓	**0.9927**

HR, the CH, and the corner feature of the ST, to investigate the potential benefits and choice of design. The baseline method for comparison is the original DNA-Net model, which does not introduce any prior target feature information as a guide. The ablation experiments are uniformly trained on the NUDT-SIRST dataset and tested on the single-frame dataset.

Table II lists the control group settings and the corresponding AUC results in the ablation experiment. It can be seen that the detection performance of the model has been greatly improved after introducing three prior target features to guide the dense nested network. On the one hand, when HRs are not considered, it may lead the network to focus on some potential clutter information in the background, leading to an increase in false alarms, as shown in Fig. 17. Therefore, it is necessary to introduce HRs to improve the model's ability to distinguish targets. On the other hand, when CHs are not considered, the model will be interested in strong edge corners under the influence of high-order Riesz and ST corner features, resulting in pseudo-targets appearing at some edge corners, and it is difficult to extract features of weak target areas, as shown in Fig. 18.

In summary, the ablation experiment first proves the effectiveness of our PGDN-Net model, and second, compared to the original DNA-Net model, PGDN-Net model demonstrates more significant advantages.

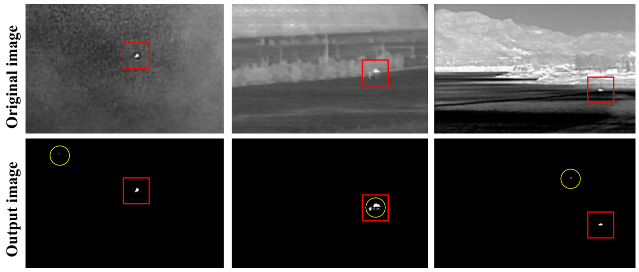

Fig. 17. Output results of PGDN-Net1. The red box indicates that the real target is detected and the yellow circle marks the clutter and false target points present in the background.

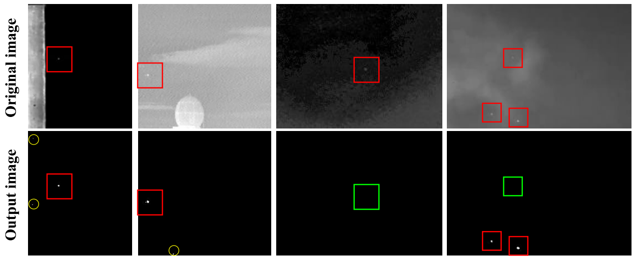

Fig. 18. Output results of PGDN-Net2. The red box indicates that the real target is detected, the green box indicates that the real target is lost, and the yellow circle marks the clutter and false target points present in the background.

V. CONCLUSION

This article proposes a PGDN-Net model, aiming to solve the inherent problem of lack of prior features in IRSTD tasks and the practical problem of imbalanced positive and negative sample information in deep learning networks. The PGDN-Net model achieves effective integration of low-level detail features and high-level semantic features in the network through a dense nested structure. It introduces HRs, CHs, and corner features of ST features to guide the network model to learn target features at different depths and ensure the preservation and recovery of target features, thereby alleviating the problem of few characteristics, hard to extract, and difficult to learn features for infrared small targets. At the same time, under the action of the attention mechanism module, the attention of the entire network is focused on the interested target area, effectively solving the problem of unbalanced positive and negative sample information and avoiding erroneous learning of background features in the network. The key point of PGDN-Net is to combine traditional methods with deep learning methods to achieve robust IRSTD. Experiments on multiple datasets have shown that it achieves more outstanding comprehensive detection performance, with excellent target enhancement and background suppression effects. Moreover, the detection results in various scenarios effectively demonstrate the generalization and robustness of the proposed PGDN-Net model.

References

[1] T. Ma, G. Guo, Z. Li, and Z. Yang, "Infrared small target detection method based on high-low-frequency semantic reconstruction," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[2] 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.

[3] 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, Jun. 2023, Art. no. 5612912.

[4] Y. Li, Z. Li, J. Li, J. Yang, and A. Siddique, "Robust small infrared target detection using weighted adaptive ring top-hat transformation," *Signal Process.*, vol. 217, Apr. 2024, Art. no. 109339.

[5] T. Ma et al., "MDCENet: Multi-dimensional cross-enhanced network for infrared small target detection," *Infr. Phys. Technol.*, vol. 141, Sep. 2024, Art. no. 105475.

[6] D. Zhou and X. Wang, "Robust infrared small target detection using a novel four-leaf model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1462–1469, 2024.

[7] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, "A robust infrared small target detection algorithm based on human visual system," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168–2172, Dec. 2014.

[8] C. Liu, F. Xie, L. Qiu, H. Ji, and Z. Shi, "Infrared small target detection based on monogenic signal decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5004016.

[9] Z. Qiu, Y. Ma, F. Fan, J. Huang, and M. Wu, "Adaptive scale patch-based contrast measure for dim and small infrared target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[10] A. Ciocarlan, S. Le Hegarat-Mascle, S. Lefebvre, A. Woiselle, and C. Barbanson, "A contrario paradigm for yolo-based infrared small target detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 5630–5634.

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

[12] 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.

[13] W. Duan, L. Ji, S. Chen, S. Zhu, and M. Ye, "Triple-domain feature learning with frequency-aware memory enhancement for moving infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5006014.

[14] S. Chen, L. Ji, J. Zhu, M. Ye, and X. Yao, "SSTNet: Sliced spatio-temporal network with cross-slice ConvLSTM for moving infrared dim-small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5000912.

[15] 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, Oct. 1999.

[16] 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, Jun. 2010.

[17] 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.

[18] S. Qi, G. Xu, Z. Mou, D. Huang, and X. Zheng, "A fast-saliency method for real-time infrared small target detection," *Infr. Phys. Technol.*, vol. 77, pp. 440–450, Jul. 2016.

[19] H. Yao, L. Liu, Y. Wei, D. Chen, and M. Tong, "Infrared small-target detection using multidirectional local difference measure weighted by entropy," *Sustainability*, vol. 15, no. 3, p. 1902, Jan. 2023.

[20] 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.

[21] 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.

[22] 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. (ICCV)*, Oct. 2019, pp. 8508–8517.

[23] 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. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8010–8024, 2022.

[24] M. Shi and H. Wang, "Infrared dim and small target detection based on denoising autoencoder network," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1469–1483, Aug. 2020.

[25] H. Fang, M. Xia, G. Zhou, Y. Chang, and L. Yan, "Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[26] 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. (CVPR)*, Jun. 2022, pp. 877–886.

[27] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013.

[28] 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.

[29] 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, p. 3200, Aug. 2021.

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

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

[32] Q. Hou, L. Zhang, F. Tan, Y. Xi, H. Zheng, and N. Li, "ISTDU-net: Infrared small-target detection U-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[33] Y. Dai, X. Li, F. Zhou, Y. Qian, Y. Chen, and J. Yang, "One-stage cascade refinement networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5000917.

[34] T. Soni, J. R. Zeidler, and W. H. Ku, "Performance evaluation of 2-D adaptive prediction filters for detection of small objects in image data," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 327–340, Jul. 1993.

[35] Y. Zhao, H. Pan, C. Du, Y. Peng, and Y. Zheng, "Bilateral two-dimensional least mean square filter for infrared small target detection," *Infr. Phys. Technol.*, vol. 65, pp. 17–23, Jul. 2014.

[36] 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.

[37] 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.

[38] C. Xia, X. Li, L. Zhao, and R. Shu, "Infrared small target detection based on multiscale local contrast measure using local energy factor," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 157–161, Jan. 2020.

[39] Y. Qin, L. Bruzzone, C. Gao, and B. Li, "Infrared small target detection based on facet kernel and random Walker," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7104–7118, Sep. 2019.

[40] Y. Chen, G. Zhang, Y. Ma, J. U. Kang, and C. Kwan, "Small infrared target detection based on fast adaptive masking and scaling with iterative segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[41] M. Nasiri and S. Chehresa, "Infrared small target enhancement based on variance difference," *Infr. Phys. Technol.*, vol. 82, pp. 107–119, May 2017.

[42] Y. He, C. Zhang, T. Mu, T. Yan, Y. Wang, and Z. Chen, "Multiscale local gray dynamic range method for infrared small-target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 10, pp. 1846–1850, Oct. 2021.

[43] 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.

[44] 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," *Infr. Phys. Technol.*, vol. 81, pp. 182–194, Mar. 2017.

[45] 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, p. 1821, Nov. 2018.

[46] P. Zhang, L. Zhang, X. Wang, F. Shen, T. Pu, and C. Fei, "Edge and corner awareness-based spatial–temporal tensor model for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10708–10724, Dec. 2021.

[47] Y. He, M. Li, J. Zhang, and Q. An, "Small infrared target detection based on low-rank and sparse representation," *Infr. Phys. Technol.*, vol. 68, pp. 98–109, Jan. 2015.

[48] M. Zhao, W. Li, L. Li, P. Ma, Z. Cai, and R. Tao, "Three-order tensor creation and Tucker decomposition for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5000216.

[49] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1442–1450.

[50] J. Ma, H. Guo, S. Rong, J. Feng, and B. He, "Infrared dim and small target detection based on background prediction," *Remote Sens.*, vol. 15, no. 15, p. 3749, Jul. 2023.

[51] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 949–958.

[52] 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, pp. 1–5, 2022.

[53] K. Zhao and W. Xiong, "Exploring region features in remote sensing image captioning," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 127, Mar. 2024, Art. no. 103672.

[54] Y. Wei, L. Li, and S. Geng, "Remote sensing image captioning using hire-MLP," in *Proc. 4th Int. Conf. Comput. Vis., Image Deep Learn. (CVIDL)*, May 2023, pp. 109–112.

[55] Z. Ni, Z. Zong, and P. Ren, "Incorporating object counts into remote sensing image captioning," *Int. J. Digit. Earth*, vol. 17, no. 1, Dec. 2024, Art. no. 2392847.

[56] L. Wietzke, O. Fleischmann, and G. Sommer, "2D image analysis by generalized Hilbert transforms in conformal space," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, D. Forsyth, P. Torr, and A. Zisserman, Eds., Berlin, Germany: Springer, Jan. 2008, pp. 638–649.

[57] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3136–3144, Dec. 2001.

[58] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Riesz pyramids for fast phase-based video magnification," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2014, pp. 1–10.

[59] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 321–324.

[60] C. Liu, F. Xie, X. Dong, H. Gao, and H. Zhang, "Small target detection from infrared remote sensing images using local adaptive thresholding," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1941–1952, 2022.

[61] M. Brown, R. Szeliski, and S. Winder, "Multi-image matching using multi-scale oriented patches," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 510–517.

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

[63] X. Kong, C. Yang, S. Cao, C. Li, and Z. Peng, "Infrared small target detection via nonconvex tensor fibered rank approximation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5000321.

[64] B. Hui et al., "A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background," *Sci. Data Bank*, vol. 5, p. 12, Oct. 2019.

[65] 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.

[66] C. Gao, L. Wang, Y. Xiao, Q. Zhao, and D. Meng, "Infrared small-dim target detection based on Markov random field guided noise modeling," *Pattern Recognit.*, vol. 76, pp. 463–475, Apr. 2018.

[67] M. Zhang, R. Zhang, J. Zhang, J. Guo, Y. Li, and X. Gao, "Dim2Clear network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5001714.

[68] S. Xu et al., "HCF-net: Hierarchical context fusion network for infrared small object detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2024, pp. 1–6.

[69] 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, 2024, Art. no. 5002615.

Chang Liu received the Ph.D. degree in pattern recognition and intelligent systems from Beihang University, Beijing, China, in 2024.

He is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include infrared small target detection, remote sensing image analysis, image quality assessment, and image quality enhancement.

Xuedong Song received the master's degree in pattern recognition and intelligent systems from Beihang University, Beijing, China, in 2021.

His research interests include remote sensing image processing and small object recognition.

Dianyu Yu received the B.S. degree in detection, guidance, and control techniques from Beihang University, Beijing, China, in 2023, where he is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the Image Processing Center, School of Astronautics.

His research interests include infrared small target detection and multi-modal learning.

Linwei Qiu received the B.S. degree in detection, guidance, and control techniques from Beihang University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the School of Astronautics.

His research interests include low-level vision tasks and medical image analysis.

Fengying Xie (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, in 2009.

She was a Visiting Scholar with the Laboratory for Image and Video Engineering, The University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. She is currently a Professor with the Image Processing Center, School of Astronautics, Beihang University. Her research interests include biomedical image processing, remote sensing image understanding and application, image quality assessment, and object recognition.

Yue Zi received the Ph.D. degree in pattern recognition and intelligent systems from Beihang University, Beijing, China, in 2023.

He is currently a Lecturer with the School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha, China. His research interests include remote sensing image processing, target detection, semantic segmentation, and deep learning.

Zhenwei Shi (Senior Member, IEEE) is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing, China. He has authored or co-authored over 200 scientific articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and the IEEE International Conference on Computer Vision (ICCV). His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Prof. Shi serves as an Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition*, *ISPRS Journal of Photogrammetry and Remote Sensing*, and *Infrared Physics and Technology*. His personal website is http://levir.buaa.edu.cn/.