Multiscale Progressive Fusion Filter Network for Infrared Small Target Detection

Pengfei Zhang, Zhile Wang, Guangzhen Bao, Jianming Hu, *Student Member, IEEE*, Tianjun Shi, *Graduate Student Member, IEEE*, Guanjie Sun, and Jinnan Gong

***Abstract*— Infrared small target detection is widely used in remote sensing fields. However, the application scenes of space-based remote sensing imaging often lead to problems such as small target scale, weak energy, and serious influence by strong clutters. At present, traditional methods are often difficult to adapt to the change of target scale. And deep-learning methods often difficult to extract small target features, and the change in imaging characteristics also brings challenges to the generalization ability. To complement each other's advantages, we propose an infrared small target detection method that combines the traditional methods with the deep-learning methods. First, we construct a multistage feature extraction network for guiding the typical multiscale traditional filtering results to progressively fuse. Second, we propose a multiscale attention supervision (MAS) module to adjust the semantic consistency of different stages, improving the network generalization ability. Next, a dynamic weight convolution (DWC) module is utilized to obtain the optimal distribution of grayscale in the neighborhood. Finally, we use the background modeling results to suppress the background, effectively weakening the influence of background clutters and enhancing the target contrast. Experimental results show that the proposed method has good detection results for targets with different scales and signal-to-clutter ratios in a variety of complex scenes. Compared with the typical methods, our method has better detection performance and generalization ability.**

***Index Terms*— Attention mechanism, complex background, infrared small target detection, multiscale information, spatial filtering.**

I. Introduction

BECAUSE most objects have the ability to emit, reflect, and absorb electromagnetic waves, their thermal radiation characteristics could be collected by infrared imaging systems and be different due to environments, specific heat, and states. The infrared remote sensing imaging system [1], [2] has the characteristics of lightweight, small size, and strong anti-interference ability. It has important research significance and engineering application value in sky threat target perception, fire warning, astronomical planet detection, night security, and other aspects [3], [4], [5], [6]. However, in the practical application scenes of space-based infrared remote sensing, the target becomes small and weak due to the long detection distance, which means it only occupies less than 80 pixels (9×9) in most situations and its signal-to-clutter ratios could be less than 6 [7]. Besides, the change in orbit height, atmospheric radiation absorption, and the target's own radiation make the scale and energy of the target change dynamically, and targets are more likely to be submerged in the background. In summary, infrared small target detection has become a challenging task in the field of remote sensing image processing.

At present, single-frame infrared small-target detection methods can be divided into three categories: local information-based methods, background suppression methods, and deep-learning-based methods [8]. The first category of methods makes use of the saliency of the target, such as the intensity, contrast, and gradient characteristics. The typical one is the local contrast method (LCM) proposed by Chen et al. [9] based on the human visual system (HVS). After that, Han et al. [10] developed a multiscale relative LCM (RLCM) to adapt different target sizes in actual scenes. Liu et al. [11] proposed a novel IR small target detection method utilizing halo structure prior-based LCM (HSPLCM). However, such extraction methods require the target to have strong energy or large contrast and gradient with its surroundings. When the target energy is weak, the contrast is low, or there are other strong clutters that are similar to the target characteristics in the image, the performance of the algorithm will be degraded. At the same time, most of these methods adopt block processing requiring the target size to be the same or slightly smaller than the block size, which has certain limitations for the detection of variable scale targets.

The background suppression methods apply filters or use the sparsity of the target and low rank of the background to build a background model. Traditional filtering methods include Top-hat [12], Max-Mean [13], and Max-Median [13]. For image data structure-based methods, Gao et al. [14] developed a novel IR patch-image (IPI) model based on the nonlocal self-correlation property. Dai and Wu [42] proposed a reweighted infrared patch-tensor (RIPT) model to extend the IPI model to patch-tensor space and Zhang and Peng [43] proposed a nonconvex approach based on the partial sum of the tensor nuclear norm (PSTNN) to better approximate the tensor rank.

Manuscript received 3 February 2023; revised 24 May 2023, 18 October 2023, and 11 November 2023; accepted 6 December 2023. Date of publication 18 December 2023; date of current version 29 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62101160 and in part by the National Key Laboratory of Air-Based Information Perception and Fusion under Grant 20220001077001. *(Corresponding author: Jinnan Gong.)*

Pengfei Zhang, Zhile Wang, Guangzhen Bao, Jianming Hu, Tianjun Shi, and Jinnan Gong are with the Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: 22B921003@stu.hit.edu.cn; wangzhile@hit.edu.cn; bgz@stu.hit.edu.cn; hjm1007491571@163.com; 22B921004@stu.hit.edu.cn; gongjinnan@hit.edu.cn).

Guanjie Sun is with the China Academy of Space Technology, Beijing 100094, China (e-mail: sunguanjie1991@163.com).

Digital Object Identifier 10.1109/TGRS.2023.3343496

1558-0644 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

In order to enhance the ability to detect targets and exploit more spatial and structural information, Zhao et al. [15] developed a novel method using multiple morphological profiles (MMPs) and an effective method based on three-order tensor creation and Tucker decomposition (TCTD) [16]. However, data structure-based methods are often difficult to apply in real-time due to complex calculations. Differentiating modeling results from the original image, it can enhance the target features and reduce the fluctuation of background clutters. However, due to the certain size of the filter window used in the spatial filtering method, the background clutters residue is serious when the window is large, while the target is easily estimated as the background when the window is small, resulting in missed detection. At the same time, such methods based on artificial features often have limited background suppression ability for strong clutter fluctuations.

The deep-learning method of small target detection is mainly divided into two categories. One of them detects the target directly, such as ALCNet [17], DNA-Net [18], LPNet [19], and IAANet [20]. The other type is based on generative adversarial network data augmentation methods, such as MDvsFA-cGAN [21] and IRSTD-GAN [22]. By transforming image features into higher level and more abstract expressions, deep-learning methods perform better compared to traditional detection methods. However, infrared small targets contain almost no texture information, which makes it difficult for convolutional networks to extract features and makes the features easily lost during information transfer. And the seriously unbalanced proportion of positive and negative samples also affects the training of the network. Under the condition of space-based remote sensing observation, the observation attitude, imaging range, ground object scenes, and other factors, which change significantly over time, will obviously affect the imaging results. These also bring serious challenges to the generalization ability and robustness of deep-learning methods.

To obtain better detection performance and solve the limitations of the above typical methods, this article proposes a multiscale progressive fusion filter network named MPFFNet. This network is constructed based on the combination of deep learning and traditional filters. Instead of directly generating the background result, it extracts the image features to generate the weight, which guides the fusion of the corresponding traditional filtering results. In the aspect of traditional spatial filtering, it provides additional robust features for the network while generating the background filtering results. And in the aspect of deep learning, it replaces the hand-crafted features in the traditional method, so as to obtain more precise background modeling results with more generalization ability. By focusing more on background features rather than target features, our method reduces the influence of target feature sparsity and unbalanced proportion of positive and negative samples in deep learning. The combination of traditional methods and deep-learning methods also helps to improve the interpretability and generalization of our method.

The main contributions of this work are summarized as follows.

1) A multiscale progressive fusion filter background suppression method based on the combination of traditional methods and deep learning is proposed, effectively improving the adaptability to targets of different scales and images of different resolutions.
2) A multiscale attention supervision (MAS) module is used to constrain the background modeling results and narrow the semantic differences transmitted across stages between different scale features, strengthening thus the generalization ability and robustness of the algorithm under heterogeneous data.
3) A dynamic weight convolution (DWC) module utilizes the spatial correlation of backgrounds and adaptively changes the convolution weights according to the background detail features, effectively improving the accuracy of background modeling and the ability of target extraction.

The rest of this article is organized as follows. Section II introduces the related work and existing problems of space-based infrared small target detection. Section III illustrates the rationale and details of the proposed method. In Section IV, we compare the performance of the proposed method with typical methods and present experimental results on two datasets. Finally, we summarize and conclude in Section V.

II. Previous Related Research

In recent years, there have been many developments in infrared weak and small target detection, the most notable of which is the local contrast target detection method based on the HVS. The LCM takes advantage of the selective attention feature of human vision, that is, in a complex background, salient regions can quickly attract human attention. For this, researchers have proposed many classical algorithms, such as the improved LCM method (ILCM) [23], relative local contrast measure method (RLCM) [10], the multiscale local contrast measure method (MLCM) [24], the absolute directional mean difference method (ADMD) [25], and the multiscale patch-based contrast metric method (MPCM) [26]. The above methods have proved their excellent performance in many fields. But in the field of space-based infrared remote sensing, due to the long detection distance and obvious atmospheric absorption, the target energy is weaker and its significance is greatly reduced, which has a significant impact on the detection. At the same time, the above-mentioned methods usually pay more attention to the characteristics of the target instead of the background. For the situation with a large field of view and complex imaging background in the field of space-based infrared remote sensing imaging, it is easy to detect the background clutters with high contrast as the targets, which leads to a large number of false alarms.

Therefore, we put our sight back to the more classical methods in the field of space-based infrared remote sensing target detection, that is, the background suppression methods based on traditional spatial filtering. At the same time, we also research the deep-learning target detection methods that have

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

performed well in the field of image processing in recent years. We briefly explain the advantages and disadvantages of these two methods and propose a new target detection method that combines the advantages of the two methods, trying to tap new potentials from the classic methods of traditional spatial filtering.

A. Traditional Spatial Filtering Methods

For the early traditional spatial filtering methods, such as median filter, mean filter, and Tophat, their application scenes are relatively limited, which usually require the background to be relatively flat. Otherwise, the background details may be lost. To solve such problems, researchers have proposed many new filtering methods, such as max-median filter [13], minimum gradient median filter (MGMF) [27], 2-D least mean square (TDLMS) [28] error filter, and bilateral filter (BF) [29]. In these methods, the optimal filtering parameters are selected based on hand-crafted features, such as the gradient, gray level, nonlocal self-correlation, and distance from the center pixel to better retain details such as background edges and features.

However, such methods often only extract the shallow information of the image, so the performance has disadvantages compared with deep-learning methods. A single traditional airspace filtering method usually only considers one or a few preset application scenes in principle design. While space-based infrared remote sensing imaging, its field of view usually includes plains, mountains, coasts, oceans, islands, stratus clouds, cirrus clouds, and other scene types at the same time, the grayscale distribution of the background is quite different under different observation time and angle conditions. Therefore, for such complex background applications, although a single traditional filtering method can work normally, it is often difficult to have good performance for various types of backgrounds. At the same time, because the size of the filter window is usually fixed, it is difficult to adapt to the situation where the target scale changes.

The advantage of traditional spatial filtering methods is their strong robustness. The design criterion of traditional spatial filters is based on the universal characteristic of background spatial correlation. That is to say, even if the input is other images that are completely unrelated to remote sensing images, as long as the image has a certain spatial correlation (which is also the common feature in most real images), traditional spatial filters work fine. For training-based deep-learning methods, if the input image is quite different from the data distribution type in the training set, it may get completely unexpected results. At the same time, because traditional spatial filters are based on spatial correlation, they pay more attention to the background characteristics, so they can better suppress the generation of false alarms in complex backgrounds. Varying from the aforementioned works, we provide new insights on using convolutional neural networks for replacing the hand-crafted features in the traditional methods, to obtain better performance and retain the benefits of traditional methods.

B. Deep-Learning Methods

For the deep-learning methods of image processing, the current mainstreams are the convolutional neural networks. Their typical representatives are YOLOv4 [30], Faster R-CNN [31], and so on. Convolutional neural networks have the characteristics of good real-time, excellent performance, and the ability to extract deep-level features. In remote sensing image segmentation tasks, some deep-learning methods [39], [40], [41] also demonstrate excellent performance advantages. In the research of convolutional networks, the multistage network structure [32], [33], [34], [35] is proposed to divide the task into multiple stages. Each stage is processed by a lightweight subnetwork, so as to realize the function gradually and show better performance on advanced tasks such as image restoration. The attention mechanism is to enhance the key information by extracting the dependence of features in the spatial and channel dimensions [36], [37], which greatly enhances the performance and stability of the network. Deep-learning methods have proved their effectiveness in image restoration, super-resolution, classification, and segmentation. Compared with most traditional methods, they have incomparable advantages in performance.

In the field of target detection, the convolutional neural network usually focuses on the target itself, which uses a multilayer network to extract the feature of the target to detect. This approach has shown its superior performance in area target detection. But for infrared small targets, their size is often smaller than 9×9 pixels. And in extreme cases, it may be smaller than 3×3 pixels. Therefore, the infrared small targets almost do not contain texture features, which is easy to causes the loss of small target features in the process of multilayer downsampling feature extraction. For infrared small targets, the proportion of pixels in the image is often less than 0.1%. The large difference in the number of positive and negative samples also makes network training difficult to converge. It also is difficult to obtain enough datasets to support the generalization of the network. When the altitude of the satellite orbit changes, the observation pointing changes, and the infrared spectrum of the camera changes, the real input image may be quite different from the image data distribution of the training set, resulting in network performance degradation or even completely wrong results. Due to the "black-box" problem and poor interpretability of the deep-learning methods, it is sometimes difficult to be applied in fields such as aviation, aerospace, and so on. Therefore, our method focuses more on background features and introduces adjustable traditional robust features, reducing the influence of target feature sparsity and improving the interpretability of the method.

III. Proposed Method

A. Method Overview

In view of the above problems and facing the characteristics of dynamic changes in space-based remote sensing application scenes, we innovatively propose an MPFFNet. It utilizes the neural networks to guide the fusion of multitype and multiscale traditional filtering results, obtaining better background suppression performance and better generalization ability in heterogeneous data. The structure of our method refers to a multistage processing network MPRNet [32], which has an

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

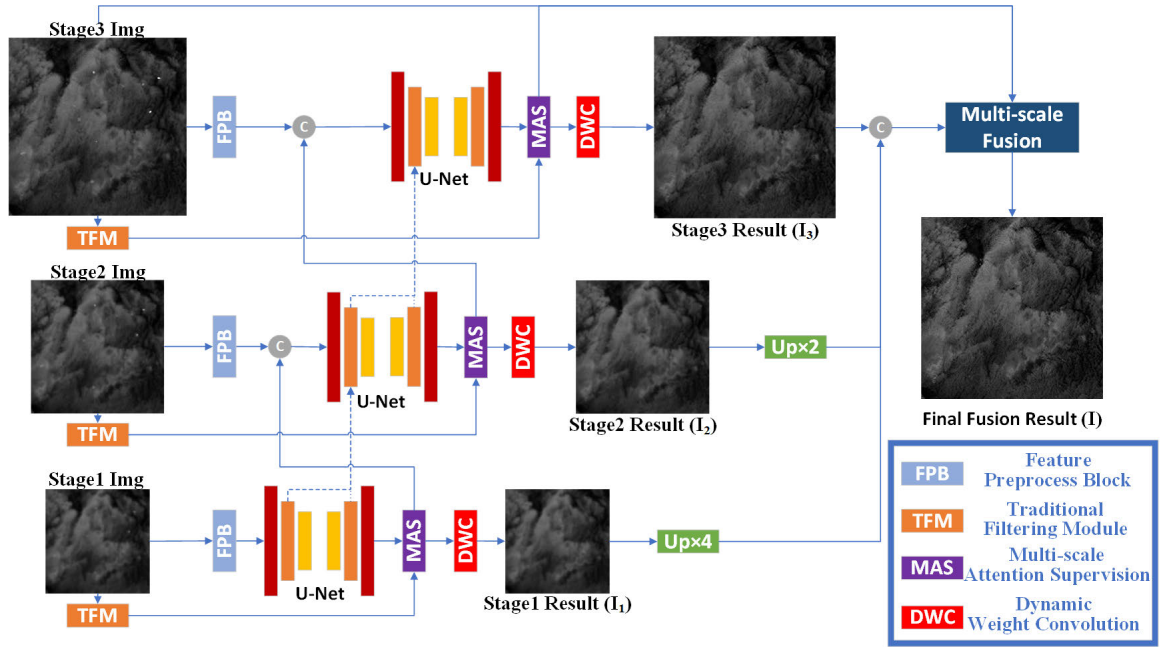


Fig. 1. Overview of MPFFNet. The subnetwork (U-Net) is used to extract the features of the image with different scales. MAS module and the DWC module are proposed to effectively improve the suppression ability of the complex background. A multiscale fusion module is performed to obtain the background modeling results that adapt to the changes in target scale and background resolution. Finally, we use the threshold segmentation method to extract the targets from the difference image of the final background modeling result and the original image.

excellent performance in image restoration. We also modify it to better combine with traditional methods and adapt to the application background of space-based infrared small target detection.

The structure is divided into three stages corresponding to three downsampling scales, which helps to better adapt to images with different resolutions and targets with different scales, improving the accuracy of background modeling and enlarging the receptive field. The specific network structure of MPFFNet is shown in Fig. 1. The traditional filtering module (TFM) extracts the robust background features by various classical spatial filters. The subnetwork extracts the features of the image with different scales to improve the adaptability to the various targets and backgrounds. Jump connections are used between the feature extraction subnetworks at each stage, which simplifies the information flow and makes the overall network convergence process more stable. To effectively improve the suppression ability of the complex background, we propose the MAS module and the DWC module, which will be focused on in Section III-C and III-D as the main innovation. Finally, a multiscale fusion module is performed to obtain the background modeling results that adapt to the changes in target scale and background resolution. After the background is suppressed by the difference between the original image and the background modeling result, we use constant false alarm rate (CFAR) detection to get the final target detection results.

B. TFM and Subnetwork Structure

As shown in Fig. 2, the TFM is a collection of image-filtering results extracted by various classical spatial filters. It provides input for background modeling results fusion and robust features based on traditional filtering for the

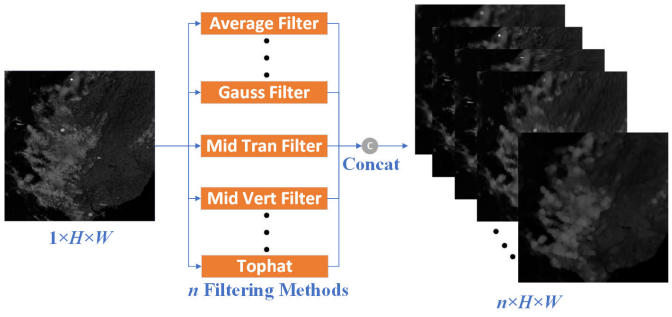


Fig. 2. Design of the TFM. The module provides image-filtering results and robust features based on traditional filtering. And the traditional filter types can be flexibly changed to adapt to different application scenarios.

MAS module. Besides, the traditional filter types can be flexibly optimized and selected according to the characteristics of the background image to improve the background suppression performance.

Background feature extraction subnetwork is a classic encoding-decoding structure based on the U-Net. The images at each stage are input into the subnetwork structure after preliminary feature processing and channel adjustment through the feature preprocessing block (FPB) composed of three layers of simple convolution. To better extract features and focus on target regions, the subnetwork adopts spatial channel attention block (SCAB) based on CBAM [37] as the base convolution module. The overall multistage network and the intrinsic encoder–decoder subnetwork are constructed based on the feature pyramid structure but are mirror-symmetrical in structure (as shown in Fig. 3). The former extracts image features from top to bottom, focusing on the recovery of spatial details. While the latter extracts image features from bottom

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

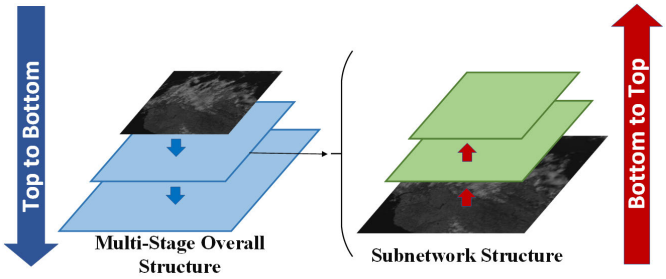


Fig. 3. Feature pyramid nested structure. This interesting structure helps to fuse semantic information and spatial information.

to top, focusing on semantic information extraction. This interesting structure helps to fuse the semantic information and spatial information and supplement shallow information while extracting deep-level features to achieve progressive background detail recovery.

C. Multiscale Attention Supervision

As we all know, among the feature maps of different scales, the large-scale feature maps usually contain more texture details, and the small-scale feature maps have richer semantic information. Among them, the semantic features help to describe the difference between the target and the background, while the texture details help to restore the edge features of the background to improve the modeling accuracy. For small infrared target detection, the target distribution is usually extremely sparse, often occupying less than 0.1% of the image. The background occupies the vast majority, which causes the network to focus more on the recovery of background texture details rather than the extraction of deeper semantic information. Therefore, when the network starts from small-scale fuzzy image features and gradually generates large-scale clear image features, the feature map often loses information due to the large semantic gap in the cross-scale feature transfer. To reduce the semantic difference between features at different scales and the loss of information in feature transfer, we propose the MAS module. The design of MAS is shown in Fig. 4.

MAS first obtains the input features $F_{\text{in}} \in \mathbb{R}^{C\times H\times W}$, passes them through the classic bottleneck convolution block (BCB) and the Softmax activation function and generates the fusion weight of the filtering result $W_F \in \mathbb{R}^{n\times H\times W}$. Here, n represents the number of channels, which is consistent with the number of traditional filtering results. H and W represent the spatial dimension of the image. The fusion weight W_F can be written as

$$W_F = \text{Softmax}(\text{BCB}(F_{\text{in}})). \tag{1}$$

Each value on the channel dimension of W_F represents the contribution ratio of the corresponding traditional filtering result in the current pixel position, and the sum of all values of each channel is 1. By artificially adjusting the values of W_F, we can introduce additional subjective tendencies into the fusion results, which is equivalent to adding the man-in-the-loop technique to the system. By limiting the upper and lower limits of filter weight, we can use the performance

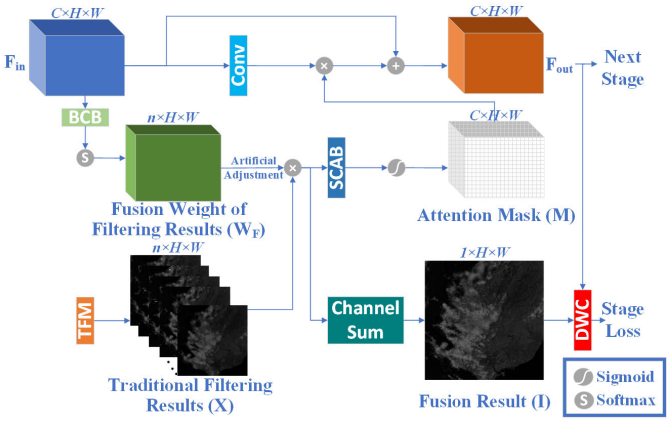


Fig. 4. Design of the MAS. The module acts as a decoder, decoding the implicit high-dimensional feature of the image into explicit fusion weights. MAS also generates an attention mask to modulate the input features, reducing the semantic difference between features at different scales.

prior knowledge of traditional filtering results in different imaging scenarios to assist in background modeling, which can improve the network's modeling adaptability to different scenarios and help enhance its generalization ability. The adjusted fusion weight W_F is multiplied element by element with the image-filtering results $X \in \mathbb{R}^{n\times H\times W}$ from the output of the TFM, and then summed along the channel dimension to obtain the fusion result of the corresponding scale $I \in \mathbb{R}^{1\times H\times W}$ at the current stage. I is expressed as

$$I = \sum_{i=1}^{n} \text{multiply}\left(X^i, W_F^i\right) \tag{2}$$

where X^i and W_F^i represent the value of X and W_F in the ith channel.

After being processed by the subsequent DWC module, the fusion results of each stage and scale are supervised and constrained by the real background image (the ground truth) and the loss function. In terms of feature transfer, the fusion weight W_F and the traditional filtering results X are processed by SCAB to generate an attention mask $M \in \mathbb{R}^{C\times H\times W}$, which is used to modulate the input features F_{in}. The attention mask M and the output feature F_{out} can be formulated as

$$M = \text{Sigmoid}\left(\text{Softmax}\left(\text{multiply}\left(X^i, W_F^i\right)\right)\right) \tag{3}$$

$$F_{\text{out}} = F_{\text{in}} + \text{multiply}(\text{conv}_{1\times 1}(F_{\text{in}}), M) \tag{4}$$

where $\text{conv}_{1\times 1}$ denotes the 1×1 convolutional transformation.

Finally, the feature F_{out} adjusted by the attention mask M is passed to the next stage network for processing. In this process, MAS acts as a decoder, decoding the implicit and uninterpretable high-dimensional feature of the image F_{in} into explicit and understandable fusion weights W_F. For the practical application of remote sensing imaging, there are often problems such as a lack of training data or significant differences between training data and real data. Therefore, we can introduce the man-in-the-loop technique to manually supervise the fusion results, that is, to make the output fusion results pay more attention to certain filters and ensure the lower limit of performance under heterogeneous data, which

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

helps the algorithm better adapt to untrained remote sensing imaging scenes. As the training data gradually enriches, this restriction will be lifted and the optimal fusion weight will be obtained through training. This enhances the controllability, flexibility, and interpretability of the algorithm. As stable and robust image features input, traditional filtering results help the network improve its generalization ability on remote sensing datasets with limited samples.

In terms of feature transfer, the module modulates the feature output of the current stage by an attention mask M, which alleviates the loss of cross-stage transfer and makes network optimization more stable. At the same time, the adjustment of feature output by spatial and channel attention also helps to improve the ability of the network to distinguish targets from different backgrounds (e.g., clouds, mountains, and oceans).

D. Dynamic Weight Convolution

According to the research on traditional filtering methods, the pixel gray value of each position in the filtering result is obtained by the gray distribution statistics in the neighborhood, and the weights of different positions are often biased. For example, the max-median filter selects the direction with the largest median value for statistics, the MGMF selects the direction with the smallest gradient for statistics, and the BF adjusts the statistical weight based on the correlation between grayscale and distance to better retain the edge and texture details. These methods take advantage of the difference in spatial correlation between the target and the background. Our method is based on the fusion of traditional filtering results of different types and scales, it only considers the channel interaction of the different filtering results but ignores the spatial correlation. Therefore, our method lacks subjective processing of the gray distribution in the neighborhood, which limits the expression ability of the background modeling to a certain extent. For this reason, we propose the DWC module, which uses deep learning to obtain the optimal distribution of grayscale in the neighborhood and dynamically change the convolution parameter distribution with the background type. Through neighborhood gray-scale modulation, our method can obtain more precise background modeling results.

To achieve the best neighborhood grayscale modulation effect, we need to generate independent convolutional weights for each pixel in the image separately. However, if traditional convolution wants to achieve this, it needs to use different kernels to convolute the image n^2 times, which introduces a time complexity of $O(n^2)$, and then extract the corresponding components of each pixel. In order to optimize the calculation efficiency, we transform the spatial 2-D matrix distribution within the neighborhood window into a 1-D vector distribution on the channel, achieving dimensionality reduction of operations. This preprocessing method shown in Fig. 5(a) transforms the spatial convolution operation on the 2-D plane into the channel-weighted summation operation in the 3-D space to meet the requirements of GPU parallel operation, thus reducing the time complexity from $O(n^2)$ to $O(1)$. Let the size of the predesigned convolution kernel be $2 \times B + 1$, and 3×3 is taken as an example in the figure. First, the background modeling result is expanded with a width of 1, and the image size becomes $1 \times (H + 2) \times (W + 2)$. Then, the expanded image is translationally extracted in the range of the original image size, and the results of each extraction are stacked in the channel dimension. Finally, we obtain a $9 \times H \times W$ region block, the 1-D channel value at any position in the region block is the same as the 2-D neighborhood value at the corresponding position of the original background modeling result. Thus, the convolution operation in the spatial domain can be equivalently converted into the addition operation along the channel dimension [as shown in Fig. 5(b)].

As shown in Fig. 6, the features of each stage are input to the DWC module, and the convolution weights are generated through the classic bottleneck convolution. The number of channels is adjusted to match the predesigned convolution kernel size. The 1-D vector of the channel at any position of the convolution weights represents a corresponding convolution kernel. The convolution weights are activated by Softmax to ensure that the sum of the convolution kernel values is 1. For each pixel position, the DWC module dynamically generates a corresponding convolution kernel based on its neighborhood feature. Finally, the convolution weights and the region block are multiplied element by element and summed along the channel dimension to output the background modeling results. Subsequent experiments show that the DWC module effectively improves the accuracy of background modeling and the ability to extract weak targets.

E. Multiscale Fusion

Due to the spatial correlation of the image background, it is less sensitive to the downsampling operation. On the contrary, the target features are easily lost after downsampling and filtering. Therefore, the multistage image processing of different scales is beneficial to distinguishing target features from background features and removing targets from background modeling results. As shown in Fig. 7, by fusing the background modeling results of different stages and different scales, the module enhances the ability to retain the background information, thus achieving a better background suppression effect.

For the background modeling results of each stage, the real background images at the same scale are used to progressively supervise the features and modeling results from blurry to clear. It helps to ensure the effectiveness and optimality of the features in each stage. Considering the balance between the overall image modeling accuracy, texture detail restoration accuracy, and the target sparseness influence, we design the following loss functions to optimize and constrain MPFFNet:

$$L = \sum_{S=1}^{3} \left[L_{\text{Char}}(I_S, B_S) + \lambda_1 L_{\text{edge}}(I_S, B_S) \right] + \lambda_2 L_{\text{tar}}(I, B) \quad (5)$$

where I_S represents the background modeling result of a certain stage, B_S represents the real background image of the corresponding scale, and λ_1 and λ_2 represent the relative contribution of edge loss and target, which are set to 0.05 [32].

L_{Char} and L_{edge} are Charbonnier loss and edge loss, which, respectively, represent the modeling accuracy of the overall image and texture details. L_{tar} is the focus loss of the target

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

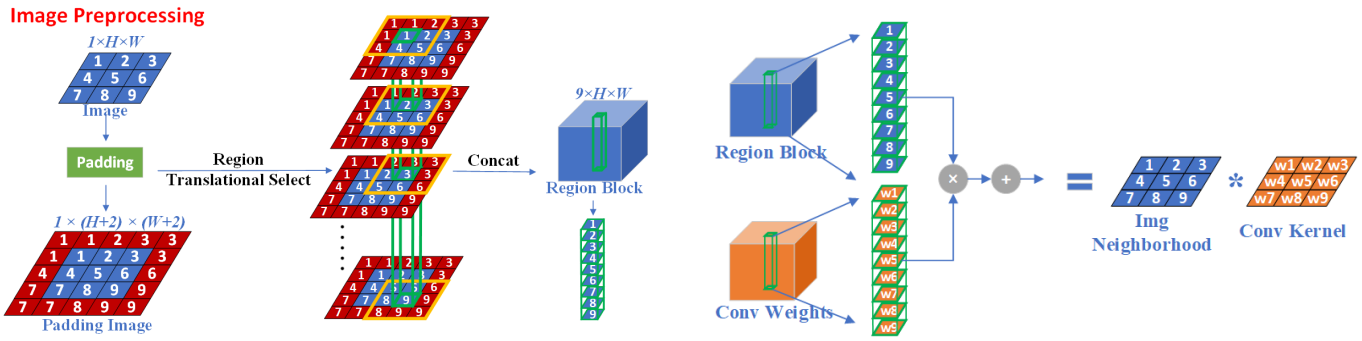


Fig. 5. (a) Design of the image preprocessing. Its processing time is related to the size of the predesigned convolution kernel. Take 3 × 3 as an example, the region selection steps only need nine times. (b) Spatial dimension convolution is equivalent to the channel dimension addition.

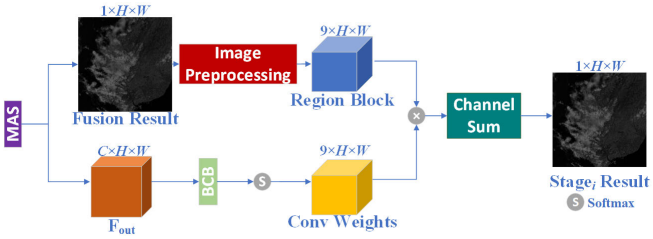


Fig. 6. Design of the DWC. The module modulates the image based on the gray distribution in the neighborhood to obtain more precise background modeling results.

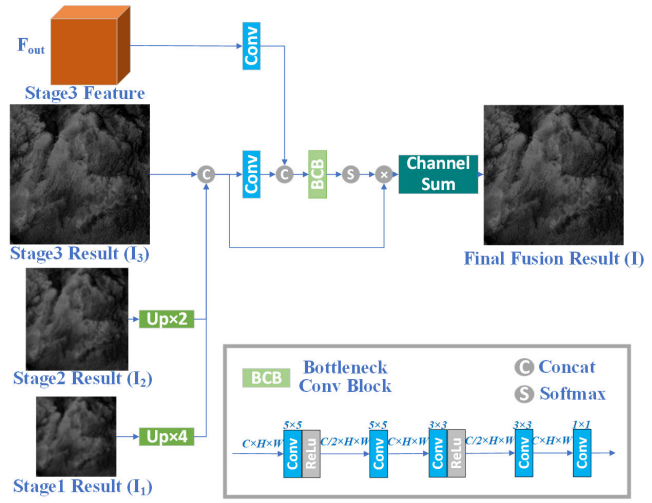


Fig. 7. Design of the multiscale fusion module. The module fuses the background modeling results of different stages and different scales to adapt to the changes in the target scale and achieve a better background suppression effect.

area, which is used to improve the target extraction ability, defined as

$$L_{\text{Char}} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(I(i)-B(i))^2+\varepsilon^2} \tag{6}$$

$$L_{\text{edge}} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(\Delta I(i)-\Delta B(i))^2+\varepsilon^2} \tag{7}$$

$$L_{\text{tar}} = \frac{1}{n_{\text{tar}}}\sum_{i=1}^{n_{\text{tar}}}\sqrt{(I_{\text{tar}}(i)-B_{\text{tar}}(i))^2+\varepsilon^2} \tag{8}$$

where Δ is the Laplacian operator, I_{tar} represents the background modeling result of only the target area, B_{tar} represents the corresponding real background image area, and n and n_{tar} are the number of pixels in the entire image and the target area.

Background suppression and target enhancement are realized through image differences between the final background modeling result obtained by MPFFNet and the original image. The difference image is divided into grid blocks, each block size is 64 × 64 pixels, and the segmentation threshold is calculated based on the CFAR criterion to segment and extract the target, which is defined as

$$\text{th} = \mu + k \cdot \sigma \tag{9}$$

where th is the segmentation threshold of the current image block, μ is the average value of the image block, σ is the standard deviation of the image block, and k is a constant.

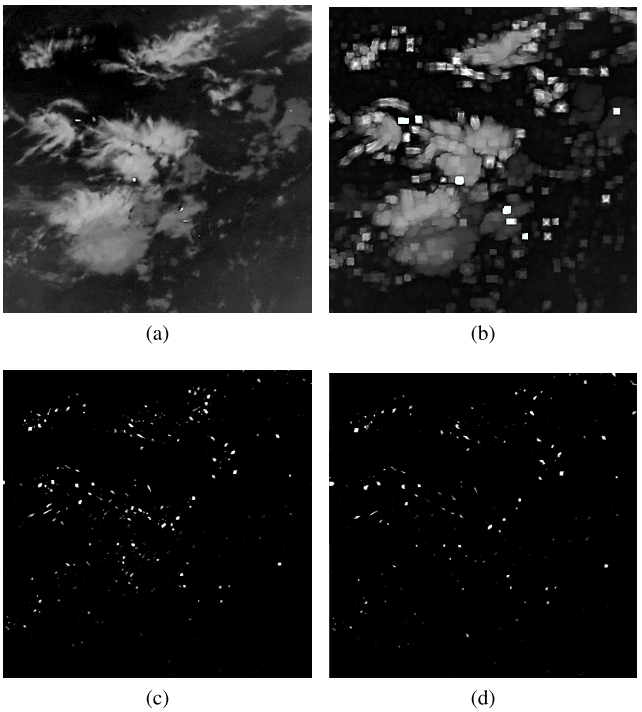
According to the above formula, we can get the final detection results.

IV. Experimental Results

In this section, we introduce the experimental settings, including the dataset, evaluation metrics, and some implementation specifics. Then, we will evaluate the algorithm in three aspects. First, the performance of the algorithm is evaluated on the infrared satellite remote sensing image dataset. Second, the generalization ability of the algorithm is evaluated on the SIRST dataset with large data differences. Finally, ablation experiments are conducted to analyze the impact of each functional module and verify the effectiveness of the proposed module.

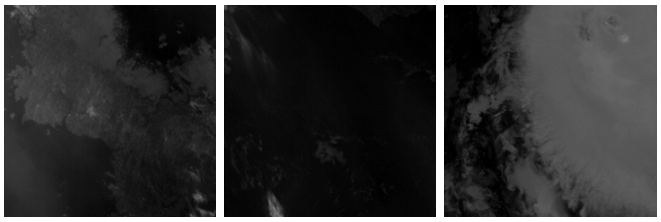
We present the comparison results of several typical methods and the proposed method to validate the performance of the algorithms. Among them, the traditional spatial filtering methods we choose are MGMF [27], BF [29], IPI [14], RIPT [42], and PSTNN [43]. The LCMs are not background modeling algorithms and introduce a large number of background clutter false alarms (as shown in Fig. 8), therefore they do not participate in the comparison. For the deep-learning algorithms, MPRNet [32], which performs well in image restoration tasks, is selected as a background modeling algorithm for comparison. We also select ALCNet [17] and

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c) (d)

Fig. 8. LCM results. (a) Original image. (b) MLCM [24]. (c) MPCM [26]. (d) ADMD [25].



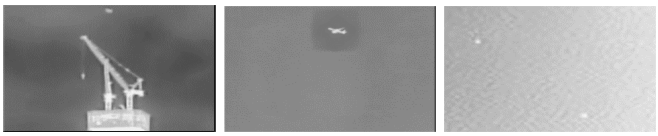
(a) (b) (c)

Fig. 9. Infrared remote sensing images. (a) Land infrared image. (b) Sea infrared image. (c) Cloud infrared image.

MDvsFA-cGAN [21] as state-of-the-art CNN-based methods for comparison.

A. Datasets and Evaluation Metrics

The training dataset is based on Landsat8 infrared spectrum images, it has ten thousand 512 × 512 pixels images obtained by random cropping, scaling, and rotation. The resolution of the training dataset is about 30–300 m. The test dataset is divided into two parts: the simulation images of the infrared staring remote sensing satellite with a large field of view and the public dataset of infrared small targets. Among them, the scale of the simulation image is about 3 × 3 k pixels, and the spatial resolution is about 800 m. As shown in Fig. 9, the imaging field of view contains a variety of background types such as clouds, land, ocean, and coast, which differs from the training dataset mainly in resolution and spectrum. We construct this test dataset with different target scales and signal-to-clutter ratios by randomly selecting regions in the image. The test image scale is 512 × 512 pixels. Morphological targets with various scales are added through the simulation method.



(a) (b) (c)

Fig. 10. SIRST images. (a) Scene type difference. (b) Target morphological difference. (c) Imaging noise difference.

Single-frame infrared small target (SIRST) [38] is selected as the public dataset of infrared small targets. This dataset is proposed by the Nanjing University of Aeronautics and Astronautics, mainly including the ground-based infrared camera and UAV aerial photography results. This dataset is different from satellite remote sensing imaging results in scene type, target morphology, imaging noise, and so on (as shown in Fig. 10), which is used to test the generalization ability and robustness of the algorithm on heterogeneous datasets. The target scale is generally distributed in the range from 5 × 5 pixels to 9 × 9 pixels, and the average signal-to-clutter ratio is about 20.

The evaluation metrics focus on two aspects: background suppression performance and target detection performance. The background suppression performance is evaluated by the background suppression factor (BSF) and the target signal-to-clutter ratio gain (SCRG), which, respectively, characterize the background modeling accuracy and target contrast enhancement ability. In statistics, to prevent the infinite value of the signal-to-clutter ratio caused by an image standard deviation of 0, the lower limit of standard deviation is set to 0.05. The BSF and SCRG are defined as

$$\mathrm{BSF} = \frac{(\sigma_B)_{\mathrm{in}}}{(\sigma_B)_{\mathrm{out}}} \tag{10}$$

$$\mathrm{SCR} = \frac{\mathrm{eng}}{\sigma_B}, \quad \mathrm{SCRG} = \frac{\mathrm{SCR}_{\mathrm{out}}}{\mathrm{SCR}_{\mathrm{in}}} \tag{11}$$

where SCR is the target signal-to-clutter ratio, eng is the target's energy, and σ_B is the image regional standard deviation (i.e., the image except the target region). We zero pixels with grayscale values less than 3 in the different images to reduce the impact of weak residual noise.

We choose the commonly used true positive rate (TPR) and the false positive rate (FPR) to evaluate the detection performance, constructing the characteristic change curve and the ROC curve for comparative evaluation. The TPR and FPR are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N_{\mathrm{img}}} \tag{13}$$

where TP, FP, and TN represent the true positive, false positive, and true negative, respectively, and N_{img} is the total number of pixels in the image.

B. Experimental Settings

Our MPFFNet is implemented using the PyTorch framework and trained end-to-end on an Nvidia RTX 2080 GPU. The training dataset has ten thousand 512 × 512 pixels images

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

based on the Landsat8 infrared spectrum, and all test networks did not use any images on the two test datasets for training. In the comparative experiments, we set the batch size to 2 and trained the model with 40 epochs. We use Adam optimizer with the initial learning rate of 1×10^{-4}, which is steadily decreased to 1×10^{-6} using the cosine annealing strategy. The random number seed of the model is fixed in the ablation experiment, and we initialize the weights and bias of our model using the Xavier method.

C. Algorithm Performance Experiments

To evaluate the detection performance of the algorithm on the infrared remote sensing image dataset, we carry out experiments on the test sets containing different scale targets and different SCR targets. Among them, in the test sets of different scales, the target scale ranges from 1×1 pixels to 9×9 pixels, and the SCR is 6. In the test sets of different SCRs, the SCR ranges from 3 to 10. In addition to Gaussian morphological targets of different scales, there are also real template morphological targets with scales larger than 11×11 pixels and with obvious tailing, which are more difficult to extract. The size of each test set is 100, and the number of targets is 10 in each image. The background suppression results for the target scale 5×5 pixels test set are shown in Fig. 11, which are obtained from the residuals between the original image and the background modeling results of each method. The red boxes are the areas where the real targets are located. In the table, bold font represents the maximum value, and underlined font represents the second largest value.

The experimental results of the target background suppression performance on the infrared remote sensing image dataset are shown in Table I. In the table, because ALCNet and MDvsFA-cGAN are not background suppression algorithms, BSF and SCRG metrics cannot be counted. In traditional methods, MGMF and BF have poor background suppression performance due to a large amount of residual background clutters. For RIPT and PSTNN, although low-rank tensor recovery methods have high BSF and SCRG, they often suffer from target loss for weak targets (as shown in Fig. 11 and Table I), which is not conducive to target detection. The above traditional methods also have obvious performance degradation with the increase of target scale. Our method achieves a good balance between suppressing the background and retaining the target energy, BSF and SCRG are close to the level of low-rank tensor recovery methods, and the target detection rate is significantly higher than other traditional methods. In terms of deep-learning methods, ALCNet and MDvsFA-cGAN, which directly detect targets, have significant performance degradations for extremely small targets, while MPRNet and our method based on background suppression perform well. For 9×9 scale targets, because our method is based on the fusion of traditional filtering results, it is affected by the performance degradation of corresponding traditional methods, and the evaluation metrics are lower than some other deep-learning methods.

In terms of target detection performance, as shown in Fig. 12(a) and (b). Under the condition of the FPR of 1×10^{-5}, the performance of our method does not decrease significantly with the increase of the target scale, which shows the ability to adapt to different scale targets. Our method can also detect targets with SCR lower than 6, which shows the ability to detect infrared weak targets. According to the ROC curve [as shown in Fig. 12(c)], our method shows good detection performance, which is slightly better than the results of MPRNet by about 3%–6%, better than the results of ALCNet and MDvsFA-cGAN by about 10%–20%, and significantly better than the results of other traditional methods by about 20%–30%.

Overall, on the infrared remote sensing image dataset, our multiscale method based on the combination of traditional methods and deep learning is significantly better than other traditional methods. Although our method is limited by the need to use the traditional filtering results for background modeling, it still achieves the same performance as the pure deep-learning method of MPRNet, and even outperforms MPRNet in some aspects, showing the performance potential. For ALCNet and MDvsFA-cGAN, they are difficult to directly extract target features in the case of extremely small targets, and their performances are degraded. Through experimental verification, our method can effectively suppress the background clutters while enhancing the target SCR and improving the target detection performance, which reflects the adaptability to targets of different scales and intensities in complex scenes.

D. Generalization Ability Experiments

Because the infrared satellite remote sensing imaging results will vary greatly with factors such as orbit height, attitude angle, spectrum selection, and imaging area, the algorithm needs to have certain adaptability to a variety of imaging results. For training-based deep-learning methods, because the amount of training dataset is limited and the type is often different from the actual detection image, the generalization performance of the algorithm on different datasets is also an important evaluation metric. In order to evaluate the robustness and generalization ability of the algorithm on heterogeneous data, we select the SIRST public dataset for experiments. The background suppression results of different methods are shown in Fig. 13, which are obtained from the residuals between the original image and the background modeling results. The red boxes are the areas where the real targets are located. In the table, bold font represents the maximum value, and underlined font represents the second largest value.

Because the resolution, scene, and noise type of the SIRST dataset are quite different from the training dataset, for the pure deep-learning method of MPRNet, the use of heterogeneous data has a great impact on its performance. It is difficult to distinguish the target from the background, which leads to the disappearance of the target in the background suppression results. As shown in Fig. 14, in some cases, MPRNet may even produce a large number of unnatural background residues. Although the performance of our method has declined, it still extracts the target effectively.

The experimental results of the target background suppression performance on the SIRST dataset are shown in Table II. For RIPT and PSTNN, although the background suppression indicators are relatively high, there is still a significant problem

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

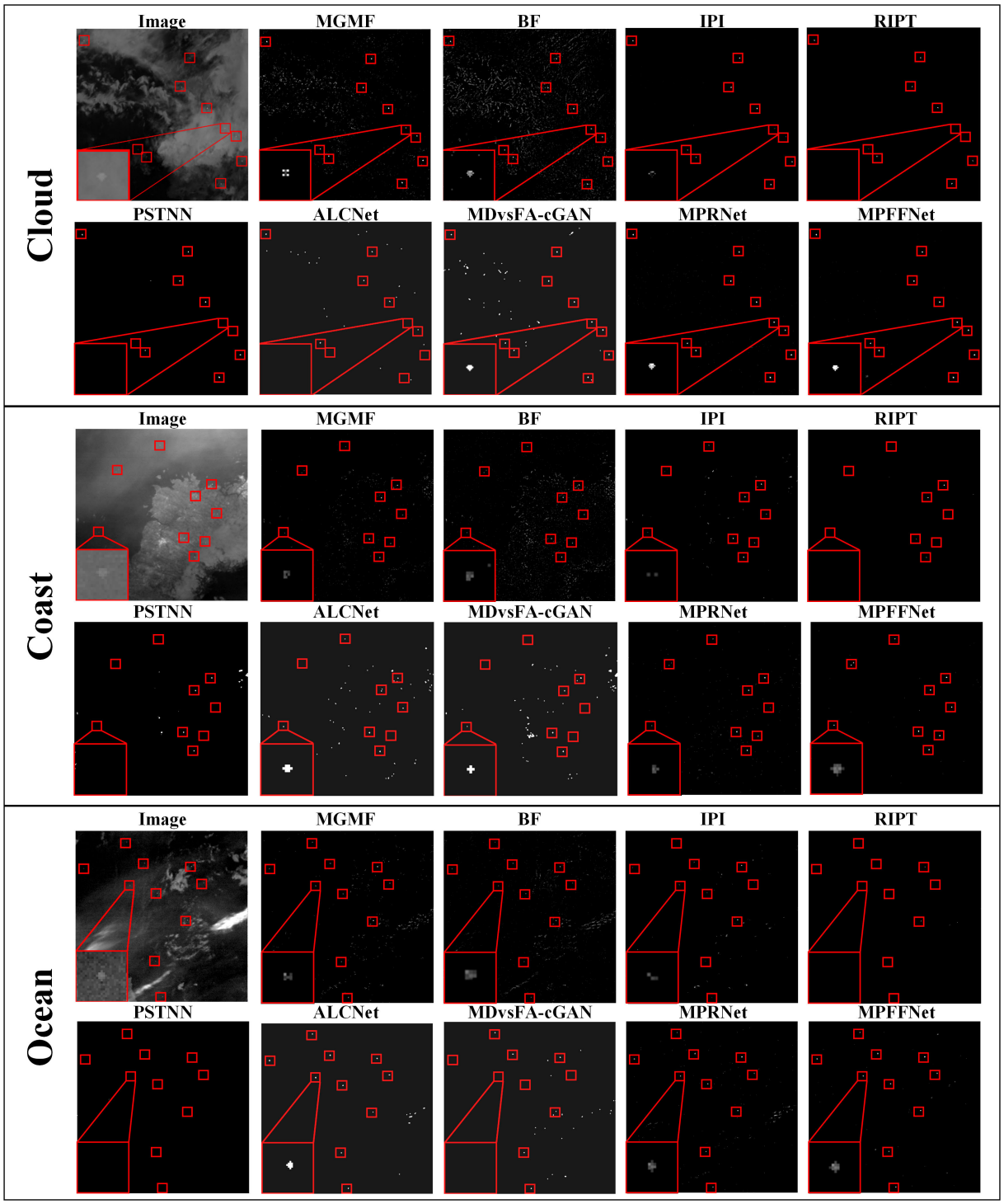


Fig. 11. Background suppression results on the infrared remote sensing image dataset (ALCNet and MDvsFA-cGAN detection results are binary masks). Our method achieves a good balance between suppressing the background and retaining the target energy.

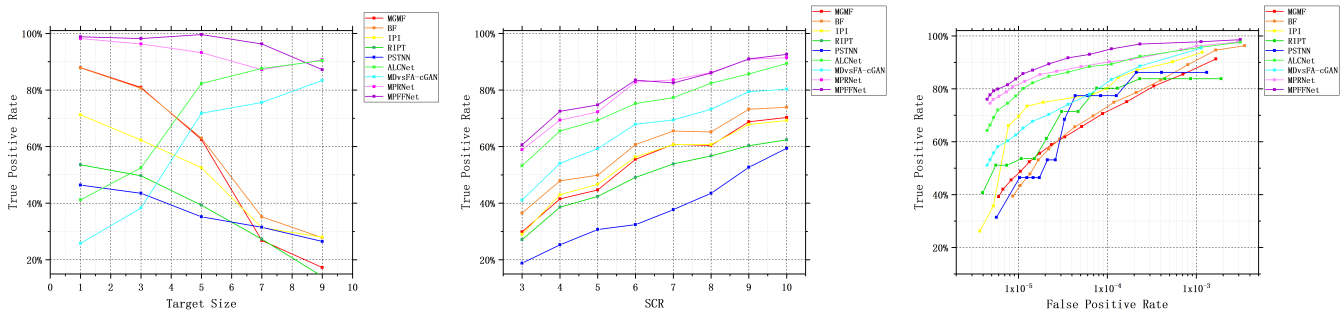
of target loss. According to the above statistical results, our method still has good performance on the SIRST dataset. Although the BSF and SCRG values decreased due to the use of heterogeneous datasets, the target detection rate is still significantly higher than other methods. In terms of target detection, since the average SCR of the target in the SIRST dataset is about 20, its brightness and contrast are higher than the infrared remote sensing dataset. Because there are fewer complex clutters in the scenes, the traditional methods can also easily separate the targets from the background and have a good performance on this dataset. For pure deep-learning methods of ALCNet, MDvsFA-cGAN, and MPRNet, the wide variations between the training dataset and test dataset result in significant performance drops, while our method maintains a performance level comparable to the traditional methods on the heterogeneous dataset.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

TABLE I

EVALUATION METRICS FOR DIFFERENT SCALES OF TARGETS ON THE INFRARED REMOTE SENSING IMAGE DATASET

Method	1×1			3×3			5×5			7×7			9×9		
	BSF	SCRG	TPR	BSF	SCRG	TPR	BSF	SCRG	TPR	BSF	SCRG	TPR	BSF	SCRG	TPR
MGMF	37.40	18.40	87.9%	39.19	23.75	80.9%	36.70	19.36	62.5%	40.86	8.71	26.8%	37.49	6.21	17.3%
BF	23.07	8.57	87.8%	23.43	7.67	80.6%	22.53	5.02	63.0%	42.84	3.69	35.2%	22.69	3.34	27.8%
IPI	149.34	36.36	71.2%	220.07	39.51	62.3%	156.69	24.49	52.5%	158.97	9.41	31.4%	146.50	8.82	28.0%
RIPT	**388.64**	**177.91**	53.6%	**325.63**	**162.11**	49.7%	207.14	**141.95**	39.3%	199.94	89.49	27.3%	170.22	46.06	14.1%
PSTNN	72.98	92.69	46.4%	170.02	81.71	43.5%	**310.97**	96.45	35.2%	293.27	80.19	31.5%	221.08	53.21	26.5%
ALCNet	–	–	41.2%	–	–	52.5%	–	–	82.3%	–	–	87.6%	–	–	90.4%
MDvsFA-cGAN	–	–	25.8%	–	–	38.3%	–	–	71.8%	–	–	75.6%	–	–	83.4%
MPRNet	141.17	61.66	98.2%	142.03	55.98	96.3%	137.64	62.07	93.2%	151.85	55.55	87.2%	133.89	60.12	**90.7%**
MPFFNet(Ours)	265.78	102.58	**98.8%**	287.05	99.77	**98.2%**	287.38	113.32	**99.6%**	**309.62**	**97.19**	**96.3%**	**280.77**	**90.36**	87.2%



(a) (b) (c)

Fig. 12. Target detection results on the infrared remote sensing image dataset. (a) Impact of target scale. (b) Impact of target SCR. (c) ROC comparison on the infrared remote sensing image dataset.

TABLE II

EVALUATION METRICS ON THE SIRST DATASET

Method	MGMF	BF	IPI	RIPT	PSTNN	ALCNet	MDvsFA-cGAN	MPRNet	MPFFNet
BSF	231.97	174.47	**393.45**	217.69	276.03	–	–	263.83	220.38
SCRG	13.74	12.83	37.99	131.19	**134.56**	–	–	5.88	50.49
TPR	78.2%	88.5%	92.5%	82.2%	78.5%	33.6%	16.7%	14.2%	**97.5%**

According to the statistics and ROC curve (as shown in Fig. 15), the TPR of our method is about 97% with the FPR of about 1×10^{-5}, it has strong robustness. On the other hand, the pure deep-learning methods show serious performance degradations on this dataset, and the TPR is only about 14%–34% with the FPR of about 1×10^{-5}, which is seriously affected by heterogeneous data. It should be emphasized that the serious performance degradation of such pure deep-learning methods is only due to the use of heterogeneous data for testing. After training with samples from the SIRST dataset, their performance rapidly improves. For large-scale targets, the performance of ALCNet is even better than our method.

Through comparison, it is proved that our method based on the combination of traditional methods and deep learning has better adaptability, stronger generalization ability, and robustness on heterogeneous datasets. It can meet the target detection requirements under different types of imaging conditions.

E. Consumption Discussion

In order to evaluate the practicability of each method, we compared our method to several competitive methods in terms of inference time. As shown in Table III. Among them, methods based on low-rank recovery are relatively slow and may be unsuitable for actual applications. MPRNet and our method have high computational efficiency, and their inference time is shorter than most background suppression methods. Due to the use of the multiscale progressive structure in our method, its efficiency is inferior to ALCNet and MDvsFA-cGAN, but the performance advantages of our method cannot be ignored. At present, the frame rate of remote sensing imaging is usually around 1–5 Hz. Compared to other methods, our method can achieve a higher detection probability while meeting real-time requirements.

TABLE III

AVERAGE INFERENCE TIME (IN SECONDS) OF DIFFERENT METHODS ON TWO DATASETS

Method	Infrared remote sensing dataset		SIRST dataset	
	TPR	Time	TPR	Time
MGMF	62.5%	0.56	78.2%	0.16
BF	63.0%	1.86	88.5%	0.52
IPI	52.5%	322.56	92.5%	48.53
RIPT	39.3%	2.96	82.2%	0.48
PSTNN	35.2%	0.88	78.5%	0.12
ALCNet	82.3%	0.02	33.6%	0.007
MDvsFA-cGAN	71.8%	**0.01**	16.7%	**0.005**
MPRNet	93.2%	0.21	14.2%	0.04
MPFFNet	**99.6%**	0.25	**97.5%**	0.05

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

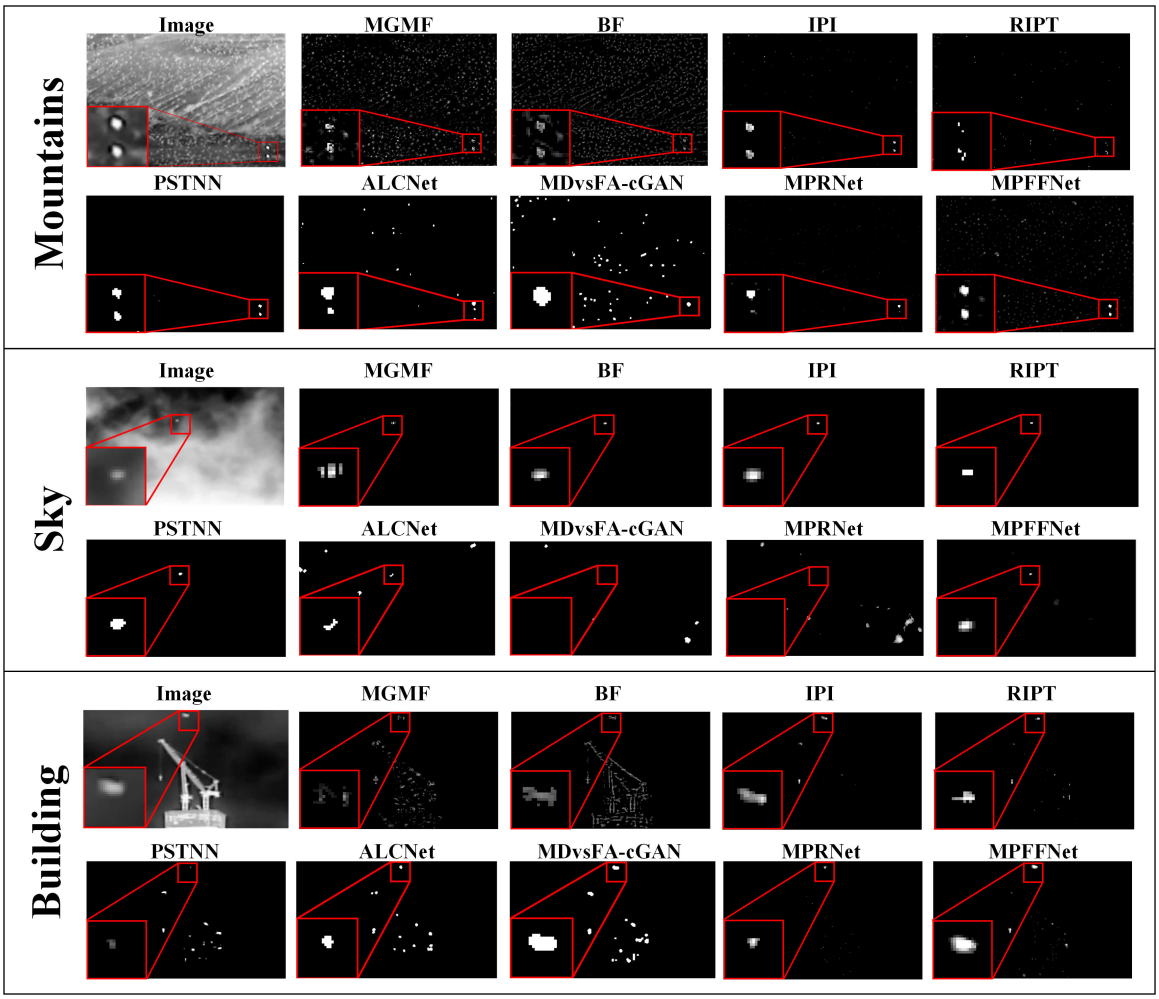


Fig. 13. Background suppression results on the SIRST dataset (ALCNet and MDvsFA-cGAN detection results are binary masks). Compared with pure deep-learning methods, our method based on the combination of traditional methods and deep learning has a stronger generalization ability on heterogeneous data.

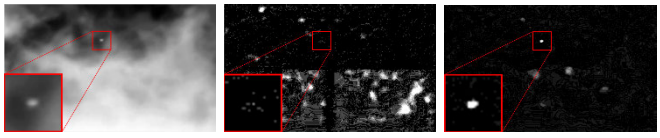


Fig. 14. Pure deep-learning methods may produce unnatural background residues. (a) Original image of SIRST dataset. (b) Result of MPRNet. (c) Result of MPFFNet (Ours).

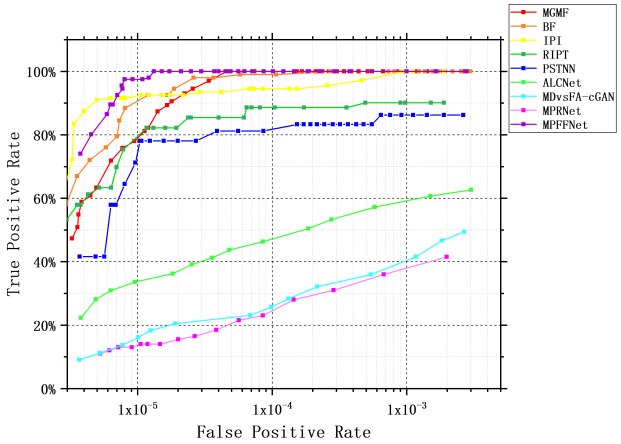


Fig. 15. ROC comparison on the SIRST dataset.

F. Ablation Experiments

In order to evaluate the effectiveness of the algorithm modules proposed in this article, the proposed modules are configured on the basic architecture. Their contribution to the final results is analyzed by comparing the background suppression performance of these modules before and after use. The experimental results are as follows.

As shown in Table IV, both MAS and DWC have a certain gain on the background suppression performance of the algorithm, and the influence of MAS is more obvious. Compared with the basic architecture, for the network configured with two modules at the same time, under the condition of the FPR of 1×10^{-5}, BSF, SCRG, and TPR have gains of 12.71, 4.75, and 1.1%, respectively, on the infrared remote sensing image dataset, and 27.82, 2.24, and 2.6%, respectively, on the SIRST dataset. Because the background noise of the

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
ABLATION EXPERIMENTAL RESULT ON THE INFRARED REMOTE SENSING IMAGE DATASET AND SIRST DATASET

Module combination	Infrared remote sensing dataset			SIRST dataset		
	BSF	SCRG	TPR	BSF	SCRG	TPR
MPFFNet(baseline)	249.76	98.36	98.3%	178.92	45.58	94.3%
MPFFNet+MAS	259.34	102.04	99.0%	195.51	47.44	96.1%
MPFFNet+DWC	253.12	100.03	98.7%	203.46	46.69	95.7%
MPFFNet+MAS+DWC	**262.47**	**103.11**	**99.4%**	**206.74**	**47.82**	**96.9%**

SIRST dataset is more significant, the algorithm modules on this dataset show a more obvious clutter suppression effect.

The experiments show that the algorithm after adding the modules can better suppress the background clutter false alarms and effectively improve the target contrast and detection performance, which proves the effectiveness of MAS and DWC.

V. CONCLUSION

In this work, we propose a multiscale infrared small target detection method based on the combination of traditional methods and deep learning. First, by constructing a multistage feature extraction network based on the feature pyramid nested structure, we progressively realize feature extraction and information fusion. It effectively improves the detection ability of targets with different scales. Second, a MAS mechanism is introduced to solve the problem of information loss in the cross-stage transfer of different scale features. By guiding the fusion of various typical traditional filtering results through the network, it gets rid of the performance limitation of hand-crafted features in traditional methods. Besides, it introduces additional robust image features to effectively improve the ability to suppress complex backgrounds and the generalization ability of the network. Finally, a DWC module is proposed, which realizes the parallel implementation of dynamic convolution through image preprocessing. It enhances the representation of background spatial correlation, thereby effectively improving the performance of background suppression and weak target extraction. As proved by experiments, our multiscale method has achieved a good balance in background suppression and target extraction. On heterogeneous datasets, compared with the pure deep-learning methods with serious performance degradation, our method shows better generalization ability and robustness. At the same time, the interpretability, controllability, and flexibility of this method are also helpful for applications in aerospace, aviation, and other fields.

However, the proposed method is still a traditional spatial filtering method in essence. Because the background modeling results of our method depend on traditional filtering results, it is also limited by traditional filtering performance. Due to the difficulty in obtaining infrared remote sensing imaging data and the significant differences in satellite imaging at different attitudes and orbital heights, our method pays more attention to the improvement of the generalization capacity on heterogeneous data. When using heterogeneous data, our method has a significant performance advantage compared to pure deep-learning methods. However, when trained and tested on the same type of dataset, the performance advantage of our method is no longer significant. Meanwhile, our method is more inclined to detect extremely small targets due to structural characteristics, so as the target scale increases, our method performance will gradually weaken compared to pure deep-learning methods such as ALCNet.

In future work, we will consider further research on the fusion weight of the network output to simplify and abstract it to construct artificial features with similar performance. This work can also help to guide the construction of new traditional filtering methods.

REFERENCES

[1] R. Hong, C. Xiang, H. Liu, A. Glowacz, and W. Pan, "Visualizing the knowledge structure and research evolution of infrared detection technology studies," *Information*, vol. 10, no. 7, p. 227, Jul. 2019.

[2] S. Doshvarpassand, C. Wu, and X. Wang, "An overview of corrosion defect characterization using active infrared thermography," *Infr. Phys. Technol.*, vol. 96, pp. 366–389, Jan. 2019.

[3] P. Barmpoutis, P. Papaioannou, K. Dimitropoulos, and N. Grammalidis, "A review on early forest fire detection systems using optical remote sensing," *Sensors*, vol. 20, no. 22, p. 6442, Nov. 2020.

[4] E. Neinavaz, M. Schlerf, R. Darvishzadeh, M. Gerhards, and A. K. Skidmore, "Thermal infrared remote sensing of vegetation: Current status and perspectives," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, Oct. 2021, Art. no. 102415.

[5] S. Jiang, J. Hu, X. Zhi, W. Zhang, D. Wang, and X. Sun, "Local adaptive prior-based image restoration method for space diffraction imaging systems," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601610.

[6] S. Jiang, X. Zhi, W. Zhang, D. Wang, J. Hu, and C. Tian, "Global information transmission model-based multiobjective image inversion restoration method for space diffractive membrane imaging systems," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607612.

[7] W. Zhang, M. Cong, and L. Wang, "Algorithms for optical weak small targets detection and tracking: Review," in *Proc. Int. Conf. Neural Netw. Signal Process.*, vol. 1, 2003, pp. 643–647.

[8] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 87–119, Jun. 2022.

[9] 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.

[10] 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.

[11] J. Liu, H. Wang, L. Lei, and J. He, "Infrared small target detection utilizing halo structure prior-based local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[12] J. Rivest, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, pp. 1886–1893, Jul. 1996.

[13] 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.

[14] 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.

[15] M. Zhao, L. Li, W. Li, R. Tao, L. Li, and W. Zhang, "Infrared small-target detection based on multiple morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6077–6091, Jul. 2021.

[16] 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.

[17] 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.

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

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.

[19] 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.

[20] 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.

[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. (ICCV)*, Oct. 2019, pp. 8508–8517.

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

[23] 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.

[24] J. Han, S. Liu, G. Qin, Q. Zhao, H. Zhang, and N. Li, "A local contrast method combined with adaptive background estimation for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1442–1446, Sep. 2019.

[25] S. Moradi, P. Moallem, and M. F. Sabahi, "Fast and robust small infrared target detection using absolute directional mean difference algorithm," *Signal Process.*, vol. 177, Dec. 2020, Art. no. 107727.

[26] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, Oct. 2016.

[27] J. Gong, Q. Hou, W. Zhang, and X. Zhi, "Non-local and nonlinear background suppression method controlled by multi-scale clutter metric," *Infr. Phys. Technol.*, vol. 71, pp. 18–27, Jul. 2015.

[28] M. M. Hadhoud and D. W. Thomas, "The two-dimensional adaptive LMS (TDLMS) algorithm," *IEEE Trans. Circuits Syst.*, vol. 35, no. 5, pp. 485–494, May 1988.

[29] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, 1998, pp. 839–846.

[30] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[32] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14816–14826.

[33] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3932–3941.

[34] Y. Zheng, X. Yu, M. Liu, and S. Zhang, "Residual multiscale based single image deraining," in *Proc. BMVC*, 2019, p. 147.

[35] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5971–5979.

[36] H. Zhao et al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[38] 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.

[39] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[40] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607315.

[41] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[42] 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.

[43] 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.

Pengfei Zhang received the M.E. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2022, where he is currently pursuing the Ph.D. degree.

His research interests include remote sensing image processing, weak target detection, and remote sensing image restoration.

Zhile Wang received the Ph.D. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2004.

He is currently a Full Professor at HIT. His research interests include optical image processing, optical system simulation testing technology, and optical synthetic aperture imaging theory.

Guangzhen Bao received the M.E. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2022, where he is currently pursuing the Ph.D. degree.

His research interests include remote sensing image acquisition and processing and optical target detection and identification.

Jianming Hu (Student Member, IEEE) received the M.Eng. and Ph.D. degrees in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017 and 2022, respectively.

He was a Visiting Scholar at the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, in 2019. Currently, he is an Assistant Professor at HIT. His research interests include remote sensing image processing, image characteristic analysis, and sea-aero target detection and identification.

Tianjun Shi (Graduate Student Member, IEEE) received the M.E. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2022, where he is currently pursuing the Ph.D. degree.

His research interests include remote sensing image processing and salient target detection and identification.

Guanjie Sun received the master's degree in communication engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2016.

His research interests include optical image processing and digital signal processing.

Jinnan Gong received the Ph.D. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2016.

He is currently an Assistant Professor at HIT. His research interests include optical image processing, intelligent information processing technology, and weak target detection.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 18,2024 at 02:16:18 UTC from IEEE Xplore. Restrictions apply.