Multibranch Mutual-guiding Learning for Infrared Small Target Detection

Qiang Li, *Member, IEEE,* Wei Zhang, Wanxuan Lu, and Qi Wang, *Senior Member, IEEE*

***Abstract*—At present, many infrared target detection approaches focus on designing modules that address the two key characteristics of targets: their weak signals and small size. However, these approaches often fail to fully leverage guided learning for weak and small target content, resulting in suboptimal detection performance, particularly in terms of shape preservation and target positioning. To tackle this challenge, this paper proposes a multi-branch mutual-guiding learning network (MMLNet) that enhances the accuracy of infrared target detection, even in the absence of clear morphological and textural features in images. The method consists of three branches: edge, positioning, and detection, each of which is designed with a specialized module from a unique perspective. In the detection branch, we introduce a multi-dimensional lossless encoder optimized through a downsampling strategy and multi-level feature fusion to mitigate feature loss in small targets. In the positioning branch, a target positioning strategy is proposed to explicitly identify candidate targets from the image by means of a learnable multi-kernel pattern. In the edge branch, a simple architecture is adopted to enhance the ability of the model to preserve the target shape. To effectively utilize the knowledge of different branches, a mutual-guiding fusion module is developed to adjust information within and between branches. The manner adaptively utilizes the specific knowledge from each input branch. Experiment results demonstrate that the proposed method achieves comparable performance, and the visualization results show the advantages of our method in shape preservation and positioning of the targets. Our code is publicly available at https://github.com/qianngli/MMLNet.**

***Index Terms*—Infrared image, small target detection, shape preservation, target positioning, mutual-guiding fusion.**

I. Introduction

INFRARED target detection is a crucial task in image processing, and plays an essential role in various fields, including military, security, autonomous driving, etc. Unlike visible imaging, infrared imaging technology effectively captures target information in challenging environmental conditions, such as low light. This capability makes infrared target detection particularly valuable for night reconnaissance and concealed scenarios. Consequently, research in infrared target detection holds significant value for applications.

In practice, infrared target detection continues to encounter numerous challenges, particularly in detecting small targets. On the one hand, the quality of infrared images is heavily influenced by environmental factors such as lighting conditions. These factors can diminish the contrast between the target and background. Moreover, due to their small size, targets are often more susceptible to background interference and noise. It complicates the detection process, negatively impacting detection accuracy. On the other hand, the characteristics of small targets in infrared image may resemble those of other targets within complex backgrounds, which increases the likelihood of both false and missed detections. Therefore, how to achieve efficient and accurate small target detection under such complex conditions has become a critical focus in current research.

Benefiting from powerful representation of deep learning [1], [2], various solutions have been proposed to address these challenges. Many studies focus on modifying model architectures to enhance their detection ability for small targets, specifically for issues related to contrast enhancement and feature fusion. As for contrast enhancement, researchers emphasize the importance of the clear distinction between the target and background. For instance, Jiang et al. [3] propose a flexible window local contrast measure that dynamically adjusts the shape of the saliency measurement window through flexible region growing. Similarly, Chuang et al. [4] develop a multi-scale local contrast learning module. It extracts and fuses local contrast information from feature maps at different scales. Dai et al. [5] modularize a conventional local contrast measure into a nonlinear feature refinement layer. Furthermore, Chen et al. [6] obtain the local contrast map of the input image using dissimilarity between the current location and its neighboring areas. This method effectively enhances target signals while simultaneously suppressing background clutter. Different from above methods, Zhang et al. [7] introduce a double dilate contrast measure that increases the distinction between the target and background. These contrast modules are effective in scenes where the background is distinctly different from the target. However, when the difference between the infrared target and the background is subtle, particularly in situations where the surrounding background exhibits minimal gradient change. These strategies have shown to be ineffective and fail to produce meaningful results. Therefore, relying solely on contrast enhancement may not be the most effective approach in such scenarios.

This work was supported in part by the Key Laboratory of Target Cognition and Application Technology under Grant 2023-CXPT-LC-005, and in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-YBXM-130. (*Corresponding author: Qi Wang*)

Qiang Li and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liqmges@gmail.com, crabwq@gmail.com)

Wei Zhang is with the School of Computer Science, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zhangwei707@mail.nwpu.edu.cn)

Wanxuan Lu are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: luwx@aircas.ac.cn)

In contrast to the aforementioned strategies, some researchers have developed models that focus on feature fusion. For instance, Chen et al. [8] propose an attention-enhanced feature fusion network that integrates features from different layers. Meanwhile, it explores the relationships between low-level and high-level features. Guo et al. [9] introduce a dual-encoder model with distinct inputs to capture more effective small target feature information by concatenating multiscale features from various layers of the decoder. Shi et al. [10] present a detection module that captures both low-level structural and textural features as well as high-level semantic features. Additionally, Li et al. [11] design a moderately dense adaptive feature fusion module that interconnects all internal features. Dai et al. [12] also propose an asymmetric contextual modulation module. The network embeds low-level contexts with fine details into high-level features, and enhance the characteristics of infrared small targets. Similar studies are [13], [14], etc. Feature fusion is indeed an effective method for extracting semantic features of infrared targets at various levels. However, these approaches overlook a critical aspect, i.e., the targets themselves are small. In the coding process, we not only focus on the feature learning of the model, but also develop strategies to prevent the loss of target features during feature extraction. Clearly, existing research lacks a more exploration of this aspect.

With respect to the above descriptions, it can is clear that there are still many challenges in infrared small target detection task. We believe that infrared small target detection involves three elements: positioning, feature retention, and edge preservation. First, accurate positioning of the target in the image is essential, especially for extremely small targets. Second, it is crucial to preserve the target features during the extraction process. Finally, the edges of the target must be well maintained after detection. Therefore, how to preserve the target shape well while reducing the feature loss is the key to accurate target detection. To achieve this end, this paper proposes a multi-branch mutual-guiding learning network (MMLNet) for infrared small target detection. The method consists of three branches: edge, positioning, and detection, where each branch builds specific modules from a unique perspective to explore the knowledge associated with them. Additionally, a mutual-guiding fusion (MGLF) module is introduced to adaptively adjust the input information across the three branches, so as to achieve infrared target detection more accurately. In summary, the contributions of the proposed approach can be summarized as follows:

- We propose a multi-dimension lossless encoder to alleviate the feature loss of targets by using downsampling strategy and multi-level feature fusion. In the downsampling process, we convert the spatial features with two frequency component in Haar wavelet into channel dimension, and realize the lossless transmission of information. In the multi-level feature fusion process, the adjacent-scale and cross-scale features are cascaded through the local and global perspectives. The manner enhances the representation of high-resolution features with low-resolution context.
- We propose a candidate target positioning branch to assist the model to locate the targets more accurately. Using prior knowledge that infrared targets typically appear obvious high-brightness textures, we introduce a learnable multi-kernel module that actively analyzes targets with different sizes to identify candidate targets. This module can highlight regions that are brighter than those surrounding the contour of the original image, effectively suppressing background noise.
- To make full use of the knowledge within and between branches, a MGLF module is developed to adaptively adjust the information of edge, detection, and location accordingly. It explicitly utilizes the specific knowledge from each input so as to improve the flexibility of fusion.

The rest of this paper is organized as follows: Section II introduces related infrared small target detection methods. Section III describes the proposed method in detail. Section IV analyses and discusses the experimental results. Finally, the conclusion is given in Section V.

II. Related Work

Infrared small target detection has an extensive history, and we briefly review the main works related to feature loss and shape preservation of small targets.

A. Feature Loss of Small Target

Feature loss presents a significant challenge in infrared small target detection, primarily due to low contrast, high noise levels, and the inherently small size of the targets. Now many researchers have focused on alleviating feature loss through feature enhancement [15]–[17]. For instance, Liu et al. [18] propose a deep denoiser regularization with low-rank prior to remove noise, thereby improving the visibility of small targets in infrared images. While the approach effectively reduces noise, it may mistakenly classify targets as noise when the contrast between the target and the background is minimal. To improve the contrast, many methods are proposed. Shao et al. [13] adopt a three-layer local contrast structure, and the difference-ration form is introduced to enhance the target and suppress the complex backgrounds. Wang et al. [17] develop the local area enhancement attention to perform fine-grained local contrast calculations. The above methods work well for situations where the contrast is slightly higher. However, when the contrast is not obvious, these methods struggle to perform effectively. Other researchers have instead focused on feature extraction [19]–[21]. For example, Chen et al. [22] design a context-aware pyramid network to preserve the loss of detailed target information, and dynamically capture multiscale information. Li et al. [23] propose a dense nested attention network. Xu et al. [24] introduce a feature extraction module for representing multiscale and multilevel features to adaptively extract features at various levels. Similarly, Chen et al. [25] design a local patch network with global attention. Despite these advancements, feature extraction for very small targets inevitably leads to feature loss. In our view, the aforementioned methods still lack a comprehensive analysis from multiple perspectives, even with contrast enhancement or multi-scale fusion strategy. This gap leads to less effective module designs during feature extraction, hindering the overall detection performance.

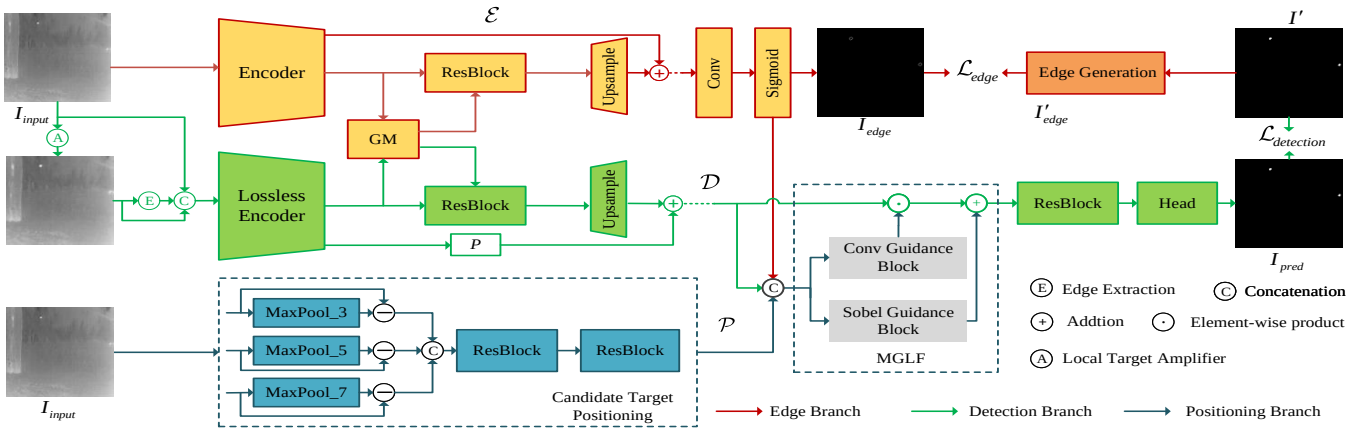


Fig. 1. Overview of the proposed multi-branch mutual-guiding learning network (MMLNet) for infrared small target detection. As for edge branch, local target amplifier, etc, they are described in detail in study [29].

B. Shape Preservation of Small Target

Shape preservation poses a notable difficulty in infrared small target detection, as current methods often struggle to preserve the complete shape of small targets. For instance, Lin et al. [26] propose a shape-reconstructable decoder that processes the edge map of the input infrared image, then optimize two shape-related consistencies simultaneously. Similarly, Lin et al. [27] develop a unified framework with shape-guided decoder. It hierarchically fuses decoder representations and edge information through cascaded ResNet blocks to reconstruct contours. Additionally, Zhang et al. [28] introduce an infrared shape network that employs a Taylor finite difference edge block, where Li et al. [29] also adopt this manner to design the model. In contrast, Ma et al. [30] design a target edge enhancement loss to sharpen the edges of infrared targets. Wang et al. [31] leverages multi-scale fusion to improve the gray response of targets and enhance contrast with the background, so as to preserve scale information. Later, Zhao et al. [32] incorporate high-frequency directional information from the original image into the network. It guides the model to notice detailed aspects such as target edges and shapes. Currently, these methods indeed preserve the shape for some targets with large size. When it comes to very small targets, these are simply impossible to deal with. Therefore, the key lies in accurately locating the targets within the image. By combining these strategies, this not only becomes better at focusing on the targets, but also enhances its ability to preserve their contours.

III. The Proposed Method

A. Motivation and Overview

With respect to infrared small target detection, we face challenges such as high noise, small target size, etc. These factors greatly complicate the detection compared to traditional target task. Despite numerous research efforts have been proposed to tackle these issues, many approaches struggle to effectively detect very small targets. A primary reason for this shortcoming is the insufficient attention given to three crucial aspects, i.e., positioning, feature retention, and edge preservation. Specifically, Accurate positioning is vital for enabling the model to recognize targets by directing attention to relevant areas. Moreover, to facilitate effective feature extraction, it is essential to maintain the inherent characteristics of these small targets throughout the process. Importantly, combining specific modules that impose the necessary constraints better preserves the target shape. In summary, when designing a model for infrared small target detection, it is vital to thoroughly consider positioning, feature retention, and edge preservation, which can achieve more accurate detection results.

To achieve this end, we follow the existing study [29] that combines detection and edge branch, and propose a multi-branch mutual-guiding learning network (MMLNet) for infrared small target detection. The overall flowchart is shown in Fig. 1. The network comprises three branches: edge, positioning, and detection, which solve corresponding problem from different perspectives. As for detection branch, we propose a multi-dimensional lossless encoder to alleviate feature loss of targets via downsampling strategy and multi-level feature fusion. The extracted features are then analyzed by a typical encoder to obtain high-level information. With respect to edge branch, we incorporate a simple encoder-decoder structure that allows the model to transmit target edge by interaction. Here, the process is performed using global average pooling (GAP) and multilayer perceptron (MLP), which is defined as GM. To accurately locate targets, a candidate target positioning branch is developed. It introduces a learnable multi-kernel module that analyze targets with different sizes to obtain candidate targets. Subsequently, feature extraction is executed on the aggregated image by stacking several convolution blocks, i.e., ResBlock. Based on the feature from the three branches, a MGLF module is developed to fully exploit the knowledge both within and between branches. It adaptively adjusts edge, detection, and position information. Finally, through the combination of these strategies and branches, infrared small targets are effectively detected.

B. Multi-Dimension Lossless Encoder

Small targets occupy fewer pixels in an image, which can lead to their loss during feature extraction. Similar to the encoder proposed by [29], we design a multi-dimensional

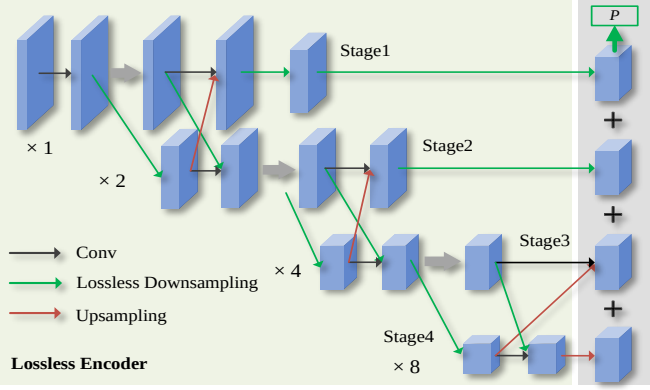


Fig. 2. Architecture of the multi-dimensional lossless encoder.

lossless encoder to alleviate this issue in the coding process, as shown in Fig. 2. Unlike this architecture, we add downsampling strategy and cascades adjacent-scale and cross-scale features within the encoder.

During the coding process, feature sizes typically need to be reduced through pooling operations to decrease complexity and memory consumption. While this process extracts the most significant features and suppresses noise, it can also cause small targets to nearly disappear from deep feature extraction. To address this trouble, we follow the study [33], and first utilize the Haar wavelet transform to the middle features $X \in \mathbb{R}^{C\times W\times H}$, generating four components. Here, C, W, and H denote the number of channel, width, and height of the feature map, respectively. The components has a low-frequency component and high-frequency component that contains the horizontal, vertical, and diagonal directions, i.e.,

$$X_L, X_H, X_V, X_D = WT(X), \tag{1}$$

$$X' = [X_L; X_H; X_V; X_D], \tag{2}$$

where $WT(\cdot)$ denotes the Haar wavelet transform function. Each component has a size of $W/2 \times H/2$ by this transformation. Then four components are combined to form a new feature map X'. Its number of channels is increased fourfold, which means that the process encodes spatial information to channel dimension effectively. This operation realizes the lossless transmission in the process of downsampling, and avoids the loss of image texture and details. Similarly, whenever downsampling is involved in the detection branch, we implement this operation to ensure that features are not easily lost.

Multi-scale features effectively capture contents across various sizes and details. It can enhance the feature representation of model when addressing complex scenes. To leverage this capability, we design a multi-level feature fusion strategy that cascades adjacent-scale and cross-scale features within the encoder to achieve both local and global feature fusion. Inspired by the study [29], local features from two adjacent stages are integrated through downsampling and upsampling. The approach enables the model to exploit subtle changes between adjacent levels, similar to the architecture of HRNet [34]. It increases sensitivity to local details, allowing for more precise capture of target information. In contrast, the cross-scale integration aggregate features from different stages through global aspect. The pattern can synthesize information across different scales and improve the comprehension of a wide range of contexts, which is crucial for identifying similar targets and relationships within intricate scenarios. When combined with a downsampling strategy, this multi-level feature fusion approach effectively utilizes low-resolution context to boost the representation of high-resolution features. It makes the model more adaptable to diverse scenarios.

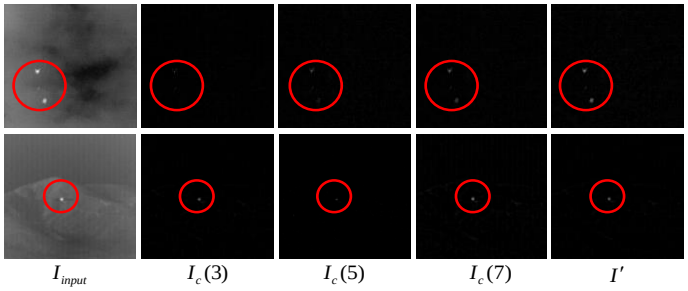


Fig. 3. Some intermediate results under different kernels.

C. Candidate Target Positioning

Several factors, such as similar features in the background, can hinder accurate target positioning, leading to missed and false detections. The key of this issue lies in the ability of the model to effectively locate the targets. To address this challenge, we design a simple yet efficient module for identifying candidate targets. Concretely, infrared targets typically appear obvious high-brightness textures in the image, and other textures in the image are generally smoother. To leverage this distinct prior knowledge, the max pooling with different kernel sizes is exploited to weaken unnecessary details. Meanwhile, it reduces the brightness of potential targets. Then, we utilize the original image I_{input} and subtract the processed image, obtaining the potential targets present in the original image, i.e.,

$$I_c(i) = I_{input} - M(I_{input}, i), \qquad i = 3, 5, 7, \tag{3}$$

where $M(\cdot, i)$ represents the max pooling operation with different kernel sizes. This operation acts as a learnable nonlinear filter that analyzes targets of various sizes based on the input image. It highlights regions that are brighter than the contour in original image, effectively suppressing background noise. On this basis, the filtered images I_c are then cascaded to form an aggregated image I', where some intermediate results are shown in Fig. 3. With background interference minimized, feature extraction is executed on the aggregated image by stacking several convolution blocks.

D. Mutual-Guiding Learning Fusion

As for the deep feature fusion from multiple branches, a straightforward approach is to concatenate the features generated by the three branches along the channel dimension. While this method allows for the calculation of correlations among the three observed features, it does not fully utilize the unique insights inherent in each branch. To make more effective use

of the intra-branch and inter-branch knowledge, we propose a MGLF module. This module aims to optimize the integration of features, ensuring that the specific information from each branch is leveraged to its fullest potential. The module specifically consists of three distinct inputs: the detection branch $\mathcal{D}$, the positioning branch $\mathcal{P}$, and the edge branch $\mathcal{E}$. As illustrated in Fig. 1, the module first utilizes the deep features from these three parallel branches. It then employs mutual guidance blocks to transform each feature based on the other two, thereby achieving more efficient feature fusion.

Let F_d, F_l, and F_e denote the three features generated by respective branches. Taking F_d as an example, the mutual guidance block linearly transforms F_d as

$$F'_d = F_d \odot P + Q, \tag{4}$$

where F'_d represents the transformed F_d, $\odot$ denotes the element-wise product, and P and Q represent the guidance parameters. This transformation adjusts each knowledge for F_d, enhancing the flexibility of feature fusion. Then, P and Q are computed using two independent guidance blocks, respectively. They are expressed as

$$P = \mathcal{F}_a([F_d; F_l; F_e]; \theta_a), \tag{5}$$

$$Q = \mathcal{F}_b([F_d; F_l; F_e]; \theta_b), \tag{6}$$

where $\mathcal{F}_a(\cdot;\theta_a)$ and $\mathcal{F}_b(\cdot;\theta_b)$ denote the guidance blocks for P and Q, respectively. In the two guidance, the Sobel operator is employed for convolutions in both horizontal and vertical directions for $\mathcal{F}_b$, while stacking M convolution blocks for $\mathcal{F}_a$. Similarly, we apply the same $\mathcal{F}_a$ and $\mathcal{F}_b$ to transform F_d, F_l, and F_e according to equation 4. For the guidance calculations of F_l and F_e, we concatenate $[F_l; F_d; F_e]$ and $[F_e; F_d; F_l]$, respectively. By leveraging the specific guidance from each branch, this module not only improves the flexibility of the feature transformations, but also facilitates a richer feature representation.

E. Loss Function

After constructing the model, the next step is to design the corresponding loss function. We follow work [28], and build two loss function terms to optimize the model parameters, including edge loss and detection loss. Specifically, we adopt an existing edge detector on the ground-truth I' to generate the corresponding edge image I'_{edge}. Then, Dice loss [35] and binary cross entropy (BCE) are introduced to formulate the edge loss term. They are formulated as

$$\mathcal{L}_{edge}^{Dice} = 1 - \frac{2|I_{edge} \cap I'_{edge}|}{|I_{edge}| + |I'_{edge}|}, \tag{7}$$

$$\mathcal{L}_{edge}^{BCE} = -[I'_{edge} \log(I_{edge}) + (1 - I'_{edge}) \log(1 - I_{edge})], \tag{8}$$

$$\mathcal{L}_{edge} = \mathcal{L}_{edge}^{Dice} + \lambda \mathcal{L}_{edge}^{BCE}, \tag{9}$$

where I_{edge} denotes the output generated by edge branch, λ is a balanced factor that we empirically set to $\lambda = 10$. For the detection loss term, we construct it based on the detection result I_{pred} and the ground-truth I', i.e.,

$$\mathcal{L}_{dection} = -[I' \log(I_{pred}) + (1 - I') \log(1 - I_{pred})]. \tag{10}$$

Finally, the total loss function for the design model is defined as

$$\mathcal{L} = \mathcal{L}_{dection} + \mathcal{L}_{edge}. \tag{11}$$

IV. Experiments

This section first describes the dataset, implementation details, etc. Then, the main module and branch of the model are discussed and analyzed from various perspectives. Finally, we compare the performance against existing methods using both quantitative and qualitative aspects.

A. Datasets

To analyze the proposed method, three datasets are utilized, including IRSTD-1k[1], SIRST Aug[2], and MDFA[3]. The IRSTD-1k dataset is collected from real-world conditions at long distances and consists of 1,001 images, each with a resolution of 512 × 512 pixels. These images are split into a training set of 800 images and a test set of 201 images. Due to the limited number of images, we add the training data by randomly selecting 5 patches from each image. The SIRST Aug dataset is created by augmenting the original SIRST dataset with 427 images, resulting in 8,525 images for training and 545 for testing. Each image in this dataset has a resolution of 256 × 256 pixels. In contrast, the MDFA dataset includes both real infrared images and synthetic ones, where training and test set contains 9,978 images and 100 images, respectively. Unlike the above datasets, each image in the training set has a fixed resolution of 128 × 128 pixels, while images in the test set has various resolutions. For all three datasets, we employ techniques such as noise addition, flipping, rotation, and other operations to augment the training images.

B. Evaluation Metrics

To quantitatively assess methods, we utilize five metrics. They are defined as

Intersection over Union (IoU):

$$IoU = \frac{TP}{T + P - TP}, \tag{12}$$

Normalized Intersection over Union (nIoU):

$$nIoU = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{TP(i)}{T(i) + P(i) - TP(i)} \right). \tag{13}$$

In two equations, N denotes the total number of images in the test set. $TP(i)$, $P(i)$, and $T(i)$ represent the number of true positive pixels, predicted positive pixels, and ground truth pixels, respectively.

Area Under Curve (AUC): AUC measures the area beneath the Receiver Operating Characteristic (ROC) curve, which is derived from two primary metrics, i.e.,

$$FPR = \frac{FP}{FP + TN}, \tag{14}$$

[1]https://github.com/RuiZhang97/ISNet
[2]https://github.com/Tianfang-Zhang/SIRST-Aug
[3]https://github.com/wanghuanphd/MDvsFA_cGAN

$$TPR = \frac{TP}{TP + FN}. \tag{15}$$

Probability of Detection (P_d): The P_d metric evaluates performance at the target level by calculating the proportion of correctly identified targets N_{pred} to the total number of targets N_{all}. It is denoted as

$$P_d = \frac{N_{pred}}{N_{all}}. \tag{16}$$

False-Alarm Rate (F_a): Similarly, F_a is another target-level metric that quantifies the ratio of incorrectly predicted target pixels N_{false} to the total number of pixels in the image N_{all}, i.e.,

$$F_a = \frac{N_{false}}{N_{all}}. \tag{17}$$

C. Implementation Details

As for the network setup, we employ the AdaGrad optimizer to train the designed network, with a weight decay set at 10^{-4}. The model starts with an initial learning rate of 0.02, which is halved after every 30 epochs. The batch size is set to 16 throughout the experiments. Moreover, to ensure that the input images contain the targets for detection, we specify the input sizes as 480×480 for the IRSTD-1k dataset, 256×256 for the SIRST Aug dataset, and 128×128 for the MDFA dataset. Note that bilinear interpolation is utilized to transform the size of the feature maps during the upsampling process. All experiments are conducted using the PyTorch framework on an NVIDIA GeForce GTX 3090 GPU.

D. Model Analysis

In this section, we analyze the proposed method from many aspects to verify its effectiveness, including multi-dimension lossless encoder, candidate target positioning, MGLF, and multi-branch.

1) Study of Multi-Dimension Lossless Encoder: As for infrared small target detection, traditional encoder often leads to the feature loss for small target. To alleviate this trouble, this paper proposes a new multi-dimension lossless encoder, which incorporates an advanced downsampling strategy and multi-level fusion. This section aims to evaluate the effectiveness of these strategies. Here, we add a traditional encoder with the typical pooling as a baseline and examine various combinations to observe changes in model performance, as shown in Table I. The table reports the model performance with different configurations. Concretely, the adjacent-scale and cross-scale features are cascaded through local and global perspectives after the introduction of multi-level fusion. It significantly improves the model performance. In addition, the lossless downsampling strategy utilizes Haar wavelet to process intermediate features, where it includes low frequency components and high frequency components. This manner effectively encodes spatial information to the channel dimension without losing information. The numerical results in the table prove it well in performance improvement. Overall, these combinations enable the model to learn richer target features, which has great significance for the overall optimization and learning of the model.

TABLE I
STUDY OF MULTI-DIMENSION LOSSLESS ENCODER.

Components	Metrics	IRSTD-1K	SIRST Aug	MDFA
Baseline	IoU	0.6653	0.7604	0.4834
	nIoU	0.6769	0.7308	0.4765
	AUC	0.8927	0.9449	0.8971
Multi-level Fusion	IoU	0.6681	0.7611	0.4870
	nIoU	0.6772	0.7231	0.4963
	AUC	0.8935	0.9405	0.8458
Lossless Downsampling	IoU	0.6693	0.7623	0.4915
	nIoU	0.6771	0.7238	0.4931
	AUC	0.8959	0.9455	0.8865
Proposed module	IoU	0.6721	0.7660	0.5133
	nIoU	0.6803	0.7316	0.5126
	AUC	0.9008	0.9468	0.8664

TABLE II
STUDY OF LEARNABLE MULTI-KERNEL MODULE.

Dataset	Metrics	Baseline	Single Kernel	Multi-Kernel
IRSTD-1K	IoU	0.6663	0.6667	0.6686
	nIoU	0.6754	0.6765	0.6789
	AUC	0.8979	0.8909	0.8881
SIRST Aug	IoU	0.7506	0.7580	0.7602
	nIoU	0.7246	0.7269	0.7295
	AUC	0.9393	0.9396	0.9433
MDFA	IoU	0.4748	0.4849	0.4925
	nIoU	0.4808	0.4857	0.4959
	AUC	0.8065	0.8584	0.8630

2) Study of Learnable Multi-kernel Module: This section aims to verify the effectiveness of the learnable multi-kernel module within candidate target positioning branch. We select EGPNet [29] as the baseline and examine the changes in performance by introducing both single kernel and multi-kernel pooling. The results are presented in Table II. Overall, the performance of the model improves after the addition of these strategies. In fact, the pooling process helps to weaken unnecessary details in a learnable manner. The processed image is subtracted from the original image to highlight potential targets. When the multi-kernel analysis is introduced, the model can enhance its flexibility. Combined with single kernel and multi-kernel pooling, the model can comprehensively understand the input data, leading to more accurate target positioning. The results of this table show that the module is effective in improving the performance of target detection.

3) Study of Mutual-Guiding Learning Fusion: To facilitate deep feature fusion from the three branches, a mutual guiding learning fusion (MGLF) module is designed. To evaluate the effectiveness of this fusion, we compare it with a traditional concatenation way. Table III presents the results of various combinations within the MGLF. It is evident that both traditional concatenation and our proposed fusion strategy yield improvements across most metrics. Here, the direct concatenation merges the features generated by the three branches along the channel dimension. It allows for the calculation of correlations among the observed features. However, it does not fully leverage the specific knowledge contained within each branch. In contrast, our proposed fusion strategy effectively utilizes

TABLE III
STUDY OF MUTUAL-GUIDING LEARNING FUSION.

Dataset	Metrics	w/o MGLF	Concat	MGLF
IRSTD-1K	IoU	0.6720	0.6704	0.6721
	nIoU	0.6801	0.6840	0.6803
	AUC	0.8921	0.8965	0.9008
SIRST Aug	IoU	0.7627	0.7638	0.7660
	nIoU	0.7348	0.7351	0.7316
	AUC	0.9452	0.9572	0.9468
MDFA	IoU	0.4974	0.5036	0.5133
	nIoU	0.4942	0.5023	0.5126
	AUC	0.8474	0.8751	0.8664

TABLE IV
STUDY OF MULTI-BRANCH WITH SETTING DIFFERENT COMBINATIONS.

Dataset	Metrics	Baseline	Edge	Positioning	All
IRSTD-1K	IoU	0.6582	0.6662	0.6661	0.6686
	nIoU	0.6634	0.6800	0.6725	0.6789
	AUC	0.8895	0.8962	0.8950	0.8881
SIRST Aug	IoU	0.7529	0.7613	0.7617	0.7602
	nIoU	0.7194	0.7259	0.7242	0.7295
	AUC	0.9284	0.9342	0.9402	0.9433
MDFA	IoU	0.4847	0.4925	0.4892	0.4925
	nIoU	0.4892	0.4996	0.4844	0.4959
	AUC	0.8550	0.8786	0.8311	0.8630

TABLE V
QUANTITATIVE EVALUATION ON IRSTD-1K DATASET. THE BOLD AND UNDERLINE INDICATE THE BEST AND SECOND PERFORMANCE.

Methods	IoU	nIoU	AUC	Fa	Pd
TopHat (GRSL'18)	0.1388	0.2825	0.6303	24.7	0.7542
ALCNet (TGRS'21)	0.6687	0.6665	0.8971	**8.5**	0.9226
ACM (WACV'21)	0.6339	0.6064	0.8932	<u>10.2</u>	0.9091
ISNet (CVPR'22)	0.6538	0.6372	0.8879	18.0	0.9226
AGPCNet (TAES'23)	0.5602	0.5344	0.8356	17.1	0.9150
DNANet (TIP'23)	<u>0.6714</u>	0.5942	**0.9216**	12.0	0.9252
EGPNet (TGRS'24)	0.6662	<u>0.6800</u>	0.8962	24.2	**0.9495**
MMLNet	**0.6721**	**0.6804**	<u>0.9009</u>	14.0	<u>0.9428</u>

TABLE VI
QUANTITATIVE EVALUATION ON MDFA DATASET.

Methods	IoU	nIoU	AUC	Fa	Pd
TopHat (GRSL'18)	0.2438	0.2927	0.7589	109.9	0.7426
ALCNet (TGRS'21)	0.3311	0.3512	0.8264	560.7	0.6429
ACM (WACV'21)	0.4312	0.4169	0.8082	197.9	0.5500
ISNet (CVPR'22)	0.4376	0.4248	0.8648	355.1	0.7286
AGPCNet (TAES'23)	0.4339	0.4152	0.8603	107.0	0.8417
DNANet (TIP'23)	0.4324	0.4075	<u>0.8676</u>	57.8	0.8201
EGPNet (TGRS'24)	<u>0.4925</u>	<u>0.4996</u>	**0.8786**	<u>37.7</u>	<u>0.8929</u>
MMLNet	**0.5141**	**0.5126**	0.8665	**16.1**	**0.9071**

both intra-branch and inter-branch knowledge to facilitate mutual guidance. Furthermore, the Sobel operator is employed to further extract target edges, which enhances the ability of the model to preserve target shapes. The combinations result in a particularly superior performance improvement on the MDFA dataset. Therefore, this study demonstrates that MGLF offers substantial advantages in enhancing feature fusion.

4) Study of Multi-Branch: To accurately detect small infrared targets, we add two branches to the baseline detection network, i.e., the positioning branch and the edge branch. These additions are specifically designed to improve focus on both the positioning and shape of the targets. This section analyzes the impact of these branches on model performance. As illustrated in Table IV, the introduction of these branches results in a significant improvement in performance. During the design process, we carefully consider the current challenges in infrared target detection, and all branches reflects these considerations. The combination of any branch with the detection framework yields better results in detecting targets within complex scenes. With the incorporation of these two branches, the model achieves superior results. Note that MGLF is not added this architecture. This further demonstrates the positive contribution of these branches to the detection task.

E. Performance Comparison with Existing Approaches

This section makes a comprehensive comparison between seven existing methods and proposed MMLNet in both quantitative and qualitative aspects. They are include TopHat [36], ALCNet [37], ACM [12], ISNet [28], AGPCNet [38], DNANet [23], and EGPNet [29].

1) Quantitative Evaluation: Tables V-VII present the results of the quantitative evaluation across three public datasets. Among these methods, the TopHat is traditional technology and relies heavily on manual features, resulting in poor performance. Consequently, traditional methods such as TopHat do not perform nearly as well as those based on deep learning. As the models based on deep learning, AGPCNet utilizes attention mechanism, contextual information, and feature fusion to explore targets. However, its results are quite unstable in complex scenarios, particularly on the IRSTD-1k and SIRST Aug datasets. ACM and DNANet employ the multi-level feature integration strategy that alleviates the feature loss associated with small targets. Hence, they yield relatively favorable results on some evaluation metrics. In contrast, ISNet focuses on preserving the shape of targets by integrating edge to maintain the contour of small targets. However, it does not effectively utilize this edge information during feature extraction, obtaining the overall mediocre performance. Notably, the proposed method enhances EGPNet by improving the encoder, and integrates cross-scale features within the encoder. Additionally, a positioning branch is introduced to calculate candidate target positions. On this basis, a MGLF module is applied across multiple branches. This design allows MMLNet to comprehensively address positioning, feature retention, and edge preservation. Figs. 4 and 5 clearly demonstrate that our method outperforms competitors across most metrics. Furthermore, the running time of each algorithm on the IRSTD-1k dataset are detailed in Table VIII. The table shows that the running time of the proposed method is indeed the most time-consuming, but it is not much different from other methods. Overall, the proposed approach attains a favorable tradeoff between the performance and running time.

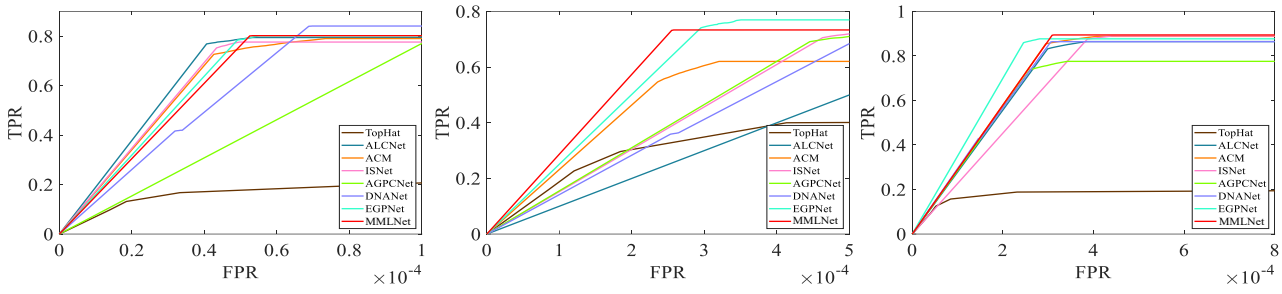


Fig. 4. ROC curves of different methods on three dataset. The figures from left to right represent the results on IRSTD-1k, MDFA, and SIRST Aug datasets, respectively.

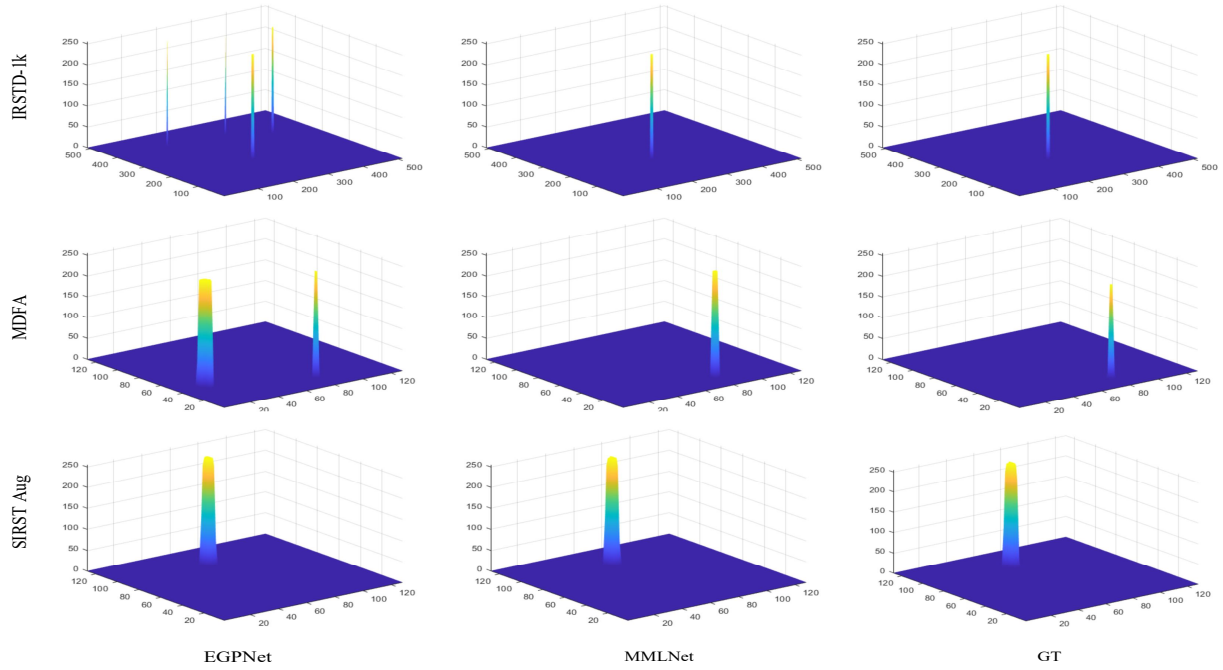


Fig. 5. 3D visualization results of different methods on three dataset. This figure can reflect two aspects: positioning accuracy and target size. Here, the visual size varies with different target sizes.

TABLE VII
QUANTITATIVE EVALUATION ON SIRST AUG DATASET.

Methods	IoU	nIoU	AUC	Fa	Pd
TopHat (GRSL'18)	0.1399	0.2394	0.6065	88.1	0.8377
ALCNet (TGRS'21)	0.7189	0.6806	0.9319	181.5	0.9202
ACM (WACV'21)	0.7357	0.6924	0.9429	302.0	0.8968
ISNet (CVPR'22)	0.7235	0.7075	0.9443	162.6	0.9752
AGPCNet (TAES'23)	0.6523	0.6781	0.8847	44.9	0.9312
DNANet (TIP'23)	0.7414	0.6891	0.9325	265.6	0.9381
EGPNet (TGRS'24)	0.7613	0.7259	0.9342	**19.2**	0.9904
MMLNet	**0.7660**	**0.7316**	**0.9469**	49.6	**0.9917**

2) Qualitative Evaluation: To qualitatively evaluate the performance of our method, we select two images from each dataset, with the detection results presented in Fig 6. Considering that the target itself is very small and difficult to observe, we employ different colors and shapes to label the detection results in the images. Overall, the proposed method demonstrates the better detection effect in terms of false detection, missed detection, and shape retention, particularly for very small targets. Specifically, many competitors exhibit significant false and missed detections on the IRSTD-1k and MDFA datasets, as illustrated in image (2). Different from this condition, all methods show good target detection in images (5) and (6) on the SIRST Aug dataset. However, there are still slight differences in shape retention. Interestingly, the detection results achieved by our proposed method are much closer to the ground truth (GT). The analysis reveals the instability of the comparison methods when dealing with scene changes. In summary, the proposed method can address various challenges, such as complex backgrounds and diverse target shapes, resulting in superior visual results.

V. CONCLUSIONS

This paper presents a multi-branch mutual-guiding learning network (MMLNet) for infrared small target detection, which enhances detection performance through three key

TABLE VIII
TIME ANALYSIS FOR EXISTING INFRARED TARGET DETECTION APPROACHES ON IRSTD-1K DATASET.

Time Analysis	TopHat	ALCNet	ACM	ISNet	AGPCNet	DNANet	EGPNet	MMLNet
Time (s)	0.16	0.58	0.28	1.0	0.51	0.23	0.58	0.71

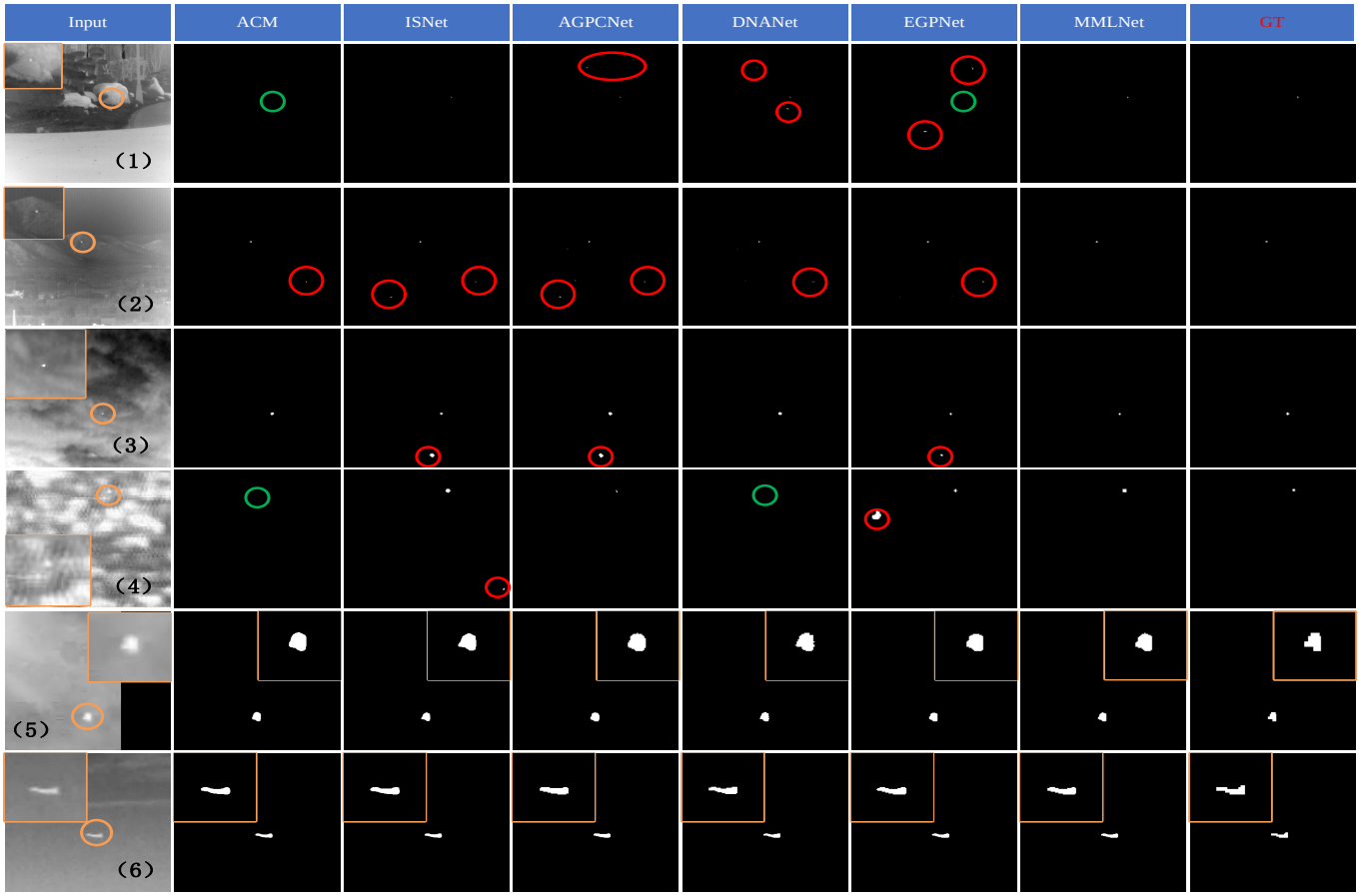


Fig. 6. Qualitative results achieved by different infrared target detection methods. The red indicates false detection, and the green indicates missed detection.

components: a multi-dimension lossless encoder, a candidate target positioning module, and a MGLF module. The multi-dimension lossless encoder effectively reduces feature loss by employing a downsampling strategy and multi-level feature fusion, ensuring that essential image details are preserved. The candidate target positioning module utilizes prior knowledge of infrared targets to accurately identify potential target regions. It reduces background noise interference. Furthermore, the MGLF module facilitates information exchange between branches, which optimizes feature integration for improved detection accuracy. Experimental results demonstrate that our approach obtains the favorable performance. Future work will focus on analyzing similar infrared targets in two images. This research aims to better understand their context by comparing and synthesizing the target features in different images. This approach can identify potential patterns and differences, which will help improve the accuracy of infrared target detection.

REFERENCES

[1] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 7252–7263, 2022.

[2] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–12, 2022.

[3] Y. Jiang, Y. Xi, L. Zhang, Y. Wu, F. Tan, and Q. Hou, "Infrared small target detection based on local contrast measure with a flexible window," *IEEE Geosci. Remote Sens. Lett.*, pp. 1–1, 2024.

[4] C. Yu, Y. Liu, S. Wu, Z. Hu, X. Xia, D. Lan, and X. Liu, "Infrared small target detection based on multiscale local contrast learning networks," *Infrared Phys. Technol.*, vol. 123, pp. 104107, 2022.

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

[6] 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 Sensing*, vol. 52, no. 1, pp. 574–581, 2014.

[7] Y. Zhang, Z. Li, A. Siddique, A. Azeem, and W. Chen, "A real-time infrared small target detection based on double dilate contrast measure," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, pp. 1–16, 2024.

[8] F. Chen, H. Wang, Y. Zhou, T. Ye, and Z. Fan, "DSDANet: Infrared dim small target detection via attention enhanced feature fusion network," in *International Conference on Intelligent Computing*. Springer, 2024, pp. 219–235.

[9] T. Guo, B. Zhou, F. Luo, L. Zhang, and X. Gao, "DMFNet: Dual-encoder multistage feature fusion network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–14, 2024.

[10] Q. Shi, C. Zhang, Z. Chen, F. Lu, L. Ge, and S. Wei, "An infrared small target detection method using coordinate attention and feature fusion," *Infrared Phys. Technol.*, vol. 131, pp. 104614, 2023.

[11] C. Li, Y. Zhang, Z. Shi, Y. Zhang, and Y. Zhang, "Moderately dense adaptive feature fusion network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–12, 2024.

[12] 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.*, 2021, pp. 950–959.

[13] S. Yu, K. Xu, M. Ma, C. Chen, and D. Wang, "Robust infrared small target detection with multi-feature fusion," *Infrared Phys. Technol.*, vol. 139, pp. 104975, 2024.

[14] X. Tong, S. Su, P. Wu, R. Guo, J. Wei, Z. Zuo, and B. Sun, "MSAFFNet: A multiscale label-supervised attention feature fusion network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–16, 2023.

[15] Y. Xu, M. Wan, X. Zhang, J. Wu, Y. Chen, Q. Chen, and G. Gu, "Infrared small target detection based on local contrast-weighted multidirectional derivative," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–16, 2023.

[16] X. Zhang, J. Ru, and C. Wu, "Infrared small target detection based on gradient correlation filtering and contrast measurement," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–12, 2023.

[17] H. Wang, J. Liu, Y. Liu, and H. Sun, "Hierarchical interactive learning network for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[18] T. Liu, Q. Yin, J. Yang, Y. Wang, and W. An, "Combining deep denoiser and low-rank priors for infrared small target detection," *Pattern Recognit.*, vol. 135, pp. 109184, 2023.

[19] F. Zhang, S. Lin, X. Xiao, Y. Wang, and Y. Zhao, "Global attention network with multiscale feature fusion for infrared small target detection," *Opt. Laser Technol.*, vol. 168, pp. 110012, 2024.

[20] K. Wang, X. Wu, P. Zhou, Z. Chen, R. Zhang, L. Yang, and Y. Li, "AFE-Net: Attention-guided feature enhancement network for infrared small target detection," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, 2024.

[21] Y. Huang, X. Zhi, J. Hu, L. Yu, Q. Han, W. Chen, and W. Zhang, "FDDBA-NET: Frequency domain decoupling bidirectional interactive attention network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, 2024.

[22] X. Chen, J. Li, T. Gao, Y. Piao, H. Ji, B. Yang, and W. Xu, "Dynamic context-aware pyramid network for infrared small target detection," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 13780–13794, 2024.

[23] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023.

[24] H. Xu, S. Zhong, J. Zhang, and X. Zou, "Multiscale multilevel residual feature fusion for real-time infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–16, 2023.

[25] F. Chen, C. Gao, F. Liu, Y. Zhao, Y. Zhou, D. Meng, and W. Zuo, "Local patch network with global attention for infrared small target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 5, pp. 3979–3991, 2022.

[26] F. Lin, K. Bao, Y. Li, D. Zeng, and S. Ge, "Learning contrast-enhanced shape-biased representations for infrared small target detection," *IEEE Trans. Image Process.*, vol. 33, pp. 3047–3058, 2024.

[27] F. Lin, S. Ge, K. Bao, C. Yan, and D. Zeng, "Learning shape-biased representations for infrared small target detection," *IEEE Trans. Multimedia*, vol. 26, pp. 4681–4692, 2024.

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

[29] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-guided perceptual network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–10, 2024.

[30] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, 2020.

[31] Y. Wang, P. Jiang, and N. Pan, "Infrared small target detection based on local significance and multiscale," *Digit. Signal Prog.*, vol. 155, pp. 104721, 2024.

[32] J. Zhao, Z. Shi, C. Yu, and Y. Liu, "Multi-scale direction-aware network for infrared small target detection," *CoRR*, vol. abs/2406.02037, 2024.

[33] G. Xu, W. Liao, X. Zhang, C. Li, X. He, and X. Wu, "Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation," *Pattern Recognit.*, vol. 143, pp. 109819, 2023.

[34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.

[35] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and C. M. Jorge, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 240–248.

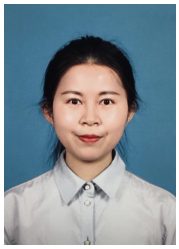
[36] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognit.*, vol. 43, no. 6, pp. 2145–2156, 2010.

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

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

Qiang Li is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Wei Zhang is pursuing a Ph.D. in computer science and technology at the School of Computer Science and the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, remote sensing, and 3D reconstruction.



Wanxuan Lu received the B.Sc. degree from the Beijing Institute of Technology, Beijing, China, in 2016, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2021. She is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include computer vision and remote sensing data processing.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition, and remote sensing. For more information, visit the link (https://crabwq.github.io/).