Edge-Guided Perceptual Network for Infrared Small Target Detection

Qiang Li, *Member, IEEE,* Mingwei Zhang, Zhigang Yang, Yuan Yuan, *Senior Member, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Infrared small target detection (IRSTD) plays a critical role in applications such as night navigation and fire rescue. Its primary purpose is to extract small targets from cluttered backgrounds. While deep learning-based methods have made great advancements in this field, there are still some limitations. One common issue is that the detected target shape tends to be smooth, and extremely small targets may not be effectively detected due to background interference. This paper proposes an edge-guided perception network (EGPNet) for IRSTD to alleviate this trouble. To maintain the information of small targets, EGPNet utilizes a multiscale feature progressive fusion encoder to extract features. This progressive fusion manner enhances semantic information and contextual correlation. Considering that the detected target shapes may result in smoothing effect, an edge-guided image refinement module is incorporated to improve the integrity of the target shape. Moreover, we introduce a local target amplifier to boost the visibility and representation of targets, while suppressing the clutter background interference. The experimental results illustrate that the proposed model can detect the targets with small and weak in different scenes well. Our code is publicly available at https://github.com/qianngli/EGPNet.**

***Index Terms*—Infrared image, small target detection, progressive fusion, edge guidance, target shape.**

I. Introduction

INFRARED image exhibits distinctive characteristics that make it less affected by certain environmental interferences such as low illumination and intense sunlight. It can generate relatively clear and visible contents even under complex conditions. Infrared small target detection (IRSTD) is a typical task based on infrared image, and its purpose is to detect small targets in infrared image. The technology has the ability to quickly analyze potential threats or targets, and can give timely information to make accurate judgment. This advantage provides substantial support for night navigation [1], fire rescue [2], etc.

Unlike general target detection [3], [4], IRSTD faces several challenges, including the following aspects: 1) Small size and low contrast. Infrared small targets typically exhibit small size, which results in limited information within the image. Meanwhile, it forms a low contrast with the surrounding background thermal radiation. 2) Complex background interference. The infrared image is usually collected from long distance, which can easily cause the targets in the image to be disturbed by the complex background. 3) Diversity and variability. Infrared small targets show the diversity and variability in terms of appearance and shape. The above points increase the possibility of false detection and missed detection. Moreover, these enable the algorithm needs to have a strong generalization ability, and be able to adapt to the changes in various situations.

The study of IRSTD has an extensive history. Currently, the existing techniques can generally be divided into two categories, i.e., traditional methods [5], [6], [7] and deep learning-based methods [8], [9], [10]. Traditional detectors are usually constructed by hand-crafted features, such as filtering-based methods [11], [12], local contrast-based methods [13], [14], [15], [16], and low rank-based methods [17], [18], [19], [20]. These approaches obviously have some limitations. For instance, they exhibit the favourable performance in uniform background clutter but fail to suppress complex background. When complex noise is presented in the image, these approaches demonstrate the poor stability. Importantly, they rely primarily on hand-crafted features when building models, and often need to adjust corresponding hyperparameters under various conditions. These factors restrict their adaptability across diverse targets and scenarios, leading to poor results.

Benefiting from the powerful feature representations of convolution neural networks [21] [22] [23], deep learning-based methods have demonstrated superior performance in adaptability to different targets and scenarios [24], [25], [10], [26], [27]. Here, some methods mainly design the model from the perspective of the properties and the shape of target. For instance, Dai et al. [28] incorporate local contrast prior and local attention feature modulation module into the network. It embed low-level context and detail into high-level features to enhance the features of small infrared targets. As for the shape of target, Zhang et al. [26] develop infrared shape network (ISNet) through shape reconstruction strategy to reserve the shape. Later, researchers use Transformer to design a deep learning framework that focuses on target response and background context in infrared images during the coding stage to improve detection accuracy [29]. Through the review of the existing methods, there are several challenges associated with IRSTD that existing methods struggle to address. Firstly, infrared small targets can indeed have significantly smaller

This work was supported in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-YBXM-130 and in part by the National Natural Science Foundation of China under Grant 62301385.

Qiang Li, Zhigang Yang, Yuan Yuan, 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, zgyang@mail.nwpu.edu.cn, y.yuan1.ieee@gmail.com, crabwq@gmail.com)

Mingwei Zhang is with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: dlaizmw@gmail.com) (*Corresponding author: Qi Wang*)

sizes compared to more common targets. In some cases, these small targets may only occupy a few pixels within an infrared image. As a result, many target detection methods suffer from significant information loss after multiple downsampling operations, leading to ineffective detection. For instance, some methods only focus on module design, which is not well avoided, such as [29], [30], etc. Secondly, these methods often utilize patches with small sizes as input during model training. This manner can result in situations where the target to be detected is not present in the constructed training pairs. Currently, some methods neglect this point. Consequently, these models lack sufficient target information to accurately learn the relevant features. Lastly, existing methods tend to produce smooth target shapes during detection. At present, there are few techniques to design the shape characteristics of small targets, such as ISNet [26]. Therefore, how to detect the small targets and precisely capture their shape details needs further efforts.

To alleviate the above issues, an edge-guided perception network (EGPNet) for IRSTD is proposed in our paper. The network architecture mainly contains two branches, i.e., edge extraction branch and target detection branch. In the encoding stage, we leverage a multiscale feature progressive fusion (MFPF) encoder to gradually integrate corresponding features across different scales, thereby enhancing the feature representations for small targets. In the decoding stage, EGPNet realizes the information interaction between edge features and target features by simple aggregation and separation. Moreover, an edge-guided image refinement module (EIRM) is designed, which enables the model to accurately model the shape of target. In summary, the contributions of the proposed EGPNet approach can be summarized as follows:

- We propose an EGPNet by dual-branch architecture to alleviate the challenges mentioned above in IRSTD. EGPNet can work independently for the image soft-edge reconstruction, and provides image soft-edge prior for high-quality IRSTD. Meanwhile, EGPNet utilizes edge guidance to enhance the performance of small target detection in terms of accuracy and reliability.
- A MFPF encoder is proposed to maintain small targets during deep feature extraction. It progressively exploits multiscale contents to enhances semantic information and contextual correlation. Moreover, an EIRM is designed to enlarge the candidate targets, which enables the detected target shape to be effectively preserved. To highlight small targets, a local target amplifier (LTA) is developed. The amplifier boosts the visibility of target in the image and suppresses background information. The strategy makes the model to better locate the targets.
- We verify the effectiveness of the proposed framework through model analysis. In addition, comparison experiments on several datasets demonstrate that the proposed model can detect the targets with small and weak in different scenes well.

The rest of this paper is organized as follows: Section II introduces related IRSTD methods. Section III describes the proposed EGPNet in detail. Section IV analyses and discusses the experimental results. Finally, Section V presents the conclusion.

II. Related Work

IRSTD has an extensive history, and we briefly review the main works in this section, which involves traditional methods and deep-learning methods.

A. Traditional Methods

Traditional IRSTD methods often rely on hand-crafted features to design the models. Through the review of the traditional methods, they are roughly divided into three categories, including filtering-based methods, local contrast-based methods, and low rank-based methods. Here, the filtering-based methods usually exploit morphological operations to detect small targets. The representative algorithm is Top-Hat transform. For example, Bai et al. [11], [31] analyze the IRSTD via morphological filtering, and make various improvements to the traditional Top-Hat transform. Later, Meng et al. [12] explore the characteristics of background, target and noise, and design an improved Top-Hat detector. Local contrast-based methods address IRSTD by exploring the local image contrast. Chen et al. [32] propose a local contrast measure (LCM). On this basis, Han et al. [13] present an improved LCM via contrast mechanism. This algorithm replaces the maximum value with the gray mean, which alleviates the noise effect. Deng et al. [15] design a weighted local difference measure. The approach uses modified local entropy to measure the local difference between small targets and background clutter. It can suppress background clutters and noise well. Similar methods have [33], [16], and [34]. The low rank-based methods formulates small target detection as an optimization problem. Here, the infrared image is represented as a low rank matrix and a sparse matrix, which effectively separates the targets from the background. For instance, Gao et al. [35] propose the infrared patch-image model. Inspired by this model, several improved algorithms are developed. He et al. [17] consider the noise factor and develop a model that combines low-rank and sparse representation. The model employs an overcomplete dictionary to sparsely represent targets and adds constraint on the noise component. In summary, traditional methods have some limitations in dealing with complex noise and scenarios. These methods are usually designed via hand-crafted features and may not be robust enough to background noise. Furthermore, they sometimes need to manually adjust hyperparameters to adapt to diverse targets and scenarios, which makes it more difficult to utilize.

B. Deep Learning-based Methods

IRSTD based on encoder and decoder is a typical detection technology to effectively extract small targets [36], [37], [26]. For instance, Zhang et al. [38] constructs three key modules to explore targets by attention, contextual content, and feature fusion. Inspired by the benefits of convolutional neural networks and Transformer, Lin et al. [39] combine them to fully explore global semantic information. Considering the advantage of local contrast measure, Dai et al. [28] integrates the local contrast prior and local attention feature modulation module into a feature pyramid network. Similarly, Dai et

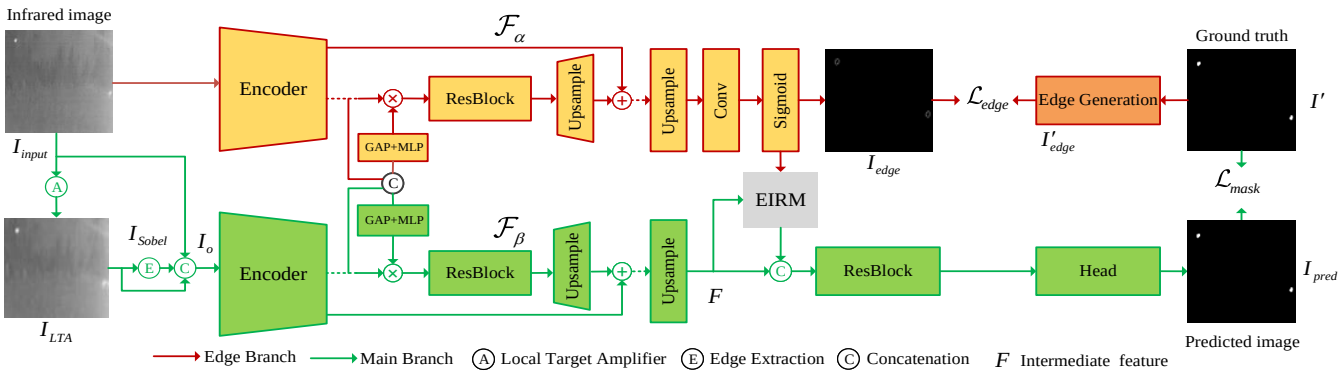


Fig. 1. Overview of the proposed Edge-Guided Perception Network (EGPNet) for IRSTD.

al. [37] also propose an asymmetric contextual modulation module to detect small targets. The method aims to embed low-level contexts with fine details into high-level features and enhance the characteristics of infrared small targets. To obtain the accurate shape of the target, Zhang et al. [26] incorporate the shape reconstruction, and develop a ISNet via shape matter strategy. Additionally, researchers also propose other IRSTD methods, such as Li et al. [30] propose a dense nested attention network, Zhou et al. [18] design a deep low-rank and sparse patch-image network, etc.

Currently, most existing deep learning-based methods commonly suffer from severe loss of target information during multiple downsampling operations, leading to ineffective target detection. Besides, some methods often utilize smaller image patches as input during model training. However, this approach may result in the inexistence of the target to be detected in the constructed training pairs. Moreover, a common issue in current networks is to produce smooth target shape in process of detection. Although ISNet can maintain the shape of small target, it does not make effective use of edges to guide target detection. Importantly, this method adds off-the-shelf edge detector during input. As a result, it inevitably has the appearance of false edges.

III. The Proposed Method

In this section, we introduce the motivation and overview of our EGPNet, and provide the details of the proposed network by several aspects.

A. Motivation and Overview

Unlike general target detection task, IRSTD aims to deal with the images collected under long distance condition, which is faced with the fact that the targets occupy only a few pixels. Although most existing methods are devoted to IRSTD task, the performance improvement is not obvious. In our opinion, there are two main reasons. On the one hand, the weak and small targets may be inundated by the noise in the image under long distance acquisition. How to highlight the target and suppress the noise is crucial for IRSTD. On the other hand, many methods often fail to adequately consider the characteristics of the target shape during model design. When the target shape exhibits irregular, the constructed models tend

Fig. 2. Illustration of two examples by amplifying targets through local target amplifier (LTA).

to produce overly smooth representation in target shape. This smoothing effect significantly impacts task that depend on the Intersection over Union (IoU) evaluation metric, particularly only a small number of pixels. These two factors lead to the poor robustness of these methods in across scenarios.

Considering the fact that the complete edge can maintain the shape of the target, an EGPNet is proposed to address the challenges for IRSTD through edge guidance. The framework is shown in Fig. 1. It can detect small targets and capture the shape information simultaneously. Given an infrared image $I_{input} \in \mathbb{R}^{W \times H}$, the purpose of IRSTD is to obtain the targets in the predicted image $I_{pred} \in \mathbb{R}^{W \times H}$, where W and H denotes the width and height of the image. Note that deep learning-based methods tend to segment small targets rather than detecting the existence of targets in a specific location. Specifically, we build the model through dual-branch asymmetric architecture, i.e., edge extraction branch $\mathcal{F}_\alpha$ and target detection branch $\mathcal{F}_\beta$. Different from previous works, EGPNet can work independently in edge extraction branch for the image soft-edge reconstruction, and provides image soft-edge prior for high-quality IRSTD, i.e.,

$$I_{edge} = \mathcal{F}_\alpha(I_{input}; \theta_\alpha), \tag{1}$$

where θ_α denotes the parameter set of the edge extraction branch. Inspired by boundary boosting algorithm [40], a simple and effective local target amplifier (LTA) in target detection branch $\mathcal{F}_\beta$ is designed to highlight the targets in the image and suppress the background information. Here, LTA aims to expand the boundary with small targets by involving local transformations of image regions through unfold function. It is almost parameter- or computation-free. The process is formulated as

$$N_i = Unfold(I_{input}) \tag{2}$$

$$I_{LTA} = Fold(Max(N_i)) \tag{3}$$

where the patches N_i with size 3×3 are obtained by slicing the image I_{input} with one step. $Fold(\cdot)$ and $Unfold(\cdot)$ aim to reshape these patches. Fig. 2 shows two examples by amplifying targets through LTA. The amplifier brightens the highlighted regions and suppresses unrelated regions in the original image. Meanwhile, Sobel operator is employed to generate the image edge I_{Sobel}. The original image I_{input} and the two processed images are fed into the branch $\mathcal{F}_\beta$ together to extract small targets, i.e.,

$$I_o = [I_{input}, I_{Sobel}, I_{LTA}], \tag{4}$$

$$F = \mathcal{F}_\beta(I_o; \theta_\beta), \tag{5}$$

where θ_β denotes the parameter set of the target detection branch. In the feature extraction stage, a multiscale feature progressive fusion (MFPF) encoder is proposed to maintain small targets during deep feature extraction, which progressively uses multiscale context information, so as to reinforce the feature representations of small targets. In the decoding stage, the information interaction between the two branches is realized through aggregation, which is attached to the corresponding branches by global average pooling (GAP) and multilayer perceptron (MLP), and their respective attributes are explicitly recalibrated. The manner is conducive to more accurate and effective feature representation for each branch. Furthermore, according to the soft-edge prior provided by edge branch, the position of candidate targets is located via edge-guided image refinement module (EIRM). The module implicitly processes the targets so that the model can accurately capture the boundary of the target. Finally, the model is optimized by edge loss $\mathcal{L}_{edge}$ and cross entropy loss $\mathcal{L}_{mask}$ to achieve target exploration. Here, we leverage Dice loss [41] and binary cross-entropy (BCE) to supervise the edge prediction, which is denoted as

$$\mathcal{L}_{edge} = \mathcal{L}_{edge}^{Dice} + \lambda \mathcal{L}_{edge}^{BCE}, \tag{6}$$

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

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

where λ is a balanced factor, and it is set to 10 in our paper. I'_{edge} is an edge image generated by off-the-shelf edge detector. With respect to $\mathcal{L}_{mask}$, it is defined as

$$\mathcal{L}_{mask} = -[I' \log(I_{pred}) + (1 - I') \log(1 - I_{pred})], \tag{9}$$

where I' is ground-truth for target detection branch. The total loss function of the proposed network is set to

$$\mathcal{L} = \mathcal{L}_{mask} + \mathcal{L}_{edge}. \tag{10}$$

B. Multiscale Feature Progressive Fusion Encoder

As mentioned earlier, the targets to be detected take up only a few pixels in most cases. When using conventional ResNet [42] to study features in the coding stage, the size of the feature map would be gradually reduced after several downsampling operations. It can cause the targets to become smaller or even invisible. This means that the model cannot effectively mine the contents of target. To overcome this limitation, a MFPF encoder is proposed, as shown in Fig. 3. Different from other strategies, this encoder maintains the UNet framework during encoding stage, and employs a progressive feature fusion technique to fuse features between adjacent scales. This approach allows the model to preserve and enhance the information related to small targets, thereby improving its ability to perceive and detect them accurately.

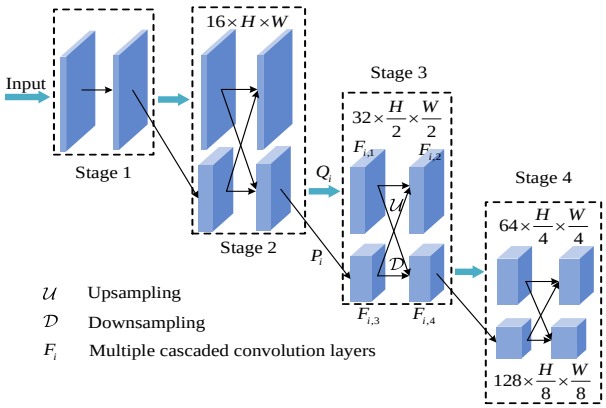


Fig. 3. Structure of the multiscale feature progressive fusion encoder.

The process first utilize the convolution operations to explore features from the input image. These features are then progressively fused with different scale features to augment semantic information and contextual correlation. The progressive feature fusion mechanism repeats these operations. It enables the model to gradually integrate features from different scales, which avoids the loss of important information. In our paper, we stack several stages in MFPF encoder to form feature extraction module. Take one of these stages, and suppose Q_i is the output of the previous $(i-1)$-th stage ($i \in \{2, 3, 4\}$) in encoder. The features P_i dowsampled by *bilinear* operation and the original features Q_i jointly are fed into i-th stage. The i-th stage is computed as

$$\begin{cases} Q_{i+1} = \mathcal{F}_{i,2}([\mathcal{F}_{i,1}(Q_i), \mathcal{U}(\mathcal{F}_{i,3}(P_i))]) \\ P_{i+1} = \mathcal{F}_{i,4}([\mathcal{F}_{i,3}(P_i), \mathcal{D}(\mathcal{F}_{i,1}(Q_i))]) \end{cases}, \tag{11}$$

where $\mathcal{F}_i(\cdot)$ denotes the multiple cascaded convolution layers of the same convolution block. $\mathcal{U}(\cdot)$ and $\mathcal{D}(\cdot)$ represent upsampling and downsampling operations. Note that the generated features Q_{i+1} also are embedded in decoder by skip connection. The model obtains richer and more comprehensive feature representations with a relatively global perspective. Compared with multiple downsamplings in traditional methods, this strategy helps to avoid the feature loss of small targets. On the whole, the MFPF incorporates a progressive feature fusion mechanism and restricts downsampling operations. It essentially represents small targets, thus improving the accuracy and reliability.

C. Edge-Guided Image Refinement Module

Despite MFPF encoder is employed in the encoding stage to alleviate information degradation, it still encounters challenges in preserving target features, particularly for target boundary.

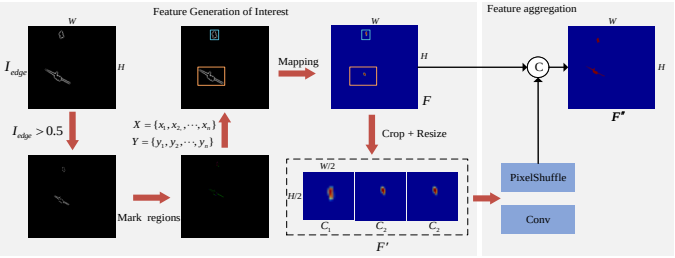


Fig. 4. Illustration of the implementation process of edge-guided image refinement module (EIRM).

It causes the smoothing effect for detected target shape. Additionally, the target to be detected takes up only a few pixels in most cases, this phenomenon can noticeably affect tasks that depend on IoU evaluation metric. Therefore, it is necessary to further analyze these targets and enrich more details. To this end, we propose an EIRM, which consists of feature generation of interest and feature aggregation. Specifically, we focus the model on this feature generation of interest. The procedures are described in **Algorithm 1**. To learn more detailed shape from the extracted feature of interest, we convolve the feature F' in feature aggregation part, which is followed by PixelShuffle layer. Finally, the upsampled feature and intermediate feature F are fused to complete lost contents, which forms a new cube F''. The module modulates the model to better pay attention to the regions of interest and refines the feature representation in shape aspect. Fig. 4 shows the illustration of the implementation process of EIRM.

Algorithm 1: Feature generation of interest

Input: Edge map $I_{edge} \in \mathbb{R}^{W\times H}$ in edge extraction branch, and intermediate feature $F \in \mathbb{R}^{W\times H}$ in target detection branch

Output: Feature of interest $F' \in \mathbb{R}^{3\times W/2\times H/2}$

Get new edge output by simply setting $I_{edge} > 0.5$;

Find the connected regions by two-pass steps, forming the candidate position coordinates $\{X, Y\}$, where $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$;

Obtain the number of connected regions τ according to $\{X, Y\}$;

if $\tau > 0$ **then**

  Generate respective confidence γ by computing average confidence within connected region;

  Select the connected regions Z with the top three confidence by descending confidence γ;

  Compute the minimum external rectangle of each region in Z and the rectangles are mapped into the corresponding location in feature F;

  Crop intermediate feature F, and obtain three patches and make the features the same size $C_m \in \mathbb{R}^{W/2\times H/2}$ $(m \in \{1, 2, 3\})$. Note that we copy any patch to meet the requirement;

  Get feature of interest F' = $[C_m]$;

else

  Copy the feature F, and obtain feature of interest by concatenation F' = $[\mathcal{D}(F), \mathcal{D}(F), \mathcal{D}(F)]$, where $\mathcal{D}(\cdot)$ represent downsampling operation;

end

IV. Experiments

Extensive experiments are conducted in this section. First, we describe the dataset, evaluation metrics, and implementation details. Then the effectiveness of model is analyze by various dimensions. Finally, the performance comparison with existing approaches is shown in quantitative and qualitative aspects.

A. Datasets

Three datasets are adopted to analyze the proposed method, i.e., IRSTD-1k[1], SIRST Aug[2], and MDFA[3]. The IRSTD-1k dataset is collected under long distance condition in the real world. The targets contain drones, vehicles, vessels, etc. The resolution of each image in this dataset is 512×512 pixels, and the total number of images is 1,001. These images are grouped into 800 and 201 in training set and test set, respectively. Since the number of images on IRSTD-1k dataset is small, we augment the given training data by randomly choosing 5 patches. The SIRST Aug dataset is obtained by augmenting SIRST with 427 images. The dataset has 8,525 images in the training set and 545 images in the test set, and the size of each image is 256 × 256 pixels. As for MDFA dataset, it contains real infrared images and synthetic images. The number in training set and test set are 9,978 and 100. Unlike above datasets, the resolution of each image in training set is fixed at 128 × 128 pixels, and the resolution of each image in the test set is different. With respect to above three datasets, we add noise, flip, rotation and other operations in training set to augment training images.

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

B. Evaluation Metrics

To quantitatively evaluate the proposed method, we apply five metrics. They are defined as

Intersection over Union (IoU):

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

Normalized Intersection over Union (nIoU):

$$nIoU = \frac{1}{N}\sum_{i=1}^{N}(TP(i)/(T(i) + P(i) - TP(i)))$$

where N represents the total number of images in test set, $TP(i)$, $P(i)$, and $T(i)$ are the number of true positive pixels, predicted positive pixels, and ground truth, respectively.

Area Under Curve (AUC): AUC is defined as a area under Receiver Operating Characteristic (ROC) curve. Here, ROC is determined by two key metrics. Here, they are obtained by

$$FPR = FP/(FP + TN), \tag{13}$$

$$TPR = TP/(TP + FN). \tag{14}$$

TABLE I
EFFECT OF TWO KEY MODULES ON NETWORK PERFORMANCE.

Architecture Type	IRSTD-1k			SIRST Aug			MDFA		
	IoU	nIoU	AUC	IoU	nIoU	AUC	IoU	nIoU	AUC
UNet Encoder Architecture	0.6289	0.6384	0.8816	0.7493	0.7178	0.9286	0.4577	0.4636	0.8038
Affine Transformation Architecture	0.6510	0.6605	0.8803	0.7603	0.7312	0.9390	0.5012	0.4812	0.8677
Only Detection Architecture	0.6582	0.6634	0.8895	0.7529	0.7194	0.9284	0.4847	0.4892	0.8550
EGPNet	0.6662	0.6800	0.8962	0.7613	0.7259	0.9342	0.4925	0.4996	0.8786

TABLE II
EFFECT OF DIFFERENT INPUT IMAGE SIZES ON NETWORK PERFORMANCE.

Metrics	IRSTD-1k			SIRST Aug			MDFA		
	128×128	256×256	480×480	128×128	192×192	256×256	64×64	96×96	128×128
IoU	0.5564	0.642	0.6662	0.7261	0.7516	0.7613	0.3851	0.4777	0.4925
nIoU	0.5785	0.6537	0.6800	0.7183	0.7245	0.7259	0.4300	0.4991	0.4996
AUC	0.8378	0.8999	0.8962	0.9463	0.9343	0.9342	0.9017	0.8527	0.8786

Probability of Detection (P_d): P_d is target-level evaluation metric. It calculates the ratio of correctly predicted targets N_{pred} and all targets N_{all}, i.e.,

$$P_d = N_{pred}/N_{all}. \tag{15}$$

False-Alarm Rate (F_a): F_a also is target-level evaluation metric. It computes the ratio of incorrectly predicted target pixels N_{false} and all image pixels N_{all}, i.e.,

$$F_a = N_{false}/N_{all}. \tag{16}$$

C. Implementation Details

With respect to experiment setups, we adopt AdaGrad optimizer to optimize the proposed method, where weight decay is set to 10^{-4}. The initial learning rate for the model is 0.02, and its value is halved every 30 epoch. The batch size is fixed at 16. To reduce the feature loss of small targets in MFPF encoder, we limit the number of downsampling operations to three. Considering that the larger patch can guarantee that the image must contain the target to be detected, the input sizes on IRSTD-1k, SIRST Aug and MDFA datasets are set to 480×480, 256×256, and 128×128 pixels, respectively. Furthermore, we set two same modules in decoder to decode the edge features and target features. As for upsample operation, the *bilinear* is utilized to upsample feature maps. All the experiments are conducted on the PyTorch framework using NVIDIA GeForce GTX 3090 GPU.

D. Model Analysis

In this section, we analyze the proposed method from many aspects to verify its effectiveness, including MFPF encoder, EIRM, input image size, and ablation study.

1) Study of MFPF Encoder: As mentioned earlier, the targets to be detected occupy rarely a few pixels in most cases. A multiscale feature progressive fusion (MFPF) encoder is proposed. It enables the model to preserve and enhance information related to small targets. To analyze its effectiveness, we replace the proposed encoder with the classical UNet architecture to extract features. To be fair, the number of downsampling in the encoder is the same as in this paper. The first row in the Table I provides the study of this module on network performance. We can notice that the performance gain proposed by our proposed encoder is intensely obvious, which is attributed to two aspects. The module integrates features between adjacent scales to implement feature exploration, which alleviates the information loss of such weak and small target. In addition, this progressive fusion manner strengthens semantic information and contextual correlation. The experiment verifies that the proposed encoder is effective.

2) Study of EIRM: Considering that the detected target shapes may result in smoothing effects, an edge-guided image refinement module (EIRM) is proposed to refine the target features and complete the integrity of the target shape. In this section, we adopt the affine transformation architecture designed in ISNet as a replacement for EIRM to validate its effectiveness. The second row in the Table I reports the impact of this module on the network performance. As one observe, our method obtains some performance improvement. However, compared with the previous MFPF encoder, the gain is not very significant. As for this affine transformation, it boosts the features from target detection branch through edge modulation, which also contributes to performance enhancement. In contrast, the proposed module focuses on the regions of interest via edge guidance and promotes the feature representations in terms of shape information by amplifying candidate regions. By doing so, EIRM owns a more notable effect on performance advance.

3) Study of Edge Extraction Branch: In target detection task, the edge information of the target is crucial for accurate detection and localization. At present, there are few methods to model the edge features of target. Inspired by the fact that edge can provide significant priors, this paper introduces a soft-edge reconstruction branch to guide the model for target detection. To validate the effectiveness of this branch, we remove it and observe the performance changes. The third row in the Table I displays the study of this branch on the network performance. The results dramatically improve after adding this branch. In our view, there are two main reasons for this. One is to achieve feature interaction between the

TABLE III
ABLATION STUDY FOR DIFFERENT COMPONENTS.

Model	Metrics	IRSTD-1k	SIRST Aug	MDFA
w/o EIRM	IoU	0.6413	0.7548	0.5092
	nIoU	0.6586	0.7273	0.4919
	AUC	0.8908	0.9334	0.8862
w/o Inter	IoU	0.6573	0.7613	0.5000
	nIoU	0.6748	0.7240	0.4937
	AUC	0.8963	0.9400	0.8306
w/o LTA	IoU	0.6468	0.7417	0.4826
	nIoU	0.6556	0.7245	0.4789
	AUC	0.8837	0.9311	0.8308
EGPNet	IoU	0.6662	0.7613	0.4925
	nIoU	0.6800	0.7259	0.4996
	AUC	0.8962	0.9342	0.8786

two branches by separation and aggregation. The other is to guide image refinement by predicted edge image. These manners reduce false positives and false negatives, and boost the robustness of model. The experimental results indicate that soft-edge reconstruction branch plays a key role in modeling the edge and shape features of target.

4) Study of Input Image Size: Since some methods usually adopt small patches as input during model training, this tends to the inexistence of targets to be detected in the constructed training data. As a result, the models lack enough information to learn the characteristics of targets accurately. To analyze the influence of input size, we set three input images with different sizes on three datasets, as shown in Table II. Interestingly, as the input image size increases, the performance of model also improves. When the number of targets to be detected in an image is very small, it is more common to encounter this phenomenon in public datasets. In such cases, larger image sizes ensure that the image contains the targets to be detected. Intuitively, the proposed EIRM can aggregate more targets and enhance the feature learning of similar contents from a global perspective. Note that large input size does not necessarily guarantee better model performance. Too large image size may greatly raise the memory footprint and extend the training time of model. Without loss of generality, larger image sizes can make the model perform better than smaller ones.

E. Ablation Study

The proposed dual-branch network incorporates several key modules to explore targets in an image. These modules include the EIRM, LTA, and information interaction (it is defined as Inter) between two branches. Although the modules have compared them with some typical architectures in the previous model analysis, the influence after deleting these modules is not been discussed. To verify this, we individually remove these modules and observe the performance degradation, as shown in Table III. The performance of this model declines to varying degrees without these modules. For example, the LTA serves to amplify the local target information in the input image. It enhances the visibility and representation of target. When the LTA is deleted, the performance of model degrades significantly, which verifies the importance of this module in feature representations. Similarly, the performance of the model shows a downward trend without EIRM and Inter. It indicates that these modules contribute to image refinement and information flow promotion, which ultimately rises the detection performance. Through the analyses of these components, it can be concluded that each component is helpful to network learning and optimization.

TABLE IV
QUANTITATIVE EVALUATION OF EXISTING IRSTD APPROACHES 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
PSTNN (ReS'19)	0.1538	0.3151	0.7222	523.5	0.7205
ALCNet (TGRS'21)	<u>0.6687</u>	<u>0.6665</u>	<u>0.8971</u>	**8.5**	<u>0.9226</u>
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	<u>0.9226</u>
AGPCNet (TAES'23)	0.5602	0.5344	0.8356	17.1	0.9150
DNANet (TIP'23)	**0.6714**	0.5942	**0.9216**	12.0	0.9252
EGPNet	0.6662	**0.6800**	0.8962	24.2	**0.9495**

TABLE V
QUANTITATIVE EVALUATION OF EXISTING IRSTD APPROACHES ON MDFA DATASET. THE BOLD AND UNDERLINE INDICATE THE BEST AND SECOND PERFORMANCE.

Methods	IoU	nIoU	AUC	Fa	Pd
TopHat (GRSL'18)	0.2438	0.2927	0.7589	109.9	0.7426
PSTNN (ReS'19)	0.2965	0.3161	0.7687	318.0	0.7986
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)	<u>0.4376</u>	<u>0.4248</u>	0.8648	355.1	0.7286
AGPCNet (TAES'23)	0.4339	0.4152	0.8603	107.0	<u>0.8417</u>
DNANet (TIP'23)	0.4324	0.4075	<u>0.8676</u>	<u>57.8</u>	0.8201
EGPNet	**0.4925**	**0.4996**	**0.8786**	**37.7**	**0.8929**

F. Performance Comparison with Existing Approaches

This section makes a comprehensive comparison between seven existing methods and proposed EGPNet in both quantitative and qualitative aspects. They are include TopHat [11], PSTNN [19], ALCNet [28], ACM [37], ISNet [26], AGPCNet [38], and DNANet [30]. Here, these codes are publicly available.

TABLE VI
QUANTITATIVE EVALUATION OF EXISTING IRSTD APPROACHES ON SIRST AUG DATASET. THE BOLD AND UNDERLINE INDICATE THE BEST AND SECOND PERFORMANCE.

Methods	IoU	nIoU	AUC	Fa	Pd
TopHat (GRSL'18)	0.1399	0.2394	0.6065	88.1	0.8377
PSTNN (ReS'19)	0.2076	0.2830	0.6078	109.0	0.4539
ALCNet (TGRS'21)	0.7189	0.6806	0.9319	181.5	0.9202
ACM (WACV'21)	0.7357	0.6924	<u>0.9429</u>	302.0	0.8968
ISNet (CVPR'22)	0.7235	<u>0.7075</u>	**0.9443**	162.6	<u>0.9752</u>
AGPCNet (TAES'23)	0.6523	0.6781	0.8847	<u>44.9</u>	0.9312
DNANet (TIP'23)	<u>0.7414</u>	0.6891	0.9325	265.6	0.9381
EGPNet	**0.7613**	**0.7259**	0.9342	**19.2**	**0.9904**

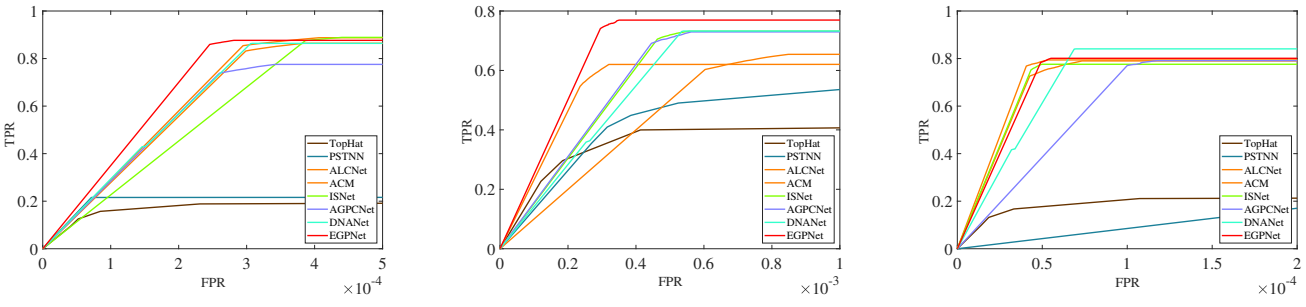


Fig. 5. ROC curves of different methods on three dataset. (a) IRSTD-1k, (b) MDFA, and (c) SIRST Aug.

TABLE VII
TIME ANALYSIS FOR EXISTING IRSTD APPROACHES ON IRSTD-1K DATASET.

Time Analysis	TopHat	PSTNN	ALCNet	ACM	ISNet	AGPCNet	DNANet	EGPNet
Time (s)	0.16	0.50	0.58	0.28	1.0	0.51	0.23	0.58

1) Quantitative Evaluation: Tables IV-VI show quantitative evaluation on three datasets. Overall, the proposed EGPNet obtains the comparable results in accuracy. Since these datasets are collected in complex scenes, traditional methods are ineffective in suppressing complex noise. In addition, they rely heavily on hand-crafted features when constructing models. Consequently, the performance of traditional methods such as TopHat and PSTNN is much lower than that of deep learning-based methods. Among these competitors based on deep learning, there is not a significant fluctuation in performance. ACM utilizes top-down global context feedback to exchange high-level semantics and subtle low-level details. Similarly, DNANet achieves the progressive interaction between high-level and low-level features. Both approaches integrate the high-level semantic features and low-level detailed features, which alleviates the feature loss of small targets. Hence, these two methods yield relatively favorable results on some evaluation metrics. ISNet introduces edge information to preserve the shape of small targets. However, it does not make effective use of edges during feature extraction to guide target detection. Its performance in various metrics is moderate. In contrast, the proposed EGPNet combines the above advantages, which takes the feature loss and shape retention of small targets into account in the model design. It not only provides image soft-edge priors, but also integrates cross-scale features in a progressive way to enhance semantic information and contextual correlation. As a result, it clearly outperforms all competitors in most metrics. This can be further supported by the ROC curves in Fig. 5. Moreover, we also supplement the real-time analysis of each algorithm on IRSTD-1k dataset in Table VII. As seen this table, our method achieves average run time with no advantage compared with others. This is also the weakness of the proposed method.

2) Qualitative Evaluation: The qualitative results of six samples on three datasets are illustrated in Fig. 6. Compared with traditional methods, EGPNet can produce accurate target location and shape segmentation. Note that only results obtained by deep learning-based methods are presented for clarity. Overall, most methods often lead to false alarm and missed detection, especially for very small targets. For instance, image(4) predicted by multiple approaches exhibits several false alarms, and image(2) fails to detect in local regions. It reveals the instability of these methods in dealing with scene changes. Although some competitors can achieve reasonable target location and shape segmentation, they can not maintain the true shape of target well. In contrast, our proposed method can well solve to the challenges such as various complex backgrounds and target shapes, so as to get better visual results.

V. CONCLUSIONS

This paper introduces an edge-guided perception network (EGPNet) for IRSTD. It improves the perception and detection ability of small targets by means of multiscale feature progressive fusion encoder, edge-guided image refinement module, and local target amplifier. EGPNet gradually fuses the features with different scales and retains the feature information of small target. Meanwhile, it leverages the edge information to restore the target shape to enhance the integrity of target. In addition, the local target amplifier is introduced to focus and improve the feature representations of target region to reduce the influence of background interference. Experiments show that EGPNet has advantages over existing methods for IRSTD. The scheme provides a more reliable and accurate solution for detection task. In the future, we will develop from two aspects. One is to build a new dataset with precise data annotation. The other is to strengthen the contextual learning for extremely small and similar contents.

REFERENCES

[1] G. Bhatnagar and Z. Liu, "A novel image fusion framework for night-vision navigation and surveillance," *Signal Image Video Process.*, vol. 9, pp. 165–175, 2015.
[2] I. Bosch, S. Gomez, L. Vergara, and J. Moragues, "Infrared image processing and its application to forest fire surveillance," in *IEEE Conf. Adv. Video Signal Based Surveill.*, 2007, pp. 283–288.

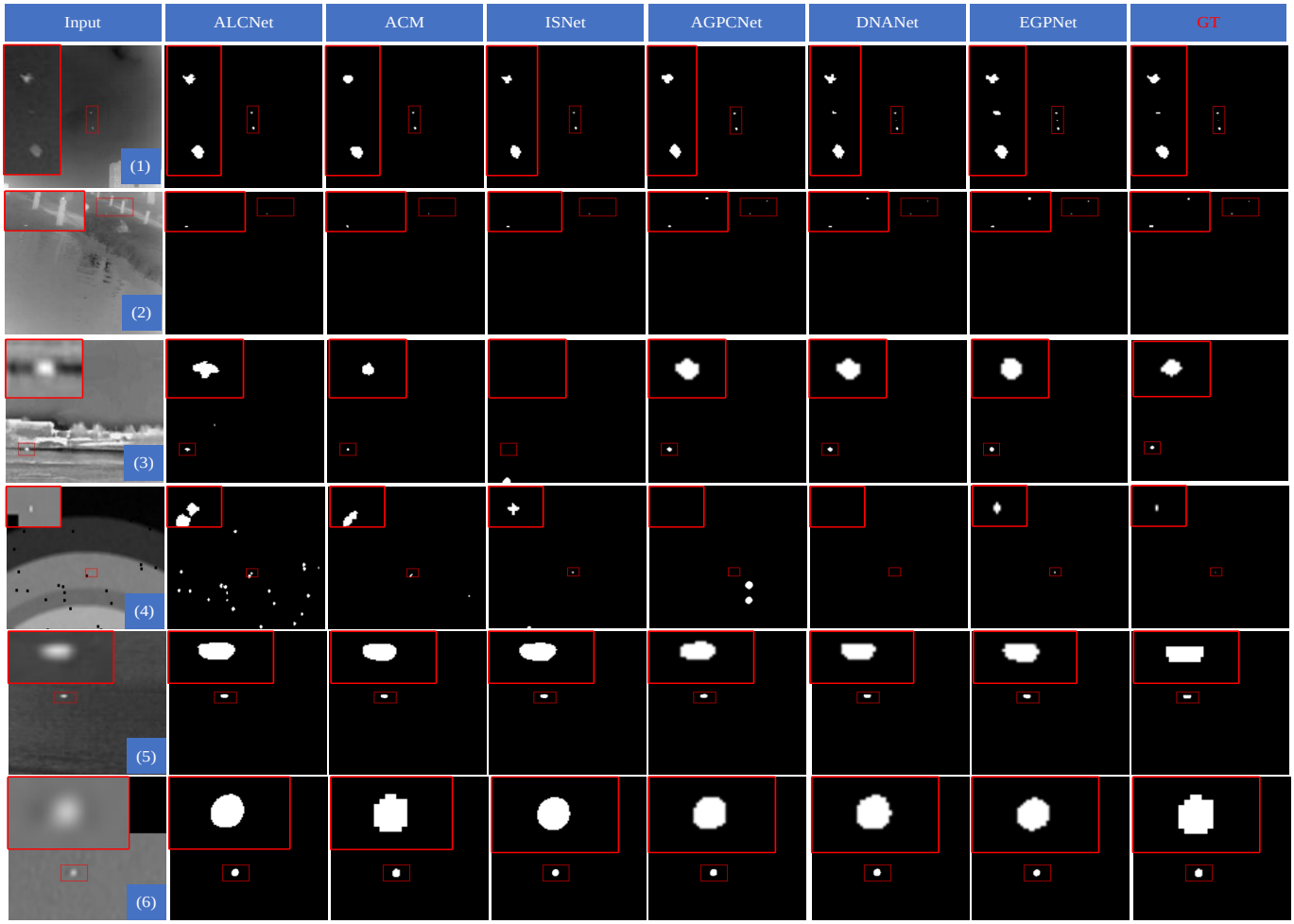


Fig. 6. Qualitative results achieved by different IRSTD methods. To provide a better visualization, the local region is enlarged in the left-top corner.

[3] H. Wang, Q. Wang, H. Zhang, Q. Hu, and W. Zuo, "CrabNet: Fully task-specific feature learning for one-stage object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2962–2974, 2022.

[4] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling knowledge from super-resolution for efficient remote sensing salient object detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–16, 2023.

[5] I. H. Lee and C. G. Park, "Infrared small target detection algorithm using an augmented intensity and density-based clustering," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–14, 2023.

[6] X. Guan, Z. Peng, S. Huang, and Y. Chen, "Gaussian scale-space enhanced local contrast measure for small infrared target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 327–331, 2020.

[7] L. Li, Z. Li, Y. Li, C. Chen, J. Yu, and C. Zhang, "Small infrared target detection based on local difference adaptive measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1258–1262, 2020.

[8] Y. Chen, L. Li, X. Liu, and X. Su, "A multi-task framework for infrared small target detection and segmentation," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–9, 2022.

[9] W. Y. Chung, I. H. Lee, and C. G. Park, "Lightweight infrared small target detection network using full-scale skip connection u-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[10] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2022.

[11] 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.

[12] W. Meng, T. Jin, and X. Zhao, "Adaptive method of dim small object detection with heavy clutter," *Appl. Optics*, vol. 52, no. 10, pp. D64–D74, 2013.

[13] 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, 2014.

[14] J. Han, S. Moradi, I. Faramarzi, H. Zhang, Q. Zhao, X. Zhang, and N. Li, "Infrared small target detection based on the weighted strengthened local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1670–1674, 2020.

[15] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Trans. Geosci. Remote Sensing*, vol. 54, no. 7, pp. 4204–4214, 2016.

[16] J. Liu, Z. He, Z. Chen, and L. Shao, "Tiny and dim infrared target detection based on weighted local contrast," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1780–1784, 2018.

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

[18] X. Zhou, P. Li, Y. Zhang, X. Lu, and Y. Hu, "Deep low-rank and sparse patch-image network for infrared dim and small target detection," *IEEE Trans. Geosci. Remote Sensing*, 2023.

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

[20] H. Zhu, S. Liu, L. Deng, Y. Li, and F. Xiao, "Infrared small target detection via low-rank tensor completion with top-hat regularization," *IEEE Trans. Geosci. Remote Sensing*, vol. 58, no. 2, pp. 1004–1016, 2019.

[21] 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.

[22] 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.

[23] Q. Li, Y. Yuan, and Q. Wang, "Multiscale factor joint learning for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–10, 2023.

[24] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 8508–8517.

[25] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–13, 2022.

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

[27] Y. Zhang, Y. Zhang, Z. Shi, R. Fu, D. Liu, Y. Zhang, and J. Du, "Enhanced cross-domain dim and small infrared target detection via content-decoupled feature alignment," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–16, 2023.

[28] 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.

[29] H. Yang, T. Mu, Z. Dong, Z. Zhang, B. Wang, W. Ke, Q. Yang, and Z. He, "Pbt: Progressive background-aware transformer for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–13, 2024.

[30] 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.

[31] X. Bai, F. Zhou, and T. Jin, "Enhancement of dim small target through modified top-hat transformation under the condition of heavy clutter," *Signal Process.*, vol. 90, no. 5, pp. 1643–1654, 2010.

[32] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 52, no. 1, pp. 574–581, 2013.

[33] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 33–37, 2017.

[34] 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, 2018.

[35] 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, 2013.

[36] X. Ying, L. Liu, Y. Wang, R. Li, G. Chen, Z. Lin, W. Sheng, and S. Zhou, "Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15528–15538.

[37] 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.

[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.

[39] J. Lin, S. Li, L. Zhang, X. Yang, B. Yan, and Z. Meng, "IR-TransDet: Infrared dim and small target detection with ir-transformer," *IEEE Trans. Geosci. Remote Sensing*, 2023.

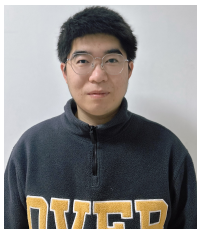
[40] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "LSNet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, 2023.

[41] 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.

[42] M. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

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.

Mingwei Zhang received the B.E. degree in automation from Zhengzhou University, Zhengzhou, China, in 2021, and the M.S. degree from the Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing image acquisition and processing.



Zhigang Yang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing, computer vision and machine learning.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

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, pattern recognition, and remote sensing.