ISPANet: A Pyramid Self-Attention Network for Single-Frame High-Resolution Infrared Small Target Detection With a Large-Scale Dataset SHR-IRST

Wenjing Wang, Chengwang Xiao, Haofeng Dou, Ruixiang Liang, Huaibin Yuan, Guanghui Zhao, Zhiwei Chen, and Yuhang Huang

***Abstract*—In recent years, the imaging resolution of infrared detection equipment has gradually increased, and higher resolution infrared detection equipment has been used in various fields. Compared with lower resolution infrared images, small targets in higher resolution infrared images have larger average target sizes and occupy more pixel points. Therefore, each pixel in the target has less impact on the pixel-level metrics. A method with good pixel-level metrics does not necessarily indicate a strong target detection capability. We aim to enhance the target-level detection performance of high-resolution infrared small target images through the following series of initiatives. First, we construct a single-frame high-resolution infrared small target dataset. Then, we propose an infrared small target pyramid self-attention network (ISPANet) according to the features of small targets in high-resolution infrared images. Finally, we propose a new loss function Focal Soft-IoU (F-SIoU). F-SIoU loss makes the network more concerned about the hard positive examples. Comparative experiments with state-of-the-art networks have shown that our proposed ISPANet has much better target detection performance in high-resolution infrared small target images. Especially in suppressing false alarms of targets, it has a significant effect, with a 2% increase in target detection probability (P_d) and a 4% increase in target accuracy (P_a). ISPANet effectively enhances the credibility of target detection results. At the same time, we also use the actual acquired dataset to test the ISPANet and obtain good results. The dataset and network will subsequently be made public on GitHub.**

***Index Terms*—Deep learning, infrared image, self-attention, small target detection.**

I. Introduction

THE infrared imaging system has characteristics as being unaffected by light conditions, having strong resistance to electromagnetic interference, adapting to various environments, simple structure, being easy to carry, etc. It has been widely used in military and civil fields, such as infrared search and tracking, low and slow small aircraft detection and identification, and autopilot and electric check. In some scenarios that need to be prejudged, the target has a far distance and weak radiation intensity, in addition, the environment around the target is complex and has interference information. At this time, the target appears as a small local bright block in the infrared image and does not contain very specific shape and texture information, which belongs to the infrared small target.

Infrared small target detection methods have been evolving over the years [1]. In the early days, small targets in infrared images were modeled as salient targets in the image or objects with high local contrast [2], [3]. Subsequently, small targets were modeled as sparse components in a low-rank background [4], taking into account their "sparse." The development of these a priori model-based methods has improved the detection ability of infrared small targets [5]. However, the local contrast and sparsity priors of the target are actually summarized by human experience, which means the model-driven approach does not adapt well to more varied and complex scenes.

In recent years, with the rapid development of deep learning (DL) and attention mechanisms in computer vision (CV) tasks, many advanced DL models have emerged in subdivision tasks, such as image classification [6], [7], target detection [8], semantic segmentation, and image change detection [9]. The data-driven infrared small target detection methods can be divided into single-frame detection methods and sequence frames detection methods. Single-frame detection methods typically focus on enhancing local attention to small targets [10] and enhancing fusion between features at different levels [11]. Sequence frames contain temporal information, and sequence frame detection methods typically use interframe difference [12], multistrategy fusion [13], or optical flow estimation [14] to detect moving targets. This article mainly studies single-frame infrared small target detection method based on DL, these data-driven methods have gained substantial performance improvement compared with traditional methods, but still have problems, such as poor adaptability, high false alarm rate, and high miss detection rate.

On the other hand, with the development of infrared detectors, the infrared detection equipment has longer detection distance

Manuscript received 12 December 2023; revised 26 February 2024; accepted 16 March 2024. Date of publication 31 May 2024; date of current version 14 June 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 42306203 and in part by the Young Star of Science and Technology in Shaanxi Province under Grant 2023KJXX-104. *(Corresponding author: Chengwang Xiao.)*

Wenjing Wang and Chengwang Xiao are with the School of Electronic Information, Central South University, Changsha 410083, China (e-mail: wwjnews22@gmail.com; xcwnews22@gmail.com).

Haofeng Dou, Ruixiang Liang, and Huaibin Yuan are with the China Academy of Space Technology (Xi'an), Xi'an 710100, China (e-mail: haofeng_dou@163.com; liangrui_xiang@163.com; 1062441196@qq.com).

Guanghui Zhao, Zhiwei Chen, and Yuhang Huang are with the School of Electronic Information and Communications (Science and Technology on Multi-spectral Information Processing Laboratory), Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: guanghui_zhao@hust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3381779

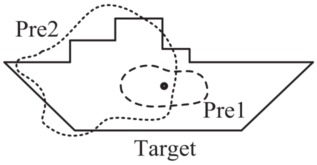


Fig. 1. Schematic representation of a target and its two different predictions. Pre2 predicts more pixel points but the prediction of the target center position is more biased. Pre1 predicts fewer pixel points but the prediction of the target center position is more accurate.

and higher imaging resolution. This poses a new challenge to infrared small target detection problem. Compared with low-resolution infrared small target images, the average size of small targets in high-resolution infrared images is larger, and the size gap between targets is larger. The output of the infrared small target detection network is a binary classification of each pixel in the input image. As shown in Fig. 1, for high-resolution infrared images, small targets occupy more pixels. A method that has good prediction ability at the pixel level does not necessarily mean good prediction ability at the target level.

Existing methods may encounter some problems in small target detection of high-resolution infrared images, such as inaccurate target prediction and localization, and severe false alarms of targets. Based on the above problems, this article mainly investigates how to improve the target-level prediction ability of small target detection methods in high-resolution infrared images. The main content of this article is as follows.

1) We construct a single-frame high resolution infrared small target (SHR-IRST) dataset with a resolution of 1280 × 1024, which is the largest single-frame infrared small target dataset at present, with better diversity in small targets and backgrounds.
2) Based on the small target data characteristics in high-resolution infrared images, we design an infrared small target pyramid self-attention network (ISPANet) network, which can maintain good pixel-level metrics while obtaining better target-level metrics, especially in suppressing the target false alarm capability (improving the credibility of target detection results).
3) We propose a new loss function, Focal Soft-IoU (F-SIoU), which can increase the supervision of hard positive examples and improve the performance of ISPANet in infrared small target detection.

II. Related Works

A. Attention Mechanism

In the last decade, attention mechanism has been introduced into CV tasks and gradually plays an important role. Hu et al. [15] argued that different channels in a feature map contain different information and proposed squeeze-and-excitation network. Based on this, Woo et al. [16] proposed a plug-and-play attention mechanism that combines spatial and channel dimensions. Self-attention mechanism [17] was first proposed in 2017 and rapidly developed in natural language processing. Wang et al. [18] were the first to introduce self-attention mechanism into CV. Some subsequent works on variations based on self-attention [19], [20] focus on increasing the speed of network and reducing the computational complexity. Huang et al. [21] proposed criss-cross attention, which recursively considers row and column attention to obtain global information. ViT [22] splits the image into multiple patches, which are encoded and then used as input to the Transformer, reducing the number of input sequences.

The development of the attention mechanism has promoted the development of CV, and some new solutions have emerged for generalized vision problems, such as target detection and semantic segmentation with good results. However, the infrared small target detection problem has its unique characteristics, and the above method cannot be transferred directly. It is necessary to design a self-attention mechanism module according to its characteristics.

B. Infrared Small Target Detection

This section introduces the related works of infrared small target detection from four aspects: datasets, network structure, evaluation metrics, and loss function.

1) Datasets: In recent years, some scholars have done a lot of work in the collection and production of infrared small target datasets, and have publicly released some datasets. Table I gives the basic information of the currently publicly available single-frame infrared small target datasets. From Table I, it can be seen that the number of real single-frame infrared small target images is small. At the same time, consider the increasing demand for high-resolution infrared small target detection. There is an urgent need to construct a large-scale and high-quality high-resolution single-frame infrared small target dataset.

2) Network Structure: With the rapid development of DL method in CV tasks, DL-based networks have also emerged for infrared small target detection.

Early researchers were more inspired by other visual tasks. Gao et al. [23] drew on the idea of image super-resolution to map low-dimensional features to a high-dimensional space before target detection. Zhao et al. [24] proposed TBC-Net, which uses the results of the classification branch to guide the feature extraction in the segmentation branching network. Then, some researchers have conducted some research on how to enhance feature fusion at different layers [25], [26], [27]. Dai et al. [28] argued that small targets are easily submerged in features at high layers of the network, so they proposed an asymmetric contextual modulation (ACM) mechanism. Based on this, Dai et al. [10] modularized the local contrast measure from the traditional approach to the network. Li et al. [11] proposed DNANet, which maintains the stabilization of small infrared targets in the deep layers through repetitive interactions within the dense nested interactive module. Zhang et al. [27] proposed AGPCNet, which includes attention guided context block that perceive pixel correlations within and between patches at specific scales. Zhang et al. [30] proposed ISNet, which extracts low-level features in the row and column directions and combines them with high-level features using the two-orientation attention aggregation module.

TABLE I
DETAILS OF THE PRESENT SINGLE-FRAME INFRARED SMALL TARGET DATASETS

Dataset Type	Dataset	Image Num	Image Size	Provided Label	Target True Class	Background Type
real	SIRST [10]	427	96×135 to 388×418	Pixel/Box	Aircraft/Drone/Ship/Vehicle	Cloud/Grass/ River
	IRSTD-1k [30]	1001	512×512	Pixel	Drone/Bird/Animal	Cloud/Building/Grass/River/Lake
synthetic	NUDT-SIRST [11]	1327	256×256	Pixel	-	Cloud/Building/Vegetation
	IRST640 [31]	1024	640×512	Pixel	-	Cloud/Building
real/ synthetic	SHR-IRST (our)	2843	1280×1024	Pixel/Box/ Center	Aircraft/Drone/Ship/Vehicle /Bird/Animal	Cloud/Building/Grass/River /Lake/Vegetation

Some other researchers have introduced a self-attention mechanism in the network [29]. Chen et al. [31] proposed IRSTFormer, whose overall structure inherits the encoder–decoder structure of the classical Transformer [17]. MTU-Net [32] is a small target method for space-based infrared ship small targets, with a multilevel ViT CNN hybrid encoder and a U-shape decoder. In addition to this, several other researchers have introduced bounding box prediction methods in the network. Wang et al. [33] proposed IAANet, a two-stage network from coarse to fine, where the coarse stage predicts the potential target regions through region proposition networks. Chen et al. [34] added a target bounding box prediction header to the U-Net network in order to enhance the performance of semantic segmentation prediction. Dai et al. [35] proposed OSCAR to predict bounding boxes of small targets in a cascade from coarse to fine. In order to solve the problem of class imbalance between the target and background, Sun et al. [36] proposed RDIAN. It uses convolutional layers with different receptive fields to capture target features in different local regions, enhancing the diversity of target features. Kou et al. [37] proposed a lightweight infrared small target segmentation network and successfully deployed it on embedded platforms.

There are still some shortcomings in the existing methods. ACM [28] is a simple network, and it is insufficient in the detection rate of infrared small targets. ALCNet [10], IAANet [33], and RDIAN [36] all designed a module that focuses more on local features of small targets, but they ignored global and deep features. This makes their predicted target false alarms more severe. DNANet [11], IRSTFomer [31], and MTU-Net [32] have improved the performance of networks through multilevel feature extraction and fusion, but there is still a problem of poor target prediction integrity, that is, they can easily predict a target as two or more targets.

3) Evaluation Metrics: Infrared small target detection networks usually use both pixel-level metrics and target-level metrics to comprehensively evaluate the performance of small target detection. Pixel-level metrics include the following.

IoU represents the ratio of intersection and union between the predicted and true results

$$\mathrm{IoU} = \frac{\mathrm{TP}}{T + P - \mathrm{TP}} \tag{1}$$

where T, P, and TP denote the true, positive, and true positive, respectively. nIoU [10] is the numerical result normalized by the IoU value of each target

$$\mathrm{nIoU} = \frac{1}{N}\sum_{i}^{N}\frac{\mathrm{TP}\left[i\right]}{T\left[i\right] + P\left[i\right] - \mathrm{TP}\left[i\right]} \tag{2}$$

where N represents the total number of targets. Precision (P) and recall (R) refer to the proportion of correctly predicted positive samples out of all predicted positive samples and all true positive samples, respectively. F1-Score (F1-P) is the harmonic mean of precision and recall

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \mathrm{F1-P} = \frac{2P\cdot R}{P+R} \tag{3}$$

where FP and FN denote the false positive and false negative, respectively.

The receiver operating characteristic (ROC) curve [38] shows the performance of the model at all classification thresholds. The precision recall (PR) [39] curve reflects the accuracy of the classifier in predicting positive examples.

P_d (probability of detection) and F_a (false-alarm rate) [4] are target-level metrics. P_d measures the ratio of the number of correctly predicted targets to the number of all targets. F_a measures the ratio of incorrectly predicted pixels to all pixels of the image

$$P_d = \frac{\#\ \text{num of true detections}}{\#\ \text{num of actual targets}} \tag{4}$$

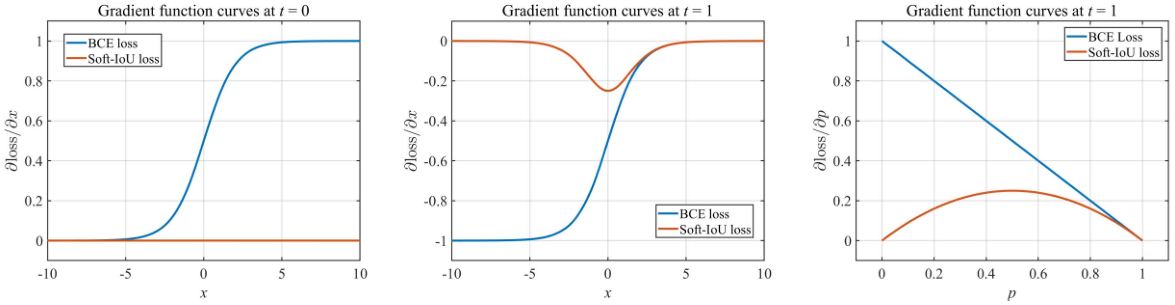
$$F_a = \frac{\#\ \text{num of false predicted pixels}}{\#\ \text{num of all pixels}}. \tag{5}$$

In fact, the existing target-level metric F_a still evaluates the prediction ability at the pixel level, which cannot reflect the false alarm rate of the target level.

4) Loss Function: The output of the infrared small target detection network is the binary classification result of each pixel of the input image. Commonly used loss functions include binary cross-entropy (BCE) loss and Soft-IoU loss [40]. BCE loss represents a measure of the difference between two probability distributions for a given set of random variables or events and is mainly used in the classification prediction task of binary classification. It is defined as

$$\mathcal{L}_{\mathrm{BCE}} = -(t\cdot\log p + (1-t)\cdot\log(1-p)) \tag{6}$$

where t is the true labeling value, with a labeling value of 1 for target (positive sample) and 0 for background (negative sample). p is the target probability predicted by the network and is defined



(a) (b) (c)

Fig. 2. Gradient function curves for BCE loss and Soft-IoU loss. (a) and (b) are the gradient function curves with respect to x for $t = 0$ and $t = 1$, respectively. (c) is a gradient function curve with respect to p for $t = 1$. (The value of gradient function is taken to absolute in (c) for ease of presentation).

as

$$p = \text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

where x is the output of the network. Soft-IoU loss measures the intersection ratio between the predicted result and true label in a soft judgment manner, which is defined as

$$\mathcal{L}_{\text{Soft}-\text{IoU}} = 1 - \frac{\sum_{i,j} p_{i,j} \cdot t_{i,j} + \lambda}{\sum_{i,j} p_{i,j} + t_{i,j} - p_{i,j} \cdot t_{i,j} + \lambda} \tag{8}$$

where i and j denote the pixel points in the image. λ is a smooth factor. More methods choose Soft-IoU loss because of the serious positive and negative sample imbalance in infrared small target images. However, we found that when using Soft-IoU loss training, the loss value did not significantly decrease after many epochs. The network needs to go through dozens of epochs before it begins to converge. This situation does not occur when using BCE loss training. In order to find the reason for the nonconvergence of Soft-IoU loss, we analyze the gradient functions and previous epochs' experimental results.

The gradient function of Soft-IoU loss according to the chain rule of the gradient is obtained as

$$\frac{\partial \mathcal{L}_{\text{Soft}-\text{IoU}}}{\partial x} = \frac{\partial \mathcal{L}_{\text{Soft}-\text{IoU}}}{\partial p} \cdot \frac{\partial p}{\partial x} = \begin{cases} \frac{\lambda p(1-p)}{(p+\lambda)^2}, & t = 0 \\ \frac{p(p-1)}{1+\lambda}, & t = 1 \end{cases} \tag{9}$$

$$\frac{\partial p}{\partial x} = \frac{e^{-x}}{(1 + e^{-x})^2} = p(1 - p). \tag{10}$$

Since the smooth factor λ in Soft-IoU loss is an infinitesimal number, the gradient function of Soft-IoU loss can be organized as

$$\frac{\partial \mathcal{L}_{\text{Soft}-\text{IoU}}}{\partial x} = \frac{\partial \mathcal{L}_{\text{Soft}-\text{IoU}}}{\partial p} \cdot \frac{\partial p}{\partial x} = \begin{cases} 0, & t = 0 \\ p^2 - p, & t = 1. \end{cases} \tag{11}$$

Due to the existence of smooth factor λ, the gradient value at $t = 0$ is not completely equal to 0, but rather a very small value. However, due to the small value, its contribution to the network is limited. The gradient function of the BCE loss according to the chain rule of the gradient is obtained as

$$\frac{\partial \mathcal{L}_{\text{BCE}}}{\partial x} = \frac{\partial \mathcal{L}_{\text{BCE}}}{\partial p} \cdot \frac{\partial p}{\partial x} = \begin{cases} p, & t = 0 \\ p - 1, & t = 1. \end{cases} \tag{12}$$

TABLE II
EVALUATION OF PIXEL-LEVEL METRICS FOR THE FIRST THREE EPOCHS OF THE NETWORK TRAINED WITH SOFT-IOU LOSS

Epoch	nIoU	IoU	R	P	F1-P
1	0. 0004	0. 0004	0. 0917	0. 0004	0. 0008
2	0. 0008	0. 0008	0. 3349	0. 0008	0. 0016
3	0.0010	0. 0013	0. 9612	0.0010	0. 0019

The gradient function curves of BCE loss and Soft-IoU loss about the network output x at $t = 0$ and $t = 1$ are shown in Fig. 2(a) and (b), respectively.

Because of the weight initialization, the network output x is close to 0 at the beginning of the training. At this stage, for a positive sample and a negative sample, BCE loss can provide the same gradient, but Soft-IoU loss only provides gradient for positive sample.

Table II represents the pixel-level metrics for the first three epochs of the network trained with Soft-IoU loss. It can be seen that after three training epochs, the recall (R) has grown to close to 1, but the precision (P) is close to 0. This shows that since there are few positive examples, the network has learned almost all the positive examples at this time. But there are still a large number of negative examples that are misjudged as positive examples. At this time, the gradient values provided by the network to both positive and negative samples are close to 0. Therefore, network exhibits gradient saturation phenomenon, and it is difficult for the network to update parameters by gradient descent. To sum up, Soft-IoU loss does not converge because it has the following deficiencies: lack of supervision of negative examples that are incorrectly predicted as positive examples.

The gradient function curves of BCE loss and Soft-IoU loss on the predicted target probability p at $t = 1$ are shown in Fig. 2(c). It can be seen that the gradient function of BCE is a one-time function, the farther the predicted value is from the correct value, the greater the gradient value provided. The gradient function of Soft-IoU is a quadratic function, the gradient value is the largest when the predicted value is 0.5, and the gradient value provided becomes smaller as the predicted value moves away from the correct value. This shows that Soft-IoU loss also has the following deficiencies: the supervision of hard positive examples is insufficient.

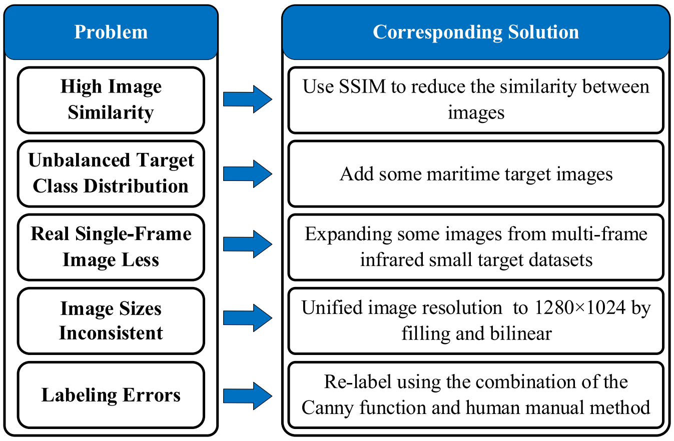

Problem
Corresponding Solution
High Image Similarity
Use SSIM to reduce the similarity between images
Unbalanced Target Class Distribution
Add some maritime target images
Real Single-Frame Image Less
Expanding some images from multi-frame infrared small target datasets
Image Sizes Inconsistent
Unified image resolution to 1280×1024 by filling and bilinear
Labeling Errors
Re-label using the combination of the Canny function and human manual method

Fig. 3. Problems in existing datasets and corresponding solutions.

III. SHR-IRST Dataset

DL networks belong to data-driven methods. Compared with traditional model-driven methods, DL networks require a large amount of diverse data to train to improve their generalization ability. In addition, with the development of infrared detection and imaging equipment, infrared detection equipment with high imaging resolution is gradually applied in various fields and the demand for high-resolution infrared small target data increases.

Based on the above reasons, we analyze, organize, and expand the existing datasets to construct an SHR-IRST dataset with a 1280 × 1024 size. Section III-A introduces the problems existing in the existing dataset and the corresponding solutions. Section III-B describes the statistical properties of the SHR-IRST dataset.

A. Problems and Solutions for Existing Datasets

Because the data in different datasets have different sources, are manually labeled by different people, and have different resolutions. Therefore, the data collation and summary work are not simple "additions." As shown in Fig. 3, first analyze the problems in the existing single-frame infrared small target dataset, and then propose corresponding solutions.

High Image Similarity: The synthetic dataset IRST640 has high image similarity and poor data diversity. Networks trained with IRST640 have poor generalization ability [31]. Structural similarity (SSIM) measures the similarity between two images, and the closer the value is to 1, the more similar the images are. So, it is used to evaluate the similarity of images in the IRST640 dataset, where images with SSIM value above 0.9 (for simple background) or 0.85 (for complex background) are considered highly similar and removed. Among them, simple images refer to images with only clouds and high-voltage power lines in the background, whereas complex images refer to images containing urban buildings and trees. Finally, 358 images are remained from 1024 images. The average SSIM of images is reduced from 0.9387 to 0.8285.

Fig. 4. Example of expanded infrared image of maritime target class.

TABLE III
Comparison of nIoU and IoU Metrics for Crop and Fill Unified Image Size Methods

Method	nIoU	IoU
crop	0.7135	0.6810
fill	**0.7727**	**0.7582**

Unbalanced Target Class Distribution: The targets in infrared small target images mainly include three categories: aerial targets, maritime targets, and land targets. Existing datasets focus more on aerial targets. There are 997 images of aerial targets (65.8%), 416 images of land targets (27.5%), and 102 images of maritime targets (6.7%) in the existing real single-frame infrared small target datasets. It can be seen that the distribution of small target categories in the existing datasets is extremely unbalanced, that is, the number of maritime target images is seriously insufficient. Therefore, some maritime target images are collected from another public dataset. A total of 204 infrared images of ship small targets with river or ocean backgrounds are expanded. As shown in Fig. 4, both single-ship images and multiship images are included.

Real Single-Frame Image Less: There is currently limited data on real single-frame infrared small targets, whereas images from real multiframe dataset can be used to expand single-frame data. To ensure data diversity, every 50 images in IRDST are extracted, and SSIM is used to retain only less similar images (SSIM values less than 0.85). Finally, 806 images were selected from IRDST, accounting for 0.56% of IRDST.

Image Sizes Inconsistent: In general, DL networks have fixed data sizes for input and output. In [4], the size of the infrared small target images is unified through direct resizing. As shown in Fig. 5(b) and (c), direct resizing will deform and distort the targets in the image, and the labels are no longer binarized. Using the SIRST dataset as a benchmark, we compare two undistorted resizing methods crop and fill. *Crop:* according to the specified aspect ratio to maximize crop image, and ensure that small targets remain intact [see Fig. 5(d)]. *Fill:* fill with 0 at the bottom or right-hand side of the image to the specified aspect ratio, then use bilinear to resize image to the specified size [see Fig. 5(e)]. As given in Table III, the fill method obtains better IoU and nIoU metrics. Therefore, fill is used to unify the images sizes in SHR-IRST.

Labeling Errors: The small target in the infrared image is weak and blurry, without clear edges. There is a deviation in the edge definition of infrared small targets during human manual annotation, which leads to label errors. Here, a semiautomatic labeling method is proposed: using the Canny edge detection

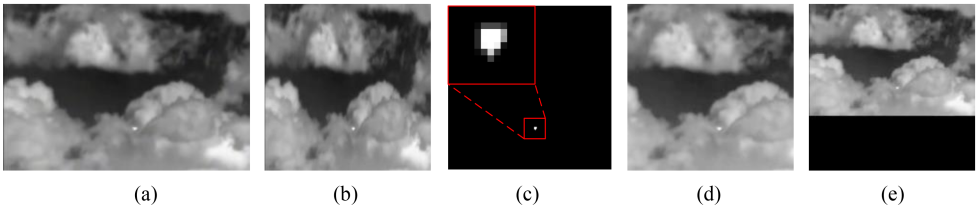


Fig. 5. Example of uniform image size method: (a) original image, (b) resized image, (c) resized label, (d) cropped image, and (e) filled image.

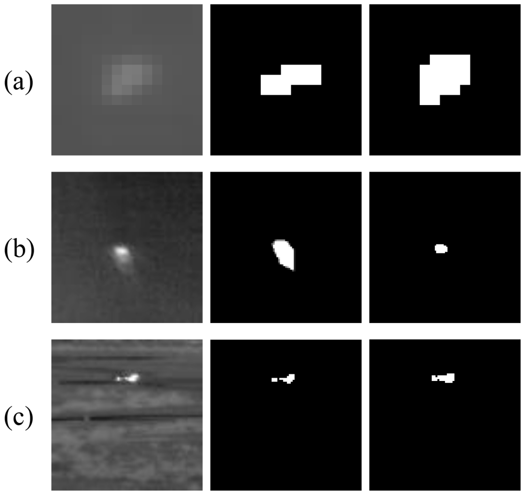


Fig. 6. Comparison of part of the relabeled images. From left to right, the real image, the original label, and the relabel. (a) Point target: Mislabeling due to poorly defined boundary. (b) Boat target: Mislabeling of the boat as a target along with the boat's reflection on the water. (c) Car target: In nonclear weather, the car's source of heat radiation is the engine and chassis. The object semantics corresponding to the target is ignored when labeling, and it is mislabeled as two independent small targets.

algorithm to determine the edge information of small targets, and manually assisting in confirming the label of small targets. This method reduces human participation in defining small target edges. However, real objects will be fully labeled in infrared images, such as drones, planes, cars, missiles, and ships, see Fig. 6. Small targets in infrared images have high local contrast [1]. Based on this, we propose an indicator to evaluate the accuracy of infrared small target labels: local pixel difference (LPD). As shown below:

$$\text{LPD} = \mu_t - \mu_b \tag{13}$$

where μ_t and μ_b represent the mean value of grayscale in the target area and background area around the target, respectively. As shown in Fig. 7, the area around the target is defined as the smallest rectangle that can wrap both labels. LPD measures the numerical representation of the difference in grayscale between a small target and its surrounding background. The larger the LPD, the more accurate the label. This is because a more accurate label should reflect the grayscale difference between the target and the surrounding background to the greatest extent possible. It is important to note that LPD is an unnormalized value, so comparisons should be made by target. Due to the difference in local contrast between different targets, if the μ_t and μ_b values of all targets are added together, the contribution of high-contrast targets to the result will be amplified.

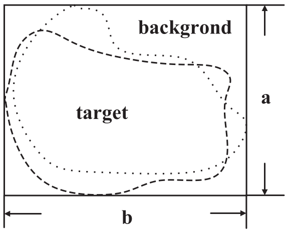


Fig. 7. Small target and its surrounding area background division method.

TABLE IV
COMPARISON OF LPD METRIC FOR SIRST AND SIRST-RE

LPD	Increase	Constant	Decrease
SIRST	79	112	306
SIRST-RE	306	112	79

We relabeled the SIRST dataset using the method combined with the Canny function to obtain relabeled SIRST (SIRST-RE) dataset. The label accuracy of SIRST and SIRST-RE are evaluated using LPD (see Table IV). There are 79 targets that decrease in LPD values because some small targets are labeled according to the target's semantics, rather than just labeling local bright blocks (see Fig. 6). The comparison results show that more labels in SIRST-RE have higher LPD. This indicates that the proposed labeling method has higher labeling accuracy.

B. Statistical Properties of SHR-IRST Dataset

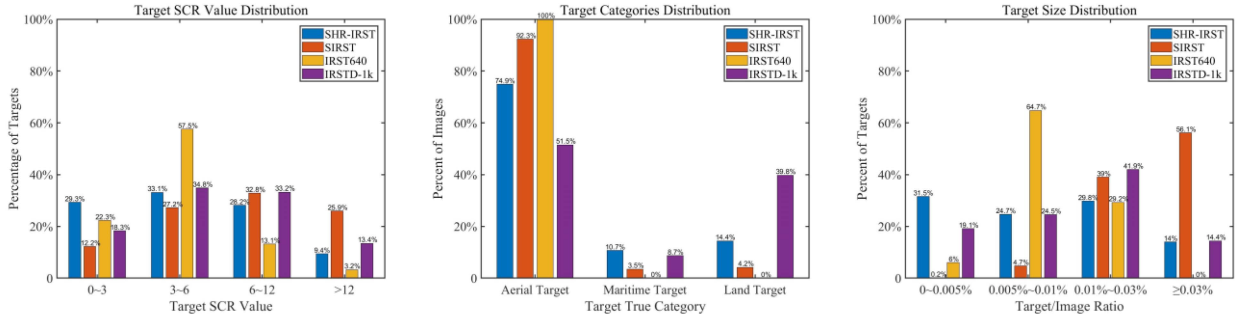
Table V presents the basic information of SHR-IRST and other mainstream single-frame infrared small target datasets. Fig. 8 shows the statistical analysis of the above datasets in terms of target signal-to-clutter ratio (SCR), target category, and target size. From Table V and Fig. 8, it can be seen that the SHR-IRST dataset has the following characteristics.

Larger Target Size: As can be seen from Table V, the average pixel number of the targets in SHR-IRST is 209.47, indicating that the targets in high-resolution images are usually larger and occupy more pixels. For larger targets, the pixel-level metrics results do not reflect the method's target prediction capability well.

More Difficult to Detect: The SCR [4] is the normalized value of the difference between the grayscale of the target and the surrounding background area, which can be used to describe the difficulty of detecting small targets. A higher SCR means that the

TABLE V
DETAIL OF THE SHR-IRST DATASET AND OTHER MAINSTREAM DATASETS

Dataset	SHR-IRST	SIRST[10]	IRSTD-1K[30]	IRST640[31]
Data Type	Real/Synthetic	Real	Real	Synthetic
Image Size	1280×1024	96×135 to 388×418	512×512	640×512
Image Num/ Target Num	2843/4020	427/533	1001/1495	1024/1662
Target Pixel Range (Average)	4–1788 (209.47)	4–330 (32.86)	1–1065 (50.11)	1–51 (27.73)
Target Size Range	2×2–32×112	2×2–14×34	1×1–56×53	1×1–9×8
(Average)	(13.16×17.50)	(5.62×6.94)	(7.69×8.74)	(6.11×6.33)
Target SCR Range (Average)	0.024–45.74 (5.63)	0.17–42.81 (9.20)	0.004–68.76 (7.13)	0.04–70.34 (4.93)



(a) (b) (c)

Fig. 8. Statistical properties of SHR-IRST and other mainstream datasets. (a) Statistical properties of target SCR value distribution. (b) Statistical properties of target class distribution. (c) Statistical properties of target size distribution.

target is easier to detect, and a lower SCR means that it is harder to detect. As can be seen from Table V and Fig. 8(a), the average SCR of SHR-IRST is smaller than that of other real single-frame datasets, indicating that target detection in SHR-IRST is more difficult and challenging.

More Balanced Category Distribution: Fig. 8(b) shows the statistical diagram of real objects category corresponding to targets in images. Data categories in SHR-IRST are more evenly distributed. We expanded images of maritime targets from 87 (8.7%) in IRSTD-1k to 304 (10.7%) in SHR-IRST by nearly three times. This greatly alleviates the problem of sample scarcity in the maritime targets category.

Smaller Target Percentage: Fig. 8(c) shows that more targets in SHR-IRST account for less than 0.005% of the entire image. It indicates that targets in the SHR-IRST dataset are more likely to be submerged in high-level features and have greater detection difficulty.

Larger target size gap: From Table V, it can be seen that the targets in SHR-IRST (1784) have a larger size gap than other datasets (50, 326, and 1064). This means that small target detection in high-resolution infrared images is more challenging than that in low-resolution images.

The IoU metric result of state-of-the-art (SOTA) methods (DNANet and ALCNet) on SHR-IRST is compared with SIRST. As given in Table VI, both DNANet and ALCNet showed a large degree of performance degradation on SHR-IRST. This shows that SHR-IRST presents a new challenge to existing methods. It is necessary to design network according to the characteristics of high-resolution infrared small target images to improve the detection ability of the network.

TABLE VI
COMPARISON OF IOU METRIC FOR SIRSTH AND SHR-IRST DATASETS BY SOTA METHODS

Method	SIRST	SHR-IRST
DNANet	0.7570	0.6168
ALCNet	0.7757	0.5846

IV. METHODOLOGY

This section provides a detailed introduction to the specific structure of the proposed small target detection network ISPANet for high-resolution infrared images and a loss function F-SIoU loss that makes the network focus more on hard positive examples.

A. *ISPANet*

Our proposed ISPANet consists of four modules (see Fig. 9): a local feature extraction (LFE) module, a multiscale receptive field self-attention (MFSA) module, a multilevel feature fusion (MFF) module, and a full conv head with bias (FCB Head).

Input a high-resolution (1280 × 1024) infrared image. First, the input image undergoes multilayer convolution operations in the LFE module to obtain feature maps with different downsampling multiples. Second, in the MFSA module, further feature extraction is carried out in larger receptive fields at different scales. With the help of the self-attention mechanism, networks can explore the connections between features at farther distances. Then, local and global feature maps, and shallow and deep feature maps are gradually upsampled and fused in the MFF

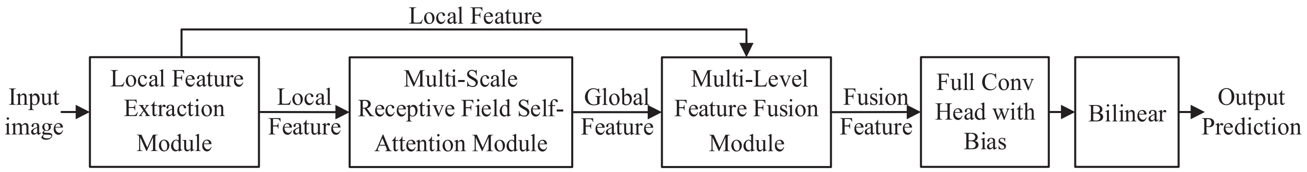


Fig. 9. Overall architecture of our ISPANet.

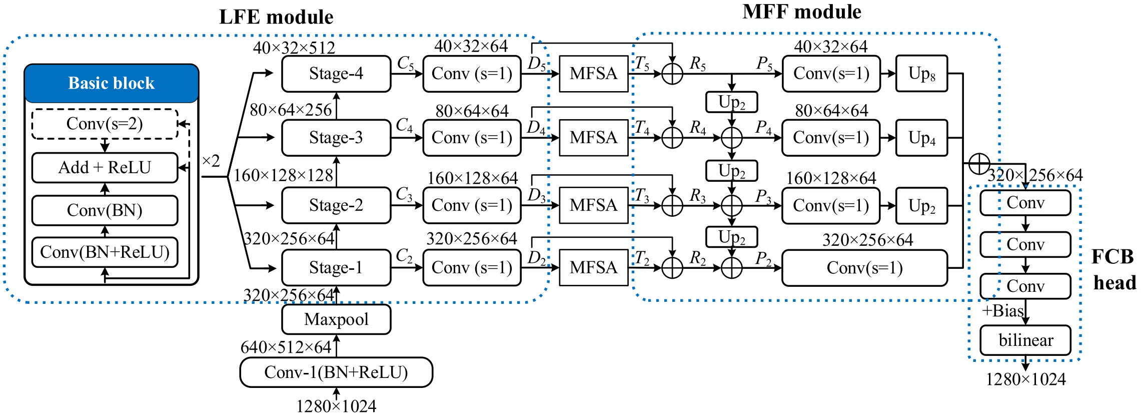


Fig. 10. Detailed network structure of ISPANet.

module. Finally, in the FCB Head, the number of channels in the fused feature map is gradually reduced to obtain a prediction map. The prediction map of the network has a step size of 4 relative to the input image. Use bilinear functions to interpolate the predicted map, and then perform threshold (0.5) judgment to obtain the prediction results for binary classification of each pixel in the input image. The following content introduces the specific structure of each module in the network.

1) LFE Module: The LFE) module consists of multiple convolutional layers and its structure is shown in Fig. 10. In the LFE module, feature extraction is carried out on the input image through a bottom–up approach to obtain feature maps ($\boldsymbol{C_2}$, $\boldsymbol{C_3}$, $\boldsymbol{C_4}$, and $\boldsymbol{C_5}$) of different scales. The step size between adjacent scale feature maps is 2. Then, the channel numbers of feature maps ($\boldsymbol{C_2}$, $\boldsymbol{C_3}$, $\boldsymbol{C_4}$, and $\boldsymbol{C_5}$) are unified to c ($c = 64$) through a 1 × 1 convolution layer, and the feature maps $\boldsymbol{D_2}$, $\boldsymbol{D_3}$, $\boldsymbol{D_4}$, and $\boldsymbol{D_5}$ are obtained. The input image has a high resolution, taking into account the network computing cost and the characteristics of small targets, the LFE module conducts a total of five downsamplings. The step size of the deepest feature map relative to the input image is 32.

2) MFSA Module: Unlike the local receptive field of convolution, the self-attention mechanism has a global receptive field and can better capture the internal correlations between features [17]. However, the computational complexity of the standard self-attention mechanism is proportional to the square of the number of input feature sequences [22]. Despite multiple times downsampling, the feature maps of high-resolution images still contain a large number of features, which will bring significant computational overhead when using the standard self-attention mechanism. Based on the characteristics of small targets in high-resolution infrared images, we design an MFSA module. In order to reduce the computational overhead in the self-attention mechanism, there are two specific network architecture designs in MFSA module. One is multiscale receptive field patch partitioning, and the other is feature scaling in spatial dimensions. Here is a specific introduction.

Small targets in infrared images have locality and sparsity, which means that features that are far from the target have less correlation with the target and are redundant features. Based on this, a multiscale receptive field patch partitioning structure is proposed in MFSA. According to the downsampling multiple of feature map $\boldsymbol{D_i}$ ($i = 2, 3, 4, 5$), different partition quantities are designed to make the subfeature map $\boldsymbol{D_{ij}}$ have different receptive fields. Specifically, the number j of subfeature maps divided by feature map $\boldsymbol{D_i}$ is shown below:

$$j = (6 - i)^2 (i = 2, 3, 4, 5). \tag{14}$$

For the deepest feature map $\boldsymbol{D_5}$, which contains more semantic information, the internal correlation between semantic features is calculated in the global receptive field. From $\boldsymbol{D_4}$ to $\boldsymbol{D_2}$, the downsampling factor decreases, the detailed and texture information in the features increases, and the correlation between long-distance features and targets weakens. Therefore, the number of subfeature maps divided increases, and the receptive field of each subfeature map decreases, allowing for a more focused exploration of the relationship between the target and surrounding features. Calculating self-attention between features at different scales of receptive fields enables better integration of the characteristics of feature maps at different scales with self-attention mechanisms, and is also more in line with the target characteristics of infrared small target images.

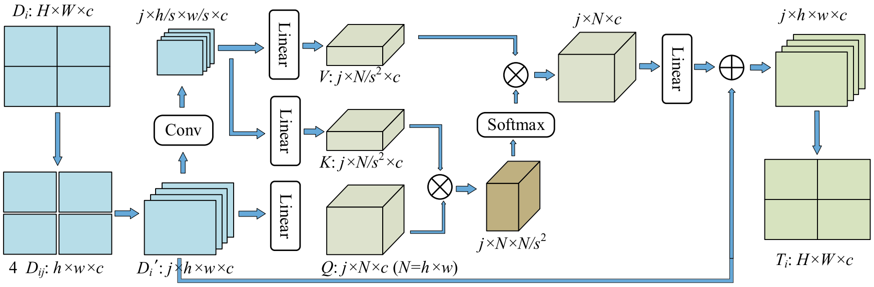


Fig. 11. Structure of MFSA module.

It is worth noting that before the self-attention calculation process, all subfeature maps (from $\boldsymbol{D_{i1}}$ to $\boldsymbol{D_{ij}}$) will be concatenated together in a new dimension. For $\boldsymbol{D_{ij}} \in \mathbb{R}^{h\times w\times c}$, obtain $\boldsymbol{D'}_i \in \mathbb{R}^{j\times h\times w\times c}$ through concat operation. This means that during the self-attention calculation process, subfeature maps $\boldsymbol{D_{ij}}$ divided by $\boldsymbol{D_i}$ share parameters with each other.

MFSA has added spatial scaling operations on the basis of the standard multihead self-attention mechanism (see Fig. 11). The process of self-attention calculation is as follows:

$$\text{Attention}(QKV) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{15}$$

where *QKV* is obtained through three sets of linear mappings of input feature sequences. However, through spatial scaling operations, the length of the input feature sequence can be scaled to $1/s^2$ (*s* is the scaling factor). After spatial scaling, *K* and *V* are generated by two sets of linear projections of the scaled feature sequence, while *Q* is generated by linear projection of the nonscaled feature sequence, as shown below:

$$\begin{aligned} Q &= \text{Linear}(\boldsymbol{D'}_i) \\ K &= \text{Linear}(\text{Conv}_{s\times s}(\boldsymbol{D'_i}))\ i = 2,3,4,5 \\ V &= \text{Linear}(\text{Conv}_{s\times s}(\boldsymbol{D'_i})) \end{aligned} \tag{16}$$

where $\text{Conv}_{s\,\times\,s}$ represents a convolution with kernel size *s* and step size *s*. After passing through the MFSA module, obtain the feature maps $\boldsymbol{T_2}$, $\boldsymbol{T_3}$, $\boldsymbol{T_4}$, and $\boldsymbol{T_5}$.

3) MFF Module: In order to better utilize the detailed information in the shallow features and the semantic information in the deep features, an MFF module is designed in the last part of the ISPANet, whose structure is shown in Fig. 10. The MFF module includes local and global feature fusion and the fusion between different scale features.

In the MFF module, the local and global features generated in the first two modules are first fused by element-by-element summation, and the feature maps generated after fusion are called $\boldsymbol{R_i}$

$$\boldsymbol{Ri} = \boldsymbol{Di} \oplus \boldsymbol{Ti}(i = 2,3,4,5) \tag{17}$$

where ⊕ represents adding elements by elements. Then, through the approach from deep layer to shallow layer, the deep features with less detailed information but rich semantic information are upsampled first. Starting from the deepest feature map $\boldsymbol{R_5}$, the spatial resolution of the feature map $\boldsymbol{R_5}$ is increased by two times by deconvolution operation without changing the number of channels. It is fused with the shallow feature map $\boldsymbol{R_4}$, which has the same space size, by adding element by element. A 3 × 3 convolution is added to the fused features to generate the final fused feature map $\boldsymbol{P_4}$. The 3 × 3 convolution layer can reduce the aliasing effect of the upsampling operation [42]. The process of generating fused feature map $\boldsymbol{P_i}$ is shown as:

$$\boldsymbol{Pi} = \begin{cases} \text{Conv}_{3\times3}(\delta(\mathcal{B}(\text{Up}_2(\boldsymbol{R_{i+1}}) \oplus \boldsymbol{R_i})))\ i = 2,3,4 \\ \text{Conv}_{3\times3}(\delta(\mathcal{B}(\text{Conv}_{3\times3}(\boldsymbol{R_i}))))\ i = 5 \end{cases} \tag{18}$$

where $\mathcal{B}$ denotes BN regularization, δ denotes ReLU, and Up_2 denotes twofold upsampling. For the fused feature map $\boldsymbol{P_i}$, the number of channels is reduced to *c*/2 by a 3 × 3 convolution. Then, the feature map $\boldsymbol{S_i}$ is obtained by upsampling the feature map to a step size of 4 relative to the input image after a layer of deconvolution. All feature maps $\boldsymbol{S_i}$ are added and fused to obtain $\boldsymbol{F}$

$$\begin{aligned} \boldsymbol{F} = &(\text{Conv}_{3\times3}(\boldsymbol{P_2})) \oplus (\text{Up}_2(\text{Conv}_{3\times3}(\boldsymbol{P_3}))) \\ &\oplus (\text{Up}_4(\text{Conv}_{3\times3}(\boldsymbol{P_4}))) \oplus (\text{Up}_8(\text{Conv}_{3\times3}(\boldsymbol{P_5}))) \end{aligned} \tag{19}$$

where Up_4 denotes fourfold upsampling and Up_8 denotes eightfold upsampling.

4) Full Conv Head With Bias: The prediction head of the network is an FCB Head, which includes three layers of convolution, gradually reducing the number of channels in the feature map $\boldsymbol{F}$ to 1 to obtain the prediction map. Then, bilinear interpolation is used to sample the predicted map to the input image size

$$\begin{aligned} X_{pre} = &\text{Bilinear}(w_1(\text{Dropout}(\delta(B(\\ &\text{Conv}_{3\times3}(\delta(\mathcal{B}(\text{Conv}_{3\times3}(\boldsymbol{F}))))))) + b) \end{aligned} \tag{20}$$

where Dropout means a Dropout layer is added after the first two $\text{Conv}_{3\times3}$ layers (dropout rate is 0.1). w_1 represents a convolution kernel of size 1, and *b* represents the initialization bias of the last convolutional layer of the network

$$b = -\log\left(\frac{1-\pi}{\pi}\right) \tag{21}$$

where $\pi = 0.01$. Because the FCB head has an initialization bias *b*, the output predictions of the network during initial training period are around π. At this point, the output predictions of

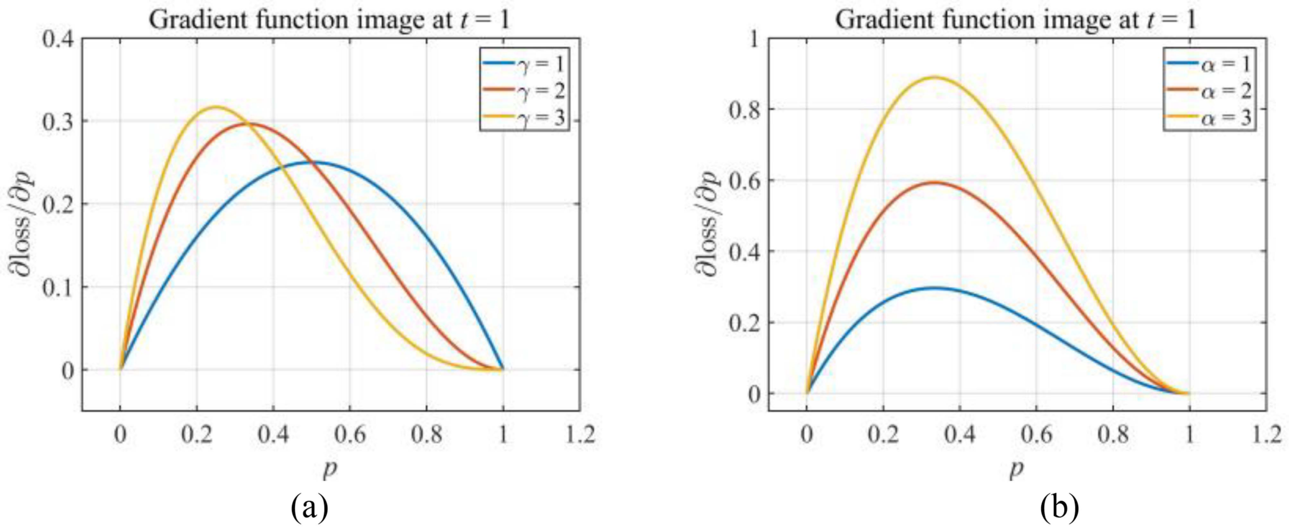


Fig. 12. Gradient function curves of F-SIoU loss in different settings. (The value of gradient function is taken to absolute in (a) and (b) for ease of presentation).

the background area (negative samples) are close to the true value 0, and the network only needs to focus on mining positive samples (target area). This can accelerate the convergence of losses during network training.

B. F-SIoU Loss

The problem with slow convergence of soft loss is that it provides gradient values close to 0 for negative examples and too small for hard positive examples (analyzed in Section II-B). The addition of bias b in the FCB head solves the problem of insufficient supervision of negative examples. In this section, F-SIoU loss is proposed, which can make the network pay more attention to hard positive examples

$$\mathcal{L}_{\mathrm{F-SIoU}} = \alpha(\mathcal{L}_{\mathrm{Soft-IoU}})^{\gamma} \tag{22}$$

where α is the adjustment factor for the peak of the gradient function, and γ is the adjustment factor for the preference of difficult sample. Solving the gradient function for the F-SIoU loss by the chain rule for gradients leads to

$$\frac{\partial \mathcal{L}_{\mathrm{F-SIoU}}}{\partial x} = \frac{\partial \mathcal{L}_{\mathrm{F-SIoU}}}{\partial p} \cdot \frac{\partial p}{\partial x} = \begin{cases} 0, & t = 0 \\ -\alpha \cdot \gamma p(1-p)^{\gamma}, & t = 1. \end{cases} \tag{23}$$

Fig. 12(a) shows the gradient function curve of the F-SIoU loss with respect to the predicted probability p shown for different values of γ when $\alpha = 1$. Fig. 12(b) shows the gradient function curve of the F-SIoU loss with respect to the predicted probability p shown for different values of α when $\gamma = 2$.

As can be seen from Fig. 12(a), hard positive examples will obtain larger gradient values when γ >1. With the increase of γ value, the weight difference between easy and hard positive examples in the loss expands further. From Fig. 12(b), it can be seen that the value of the gradient function also increases proportionally when α >1, which makes the hard positive case contribute more to the loss and accelerates the convergence of network. At the same time, when the prediction probability p is close to 0, the F-SIoU loss can provide a larger gradient value and alleviate the gradient saturation phenomenon.

V. Experiments

This section introduces experimental details and result analysis. Specifically speaking, Section V-A details the specific implementation of experiments. Section V-B presents the proposed new target-level evaluation metrics. Section V-C compares the quantitative and visual results of ISPANet with those of SOTA methods. Section V-D compares the experimental results on actual data of different methods. Section V-E compares the impact of different parameter settings for F-SIoU loss on network training results. Section V-F outlines the ablation experiments.

A. Implementation Details

During network training, adaptive gradient [43] method is used as optimizer, and Xavier method [44] is used to initialize weights and biases of network. Learning rate decay is performed using both warm-up [45] and cosine annealing [46] strategies. The total number of train iterations of ISPANet is 200 epochs, and warm-up is performed within the first five epochs. The initial and minimum learning rate is set as 0.01 and 1e-5, respectively. All networks are implemented in PyTorch [47]. The CPU of the computer that is used for training and testing is AMD Ryzen9 5950X 16-Core Processor 3.40 GHz, and the GPU is Nvidia RTX 3090. Unless otherwise noted, all the experiments are trained and tested on the SHR-IRST dataset, and the number of images in the training, validation, and testing sets is 1706, 569, and 568, respectively. The ratio of the three divisions is about 6:2:2.

The following data augmentation strategies are implemented during network training: random scaling. The scaled aspect ratio ranges from 0.58 to 1.85. The scaled length of the long side ranges from 0.5 to 2 times the original length. After random scaling, image is arbitrarily cropped to 1280 × 1024. In addition to this, the data are also enhanced by random horizontal flipping and random superimposed Gaussian noise strategies.

TABLE VII
COMPARISON OF PIXEL-LEVEL AND TARGET-LEVEL METRICS FOR ISPANET AND SOTA METHODS ON THE SHR-IRST DATASET (VALUES ARE EXPRESSED AS PERCENTAGES)

Method	P_d(↑)	M_d(↓)	P_a(↑)	F_a(↓)	F1-T(↑)	nIoU(↑)	IoU(↑)	R(↑)	P(↑)	F1-P(↑)	fps	flops(G)	params(M)
ACM	70.11	29.89	69.50	30.50	69.81	62.63	67.59	83.68	77.86	80.66	55.7	8.10	0.40
ALCNet	71.00	29.00	46.11	53.89	55.91	61.44	58.46	67.22	81.76	73.78	86.5	75.76	0.38
DNANet	73.90	26.10	66.21	33.79	69.85	62.25	61.68	68.65	**85.87**	76.30	17.2	285.64	4.70
AGPCNet	72.00	27.99	63.37	36.63	67.41	63.01	63.31	76.50	78.60	77.53	10.1	863.83	12.36
RDIAN	71.50	28.50	46.40	53.60	56.28	57.59	60.78	**84.44**	68.44	75.61	62.6	74.37	0.22
MTU-Net	76.04	23.96	69.23	30.77	72.48	66.31	67.96	78.73	83.24	80.93	40.9	126.98	8.75
ISPANet	**78.44**	**21.56**	**73.70**	**26.30**	**75.99**	**68.14**	**71.51**	81.63	85.23	**83.39**	37.3	114.15	13.62

Bold indicates best, underline indicates second best.

TABLE VIII
COMPARISON OF PIXEL-LEVEL AND TARGET-LEVEL METRICS FOR ISPANET AND SOTA METHODS ON THE SIRST DATASET (VALUES ARE EXPRESSED AS PERCENTAGES)

Method	P_d(↑)	M_d(↓)	P_a(↑)	F_a(↓)	F1-T(↑)	nIoU(↑)	IoU(↑)	R(↑)	P(↑)	F1-P(↑)
ACM	67.57	32.43	69.93	30.07	68.73	72.21	70.38	82.78	82.46	82.62
ALCNet	66.22	33.78	52.41	47.59	58.51	66.70	65.30	73.04	86.04	79.01
DNANet	68.24	31.76	65.16	34.84	66.67	68.00	67.06	73.05	**89.10**	80.28
AGPCNet	66.89	33.11	61.49	38.51	64.08	73.50	71.67	**86.73**	83.02	**84.84**
RDIAN	57.43	42.57	46.45	53.55	51.36	66.15	64.68	85.75	72.47	78.55
MTU-Net	66.22	33.78	67.83	32.17	67.01	73.61	72.14	85.66	82.04	83.81
ISPANet	**71.62**	**28.38**	**70.20**	**29.80**	**70.90**	**73.67**	**72.49**	83.43	84.68	84.05

Bold indicates best, underline indicates second best.

B. *Target-Level Evaluation Metrics*

Infrared small target detection networks [11] typically use both pixel-level and target-level metrics to comprehensively evaluate the target detection performance.

Previously, the false-alarm rate (F_a) of the target-level metrics [see (5)] represented the ratio of the wrongly predicted pixels to all pixels of the image, and did not really measure the false-alarm situation of the network from the target level. Therefore, we redefine the F_a as

$$F_a = \frac{\#\text{ num of false detections}}{\#\text{ num of predicted targets}}. \tag{24}$$

In addition, we also define the M_d (miss detection rate), the P_a (probability of accuracy), and the F1-T (F1 score) of the target-level evaluation metrics

$$M_d = \frac{\#\text{ num of not detections}}{\#\text{ num of actual targets}} \tag{25}$$

$$P_a = \frac{\#\text{ num of true detections}}{\#\text{ num of predicted targets}} \tag{26}$$

$$\mathrm{F1-T} = \frac{2P_d \cdot P_a}{P_d + P_a} \tag{27}$$

where P_d represents the probability of detection [see (4)]. When the center of the predicted target is within d pixels of the center of the real target, the target is considered to be correctly detected. The threshold d is set to 2 in this article. The *F1-T* takes both recall and accuracy into account to measure target-level prediction. A method is considered good when high values are obtained on P_d (↑), P_a (↑), and *F1-T* (↑), and low values are obtained on M_d (↓), and F_a (↓).

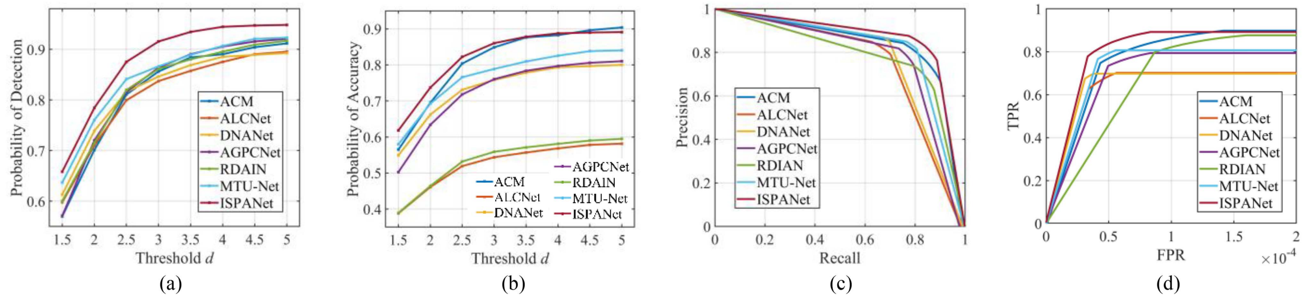
The pixel-level metrics we used include: IoU (↑), nIoU (↑), R (↑), P (↑), F1-P (↑), ROC, and PR curve. They have been explained in detail in Section II-B.

C. *Comparison With SOTA*

In this section, we compare the infrared small target detection performance of our proposed ISPANet and SOTA methods (ACM [28], ALCNet [10], DNANet [11], AGPCNet [27], RDIAN [36], and MTU-Net [32]) on the SHR-IRST dataset (see Table VII) and a mainstream SIRST dataset [10] (see Table VIII).

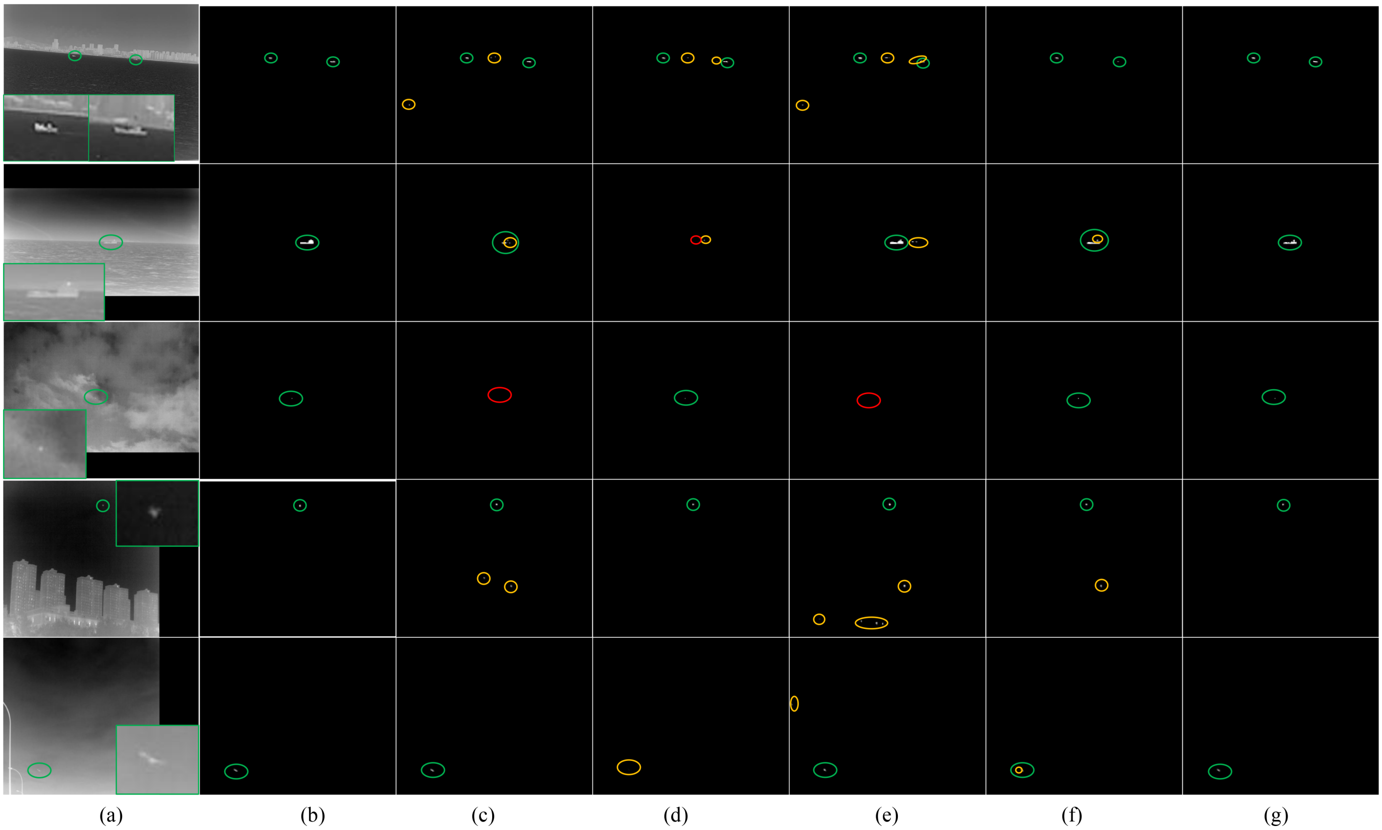
As can be seen from Table VII, ISPANet achieves the best results on almost all metrics. Compared with the suboptimal method MTU-Net, ISPANet improves by approximately 3% on all metrics. Comparing the results of Tables VII and VIII, all methods achieve better pixel-level metrics and poorer target-level metrics on the SIRST dataset. This indicates that for infrared small target detection problems, good pixel-level metrics cannot accurately reflect the target detection ability of a method. It also indicates that although a more complex and larger dataset presents new challenges for fine segmentation tasks, it can enable the model to better understand the semantics of infrared small targets. Target-level metrics are also an evaluation of whether the model accurately understands the semantic concepts of infrared small targets.

Fig. 13(a) and (b) shows the variation curves of target-level metrics P_d and P_a for all methods under different threshold d (d represents the pixel distance between GT's center and predicted target's center). When $d \geq 1.5$, ISPANet has always had a significant advantage in P_d metric. When $d \leq 4$, ISPANet has always had an advantage in P_a metric. Compared with



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

Fig. 13. Target-level P_d and P_a metrics curves for different threshold d and PR, ROC curve. (a) P_d curve. (b) P_a curve. (c) PR curve. (d) ROC curve.



(a) (b) (c) (d) (e) (f) (g)

Fig. 14. Visualization of test images for SHR-IRST dataset. The green, red, and yellow round boxes represent correctly detected target, missed target, and false alarm target, respectively. Local zoomed-in views of the corresponding regions of the real small targets are shown in box in the upper left and right corners of the images. (a) Input. (b) GT. (c) ALCNet. (d) DNANet. (e) RDAIN. (f) MTU-Net. (g) ISPANet.

suboptimal networks, the P_d and P_a metrics of ISPANet decay less with decreasing threshold d. This indicates that ISPANet has strong robustness in improving target detection recall and suppressing false alarms. Fig. 13(c) and (d) shows the PR and ROC curves for all methods. The PR curve of ISPANet is closer to the upper right-hand side corner, while its ROC curve is closer to the upper left-hand side corner, which indicates that ISPANet has superior performance.

Several images are randomly selected from the test dataset, and the visual prediction results of different methods are shown in Fig. 14. From Fig. 14, it can be seen that when there are sea surface ripples and urban building interference (Fig. 14, lines 1, 2, and 4), other methods have weak antiinterference ability, and false alarms occur. When the target being tested is a weak small target in a complex cloud background (Fig. 14, line 3), some methods may miss detection. When the grayscale distribution of the small target to be tested is uneven (Fig. 14, line 5), some methods predict it as two small targets.

Table VII also gives the image processing speed [frames per second (fps)], computational complexity [flops(G)], and model parameter count [params(M)] of different methods. Although ISPANet's model processing speed is not the fastest, it has achieved significant improvement in network performance.

From Table VII, it can also be seen that ALCNet and RDIAN achieve pixel-level metrics that are close to those of other methods, but when evaluating their target-level metrics, it is

TABLE IX
COMPARISON OF PIXEL-LEVEL AND TARGET-LEVEL METRICS FOR ISPANET AND SOTA METHODS IN THE SHAC-IRST DATASET (VALUES ARE EXPRESSED AS PERCENTAGES)

Method	P_d (↑)	M_d (↓)	P_a (↑)	F_a (↓)	F1-T (↑)	nIoU(↑)	IoU (↑)	R (↑)	P (↑)	F1-P (↑)	F_a num (↓)
ACM	73.17	26.83	66.67	33.33	69.77	51.71	56.35	67.28	77.62	72.08	9
ALCNet	73.17	26.83	56.60	43.40	63.83	55.69	54.17	63.27	82.30	71.54	12
DNANet	78.05	21.95	63.83	36.17	70.22	52.81	53.51	57.94	**85.63**	69.12	**6**
AGPCNet	73.17	26.83	49.18	50.82	58.82	48.00	52.15	65.38	72.04	68.55	22
RDIAN	80.49	19.51	22.60	77.40	35.29	47.36	50.56	**80.97**	57.38	67.16	62
MTU-Net	82.92	17.07	57.63	42.37	68.00	55.47	55.51	64.90	79.33	71.39	11
ISPANet	**85.36**	**14.63**	**76.09**	**23.91**	**80.46**	**56.81**	**58.22**	69.00	78.84	**73.59**	**6**

Bold indicates best, underline indicates second best.

Fig. 15. Equipment used in real image acquisition. (a) Thermal imaging camera. (b) UAV.

found that they exhibited severe target false alarms. This indicates that a pixel-level perspective alone cannot fully reflect the detection performance of a method for infrared small targets, and it is necessary to comprehensively consider its pixel-level and target-level predictive performance.

D. Actual Data Experiment

This section compares the detection performance of different methods on infrared small targets in the single-frame high-resolution actual collected infrared small target dataset (SHAC-IRST). The images in SHAC-IRST are actually collected by us, and they are from different sources than the images in the SHR-IRST dataset. The correlation between the images in the two datasets is relatively low, which can further evaluate the generalization ability of different methods.

The equipment used for collection includes an infrared detector, InfiRay Arrow Cruiser PH35+ [see Fig. 15(a)], which is an uncooled vanadium oxide detector with a resolution of 640 × 352, and a DJI Mini3 Pro [see Fig. 15(b)] UAV as target. In addition, the targets in the collected images also include birds and rowing. The collection environment includes various scenes from the campus of the Huazhong University of Science and Technology and Ma'anshan Forest Park. The collection angle includes looking up and horizontal observation, and the background of the image includes buildings, trees, clouds, and lakes. The actual collected images are magnified to 1280 × 704 using interpolation functions. There are a total of 80 actual collected images, including 40 infrared images with targets and 40 infrared images without targets. The no-target image contains interference similar to the imaging characteristics of infrared small targets, which are used to test the ability of different methods to suppress target false alarms. The experimental results are given in Table IX. It can be seen that ISPANet has achieved significant advantages in almost all indicators on test images with targets. ISPANet also demonstrated good suppression of false alarms on test images without targets. Fig. 16 shows visual predicted images of some test images. It can be seen from this that ISPANet tends to predict the target as a complete individual. Other methods tend to obtain more accurate pixel prediction results, which makes them more likely to predict a target as multiple small targets, leading to serious false alarms.

TABLE X
COMPARISON OF PIXEL-LEVEL METRICS FOR ISPANET TRAINED WITH DIFFERENT SETTINGS OF F-SIOU LOSS (VALUES ARE EXPRESSED AS PERCENTAGES)

α	γ	nIoU(↑)	IoU(↑)	R(↑)	P(↑)	F1-P(↑)
1	1	67.52	69.46	78.83	83.94	81.30
1	2	67.38	70.43	79.33	**85.40**	82.25
1	3	67.46	70.23	79.63	84.35	81.92
2	2	**68.14**	**71.51**	**81.63**	85.23	**83.39**
2	3	67.61	69.28	80.27	83.55	81.88

Bold indicates best, underline indicates second best.

E. Loss Function Experiment

In this section, the effects of different parameter (α and γ) settings for F-SIoU loss on network prediction performance are compared through experiments.

ISPANet is trained using F-SIoU loss with different settings, and Table X gives these pixel-level metric results. When $\alpha = 1$ and $\gamma = 1$, it represents the Soft-IoU loss. It can be seen that as γ increases, the R (Recall) gradually increases (i.e., more positive examples are predicted), indicating that the γ parameter in F-SIoU is effective in increasing the network's attention to hard positive examples. When α increases, the recall rate of the network for positive examples is further improved, which is consistent with the expected effect when the α parameter is designed. However, it can also be seen in the Table X that for the same γ value, when a larger α is set, the P (Precision) of the network decreases. Therefore, a suitable setting of α and γ parameters is important. In ISPANet, $\alpha = 2$ and $\gamma = 2$.

F. Ablation Study

In this section, the rationality and effectiveness of the network structure of ISPANet are verified through detailed ablation

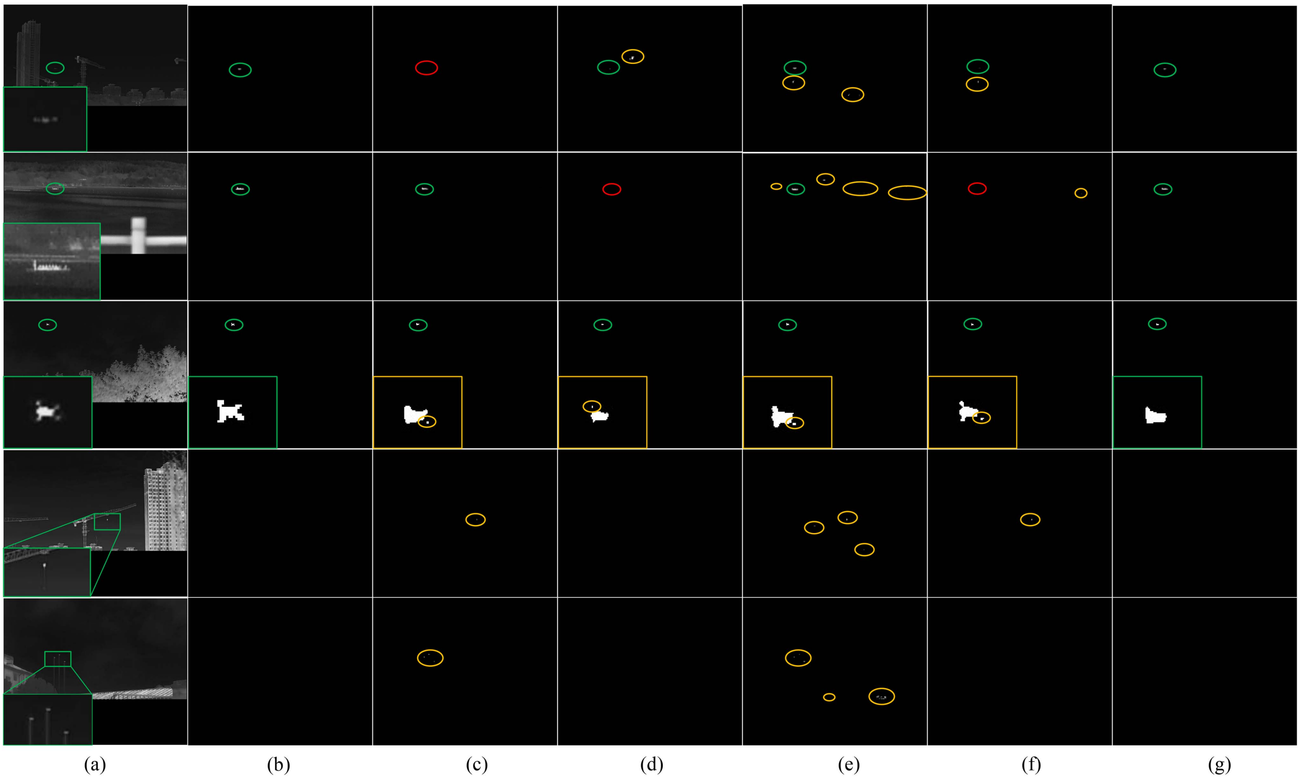


Fig. 16. Visualization of test images for the SHAC-IRST dataset. The top three lines are images with targets and the bottom two are images without targets. The green, red, and yellow round boxes represent correctly detected target, missed target, and false alarm target, respectively. Local zoomed-in views of the corresponding regions of the real small targets are shown in red box in the lower left corner of the images. (a) Input. (b) GT. (c) ALCNet. (d) DNANet. (e) RDAIN (f) MTU-Net. (g) ISPANet.

TABLE XI
NETWORK SETTINGS FOR ABLATION EXPERIMENT

LFE	MFSA	MFF	bias	Module order	Name
√		√	√	-	A
√	√	√	√	MFSA first	B(ISPANet)
√	√	√	√	MFF first	C
√	√	√		-	D

experiments (see Table XI). The network without MFSA module is named A. A is a fully convolutional network that only includes LFE module, MFF module, and FCB Head. Based on A, two networks using self-attention mechanisms for different objects are compared. Among them, the network that calculates self-attention for different layers of feature maps after feature extraction is named B (i.e., self-attention first and then feature fusion). In B, the order of four modules is LFE, MFSA, MFF, and FCB Head. B is the proposed ISPANet. The network that calculates self-attention based on the feature maps after upsampling and fusion is named C (i.e., feature fusion followed by self-attention). In C, the order of four modules is LFE, MFF, MFSA, and FCB Head. Based on A, the impact of initialization bias b set for the last layer of network in FCB Head on the network performance is compared. The network without bias b is named D. In addition, the impact of the number of feature channels c on the performance of A and B networks is also compared through experiments (c is the unified number of feature channels for different scale feature maps in the LFE module). The experimental results are given in Table XII.

From Table XII, it can be seen that as the number of channels c decreases, the network performance decreases. The decrease in target-level metrics (about 1%) is smaller than that in pixel-level metrics (about 3%). It shows that fewer feature channels are sufficient to provide accurate semantic information. As the number of channels increases, the network's ability to learn detailed information, such as shapes and contours, enhances. The impact of c on nIoU is greater than that on IoU. It shows that when the feature channels are few, the network's prediction ability for smaller targets deteriorates. Because nIoU balances the contributions of different sizes of targets, but larger targets contribute more to IoU than smaller ones. The impact of c on P is greater than that on R, indicating that as the feature channel is fewer, the network's learning ability for hard negative examples is worse.

Compare the experimental results of B, C, and A. Compared with A, the metrics of B and C have both improved, while B has more improvement. This indicates that the self-attention mechanism is beneficial for network. Calculating self-attention for downsampling feature maps first, followed by upsampling and feature fusion, can maximize the effectiveness of the self-attention mechanism.

TABLE XII
COMPARISON OF PIXEL-LEVEL AND TARGET-LEVEL METRICS FOR DIFFERENT VARIANTS OF ISPANET (VALUES ARE EXPRESSED AS PERCENTAGES)

Method	c num	P_d(↑)	M_d(↓)	P_a(↑)	F_a(↓)	F1-T(↑)	nIoU(↑)	IoU(↑)	R(↑)	P(↑)	F1-P(↑)
A	128	75.28	24.72	70.73	29.27	72.93	67.22	68.45	79.51	83.76	81.58
A	64	74.65	25.35	70.22	29.78	72.36	65.79	68.16	78.98	82.25	80.58
A	32	74.40	25.60	70.24	29.76	72.26	64.51	67.97	79.28	80.47	79.87
A	16	73.52	26.48	69.23	30.77	71.31	62.35	66.49	79.13	78.65	78.89
B	64	**78.44**	**21.56**	**73.70**	**26.30**	**75.99**	**68.14**	**71.51**	**81.63**	**85.23**	**83.39**
B	32	77.93	22.07	72.88	27.12	75.32	66.38	70.11	80.24	83.26	81.72
B	16	77.18	22.82	72.08	27.92	74.54	65.26	68.89	79.67	81.41	80.53
C	64	75.91	24.09	70.65	29.35	73.19	66.49	69.65	80.07	82.32	81.18
D	64	76.67	23.33	71.52	29.48	74.01	67.22	70.08	79.86	83.36	81.57

Bold indicates best, underline indicates second best.

Compare the experimental results of D and B. Compared with B, the metrics of D have decreased, with the largest decrease in P_a being about 3%. After adding bias b, the learning task of the network is equivalent to finding targets from background, the number of hard negative examples is reduced, and the network only needs to focus on mining positive examples. This improves the predictive ability of the network.

VI. Conclusion

This article analyzes the problems in the existing single-frame infrared small target dataset and proposes corresponding solutions. On this basis, we construct a high-resolution SHR-IRST dataset and demonstrate its superiority (with more balanced data distribution and greater detection difficulty) through statistical analysis. Due to the poor detection performance of existing methods on the SHR-IRST dataset, we propose a new network ISPANet based on the characteristics of high-resolution infrared small targets. We propose an F-SIoU loss to solve the problems of slow convergence and insufficient attention to hard positive examples in Soft-IoU loss.

We train and test different methods on SHR-IRST dataset. Compared with SOTA methods, ISPANet achieves optimal results on almost all metrics. In terms of target-level metrics, ISPANet achieves a balanced performance improvement, which is about 3% higher than suboptimal methods. This result indicates that ISPANet has good target recall and false alarm suppression capabilities, which effectively improves the credibility of the network's target detection results.

At the same time, we also use the actual acquired dataset (SHAC-IRST) to test different methods. Compared with SOTA methods, ISPANet improves pixel-level metrics by about 1%. There is a significant performance improvement in target-level metrics, especially a 10% increase in target-level accuracy (P_a).

Considering the differences in collection equipment and targets between the two different datasets, SHAC-IRST (including actual acquired data) and SHR-IRST (including actual acquired data and simulation data), it indicates that ISPANet has good generalization ability and can achieve good target detection results for different datasets.

ISPANet has achieved significant improvements in the target detection performance of infrared small targets, but when applied to small target detection in large-size infrared images, a target detection method is expected to have faster image processing speed. Therefore, in future research, efforts will be made to build a lightweight and efficient neural network that improves image processing speed while maintaining good small object detection performance.

References

[1] R. Kou et al., "Infrared small target segmentation network: A survey," *Pattern Recognit.*, vol. 143, Nov. 2023, Art. no. 109788, doi: 10.1016/j.patcog.2023.109788.

[2] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[3] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Infrared small-target detection using multiscale gray difference weighted image entropy," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 1, pp. 60–72, Feb. 2016.

[4] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[5] F. Wu, H. Yu, A. Liu, J. Luo, and Z. Peng, "Infrared small target detection using spatiotemporal 4-D tensor train and ring unfolding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5002922, doi: 10.1109/TGRS. 2023.3288024.

[6] Z. Chen, D. Hong, and H. Gao, "Grid network: Feature extraction in anisotropic perspective for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 20, 2023, Art. no. 5507105, doi: 10.1109/LGRS.2023.3297612.

[7] Z. Chen, G. Wu, H. Gao, Y. Ding, D. Hong, and B. Zhang, "Local aggregation and global attention network for hyperspectral image classification with spectral-induced aligned superpixel segmentation," *Expert Syst. Appl.*, vol. 232, Dec. 2023, Art. no. 120828.

[8] Z. Chen et al., "Global to local: A hierarchical detection algorithm for hyperspectral image target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5544915, doi: 10.1109/TGRS.2022.3225902.

[9] Z. Chen et al., "Temporal difference-guided network for hyperspectral image change detection," *Int. J. Remote Sens.*, vol. 44, no. 19, pp. 6033–6059, 2023.

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

[11] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023, doi: 10.1109/TIP.2022.3199107.

[12] C. Kwan and J. Larkin, "Detection of small moving objects in long range infrared videos from a change detection perspective," *Photonics*, vol. 8, no. 9, Sep. 2021, Art. no. 394.

[13] R. Kou, C. Wang, Y. Yu, Z. Peng, F. Huang, and Q. Fu, "Infrared small target tracking algorithm via segmentation network and multi-strategy fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612912, doi: 10.1109/TGRS. 2023.3286836.

[14] C. Kwan and B. Budavari, "Enhancing small moving target detection performance in low-quality and long-range infrared videos using optical flow techniques," *Remote Sens.*, vol. 12, no. 24, Dec. 2020, Art. no. 4024.

[15] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

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

[17] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.

[18] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[19] W. Wang et al., “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 548–558.

[20] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, “Expectation-maximization attention networks for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9166–9175.

[21] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[22] A. Dosovitskiy et al., “An image is worth 16 × 16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–21.

[23] Z. Gao, J. Dai, and C. Xie, “Dim and small target detection based on feature mapping neural networks,” *J. Vis. Commun. Image Representations*, vol. 62, pp. 206–216, Jul. 2019.

[24] M. Zhao, L. Cheng, X. Yang, P. Feng, L. Liu, and N. Wu, “TBC-Net: A real-time detector for infrared small target detection using semantic constraint,” 2019, *arXiv:2001.05852*.

[25] X. Tong et al., “MSAFFNet: A multiscale label-supervised attention feature fusion network for infrared small target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5002616, doi: 10.1109/TGRS.2023. 3279253.

[26] 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, 2023.

[27] 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, Aug. 2023.

[28] 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. 949–958.

[29] W. Wang et al., “CCRANet: A two-stage local attention network for single-frame low-resolution infrared small target detection,” *Remote Sens.*, vol. 15, no. 23, 2023, Art. no. 5539.

[30] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, “ISNet: Shape matters for infrared small target detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 867–876.

[31] G. Chen et al., “IRSTFormer: A hierarchical vision transformer for infrared small target detection,” *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3258.

[32] T. Wu et al., “MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601015, doi: 10.1109/TGRS.2023.3235002.

[33] K. Wang, S. Du, C. Liu, and Z. Cao, “Interior attention-aware network for infrared small target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013, doi: 10.1109/TGRS.2022.3163410.

[34] Y. Chen, L. Li, X. Liu, and X. Su, “A muti-task framework for infrared small target detection and segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5003109, doi: 10.1109/TGRS.2022. 3195740.

[35] Y. Dai, X. Li, F. Zhou, Y. Qian, Y. Chen, and J. Yang, “One-stage cascade refinement networks for infrared small target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5000917, doi: 10.1109/TGRS.2023.3243062.

[36] H. Sun, J. Bai, F. Yang, and X. Bai, “Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5000513, doi: 10.1109/TGRS.2023.3235150.

[37] R. Kou et al., “LW-IRSTNet: Lightweight infrared small target segmentation network and application deployment,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5621313, doi: 10.1109/TGRS. 2023.3314586.

[38] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS One*, vol. 10, Mar. 2015, Art. no. e0118432.

[39] K. A. Islam, M. S. Uddin, C. Kwan, and J. Li, “Flood detection using multi-modal and multi-temporal images: A comparative study,” *Remote Sens.*, vol. 12, no. 15, Jul. 2020, Art. no. 2455.

[40] M. A. Rahman and Y. Wang, “Optimizing intersection-over-union in deep neural networks for image segmentation,” in *Proc. Int. Symp. Vis. Comput.*, 2016, pp. 234–244.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[43] J. C. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 2121–2159, 2011.

[44] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

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

[46] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” 2016, *arXiv:1608.03983*.

[47] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

Wenjing Wang received the B.S. degree in communication engineering from the Wuhan University of Technology, Wuhan, China, in 2020, and the M.S. degree in electronic science and technology from the Huazhong University of Science and Technology, Wuhan, in 2023.

She is currently with the School of Electronic Information, Central South University, Changsha, China. Her research interests include infrared small object detection, image inversion, and deep learning.

Chengwang Xiao received the B.S. degree in communication engineering from the Northwestern Polytechnical University (NWPU), Xi’an, China, in 2014, and the M.S. and Ph.D. degrees in electromagnetic and microwave technology from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2017 and 2023, respectively.

He is currently with the School of Electronic Information, Central South University, Changsha, China. His research interests include microwave remote sensing, signal processing, and deep learning.

Haofeng Dou received the B.S. degree in electrical engineering and the Ph.D. degree in electromagnetic and microwave technology from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014 and 2020, respectively.

He is currently a Senior Engineer with the China Academy of Space Technology (Xi’an), Xi’an, China. His research interests include microwave remote sensing, antenna array, and signal processing.

Ruixiang Liang received the B.S. and M.S. degrees in electrical engineering from the Xidian University, Xi’an, China, in 2010 and 2013, respectively.

She is currently a Senior Engineer with the China Academy of Space Technology (Xi’an), Xi’an. Her research interests include reflector antenna, spot antenna, and antenna array.

Huaibin Yuan received the B.S. degree in composite materials and engineering from the Xi'an University of Aeronautics and Astronautics, Xi'an, China, in 2021.

He is currently a Senior Engineer with the China Academy of Space Technology (Xi'an), Xi'an. His research interests include fixed surface antennas and mechanical movable antennas.

Guanghui Zhao received the B.S. degree from the School of Telecommunications Engineering, Xidian University, Xi'an, China, in 2013. He started his M.S. in 2014 and switched to work for a Ph.D. in 2016. He is currently working toward the Ph.D. degree in electromagnetic and microwave technology from the Huazhong University of Science and Technology (HUST), Wuhan, China.

His research interests include microwave remote sensing and its data processing.

Zhiwei Chen received the B.S. degree in electrical engineering in 2015 from the Huazhong University of Science and Technology, Wuhan, China, where she is currently working toward the Ph.D. degree in electromagnetic and microwave technology.

Her research interests include microwave remote sensing and retrieval of sea surface temperatures from infrared and microwave radiometers.

Yuhang Huang received the B.S. degree in communication engineering from the Taiyuan University of Technology, Taiyuan, China, in 2018. He is currently working toward the Ph.D. degree in electromagnetic and microwave technology with the Huazhong University of Science and Technology, Wuhan, China.

His research interests include microwave remote sensing and signal processing.