Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.045987

ARTICLE

Infrared Small Target Detection Algorithm Based on ISTD-CenterNet

Ning Li[*], Shucai Huang and Daozhi Wei

Air and Missile Defense College, Air Force Engineering University, Xi'an, 710051, China

*Corresponding Author: Ning Li. Email: 18756053341@163.com

Received: 14 September 2023 Accepted: 08 November 2023 Published: 26 December 2023

ABSTRACT

This paper proposes a real-time detection method to improve the Infrared small target detection CenterNet (ISTD-CenterNet) network for detecting small infrared targets in complex environments. The method eliminates the need for an anchor frame, addressing the issues of low accuracy and slow speed. HRNet is used as the framework for feature extraction, and an ECBAM attention module is added to each stage branch for intelligent identification of the positions of small targets and significant objects. A scale enhancement module is also added to obtain a high-level semantic representation and fine-resolution prediction map for the entire infrared image. Besides, an improved sensory field enhancement module is designed to leverage semantic information in low-resolution feature maps, and a convolutional attention mechanism module is used to increase network stability and convergence speed. Comparison experiments conducted on the infrared small target data set ESIRST. The experiments show that compared to the benchmark network CenterNet-HRNet, the proposed ISTD-CenterNet improves the recall by 22.85% and the detection accuracy by 13.36%. Compared to the state-of-the-art YOLOv5small, the ISTD-CenterNet recall is improved by 5.88%, the detection precision is improved by 2.33%, and the detection frame rate is 48.94 frames/sec, which realizes the accurate real-time detection of small infrared targets.

KEYWORDS

Infrared small target detection; CenterNet; data enhancement; feature enhancement; attention mechanism

1 Introduction

In recent years, infrared detection equipment has been applied to military and civil applications on a large scale, and as a key technology in infrared detection systems, infrared small target detection technology has received wide attention. At present, this technology is mostly used in the fields of early warning, precision guidance, marine surveillance and fault detection. Infrared detector detection distance, susceptible to atmospheric scattering and background temperature noise and resolution is not high, so the image quality of infrared sensors than visible light sensor image quality is poor. Compared with traditional targets, infrared small targets have the following characteristics: (1) According to the standards of the International Society of Photo-Optical Instrumentation Engineers (Society of Photo-Optical Instrumentation Engineers, SPIE), infrared sensors to generate an image of the infrared small targets in the image occupies the number of pixels on the image does not exceed the entire image of 0.15%. For a $256 * 256$ standard size infrared image, the number of pixels occupied by infrared small targets will not be more than $9 * 9$. (2) There are no obvious shape features and texture features. (3) The signal-to-noise ratio (SCR) of infrared small targets is low, and they can be easily submerged in the complex background. Therefore, infrared small target detection is a serious problem.

Advanced traditional infrared small target detection algorithms mainly use a model-driven approach to separate the target from the background by enhancing the contrast between the target and the background in the local range, and then obtain the features of the target. However, the method also relies on hand-crafted features. When the size, shape and texture of the target, the signal-to-noise ratio and the background it is located in change, it is difficult for the hand-crafted features to cope with these changes, leading to a significant reduction in detection accuracy.

Different from the traditional infrared small target detection algorithm [1], the deep learning based infrared target detection algorithm adopts a data-driven approach and uses DNN (Deep Neural Network) to extract features, and the algorithm is generalized. The algorithm mainly includes two types of Anchor-base and Anchor-free. Xu et al. [2] added dilated convolution to the backbone module of YOLOv3 to enhance the feature extraction capability of the module and used a multi-scale training method to realize the inspection of the working status of high voltage lead connector in the substation. Zhao et al. [3] realized the recognition of target infrared features under complex background by constructing an infrared image dataset based on optical images and training SSD detectors. Wei et al. [4] utilized the fused image orientation gradient histogram (HOG) to first extract the target features, and then input the features into the regional convolutional neural network RCNN to complete the task of detecting small infrared targets. Zheng et al. [5] added a cross-stage localization module to DarkNet53, the backbone of YOLOv3, and added a feature fusion path aggregation module after the feature pyramid module of the detector to improve the feature extraction capability of the detector, which in turn realized accurate prediction of targets. The above papers use an Anchor-base based target detector which predicts the category and location of the target using an initially set anchor box, however, it is difficult to set the hyperparameters of the IOU thresholds of the anchor box and the bounding box, and a large number of redundant anchor boxes will cause the problem of positive and negative sample imbalance. He et al. [6,7] and Liu et al. [7,8] designed a dense nested attention network, a subpixel sampling wedge network and an image enhancement network, respectively, to realize the progressive interactive features between high-level and low-level and generate multi-scale feature maps, and the contextual information of small targets is fully applied through iterative fusion and enhancement. However, the model complexity of the method is high.

Different from it, anchor-free target detection algorithm mainly detects the corner and center points of the target box, which will not produce a large number of anchor boxes. The algorithm has a simple structure, and the detection speed is faster under the condition of guaranteeing high detection accuracy, so it is widely used in the detection of infrared small targets with high real-time requirements. The Anchor-free algorithm is originated from the DenseBox and the YOLOv1 [9], CornerNet [10] inherits the advantages of the Anchor-free algorithm and creatively converts the prediction of the target position into the prediction of the corners of the target prediction frame, which creates a new era of Anchor-free target detection algorithms. The subsequent CornerNet-Lite and ExtremeNet [11] achieve fast and accurate target prediction by replacing CornerNet's backbone network, random block, and convolutional block, respectively, to predict the four extreme and center points of the target prediction frame. The proposal of CenterNet marks the maturity of the Anchor-free target detection algorithm, and CenterNet is a new generation of the Anchor-free target detection algorithm. Maturity, CenterNet does not require anchors, rep-resents each desired object with a set of keypoints and predicts these keypoints directly from the model, which transforms the target detection problem into a standard keypoint detection problem without anchor intervention or other processing. The location of the target center point is predicted by generating a heat map of keypoints, and then attributes such as width, height, category, and orientation of the target are regressed from the feature map corresponding to the center point. Due to the high accuracy of CenterNet detection model and the obvious advantages for embedded platforms with smaller computing power, CenterNet-based infrared target detection algorithms have been widely used in the detection of ship targets, substation targets, railroad foreign object targets, road surface defects, and composite surface layups [12–15], but most of this type of methods are targeted at certain special properties of the targets. While small-scale target detection in infrared images in complex scenes contains little target information, low target intensity and lacks robust shape and texture features, so the paper proposes an infrared small target detection method based on improved CenterNet.

Firstly, for the problem that features extraction of CenterNet backbone network (e.g., ResNet, Hourglass Network, etc.) is easy to lose small target features and the number of model parameters is large, HRNet-W32 [16] was proposed to be improved as the backbone of the network, and the ISTD-CenterNet small target detection framework is constructed; secondly, for the problem that high-resolution and depth feature maps cannot coexist, the ECBAM attention module and SEM scale enhancement module are added to each stage branch of HRNet to further enhance the deep infrared small target features while obtaining high-resolution feature maps; then, in order to fully utilize the semantic information in the low-resolution feature maps, the improved receptive field enhancement module, RFE, was designed; finally, in response to the problem of poor stability and difficulty in convergence of the small target detection network, the ISTD-CenterNet infrared small target detection framework is constructed. With poor stability and difficult convergence, a loss function combining binary cross-entropy loss [17] and EIOU [18] is used to improve the detection performance and convergence speed of this detection model. The experimental results show that the improved ISTD-CenterNet infrared small target detection network realizes the real-time tracking of infrared small targets while ensuring a lower leakage detection rate and higher detection accuracy.

The paper is organized as follows. Section 2 describes the improved ISTD-CenterNet network, including the introduction of the structure of CenterNet, the improved EHRNet with the effective attention module ECBAM, the scale enhancement module SEM, the feature enhancement module RFE; Section 3 presents the enhanced dataset ESIRST and demonstrates the improvement of the performance of the network through experiments; Section 4 addresses issues related to the algorithm and looks ahead to the next step in the process; Section 5 summarizes the performance of the whole improved network and presents the conclusions of this study.

2 ISTD-CenterNet Network Structure Analysis

2.1 Introduction to CenterNet Algorithm

CenterNet first uses the feature maps generated through the encoder and decoder to output the heat map of key points, and then obtains the prediction center point of the target, and then regresses the attributes of the target's prediction box, such as width, height, category, and direction, from this center point. Compared with the Anchor-base algorithm, CenterNet does not generate a large number of anchor frames and does not require additional post-processing such as non-great suppression, the algorithm is simple in structure, and the detection speed is faster under the condition of ensuring high detection accuracy. The specific detection process is as follows.

Suppose the network inputs an infrared image $I \in R^{W \times H \times 1}$, where W and H are the width and height of the image, respectively. First, the image is down-sampled by the encoder to extract features, then, the output feature map is sent to the decoder for up-sampling, then, the output feature map is fed into the detection head to output the key point heat map, center point offset and target frame size, finally, the maximum pooling operation is performed on the key point heat map, the local peak point obtained is considered as the target center point, and the center point offset is compensated to the target center, and combined with the target frame size to output the detection results. Taking the typical encoder ResNet-18 as an example, the network detection process is shown in Fig. 1.

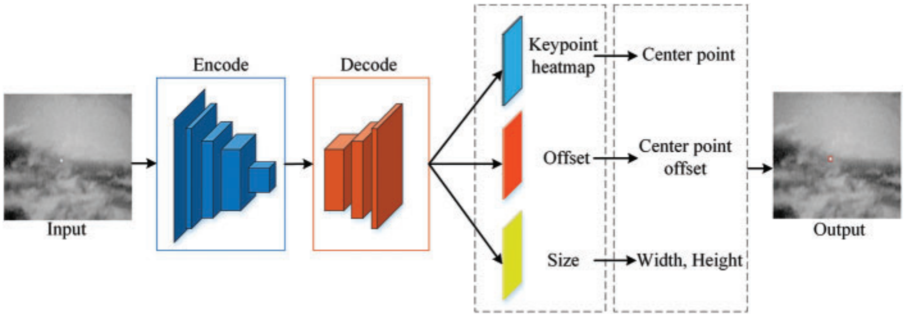


Figure 1: CenterNet target detection process

The CenterNet network takes the target's real label box center point as the target center point and the real label box width and height as the target's dimensions, and after all the target center points have been downsampled by the encoder, the target center points are projected into the feature map by a Gaussian scattering kernel to obtain the keypoint heat map $\hat{Y}_{xyc}$, which is an image consisting of the probability values of the center of mass of each pixel point of the input image.

$$\hat{Y}_{xyc} = \exp\left[-\frac{(x-\widetilde{p}_x)^2 + \left(y-\widetilde{p}_y\right)^2}{2\sigma_p^2}\right] \quad (1)$$

where $\hat{Y}_{xyc} \in [0, 1]^{\frac{W}{r} \times \frac{H}{r} \times C}$, C is the number of target types, c represents one of the target types, r is the downsampled image shrinkage times, x, y are the coordinates of the key point at any location on the heat map, σ_p is the standard deviation of the dimensions that can adapt to the size of the target area, and $\widetilde{p}_x$ and $\widetilde{p}_y$ denote the coordinates of the key point at the center of the target after downsampling. The Visualization of the detection flow of CenterNet is shown in Fig. 2.

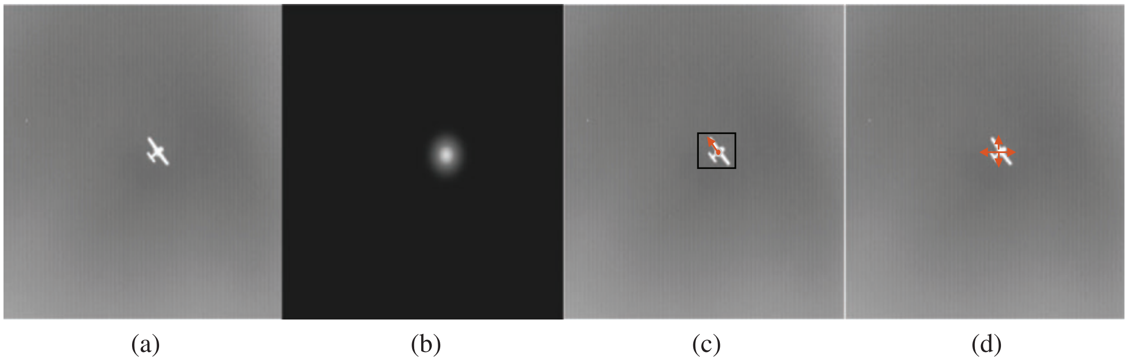


Figure 2: Visualization of the detection flow of CenterNet (a) Original image (b) Heat map of key points (c) Key point offset (d) Target size

2.2 *ISTD-CenterNet Overall Network Architecture*

The essence of single-frame image detection is to build a model that relates image information and depth information [19], that is, to predict the depth of each pixel in the image within a specified range, which requires a combination of multiple factors such as feature map resolution, multiscale information fusion, and so on. The ISTD-CenterNet proposed in the paper is presented in this context, and the overall structure is shown in Fig. 3. After the network receives the image input, it first feeds the image into the improved high-resolution EHRNet feature extraction network to get out four kinds of resolution feature maps, and then the feature fusion network using the four kinds of feature maps outputs the high-resolution feature maps for target detection. Finally, the feature maps are fed into the CenterNet detection head to get the target regression frame and complete the infrared small target detection task.

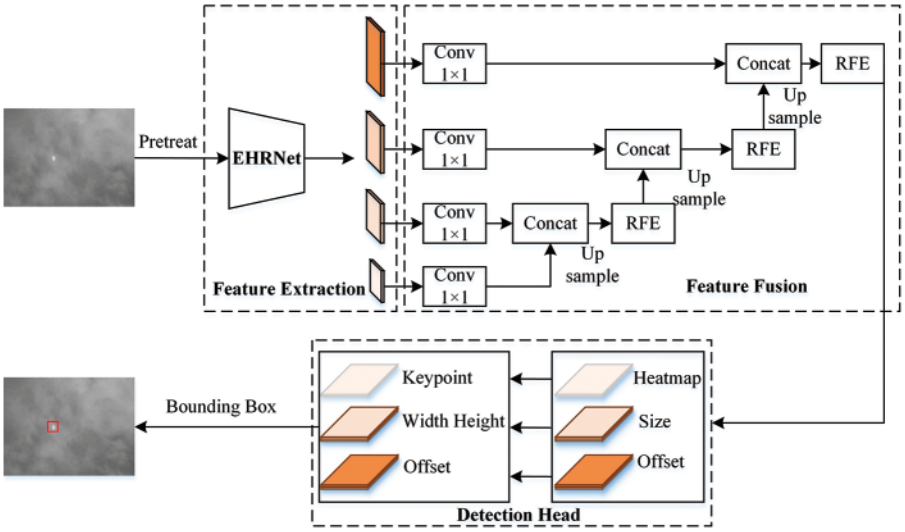


Figure 3: ISTD-CenterNet overall network architecture

2.3 *Feature Extraction Networks EHRNet*

In computer vision, the CenterNet target detection network backbone mainly consists of ResNet, Hourglass Network. The backbone network encodes and downsamples the input image to obtain low-resolution feature maps, and then the decoding module uses upsampling and other methods to recover these feature maps into high-resolution feature maps, which are finally input into the detection head for target detection. The resolution of the feature maps obtained by downsampling will be reduced, simple upsampling often loses more global information, and the use of hopping connections to fuse the low-level features step by step is prone to introducing background noise, resulting in the loss of the features of the infrared small targets as well as a reduction in the detection accuracy of the phenomenon. HRNet is different from the encoder-decoder model described above. First, HRNet proposes a parallel structure that allows the network to maintain a high-resolution representation when extracting features, rather than trying to recover from a low-resolution representation. Second, in HRNet, HRNet not only obtains a strong semantic representation, but also a more accurate resolution spatially, due to the parallelization structure, which makes it possible to fuse different features at different scales consecutively to extract each feature efficiently so that the recovery process distorts the original image as little as possible, instead of joining or adding low and high level representations of the same scale. Since infrared small targets are usually submerged in complex backgrounds with low signal-to-noise ratio, for the network, the size of the obtained features gradually decays and the semantic representation increases as the network continues to deepen, and the detection of these small targets with low false alarms requires high-level semantic representations of the entire infrared image and fine resolution prediction maps, the HRNet structure is chosen to be the backbone of the infrared small target detection network structure.

For infrared small target detection network feature extraction, the improved multi-stage high-resolution EHRNet backbone structure proposed in the paper is shown in Fig. 4, which mainly consists of a multi-scale feature extraction part. The structure refers to the convolution process between two neighboring downsampling modules as a stage, and there are four stages in this backbone network. There is one branch in stage 1, which has the highest resolution of the feature map, and starting from stage 2, a parallel branch is added in each stage, and the output of the previous stage branch is used as the input of the branch in the next stage, and the resolution of the new branch feature map is half of the lowest resolution of the feature map of the branch in the previous stage, and the number of channels is increased by a factor of two. This backbone finally obtains four scaled feature maps with channel numbers 64, 128, 256, and 512, respectively, at resolutions 1/4, 1/8, 1/16, 1/32 of the original map. The scale enhancement module SEM is inserted after the output of each stage to enhance the multiscale representation learned in each convolutional layer, and the feature maps of different resolutions output from different stages undergo the channel splicing operation to enhance the feature interactions between the branches. To reduce noise interference, the convolutional block attention module ECBAM (Convolutional Block Attention Module, CBAM) [20] is added after the spliced features to recalibrate the spliced features from the channel and spatial pixel level. In order to improve the network detection efficiency without reducing the detection accuracy, it was experimentally determined that the stage 3 modules were reduced from four to two and the stage 4 modules were reduced from two to one. Among them, Basic Block, Bottleneck module and up and down sampling are shown in Fig. 4.

2.3.1 Scale Enhancement Module SEM

For shallow networks that may not be able to detect edges at different scales and a large number of convolutional layers involving too many parameters, which makes training difficult, feature enhancement of output features at each stage is performed using SEM in order to achieve a better balance between efficiency and accuracy. For the input feature map $x \in R^{H \times W}$ with convolution filter $w \in R^{m \times n}$, the output $y \in R^{H' \times W'}$ of the dilated convolution at position (i,j) is calculated by Eq. (2).

$$y_{i,j} = \sum_{m,n}^{h,w} x\left[i + r \cdot m, j + r \cdot n\right] \cdot w\left[m, n\right] \tag{2}$$

where r is the expansion rate, which indicates the step size of the sampled input feature map, and the standard convolution can be regarded as a special case of. The above equation shows that the dilated convolution enlarges the receptive field of neurons without decreasing the feature map resolution or increasing the parameters. Replacing the original 3 × 3 convolution with a cascade of 1 × 3 and 3 × 1 convolutions not only reduces the number of parameters in the model, but also increases the nonlinearity of the network, making the network fit more accurately. Each SEM has N dilated convolutions with different expansion rates, where the expansion rate of the nth dilated convolution is set to $r_n = \max(1, r_0 \times n)$, where r_0 is the expansion rate factor. The structure of the SEM for the scale enhancement module is shown in Fig. 5.

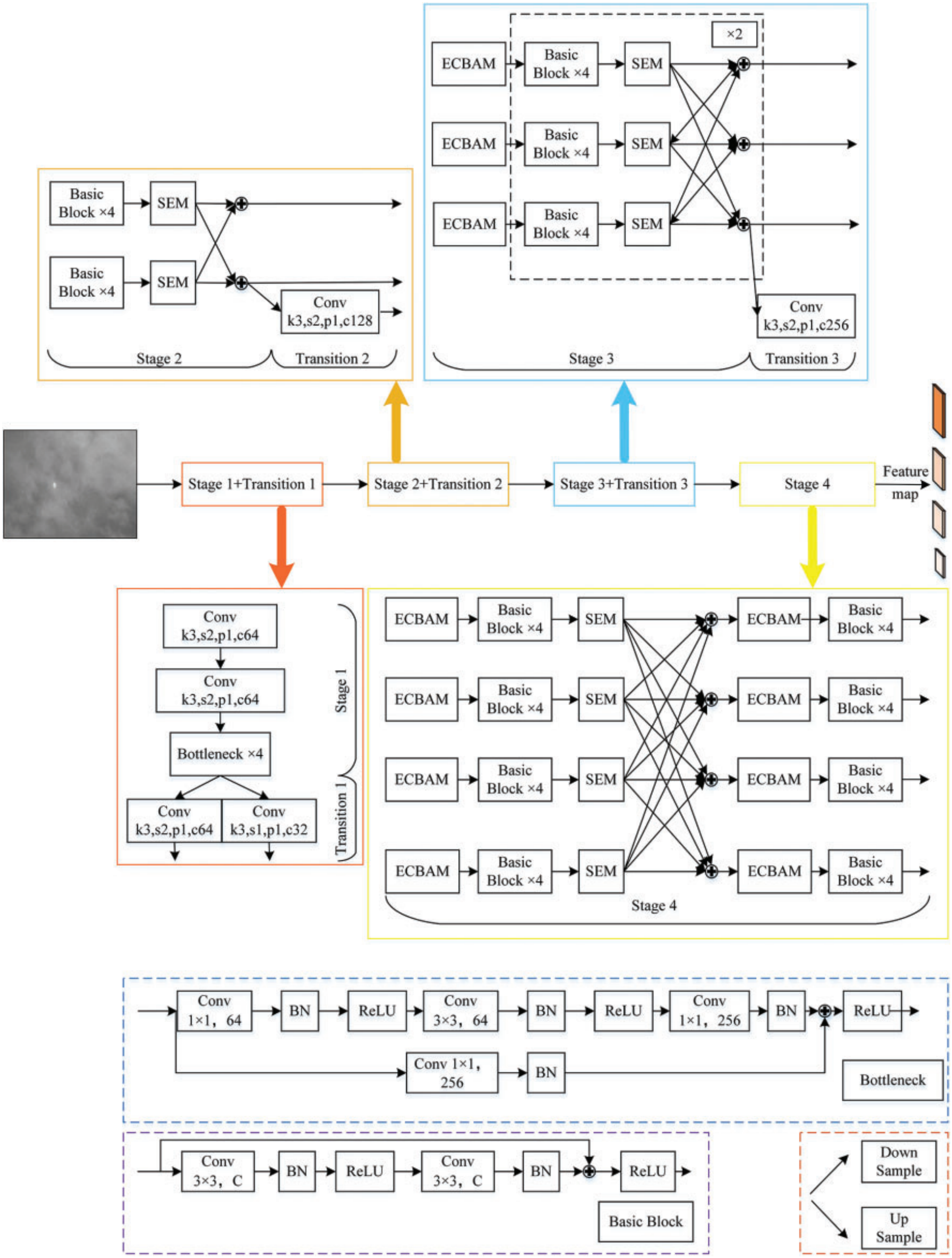


Figure 4: Improved multi-stage high-resolution EHRNet

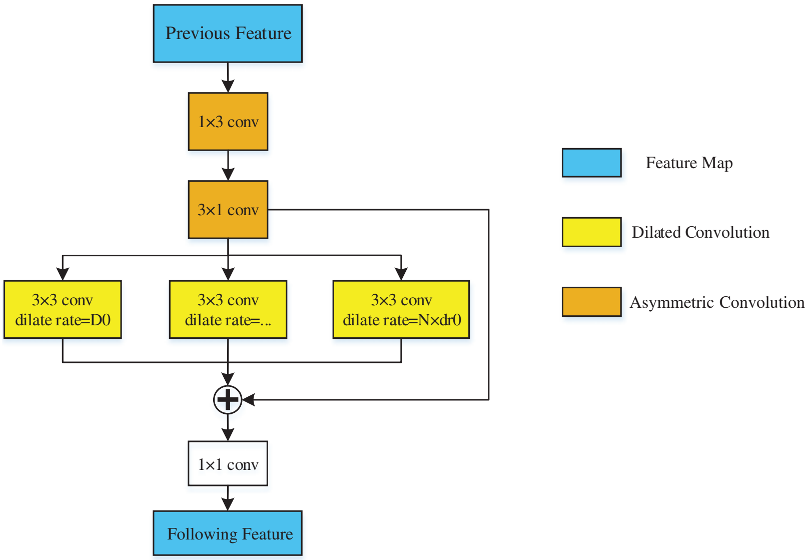


Figure 5: SEM of the scale enhancement module

2.3.2 *Effective Convolutional Attention Mechanism Module ECBAM*

Aiming at the problem that infrared images are disturbed by complex backgrounds, which leads to the weak saliency of small targets and thus makes it difficult to detect them, inspired by the cognitive mechanism of visual neural hierarchy, we embed the attention module ECBAM on the multi-stage feature transfer branch structure of HRNet, so as to enable the network to perform feature filtering on the multi-layer feature transfer branches, and thus to capture the correlations between different levels of small target features. Unlike the attention mechanisms in the literature [21], ECBAM improves the saliency of small target features under complex environments in the channel dimension and the spatial dimension at the same time. Channel dimension and spatial dimension to improve the saliency of small target features in complex environments and enhance the detection model's ability to perceive infrared target features. Among them, the channel attention mechanism adopts the improved Efficient Channel Attention [22]. In order to enhance the extraction of detailed features such as infrared target texture, reduce the number of model parameters, and extract complete and effective due to the target related contextual information, the maximum pooling operation and average pooling operation are added before the one-dimensional adaptive convolution operation of the ECA, and the feature maps obtained from the two pooling are summed up by bits in the after that, the local cross-channel interaction method in the ECA is adopted to provide each channel with more accurate network parameters to enhance the semantic features of infrared targets. The calculation is shown in Eq. (3).

$$M_c(F) = E(C1D_k(MeanPool(F)) + MaxPool(F))) \tag{3}$$

where $M_c(F)$ denotes the output feature map of the channel attention mechanism, $E(\cdot)$ is the ELU nonlinear activation function, $C1D_k(\cdot)$ denotes the one-dimensional convolution with an adaptive convolution kernel of k; $MeanPool(\cdot)$ and $MaxPool(\cdot)$ denote the mean pooling and maximum pooling, respectively.

The spatial attention mechanism utilizes the spatial relationship between features to generate a weight mask and weighted output of the position information to generate a spatial saliency feature map, which enhances the feature expression of key regions, weakens irrelevant regions such as the background, and enhances the infrared small target localization features, and the computational process is shown in Eq. (4).

$$M_s(F) = E\left(f_{7\times7}([AvgPool(F);MaxPool(F)])\right) = E\left(f_{7\times7}\left(\left[F^C_{avg}; F^C_{max}\right]\right)\right) \tag{4}$$

where $M_s(F)$ denotes the spatial attention mechanism output feature map, $f_{7\times7}$ denotes the convolution operation of 7×7, and F^C_{avg} and F^C_{max} denote the channel global-averaged pooled features, and maximum pooled features, respectively. The ECBAM structure is shown in Fig. 6.

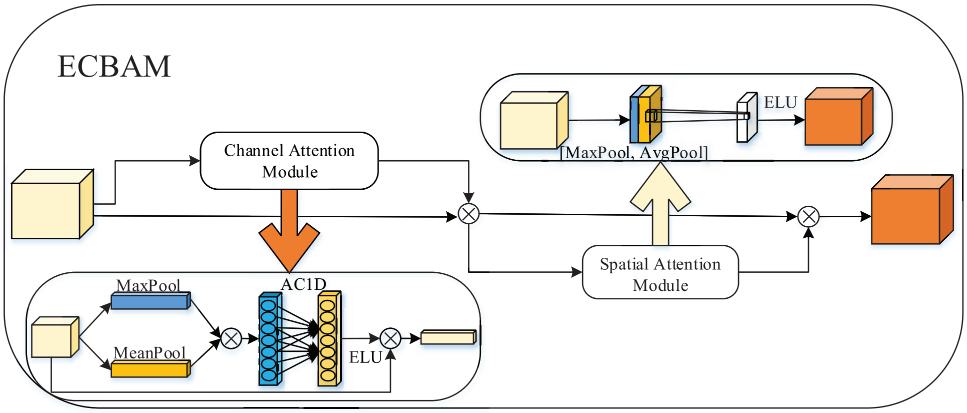


Figure 6: Improved effective convolutional attention mechanism module ECBAM

2.4 Feature-Fusion Network

Infrared small target detection network mainly uses high-resolution feature maps to achieve small target detection, because high-resolution feature maps contain rich small target shape, location and other detailed information, but the backbone network extraction capability is limited, the feature maps are not adequately processed, so that the high-resolution feature maps have less semantic information, the sense of the field is limited, and can not be better adapted to the detection of small targets, and the existence of high-level semantic information in the low resolution feature maps can help to improve the detection effect of the small targets, so the feature enhancement and fusion of the feature maps of the four scales obtained after the completion of the feature extraction. Firstly, the 1 × 1 convolution is used to align the feature maps of each scale in the number of channels, then the bilinear interpolation up-sampling method is applied to the fourth level feature map, which reduces the network learning parameters while transforming the scale of the feature maps, then the feature maps are spliced with the feature maps of the third level that have undergone 1 × 1 convolution, and then the fusion features are obtained by the RFE data enhancement module, and finally the output features of the previous step are bilinearly up-sampled, and spliced with the feature maps of the second level that have undergone 1 × 1 convolution, in this order. Finally, the output features of the previous step are bilinearly upsampled and spliced with the feature map of the second stage after 1 × 1 convolution, and then operated sequentially to make full use of the feature maps of different resolutions while enhancing the feature representation, and obtain the final high-resolution feature map with rich information of small targets.

Among them, the data enhancement module RFE is derived from RFB [23], which utilizes the human visual receptive field mechanism to pay more attention to the center point of the target, the high importance of the features at the center of the receptive field, and the low importance at the edges, and by simulating the visual receptive field enhancement mechanism, the importance of the receptive field as well as the features at the center of the receptive field is increased at the same time. The information at the center of the infrared weak target is very important, so the network using this mechanism significantly enhances the feature extraction ability of the infrared weak target, but the infrared weak target is smaller than the ordinary small target, the target occupies a smaller area of the image, the RFB is not applicable to the detection of the infrared small target, so this paper constructs the RFE feature enhancement module from the principle of the RFB. The structure of the RFE is shown in Fig. 7, and the multi-branch structure has 4 branches, and the RFE is a multi-branch structure. The multi-branch structure has 4 branches, all of which first perform 1 × 1 convolution operation on the input feature map, and initially process and adjust the number of feature map channels for subsequent processing of the feature map, among which: the 4th branch is a residual structure, which forms an equivalent mapping at the output end and retains the good feature information of the small target; the middle 2 branches are all cascaded by asymmetric convolution operations such as 1 × 3, 3 × 1, and so on, to extract more detailed feature information by different scale convolution operations. The two middle branches are subjected to asymmetric convolution operations, such as 1 × 3 and 3 × 1, and cascaded to extract more delicate small target features with different scale convolution operations; the expansion rate of the fourth branch's dilated convolution layer is set to 3, which reduces the background information in the extracted feature maps and enhances the feature effectiveness.

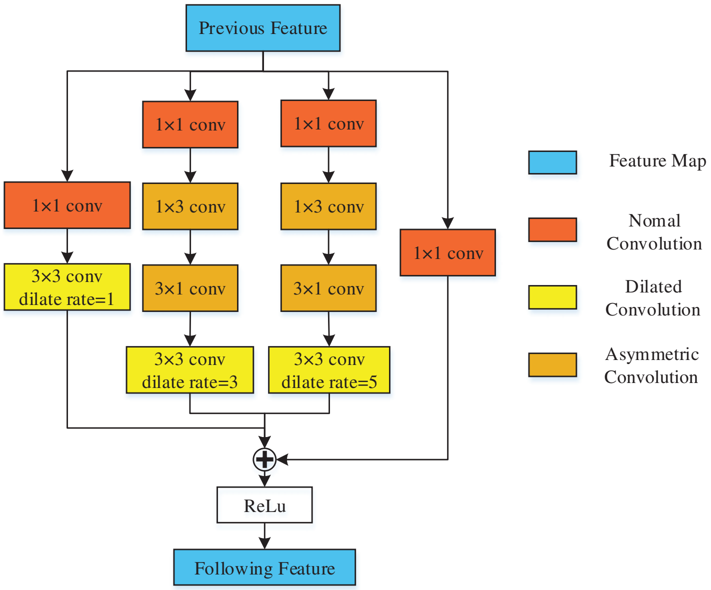


Figure 7: Feature enhancement module RFE

Combined with the experimental results, the computation process of the feature enhancement module RFE is shown in Eqs. (5)–(8).

$$W_1 = f_{DC}^{3\times3}\left[f_C^{1\times1}(PF)\right] \tag{5}$$

$$W_2 = f_{DC}^{3\times3}\left\{f_C^{3\times1}\left\{f_C^{1\times3}\left[f_C^{1\times1}(PF)\right]\right\}\right\} \tag{6}$$

$$W_3 = f_{DC}^{3\times3}\left\{f_C^{1\times3}\left\{f_C^{3\times1}\left[f_C^{1\times1}(PF)\right]\right\}\right\} \tag{7}$$

$$FF = Concat(W1, W2, W3) \oplus f_C^{1\times1}(PF) \tag{8}$$

where $f_C^{1\times1}(\cdot)$, $f_C^{1\times3}(\cdot)$ and $f_C^{3\times1}(\cdot)$ denote the convolution operations with convolution kernels of 1×1, 1×3, and 3×1, respectively, and $f_{DC}^{3\times3}(\cdot)$ in Eqs. (1)∼(3) denotes the null convolution operation with expansion rates of 1, 3, and 5, respectively; Concat denotes the feature map splicing operation; $\oplus$ denotes the feature map summing operation by bit; PF denotes the feature map input; W_1, W_2 and W_3 denote the feature maps after convolution of the 1st∼3rd paths, and FF denotes the output of the feature maps after feature enhancement.

2.5 Loss Function

In the detection part, the CenterNet lightweight detection header is utilized to obtain the target keypoint heat map (i.e., keypoint coordinates), the target frame dimensions and the target center point offset, and this lightweight structure also compensates for the effect of the backbone on the number of network parameters. In this detection network, output feature maps in Section 2.4 are utilized for detection, and the corresponding loss function consists of three components: loss of keypoint coordinates, loss of target frame dimensions, and loss of target center point offset. Since there is only one type of infrared small targets in the selected dataset, the binary cross entropy loss is chosen as the key point coordinate loss, which utilizes the logarithmic and exponential techniques to increase the numerical stability, and the EIOU is chosen as the target frame size loss, the EIOU he increases the stability of the network and accelerates the speed of convergence by calculating the difference value of the width and height of the bounding box, and at the same time, the Focal Loss is introduced to solve the sample imbalance Balance problem. In the original CenterNet network, due to downsampling the output feature map will introduce accuracy error when remapped to the original image size. Therefore, an additional offset compensation is needed for each key point, but in the network of the thesis, the size of the output feature map is the same as the size of the input feature map, therefore, there is no need to compensate the offset for the key points. So, the loss function expression is shown in Eqs. (9)–(11).

$$L = L_{center} + L_{bound} = BCE + EIoU \tag{9}$$

$$EIoU = 1 - IOU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \frac{\rho^2\left(w, w^{gt}\right)}{C_w^2} + \frac{\rho^2\left(h, h^{gt}\right)}{C_h^2} \tag{10}$$

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right) \tag{11}$$

where BCE denotes Binary Cross Entropy Loss (BCE) and EIoU denotes Focal and Efficient IOU Loss.

In summary, the ISTD-CenterNet target detection network proposed in the paper can take into account the lightweight and feature extraction capabilities, has the advantages of strong fusion of high and low layer feature information, high utilization of low layer feature information, and is more conducive to the detection of small infrared targets.

3 Experimental Results and Analysis

The hardware information used for the experiment is shown in Table 1.

Experimental parameter design: In order to obtain better experimental results, it is necessary to select appropriate hyperparameters. In the experiment, when the infrared small target detection network is trained, the initial learning rate is set to 0.001, the weight decay is set to 0.005, the momentum factor is set to 0.945, and the optimizer selects Stochastic Gradient Descent (SGD). According to the experimental results, the expansion rate factor $r_0 = 4$, and the number of dilated convolutions $K = 3$ in SEM are set. The network parameters for training under different datasets are different, and the experimental parameter settings are shown in Table 2.

Table 1: Experimental hardware parameters

Name	Configuration
Operating system	Win11
Computing platform	CUDA 11.4
CPU	Inter Core i7/16G
GPU	NVIDIA RTX3060
GPU memory size	12G

Table 2: Experimental parameter setting

Parameter	ESIRST	Terravic motion IR database	FLIR thermal database	MOTIID
Class number	1	1	3	7
Epoch	400	400	400	400
Batch size (train)	16	4	16	16
Image size	320 × 240	320 × 240	512 × 512	640 × 480
Optimizer	SGD	SGD	SGD	SGD
Initial learning rate	0.0001	0.0001	0.0001	0.0001
Momentum factor	0.945	0.945	0.945	0.945

3.1 Dataset Preprocessing

CenterNet is based on a data-driven end-to-end model, and its detection accuracy is affected by the image of the model's input dataset. Unlike optical image-based detection tasks, most infrared small target detection algorithms are trained on private datasets for validation and testing, so datasets for infrared small target detection have been scarce for a long time, and currently, IRSTD-1k [24], SIRST [25], and MFIRST [26] are the more used public datasets. SIRST, as the first publicly available large-scale dataset for infrared small targets. It selects a representative image from a sequence to explicitly construct the open single-frame dataset and annotates it with five different forms, so this dataset is more suitable for infrared small target detection.

The paper also uses this dataset, but the limited number of images in this dataset, only 427, will make it difficult for the detection network to be effectively trained [27], which in turn affects the detection effect, so the dataset images are expanded by cropping, rotating, translating, mirroring and other variations, total 4270 images. Use Labelimg and use the unified 7×7 pixel expansion frame to label the target, and finally transform the labelling into VOC data format. This increases the effectiveness of the detection network in extracting target features.

Some of the dataset images are shown in Fig. 8. The new dataset is named as ESIRST dataset and it is randomly divided according to training:validation:test = 8:1:1, i.e., 3416 images for the training set and 427 images each for the validation and test sets.

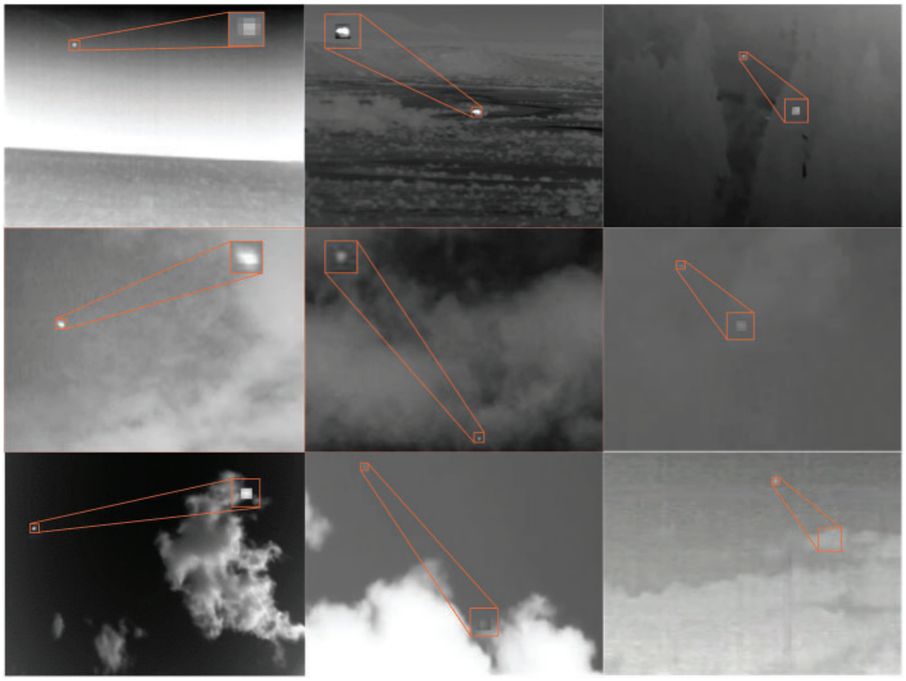


Figure 8: Images of selected datasets

To verify the generalisation ability of the proposed detection network, the paper performs comparative experiments on CSIR-CSIO Moving Object Thermal Infrared Imagery Dataset (MOTIID) and Terravic Motion IR Database.

3.2 Selection of Indicators for Evaluation

In the field of target detection, there are two important metrics, which are accuracy rate (precision rate) and recall rate (recall rate), calculated as shown in Eqs. (12) and (13).

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{12}$$

$$R = \frac{N_{TP}}{N_R} \tag{13}$$

where a is the number of real infrared small targets detected by the target detection network, c is the number of false infrared small targets detected by the target detection network, and b is the number of real infrared small targets present in the image.

In the field of early warning detection, people pay more attention to whether the detector can detect all the threatening targets, which is crucial for our protection party, so the paper chooses the recall rate as one of the evaluation indexes. On the other hand, considering the size of the infrared small targets, the deviation of some pixels may lead to a large change in the intersection and merger ratio, so the prediction is considered reasonable when the intersection and merger ratio between the predicted bounding box and the real labeling box reaches 50% [28], so the AP_{50} is chosen as one of the evaluation indexes.

3.3 Comparison Experiment

In order to verify the effectiveness of ISTD-CenterNet, the infrared small target detection network proposed in the paper, comparison experiments are conducted on the processed ESIRST dataset. The Focal Loss function proposed by Retinanet refines the model by increasing the weight occupied by the difficult samples, which is one of the difficult problems in the infrared small target detection, so one of the comparison networks chosen is the single-stage network Retinanet.

In addition, YOLOv3 and EfficientDet are classical target detection networks, and YOLOv4-tiny and YOLOv5small are advanced and efficient small target detectors. The experimental results are shown in Table 3.

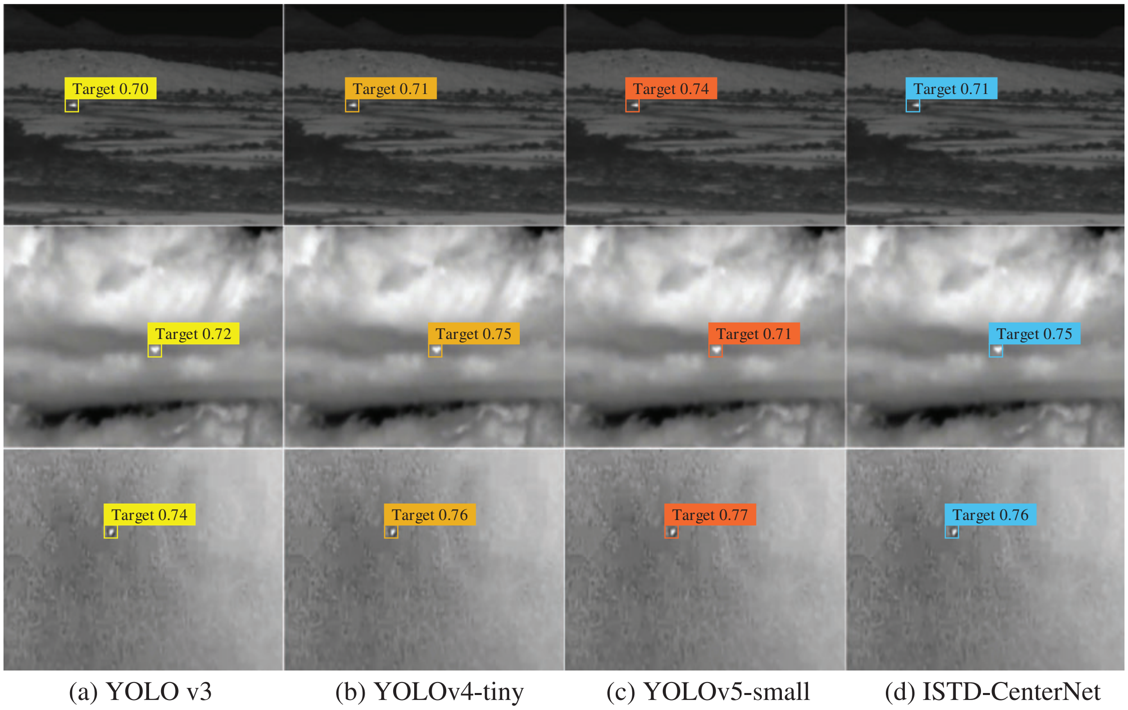
Table 3: Comparison tests under the ESIRST dataset

Method	Recall rate	AP_{50}	Model size/MB	FPS
Retinanet-Resnet50	0.6104	0.6732	98.7	10.58
CenterNet-Resnet50	0.6238	0.6973	132.6	40.91
CenterNet-HRNet	0.6142	0.6803	16.5	42.35
YOLOv3	0.8164	0.8115	114.3	16.77
EfficientDet b0	0.1437	0.1052	16.9	36.31
YOLOv4-tiny	0.7743	0.8051	**14.7**	29.41
YOLOv5small	0.7839	0.7906	18.1	30.74
ISTD-CenterNet	**0.8427**	**0.8139**	15.2	**48.94**

As can be seen from Table 3, compared with the typical single-stage network Retinanet-Resnet50, ISTD-CenterNet has superiority in detecting small infrared targets, with higher detection efficiency, faster detection speed and smaller model size. Compared with YOLOv3 with complex network structure, ISTD-CenterNet's recall and AP_{50} metrics are still higher, which indicates that the complex network is not targeted in detecting infrared small targets with fewer features, and the large model size of YOLOv3 is not conducive to integration, with slower detection speeds and poorer real-time performance. The EfficientDet b0 network structure is simple, lightweight design, faster detection speed, but can not be accurately detected for infrared small targets. CenterNet-Resnet50 is a typical anchorless frame detection network, fast detection speed, but lower detection accuracy. Compared with the advanced small target detectors YOLOv4-tiny and YOLOv5small, ISTD-CenterNet is more targeted for infrared small target detection, with a 5.88% increase in recall and a 2.33% increase in detection speed. Compared with the benchmark CenterNet-HRNet, the recall is improved by 22.85%, AP_{50} is improved by 13.36%, and the network model size is reduced by 7.88%.

In addition, the backbone of CenterNet-Resnet50 and CenterNet-HRNet was pretrained on the large datasets ImageNet and Coco, and the detection performance would be significantly degraded if this backbone was not pretrained, but the ISTD-CenterNet network was trained from scratch and still achieved better results, which indicates that in the data-sparse infrared target detection field, the proposed ISTD-CenterNet is more suitable for infrared small target detection task.

The four detection networks as in Table 3 are selected to visualize the detection results of the networks on infrared small targets, as shown in Fig. 9, from which it can be seen that the average confidence of the network of ISTD-CenterNet reaches 0.79, which is the highest among the four networks, and the detection is more accurate.



(a) YOLO v3 (b) YOLOv4-tiny (c) YOLOv5-small (d) ISTD-CenterNet

Figure 9: Comparison of four typical network detection results

In order to compare and analyze the detection effect of four different algorithms on infrared small targets in different scenarios, we chose three typical scenarios for testing, and the final visual detection effect is shown in Fig. 9. In Fig. 9, the detection box indicates the correct detection as well as the confidence level of correctly detecting a target. In the second and third sky backgrounds in Figs. 9a–9d, the energy and local contrast of infrared small targets are relatively low, resulting in small targets submerged in thick cumulus clouds or high cumulus clouds. The experimental results show that ISTD-CenterNet can effectively detect weak targets and infrared small targets with the highest confidence level without false alarms or missed detections. In the first ground background in Figs. 9a–9d, the infrared targets have strong energy and high local contrast. These results demonstrate that all four algorithms can detect small infrared targets. However, most of the algorithms do not have high detection confidence due to the effect of interference sources such as jungle on the target, and the detection confidence is generally low.

By analyzing Figs. 9a–9d, we infer the following conclusions. (1) The detection of sky background is better than ground background. (2) The higher the target energy or local contrast, the better the detection effect. (3) YOLOv4-tiny and YOLOv5-small algorithms designed for small targets have better detection effect than YOLOv3 algorithm under different complex backgrounds. (4) From Table 3 and Fig. 9, it can be seen that the proposed ISTD-CenterNet achieves good results in terms of detection accuracy and recall as well as detection speed.

3.4 Ablation Experiment

In order to verify the impact of several key modules in the ISTD-CenterNet network on the detection performance of infrared small targets compared to the CenterNet-HRNet baseline network, ablation experiments are conducted on the processed ESIRST dataset, and the results of the experiments are shown in Table 4.

Table 4: Ablation experiments under the ESIRST dataset

CenterNet-HRNet (baseline)	With HRNet reduced number of stages	With ECBAM	With SEM	With RFE	Recall rate	AP_{50}	Model size/MB	FPS
√					0.6142	0.6803	16.5	42.35
	√				0.5236	0.5741	10.1	70.87
	√	√			0.6734	0.6513	12.8	67.06
	√	√	√		0.7551	0.7432	18.2	52.17
	√	√	√	√	0.8427	0.8139	15.2	48.94

As can be seen from the tabular data, after reducing the number of HRNet stages, the network model size is reduced and the network detection speed is improved, but the recall and AP_{50} are reduced by 9.06% and 10.62%, respectively. After adding the improved ECBAM attention to increase the saliency of small target features from channel dimension and spatial dimension, the recall and AP_{50} are improved to 67.34% and 65.13% with less reduction in detection speed. Adding SEM module to the backbone to enhance the multi-scale features greatly improves the ability of the network feature extraction, and the recall and AP_{50} exceed the original network; finally, adding the improved RFE feature enhancement module to the feature fusion path increases the significance of the features at the center of the sensing field, and improves the recall as well as the AP_{50} while decreasing the size of the network model. The final results show that the ISTD-CenterNet can detect infrared small targets quickly and accurately.

3.5 Comparison of Experimental Effects on Two Infrared Datasets

In order to verify the comprehensive detection capability of the proposed infrared target detection network for small infrared targets, comparative experiments are conducted on CSIR-CSIO Moving Object Thermal Infrared Imagery Dataset (MOTIID), Terravic Motion IR Database and FLIR Thermal Database.

There are 18 sequences in MOTIID, each consisting of one or more targets entering and leaving the field of view, and each frame has an image size of 640 × 480 pixels, which includes two different types of four-wheeled vehicles (Ambassador and Innova), a three-wheeled vehicle (rickshaw), a two-wheeled vehicle (motorcycle), pedestrians, dogs, and birds, for a total of seven target types. Terravic Motion IR Database is a motion target infrared dataset with 18 sequences, 11 sequences of which are selected for outdoor motion and tracking scenarios where the target is a pedestrian and the image size of each frame is 320 × 240 pixels. Since people, bicycles, and cars make up the majority of the small targets in all three datasets and the FLIR Thermal Database has a history of industrial applications, it was chosen as the dataset for network generalization capability verification. FLIR Thermal Database is a dataset for driverless cars that collects image sequences of various streets and highways using infrared imaging instruments installed on the car. And the proposed ISTD-CenterNet network is compared with the typical network under the datasets. The experimental results are shown in Table 5.

Table 5: Comparison experiments on three infrared target datasets

Method	MOTIID			FLIR thermal database			Terravic motion IR database		
	Recall rate	mAP_{50}	FPS	Recall rate	mAP_{50}	FPS	Recall rate	mAP_{50}	FPS
Efficientdet b0	0.2765	0.3341	16.54	0.2425	0.2618	30.14	0.8364	0.8571	39.1
YOLOv5small	0.7688	0.7439	27.62	0.7456	0.7219	36.53	0.9102	0.9217	42.2
CenterNet-HRNet	0.7037	0.6854	49.22	0.6963	0.6632	53.47	0.8974	0.9157	68.1
ISTD-CenterNet	0.7746	0.7134	45.74	0.7471	0.7049	60.18	0.9635	0.9541	74.5

The experimental results show that the proposed ISTD-CenterNet has a good detection effect on the open infrared small target dataset. In the selected Terravic Motion IR Database dataset, the target is a pedestrian, and the pedestrian target pixels occupy few pixels in the whole image, and there are tabular data that shows that ISTD-CenterNet has the lowest leakage rate, the highest detection accuracy, and the fastest detection speed, which suggests that the proposed detection network has a very good detection effect on the small infrared pedestrian targets, and has a greater advantage compared to the other networks has a greater advantage. In the MOTIID and FLIR Thermal Database datasets, with more target types and different sizes, the detection advantage of ISTD-CenterNet is not obvious, and the mean average accuracy of ISTD-CenterNet is 3.05% lower compared to the advanced detection network YOLOv5small, but the leakage rate and the detection speed are more advantageous, which is more important in the early warning task. To summarize, ISTD-CenterNet has a good comprehensive detection capability for infrared small targets.

To visualize the detection of ISTD-CenterNet on these two datasets, the results are visualized as shown in Figs. 10–12.

Fig. 10 shows the detection results on the MOTIID dataset, Fig. 11 shows the detection results on the Terravic Motion IR Database dataset and Fig. 12 shows the detection results on the FLIR Thermal Database dataset. From the results, it can be seen that the ISTD-CenterNet network achieves the detection of infrared targets with high confidence on both datasets. ISTD-CenterNet network generalises well.

4 Discussion

Although the ISTD-CenterNet proposed in the paper has better results in infrared small target detection, there are still some problems. (1) The targets and contexts in ESIRST, the public dataset used in the paper, are not specific enough, and the target definitions are not clear enough, which results in a lack of clear target traction. (2) The ISTD-CenterNet proposed in the paper is biased towards algorithmic detection accuracy improvement, and the lightweight aspect is not considered enough. In Table 3, the model size of ISTD-CenterNet is larger compared with the advanced algorithm YOLOv4-tiny. Improvement, lightweight aspect is not considered enough, in Table 3, the model size of ISTD-CenterNet is bigger compared with the advanced algorithm YOLOv4-tiny, which will lead to the difficulty of terminal deployment.

Future improvements are needed. (1) Enriching the database. After clarifying the research object and background, integrate the existing dataset to refine different scenarios in order to accomplish the task in a more targeted way, if there is insufficient data, targeted data collection and artificial synthesis can be carried out, thus enriching the dataset. (2) Generate high-quality images. Multi-band fusion techniques such as fusion of infrared and visible light images, fusion of infrared and radar images, etc. supplement the amount of data information in a single infrared image, improve image quality and enhance semantic information. (3) Balance the accuracy of the algorithm and the complexity of the calculation. Infrared small target application scenarios are special, real-time requirements are high, and the follow-up should focus on building a lightweight network.

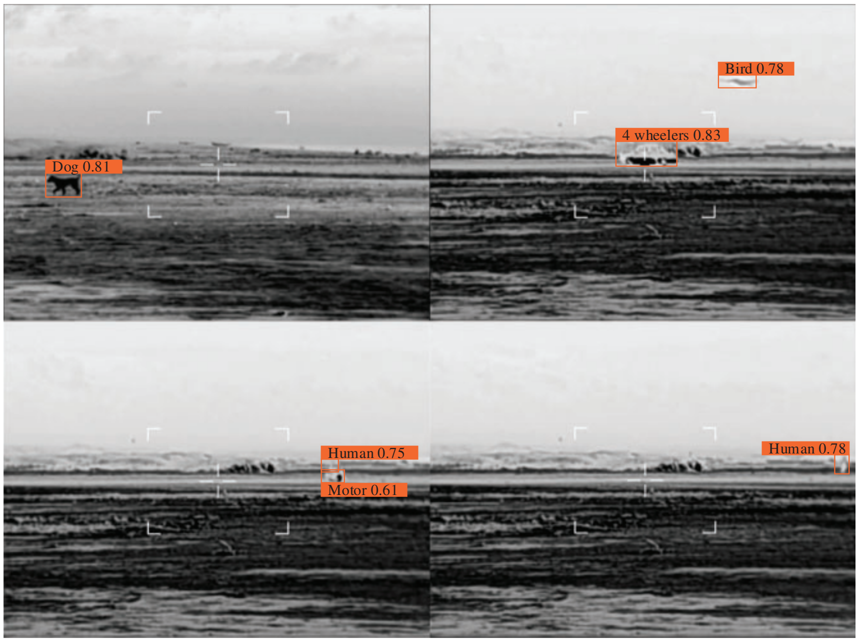


Figure 10: Detection results on the MOTIID dataset

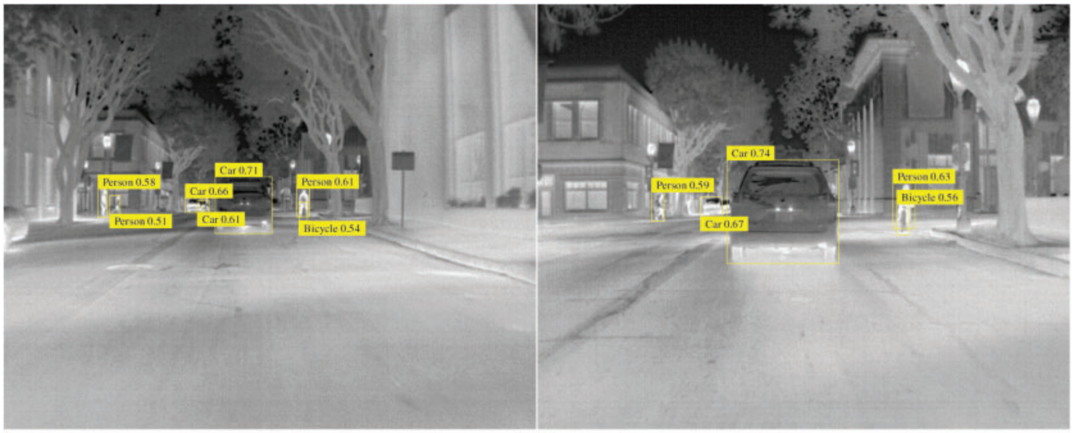


Figure 11: (Continued)

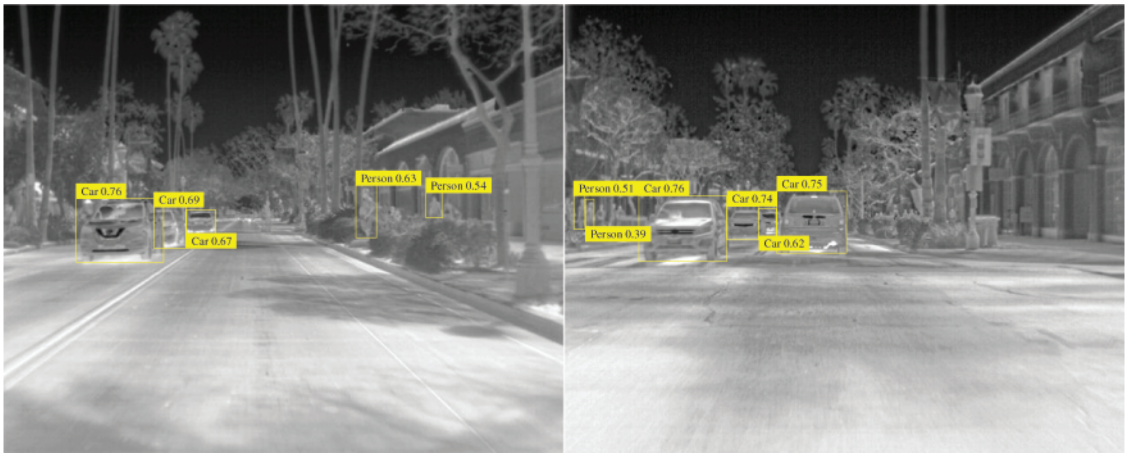


Figure 11: Detection results on the FLIR thermal database dataset

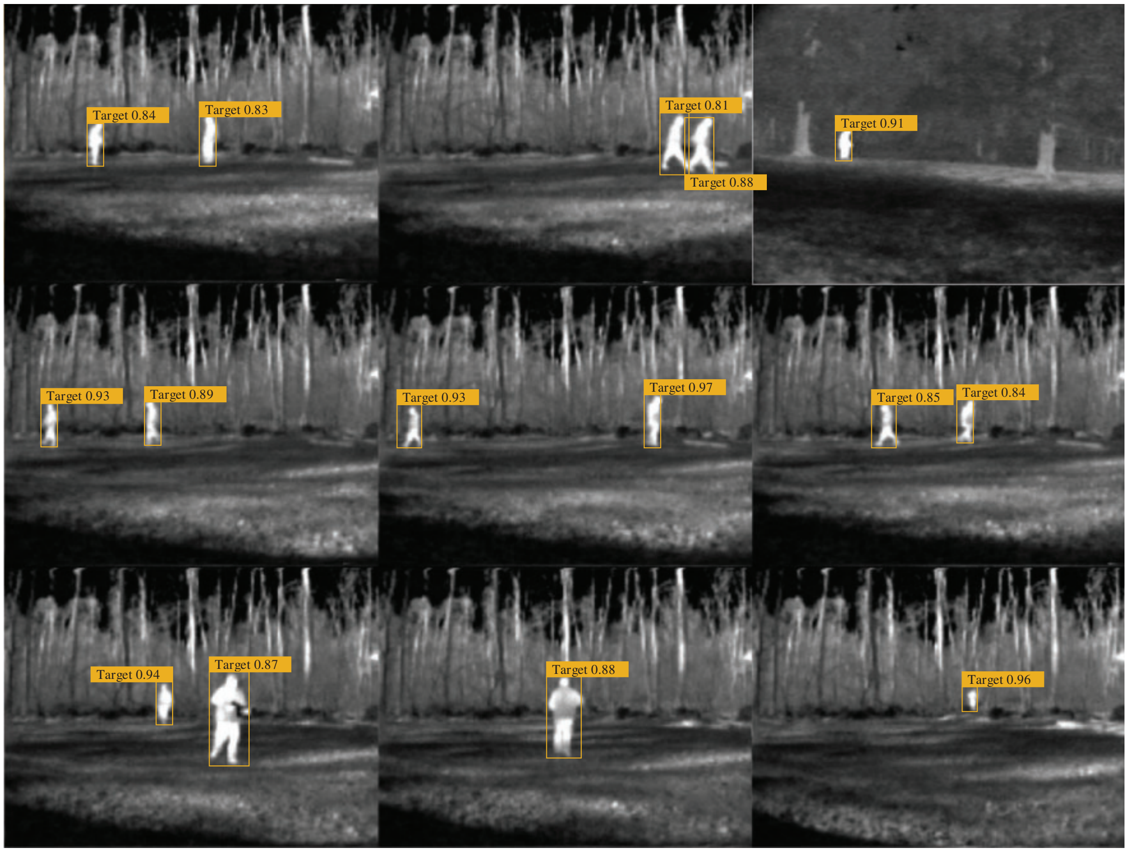


Figure 12: Detection results on the MOTIID dataset

5 Conclusion

Based on fast and accurate detection, this paper proposed an ISTD-CenterNet network for infrared small targets to tackle problems of infrared small target detections in complex environments.

The network detection speed is improved by reducing the number of stages of HRNet, the depth of the network, and the size of the net model and increasing the lightweight level of the network. In terms of detection accuracy, the feature extraction ability of the network backbone is improved by adding convolutional attention and scale enhancement modules, and in the feature fusion stage, the feature enhancement module is utilized to improve the multi-scale feature fusion effect. Several experimental results show that the ISTD-CenterNet network proposed in the paper achieves a recall rate of 84.27%, an AP50 of 81.39%, and a detection speed of 48.94 frames/s on the infrared small target dataset EIRST, which ensures a lower leakage rate and higher detection accuracy while improving the detection speed, and achieves the design objective. Meanwhile, the ISTD-CenterNet can also realize the detection of infrared small targets at a higher confidence level.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by National Natural Science Foundation of China, Fund Number 61703424.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Ning Li, Shucai Huang; data collection: Ning Li, Daozhi Wei; analysis and interpretation of results: Ning Li; draft manuscript preparation: Ning Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Ning Li, upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Zhong, H. Zhou, X. Cui, X. Cao, F. Zhang *et al.,* "Infrared small target detection based on local-image construction and maximum correntropy," *Measurement*, vol. 211, pp. 112662, 2023.

[2] Q. Xu, H. Huang, C. Zhou and X. Zhang, "Research on real-time infrared image fault detection of substation high-voltage lead connectors based on improved YOLOv3 network," *Electronics*, vol. 10, no. 5, pp. 544, 2021.

[3] X. Zhao, M. Xu, F. Wang, J. X. Yang and Z. L. Zhang, "Infrared camouflage detection method for special vehicles based on improved SSD," *Infrared and Laser Engineering*, vol. 48, no. 11, pp. 116–125, 2019.

[4] H. Wei, K. Zhang and L. Zheng, "Infrared image target detection of power inspection based on HOG-RCNN," *Infrared and Laser Engineering*, vol. 49, no. S2, pp. 20200411, 2020.

[5] H. B. Zheng, J. H. Li, Y. Liu, Y. H. Cui and Y. Ping, "Infrared target detection model for power equipment based on improved YOLOv3," *Journal of Electrotechnology*, vol. 36, no. 7, pp. 1389–1398, 2021.

[6] X. He, Q. Ling, Y. Zhang, Z. P. Lin and S. L. Zhou, "Detecting dim small target in infrared images via subpixel sampling cuneate network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[7] R. Kou, C. Wang, Z. Peng, Z. Zhao, Y. Chen *et al.,* "Infrared small target segmentation networks: A survey," *Pattern Recognition*, vol. 143, pp. 109788, 2023.

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

[9] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016.

[10] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 734–750, 2018.

[11] X. Zhou, J. Zhuo and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 850–859, 2019.

[12] C. K. Miao, S. L. Lou and W. F. Gong, "An infrared ship target detection algorithm based on improved CenterNet," *Laser and Infrared*, vol. 52, no. 11, pp. 1717–1722, 2022.

[13] H. Zheng, Y. Cui, W. Yang, J. Li, L. Ji *et al.,* "An infrared image detection method of substation equipment combining iresgroup structure and CenterNet," *IEEE Transactions on Power Delivery*, vol. 37, no. 6, pp. 4757–4765, 2022.

[14] Z. Jiao, C. Ma, C. Lin, X. Nie and A. Qing, "Real-time detection of pantograph using improved CenterNet," in *2021 IEEE 16th Conf. on Industrial Electronics and Applications (ICIEA)*, Chengdu, China, pp. 85–89, 2021.

[15] X. Wang, S. Kang and W. D. Zhu, "Surface defect detection of AFP layup based on improved CenterNet," *Infrared and Laser Engineering*, vol. 50, no. 10, pp. 210–220, 2021.

[16] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 5693–5703, 2019.

[17] Z. Bai, J. Wang, X. L. Zhang and J. Chen, "End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1330–1344, 2022.

[18] Z. Yang, X. Wang and J. Li, "EIoU: An improved vehicle detection algorithm based on vehiclenet neural network," *Journal of Physics: Conference Series*, Shanghai, China, vol. 1924, no. 1, pp. 012001, 2021.

[19] H. L. Luo and Y. F. Zhou, "Research progress on monocular depth estimation for deep learning," *Chinese Journal of Image Graphics*, vol. 27, no. 2, pp. 390–403, 2022 (In Chinese).

[20] S. Woo, J. Park, J. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 3–19, 2018.

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

[22] Q. Wang, B. Wu, P. Zhu, P. Li and W. Zuo, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, WA, USA, pp. 11534–11542, 2020.

[23] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 385–400, 2018.

[24] M. Zhang, R. Zhang, Y. Yang, H. Bai and J. Zhang, "ISNet: Shape matters for infrared small target detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 877–886, 2022.

[25] Y. Dai, Y. Wu, F. Zhou and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, HI, USA, pp. 950–959, 2021.

[26] H. Wang, L. Zhou and L. Wang, "Miss detection *vs.* false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Long Beach, CA, USA, pp. 8509–8518, 2019.

[27] S. Wang, S. Huang, S. Liu and Y. Bi, "Not just select samples, but exploration: Genetic programming aided remote sensing target detection under deep learning," *Applied Soft Computing*, vol. 145, pp. 110570, 2023.

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