Infrared Weak and Small Target Detection Algorithm Based on Deep Learning

Lei Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 2531361795@qq.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 763757335@qq.com

***Abstract*—In the infrared imaging scene where the target is at a long distance and the background is cluttered, due to the interference of noise and background texture information, the infrared image is prone to problems such as low contrast between the target and the background, and feature confusion, which makes it difficult to accurately extract and detect the target. To solve this problem, firstly, the infrared image is enhanced by combining DDE and MSR algorithm to improve the contrast and detail visibility of the image. For the RT-DETR network structure, the EMA attention mechanism is introduced into the backbone to enhance the feature extraction ability of the model by extracting context information. The CAMixing convolutional attention module is introduced into CCFM, and the multi-scale convolutional self-attention mechanism is introduced to focus on local information and enhance the detection ability of small targets. The filtering rules of the prediction box are improved, combined with Shape-IoU, and the convergence speed of the loss function in the detection and the detection accuracy of small targets are improved by paying attention to the influence of the intrinsic properties of the bounding box itself on the regression. In the experiment, the infrared weak target image dataset of the National University of Defense Technology was selected, labeled and trained. Experimental results show that compared with the original DETR algorithm, the average precision of the improved algorithm (mAP) is increased by 3.2%, and it can effectively detect infrared weak and small targets in different complex backgrounds, which reflects good robustness and adaptability, and can be effectively applied to infrared weak and small target detection in complex backgrounds.**

Keywords-RT-DETR; EMA; CAMixing; Shape-IoU

I. INTRODUCTION

As an important thermal measurement technology, infrared imaging technology uses infrared detectors to receive infrared thermal radiation in different wavelengths on the surface of the scene and convert it into images. This technique offers a variety of advantages, such as passive imaging, long range, ease of concealment, and ability to work day and night. This makes infrared imaging widely used in military, security, medical, industrial testing and other fields. However, there are also some challenges faced by infrared imaging devices in practical applications. First of all, because the imaging mechanism of infrared images is different from that of visible light, the contrast between the target object and the background is usually low, which makes it difficult to identify and detect the target. Secondly, when the background interference is strong and the target signal is weak, the signal-to-noise ratio of the infrared image is usually low, resulting in the target image often showing a small target with incomplete structure [1].

In addition, the problems of noise, scattering, and radiation inhomogeneity that are prevalent in infrared imaging further increase the difficulty of effectively detecting small targets in infrared images. With the development of computer vision technology, how to accurately and quickly detect and identify small targets in complex backgrounds has become one of the hot and difficult problems in research. The purpose of this paper is to explore and study the methods and technologies to improve the detection performance of small objects in infrared imaging, in order to provide new ideas and solutions for this field.

II. Related Works

In the early stages of micro-object detection, traditional algorithms were mainly based on filters and wavelet transforms to achieve single-frame and multi-frame detection [2]. Some commonly used techniques include median filtering, high-pass filtering, wavelet transform, and threshold segmentation. Yuan Shuai et al. [3] proposed a method to separate the target from the background by comparing the difference between the target area and the inner and outer double-layer neighborhoods, so as to enhance the local contrast of bright and weak small targets and effectively suppress complex background noise. Liu Delian et al. [4] proposed the concept of stagnation point connection, which was used as a benchmark to calculate the difference between the gray scale of each pixel and the datum to determine the target position. The improvement of these single-frame and multi-frame algorithms has the advantages of relatively simple computation and low complexity, but the detection performance is limited when the target contrast feature is not obvious in complex and changeable real scenes.

With the development of deep learning, effective and non-traditional solutions have been introduced to solve complex problems in the field of computer vision. At present, deep learning networks are mainly divided into two types in object detection: two-stage object detection algorithms (represented by R-CNN series) and single-stage object detection algorithms (represented by SSD and YOLO series). Both algorithms rely on deep convolutional neural networks (CNNs) to learn high-level features of images, capture semantic information in images, and use multi-scale strategies to detect targets at different resolutions, thereby improving detection performance [5]. Li Mukai et al. [6] introduced SEblock based on the idea of calibrating features according to weights in SENet, and improved the accuracy of YOLOv7 for small target detection to 83.97%. In view of the problems existing in the infrared image itself, Jiang Zhixin et al. [7] chose to combine the histogram equalization with the MSR shown in the image preprocessing to enhance the image, and at the same time, the loss function of the Faster-CNN network was improved, and the mAP was improved by 6.11% compared with the original network. However, these deep learning algorithms based on prior boxes still have certain difficulties in processing small targets in images, especially for small targets that are scattered and do not overlap or are occluded. The introduction of attention mechanism to improve the backbone network alleviates this problem to a certain extent, but there are still limitations of transformer decoder in feature representation. Therefore, the DETR network based on transformer architecture can effectively improve the detection ability of small targets in complex backgrounds through global modeling capabilities and encoded location information.

III. Algorithms Model

A. RT-DETR network

RT-DETR is the first real-time object detector in the DETR series and consists of four models. Different networks use different backbones, among which rtdetr-l uses HGNetv2-l as the backbone network, which has the best performance under the same conditions with fewer parameters and computational costs. The network structure is shown in Figure 1. The RT-DETR-L network structure consists of three parts: Backbone, Neck, and Decoder. The backbone network HGNetv2 consists of four HG Stage modules, each of which is mainly composed of HG Blocks. The backbone network extracts feature maps of three scales at different levels as the input of the hybrid encoder. The hybrid encoder is composed of two modules, the AIFI encoder and the CCFM feature fusion, in which the AIFI is still a multi-head transformer in essence, and only the deepest S5 feature layer is processed, and the F5 feature layer is finally output, which is used as the input of the CCFM module together with the S3 and S4 features, and the upper feature fusion and the lower feature fusion are carried out twice, respectively. For the fusion result, the anchor frame is selected through IOU-ware Query Selection, and the top-k300 is finally selected as the input into the decoder for prediction output.

B. Image preprocessing

In view of the problems of low contrast, high noise and unclear details in infrared images, the detection performance of the model can be improved by selecting an appropriate algorithm for image enhancement before the model is processed.

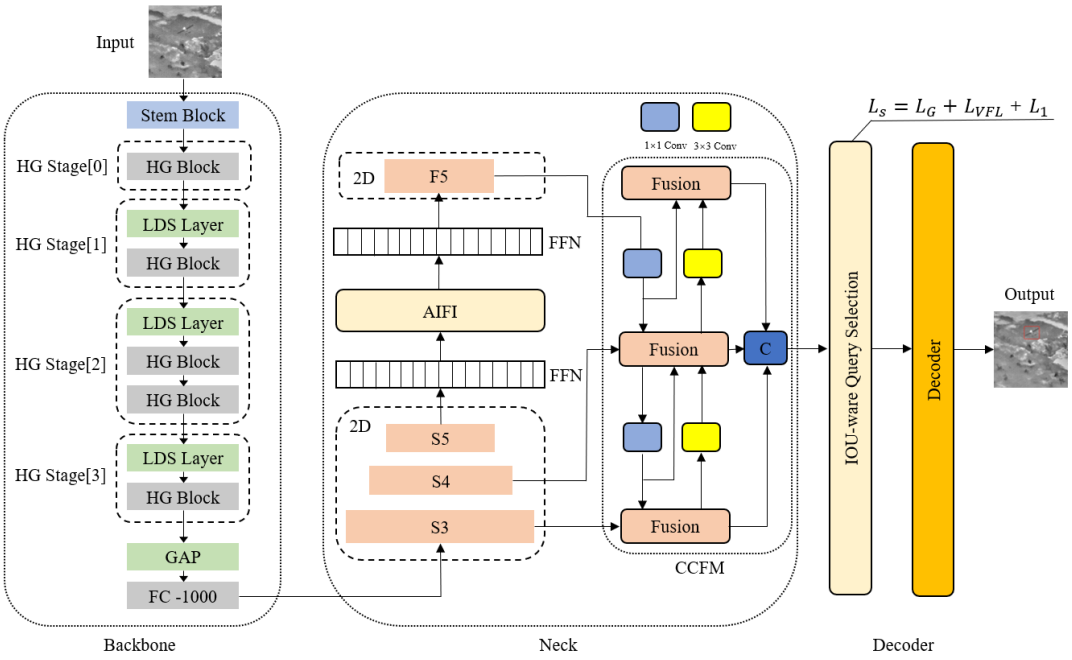

Figure 1. RT-DETR network structure

Dynamic Detail Enhancement (DDE) [8]and MSR algorithm are two commonly used image enhancement techniques in the preprocessing of infrared images. DDE separates the base layer and the detail layer of the image through the filtering algorithm, and enhances the detail layer to achieve significant enhancement of the details. However, the brightness and contrast of the base layer are limited, and it is easy to amplify the noise while amplifying the details. The multi-scale MSR algorithm improves the brightness and global contrast of the image through digital transformation and smoothing. However, the processing effect at the level of detail is limited, in complex scenes, the improvement of local contrast is not obvious. Therefore, the combination of DDE and MSR algorithm can dynamically adjust and enhance the detail and clarity of the image according to the complexity of the image content.

The detailed design process of the algorithm is as follows.

Bilateral filtering is used to decompose the original image into basic component I_B, and the output is:

$$I_B = \frac{1}{W_q}\sum_{p\in S} G_S(p) * G_r(p) * I_p \tag{1}$$

where the I_p represents the original infrared image, the G_S represents the pixel value weight, the G_r represents the spatial distance weight, and the W_q represents the sum of the weights of each pixel value in the filter window, which is used for the normalization of the weights.

The detail component is the result of subtracting the base component from the original image， The formula is shown in the following formula.

$$I_D = I_p - I_D \tag{2}$$

The basic component contains the large-scale structure and lighting information of the image, and the local details are enhanced by MSR for the basic component R_B 。

$$R_B = MSR(I_B) \tag{3}$$

Finally, the detail component is recombined with the enhanced base component to obtain the final infrared image.

C. Attention mechanisms

In order to solve the problem that small and medium-sized targets are difficult to detect in infrared image object detection, the original rt-detr-l uses HGNetv2, which is essentially still a CNN network structure, and uses HGBlock to extract features through deep convolution operations. Therefore, an improved RT-DETR network model is designed, and the EMA attention mechanism module is added to the backbone network of RT-DETR, so that it can extract multi-scale feature information that is rich in global context information and differentiated features, especially small target feature information.

Figure 2 shows how to add an EMA module.

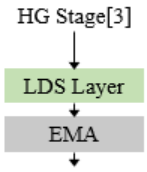

Figure 2. EMA module

At the same time, in order to suppress the interference of background and noise in the infrared image, the attention of small targets is improved. The introduction of CAMixing convolution-attention module enables CCFM to suppress the interference of irrelevant information and improve the denoising performance when multi-scale feature fusion, which is conducive to enhancing the modeling of global and local features and improving the detection rate of small targets.

Add the CAMixing module to the network, as shown in Figure 3.

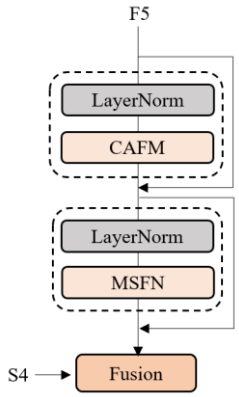

Figure 3. CAMixing module

In RT-DETR, the encoder is only used in S5, and comparative experiments are used to verify that it not only helps to significantly reduce the computational effort and improve the computational speed, but also does not cause significant damage to the performance of the model [9]. Therefore, the addition of CAMixing to the two-way downward feature fusion with S5 helps to improve the detection accuracy and reduce the amount of computation.

D. Loss Function

The original IOU-ware Query Selection uses GIoU to optimize the query selection process of the prediction box, and introduces an external rectangle to reduce the loss between the real box and the prediction box of the large target. However, for small target detection, a slight shift in the target may cause a significant change in its position information. GIoU does not directly consider the aspect ratio difference between the prediction box and the real box, and cannot fully capture the subtle changes in position information. Therefore, Shape-IoU is used instead of GIOU for small targets, and the loss is calculated by paying attention to the shape and scale of the bounding box itself, so as to optimize the detection accuracy of small targets. The regression loss calculated by Shape-IoU is shown in the following formula.

$$L_S = 1 - IOU + D^S + 0.5\Omega^S \tag{4}$$

Among them, IOU is the intersection and union ratio, S is the zoom factor, and its value is related to the size and number of targets in the dataset. and, D^S is the distance loss, which is calculated as follows:

$$ww = \frac{2 \times (w^{gt})^s}{(w^{gt})^s + (h^{gt})^s} \tag{5}$$

$$hh = \frac{2 \times (h^{gt})^s}{(w^{gt})^s + (h^{gt})^s} \tag{6}$$

$$D^S = hh \times \frac{(x_c - x_c^{gt})^s}{c^2} + ww \times \frac{(y_c - y_c^{gt})^s}{c^2} \tag{7}$$

Where w^{gt}, h^{gt} are the width and height of the GT box, x_c, y_c, x_c^{gt}, y_c^{gt} are the coordinates of the center point of the prior box and the GT box. ww and hh are calculated from the coordinates as the weight coefficients of the horizontal and vertical directions, and the loss of D^S can be adjusted by adjusting the value of s.

Ω^S is the loss of shape. It follows the formula of SIOU:

$$\Omega^S = \sum_{t=w,h} (1 - e^{-w_t})^{\theta} \tag{8}$$

Where w,h are the width and height of the prior frame, respectively, θ determines the size of the shape loss, and in order to avoid the shape loss accounting for a relatively heavy proportion of the overall loss and the position change of the prediction frame, the genetic algorithm takes the value of 4.

In the case of an infrared small target image, it is mainly composed of the background, and only a small part is occupied by the target. It is easier to learn the features of the background than the features of the target during training. Therefore, the ATFL loss function is used to replace the original VFL classification loss, which decouples the target from the background, and uses the adaptive mechanism to adjust the loss weight, forcing the model to allocate more attention to the small target features. The ATFL expression is as follows:

$$\begin{cases} -(\lambda - p_t)^{-\ln(p_t)} \log(p_t) & p_t \le 0.5 \\ -(1 - p_t)^{-\ln(p_t)} \log(p_t) & p_t > 0.5 \end{cases} \tag{9}$$

Where p_t represents the current average prediction probability value, and λ (>1) is the hyperparameter. ATFL is balanced by improving the adaptive factor in TFL $\gamma - \ln(p_t)$ .The weights of the samples [10] increase the contribution of small targets while reducing the time consumption of adjusting hyperparameters.

IV. EXPERIMENTS

A. Experimental Environment

Table I shows the experimental environment in this paper, which is based on the Ubuntu 18.04 operating system, the graphics card model is RTX2080Ti, and the memory is 16GB. The experiment basically uses the parameters recommended by RT-DETR, builds the model based on Python3 and Pytorch framework, and uses the standard SGD optimizer, with batch-size set to 8 and epochs set to 100.

TABLE I. EXPERIMENTAL ENVIRONMENT

Experimental environment	Version
CPU	IntelCorei7-11800H
GPU	NVIDIA GeForce RTX2080 Ti
Language	Python3.8
Deep Learning Framework	Pytorch1.14.0
CUDA	11.8.0

B. Dataset

The infrared aircraft small target dataset [11] used in this experiment includes a total of 22 annotated data folders, and the image content is mainly based on the ground background, sky background, multiple aircraft, aircraft distance, aircraft approaching, etc. A total of 12177 infrared images were selected, with an image resolution of

256*256, a channel count of 1, and a bit depth of 24. This dataset is widely used in tasks such as object detection and target tracking.

C. Evaluation Metrics

In this experiment, Precision (P), Recall (R) and Average Precision (AP) are mainly used as network evaluation indicators, and their mathematical expressions are as follows:

$$precision = \frac{TP}{TP + FN} \tag{10}$$

Where TP stands for True Positives, FP stands for False Positives, and Precision measures the proportion of instances that the model predicts to be positive. Used to evaluate the accuracy of instances predicted to be positive samples.

Recall, also known as recall, is the proportion of the sample predicted as positive to the predicted sample, and its mathematical expression is:

$$recall = \frac{TP}{TP + TN} \tag{11}$$

The mathematical expression for the average precision (AP) is:

$$AP = \int P(R) d_R \tag{12}$$

AP is used to evaluate the Precision-Recall curves of the model at different thresholds, and is the average value obtained by integrating the Precision-Recall curves. It measures the average accuracy of the model's predictions at different thresholds.

In addition, in order to evaluate the processing effect of the image preprocessing algorithm, PSNR and SSIM similarity were used to distinguish the similarity of the images before and after processing, the information entropy (Entropy) was used to reflect the retention degree of image information, and the average gradient (AG) and edge intensity (EME) were used to evaluate the clarity of the image.

D. Algorithm verification results

The objective evaluation method is used to evaluate the quality of the improved images, and the effectiveness of the improvement is verified compared with other algorithms.

The final enhancement is shown in Figure 4.

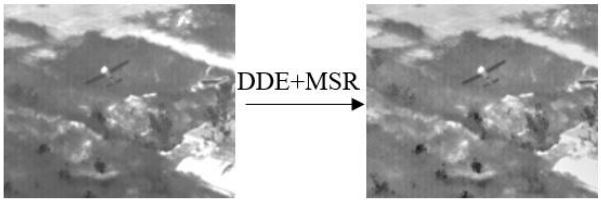

Figure 4. Image enhancement effect

Table II lists the verification results.

E. Shape-IoU parameter validation

In this paper, the Shape-IoU loss function is introduced into the network model [12]. which has a scale parameter whose size can be determined by optimizing the parameters, and the s parameter is finally set to 0.4 by comparing different parameters on the dataset. Comparative experiments are shown in Table III.

TABLE II. COMPARISON OF ALGORITHM ENHANCEMENTS

index / algorithm	PSNR	SSIM	Entropy	AG	EME
Original image			6.2850	41.9135	2.6388
SSR	28.2970	0.85522	5.7150	44.2675	2.8967
MSR	28.7772	0.8676	6.2176	44.9135	2.8932
DDE	36.0989	0.9679	6.4594	44.2206	2.6857
Bilateral filtering	34.6621	0.8395	6.3155	21.7407	1.4891
DDE+MSR	28.7581	0.8436	6.2793	46.8969	2.8882

TABLE III. COMPARATIVE EXPERIMENTS OF DIFFERENT PARAMETERS OF SHAPE-IOU

s	0.1	0.2	0.3	0.4	0.5	0.6	1.0
mAP(%)	83.5	83.4	84.9	**85.3**	84.8	83.5	83.4

F. Network comparison experiment

In order to verify the effectiveness of each improvement point of the network, based on the RT-DETR-L network, six sets of comparative experiments were carried out on the dataset. and the environment and parameter settings were uniform. The experimental results are shown in Table IV, and "√" indicates that the corresponding method was used. It can be seen from Table 4 that after adding the EMA and CAMixing modules and improving the loss function, the algorithm achieves the best detection accuracy, and the AP value is 3.2% higher than the original RT-DETR, which proves the effectiveness of the improved module in this paper. Figure 5 shows the AP variation curve of the improved method more intuitively. The detection fluctuates greatly due to the influence of the dataset, but the improved network detection accuracy is significantly stable and the accuracy increases. Figure 6 shows the detection results of the original network and the improved network, and the prediction score increases significantly, which verifies the effectiveness of the improved method and proves that the improved method is better for the detection of dense targets.

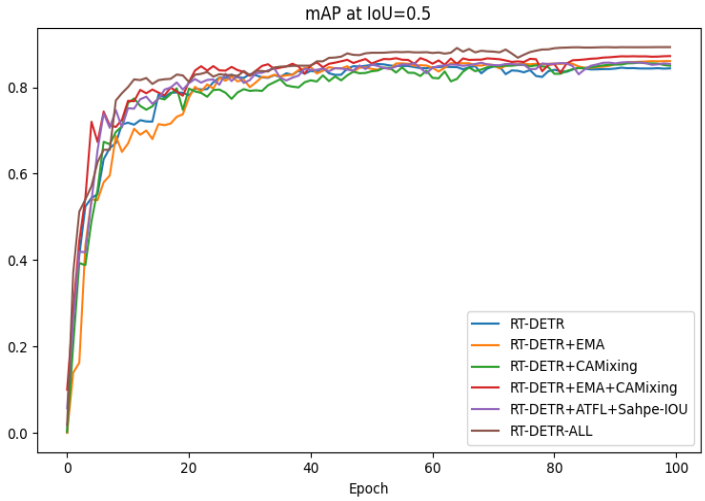

Figure 5. AP change curve

TABLE IV. COMPARES THE EXPERIMENTAL RESULTS

EMA	CMAixing	Shape-IoU	ATFL	P/%	R/%	AP/%	Param/10^6
				73.2	75.2	84.6	32.81
√				72.2	76.3	85.5	33.40
	√			74.8	75.2	85.6	34.97
√	√			74.1	75.0	86.2	35.23
		√	√	75.5	76.2	85.9	32.81
√	√	√	√	75.4	77.1	87.8	35.23

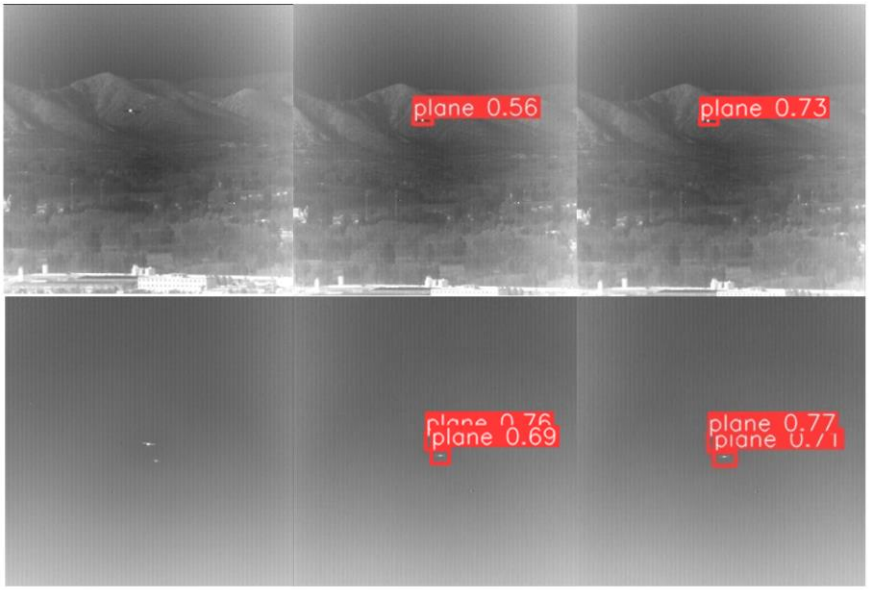

Figure 6. Improved detection results

V. Conclusion

In order to solve the challenge of target detection in infrared imaging scenes, this paper effectively enhances the contrast and detail visibility of infrared images by combining DDE and MSR algorithms, improves the RT-DETR network structure, introduces EMA attention mechanism and CAMixing convolutional attention module, and significantly improves the model's detection ability and overall detection accuracy of small targets. At the same time, the Shape-IoU and ATFL loss functions are combined to improve the regression ability of small targets under infrared conditions. Experimental results show that the improved algorithm is better than the original DETR algorithm in detecting infrared weak and small targets in complex backgrounds, and the mean average accuracy (mAP) is increased by 3.2%, showing good robustness and adaptability. However, considering the richer training data, especially the infrared images containing different weather, time and terrain conditions, it is necessary to further process the contrast between the target and the background to enhance the generalization ability of the model. In addition, real-time performance is very important in practical applications, and the detection speed can be improved by optimizing the model structure and algorithm, making it better suitable for real-time infrared object detection scenarios.

References

[1] Guo Yujie. Research on micro target detection and recognition method based on information enhancement [D]. Guangdong Technical Normal University, 2023.

[2] Liu Ying, Sun Haijiang, Zhao Yongxian. Research on infrared weak and small target detection method in complex background based on attention mechanism [J]. Computer Software and Computer Applications, 2023, 38(11):1455-1467.

[3] Yuan Shuai, Yan Xiang, Zhang Yugeng, et al. Infrared and Laser Engineering, 2022, 51(4):20221071.)

[4] Liu D, Zhang J, Dong W. Temporal profile based smallmoving target detection algorithm in infrared image sequences [J]. International Journal of Infrared and Milli-meter Waves, 2020, 28(5):373-381.

[5] LI B Y, XIAO C, WANG L G, et al. Dense nested attention network for infrared small target detection [J]. IEEE Transactions on Image Processing, 2023, 32:1745-1758.

[6] Li Mukai. Research on small-scale infrared pedestrian detection technology based on deep learning [D]. Shanghai:University of Chinese Academy of Sciences,Shanghai Institute of Technical Physics, Chinese Academy of Sciences, 2022.

[7] Jiang Zhixin. Research on infrared small target detection method at sea based on deep learning [D]. Dalian: Dalian Maritime University, 2019.

[8] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: faster and better learning for bounding box regression [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020:12993-13000.

[9] LIU X, PENG H, ZHENG N, et al. Efficientvit: Memory efficient vision transformer with cascaded attention [C]//Proceedings of group the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14420-14430.

[10] LI Y, HOU Q, ZHENG Z, et al. Large selective kernel network for remote sensing object detection [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 16794-16805.

[11] HUI B W, SONG ZY, FAN HO, et al. A dataset for infrared image dim-small aircraft target detection and track-ing under ground air background [J]. Scientific Database, Chinese Acdemy of Science, 2020, 5(3):291-302

[12] Zhou Mengran, Wang Ao. Object Detection Algorithm for Lightweight Remote Sensing Image Based on DETR[J/OL]. Journal of Chongqing Technology and Business University (Natural Science Edition). https://link.cnki.net/urlid/50.1155.N.20240328.1703.004.