Journal of Computers Vol. 35 No. 1, February 2024, pp. 145-157
doi: 10.53106/199115992024023501010

Small Object Detection in Remote Sensing Based on Contextual Information and Attention

Hua-Ping Zhou, Jie Zhang*, Ke-Lei Sun, Qiu-Fen Wen, Qi Zhao, Ying-Jie Guo

Anhui University of Science and Technology, School of Computer Science and Engineering, Huainan City, Anhui, China
13805549155@163.com, zj17355482695@163.com, klsun@aust.edu.cn

Received 16 June 2023; Revised 9 November 2023; Accepted 5 December 2023

Abstract. Many small objects, for instance vehicles and small ships, are encountered in remotely sensed images. However, small object detection has been a challenging task in remote sensing because of the problem that small objects are easily missed and influenced by the background. To address this challenge, we propose a detection method based on contextual information and attention, divided into two main parts. Firstly, for purpose of further improve the backbone network features to derive more contextual information, a multi-branch feature enhancement module is constructed to fuse multiple sensory field features to improve the ability of the backbone network to extract feature information; secondly, a new effective channel attention mechanism is proposed to reduce problems such as information confusion caused by the feature fusion process, thus reducing the influence of the background. Compared with other methods, it effectively improves the detection of small object among remote sensing images.

Keywords: small object detection, contextual information, feature enhancement, channel attention

1 Introduction

In the last few years, with the continuous progress of deep learning technology, remote sensing images object detection has gradually developed, and it is now one of the important research directions in the domain of computer vision, which has enormous application usefulness in the warlike, manufacturing, security and other sphere [1]. There are many small objects commonly found in remote sensing images. Compared with conventional object, these small objects are not easy to detect from similar backgrounds or connected object due to their small amount of information and weak features, so small object detection in remote sensing images has been a troublesome problem.

With the continuous development and application of deep learning in the field of computer vision, object detection algorithms are primarily separated into two classes, the first is the two-stage detection algorithm which is famous for detection accuracy, such as Faster RCNN [2], Mask RCNN [3], and the second is the single-stage detection algorithm which is famous for detection speed, such as SSD [4], YOLOv5 [5]. However, these superb object detection algorithms have significant hurdles when it comes to detecting small object. To address the issue of challenging small object recognition, researchers have begun to focus on the field of small object and have presented numerous outstanding approaches for small object detection.

In PANet [6], Liu et al. perform a secondary fusion of feature maps of different size scales among FPNs, which has the advantage that the FPN can also contain rich low-level features and high-level features for high-level feature maps. Lim et al. [7] arranged a contextual attention mechanism that acts to focus more on the object in the image as well as the contextual information from the object layer, which can effectively reduce the effect of noise in shallow networks, thus approving the network to concentrate more to small objects. Wang et al. [8] enhanced the SSD network with contextual information and attention mechanisms processes to increase the model ability to identify small objects. Jing et al. [9] devised an optimisation model for deep neural networks that refines the detection effect of small-size objects to some extent by reanalysis the construction of the original input data, taking into consideration about characteristics of the objects in images. Tan et al. [10] proposed the BiFPN framework in the Efficient Det meshwork, which consists of a weighted ovonic feature pyramid network with gained cross-scale connectivity to cement feature representation, adding individual weights for apiece input to preferably accomplish small object detection tasks.

* Corresponding Author

Refine feature pyramid network [11] solves the confusion effect about feature pyramids during upsampling and improves small object detection performance of optical remote sensing images by improving the building blocks and adding a constant mapping between the inputs and outputs of the same layer. Yan et al. [12] effectively improved the detection of remotely sensed weak objects through cross-level channel fusion and increased positional attention. Gong et al. [13] introduced the self-attentive mechanism into YOLOv5 to deal with the problem of the dense distribution of small objects, which enhanced the computational effort to a certain extent but the detection results were not bad.

The above work has made improvements in the enhancement of small object detection in both public and remotely sensed datasets, and although some results have been achieved, there is still a lot of room for improvement

In this paper, we propose a remote sensing small object detection method based on contextual information and attention, which can better detect remote sensing of small objects.

The main contributions of this paper are summarized as follows:

(1) Based on the Faster-RCNN architecture and use fusion ResNet and improved feature pyramid network as feature extraction network, which can enhance the ability of models to extract features from object information.

(2) A new feature enhancement module, which is constructed to improve the sensory field of the network as well as its sensitivity to small objects in remote sensing images.

(3) The problem of loss of channel information present in the feature pyramid network is solved by fusing the low-level feature maps output from the feature pyramid and weighting them with the output feature maps. An attention mechanism is also introduced to reduce the feature redundancy problem caused by multiple feature fusion, further enhancing the feature extraction capability of the model.

2 Related Models

2.1 Faster-RCNN

Faster-RCNN is an end-to-end two-stage object detection algorithm based on candidate regions, which builds on Fast-RCNN by adding a new Region Proposal Network (RPN), using a sliding window approach to generate anchor frames with corresponding aspect ratios for each feature region, then it is to output the categories of anchor frames separately and predict the bounding boxes, and finally use a non-great suppression algorithm to The prediction results are selected to obtain the desired candidate regions.

The Faster-RCNN network model consists of a feature extraction structure, a region proposal network (RPN), a ROI Pooling layer, and a classification regression layer. As shown in Fig. 1, the main process of the algorithm is to input the image to be detected, obtain the required feature maps, then beam the feature maps to the RPN to generate pre-selected boxes, pass both the pre-selected boxes and the feature map to the ROI Pooling layer, finally select the candidate boxes that best match the feature map from the pre-selected boxes, and finally send the candidate boxes to the classification regression layer to obtain the output category and regression parameters.

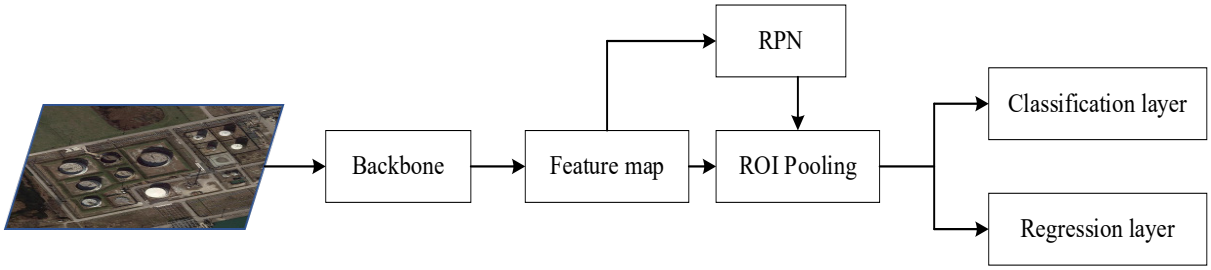

Fig. 1. Faster-RCNN model structure

The loss function for Faster-RCNN training is:

$$L\left(\{p_i\},\{t_i\}\right) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) , \tag{1}$$

The loss function of Faster-RCNN composes two functions, regression loss, and classification loss, and the regression loss can be expressed as:

$$L_{reg}(t_i,t_i^*) = smooth_{L1}(t_i - t_i^*) , \tag{2}$$

Classified losses are expressed as:

$$L_{cls}(p_i,p_i^*) = -\log\left[p_i^* p_i + (1-p_i^*)(1-p_i)\right] , \tag{3}$$

where i denotes the number of the anchor box, pi denotes the possibility of having an object in the anchor box, t_i denotes the predicted bounding box parameter, t_i* signifies the true bounding box parameter, pi* is used to distinguish whether there is an object in the anchor box, 1 is yes, 0 is no, N_{cls} and N_{reg} mean the number of classification and regression respectively, and the λ parameter values for balancing the classification loss and regression loss.

2.2 Feature Pyramid Network

Feature pyramid network is a mainstream framework that includes top-down, bottom-up, and horizontal linking operations, as shown in Fig. 2. The bottom-up process is the forward transmission of information, while the top-down process is the up-sampling of higher-level feature maps that contain more semantic information, and then connects them horizontally with the feature maps on the left, thus enhancing more feature information. As each layer of the predicted feature map contains different resolution sizes and semantic information of multiple feature strengths, it is then possible to promote the detection precision of small objects without a significant increase in detection time.

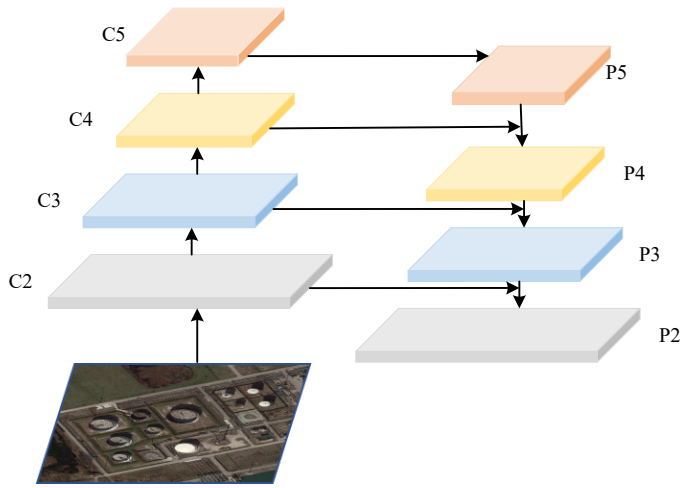

Fig. 2. Feature pyramid network structure

The backbone network of the promoted model uses ResNet50, as shown on the left side of Fig. 2, and the feature maps of several layers are obtained through convolution, noted as C2-C5. Top-down means that the feature maps owned by the higher-level feature maps are expanded to the same size as the next layer by using up-sampling, so that the information about the lower-level feature maps and the information about the higher-level feature maps are able to be combined together toward, more The laterally connected one can directly sum up P5 and C4 after downsampling, and finally eliminate the influence of information confusion brought by the upsampling process through 3x3 convolution, so as to obtain the final feature map.

3 Our Method

The framework of the feature extraction structure of the remote sensing small object method based on contextual information and attention used in this paper is shown in Fig. 3.

The steps of the algorithm are: (1) Feature extraction backbone network, since the feature extraction network of traditional Faster-RCNN is not very effective in detecting small objects, here it is replaced with ResNet50 with better feature extraction capability to enhance the detection effect. (2) Enhance the feature extraction backbone network, P5 is only obtained through C5 while the high-level feature pyramid network is mainly responsible for processing large and medium This makes the small object information easy to ignore leading to poor detection of small objects. The feature enhancement module MFEM is introduced, which consists of different scale convolutions and inflated convolutions to increase the width of the network and expand the receptive field to enhance the extraction of small objects by the network. (3) Fusing the low-level feature maps output from the feature pyramid network, weighting the output feature maps, and introducing the attention mechanism ECAM to reduce the information redundancy brought by multiple feature fusion and improve the network's detection effect on small objects.

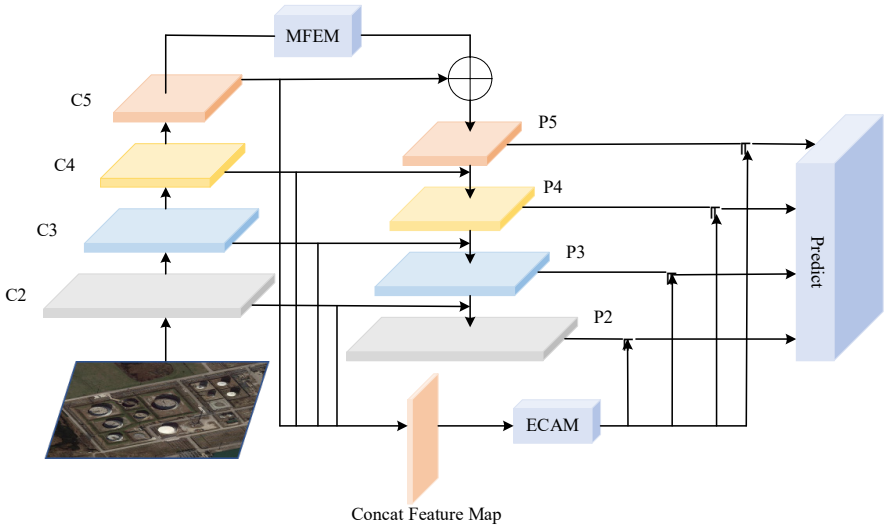

Fig. 3. Feature extraction network structure

3.1 Feature Enhancement Module

In the traditional feature pyramid network, the detection of small object usually uses the low-level feature images, due to the low-level feature images possess more useful information on location and semantics of small object. However some existing backbone networks are not very friendly to the extraction of small object information and cannot adequately process the feature images, resulting in limited sensory fields and less semantic information in the low-level feature images, which is unbeneficial to detection results of small objects.

To fully utilise global information, a new feature enhancement module, MFEM, was built to improve the perceptual field size of the network, thus further enhancing the extraction of information from small objects. The structure of MFEM is shown in Fig. 4. There are four branches in the structure The first three branches initially execute a 1x1 convolution operation on the input feature map to pre-process it and change the amount of features. For the branches with a void rate of 3 and 5, they are concatenated by regular convolution operations such as 1x3, 3x1, and so on to acquire better small object information features through different convolution operations, and the last three branches add a void convolution layer with a void rate of 1, 3, 5, respectively, so that the extracted feature map contains more contextual information; the fourth branch is sent through an adaptive pooling layer, which aims to obtain a global recptive field.

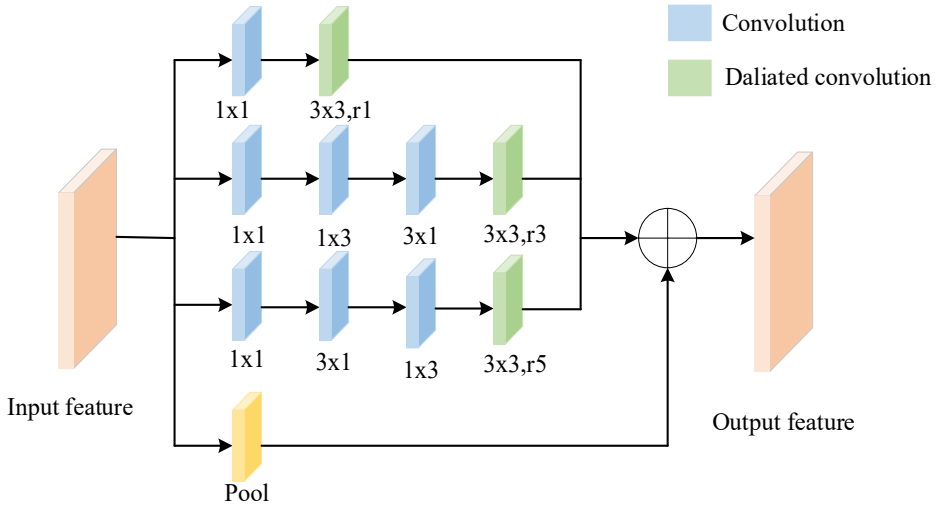

Fig. 4. MFEM structure

The context enhancement module MFEM formula can be expressed as:

$$X_1 = Disconv_{3\times3}[Conv(F)] \ , \tag{4}$$

$$X_2 = Disconv_{3\times3}\{Conv_{3\times1}\{Conv_{1\times3}[Conv_{1\times1}(F)]\}\} \ , \tag{5}$$

$$X_3 = Disconv_{3\times3}\{Conv_{1\times3}\{Conv_{3\times1}[Conv_{1\times1}(F)]\}\} \ , \tag{6}$$

$$X_4 = U\{Conv_{1\times1}[AdapPool(F)]\} \ , \tag{7}$$

$$Y = Conv_{1\times1}P\{Cat(X_1, X_2, X_3)\} + X_4 , \tag{8}$$

Where $Conv_{1x1}$, $Conv_{1x3}$, and $Conv_{3x1}$ represent normal convolution operations with convolution kernel sizes of 1x1,1x3,3x1 respectively, $Disconv_{3x3}$ represents the null convolution operation, F denotes the feature map at the beginning of the input, X_1, X_2, X_3 denotes the feature map after normal convolution and null convolution, X_4 denotes the feature map after global pooling operation AdapPool means global pooling operation, Cat means cascade operation of feature maps at different scales, + means bitwise summing operation of feature maps, U means upsampling, Y means feature maps after MFEM enhancement.

3.2 Channel Attention Module

In recent years, attention techniques have been popularly applied at domains, for instance, object detection and instance segmentation, and have demonstrated excellent performance. As a result, this paper also introduces the attention mechanism add to backbone network model, in order to improve the efficiency of small object detection. Due to the low target pixels in remote sensing images, in order to better detect small remote sensing objects and reduce the effect of information redundancy caused by multiple feature fusion. After fusing feature maps with the high-level information and the underlying information, a channel attention module is introduced to focus the model attention more on the channel aspects of the object region, so that the channel features in this part of the model can be better correlative. This reduces the impact of information redundancy on channel characteristics

The channel attention mechanism ECAM was used to alleviate the information confusion brought about by the feature fusion process, so as to better detect remote sensing small objects. As show in Fig. 5. Firstly, input feature maps were organized using maximum pooling and average pooling in several; maximum pooling can better extract feature information and enhance the acquisition of remote sensing small object information, while average pooling can well obtain the background information related to the object on the feature maps to prevent the interference of intricate background in remote sensing small object detection, after which two branches were subjected to bitwise summation operation. The local cross-channel interaction strategy and the one-dimensional convolution operation continue to be used to obtain the required channel weights to give more attention to the small object part.

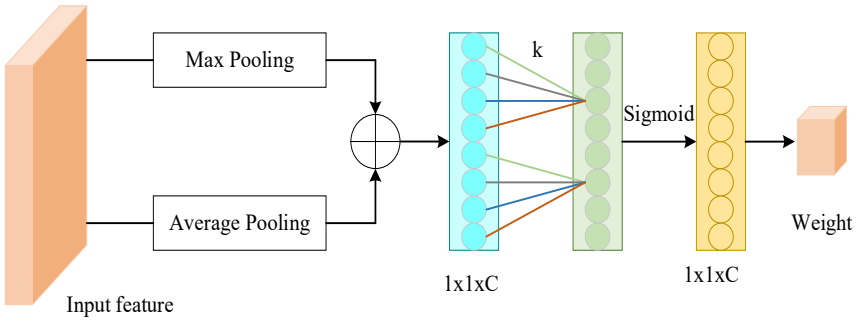

Fig. 5. ECAM structure

The ECAM formula for channel attention can be expressed as:

$$X' = Avgpool(X) + Maxpool(X) \ , \tag{9}$$

$$ECAM(X) = \ \sigma(CID_k(X')) \ , \tag{10}$$

$$R_i = ECAM(I) \odot P_i \ , \tag{11}$$

Avgpool bespeaks average pooling and Maxpool means maximum pooling, σ is the Sigmoid function, CID_k denote a one-dimensional convolution with kernel size k, ECAM () represents the channel attention function, i is the number of layers in the FPN, and R_i represents the output information after enhancing the feature pyramid network.

3.3 Data Enhancement

Objects in remote sensing images are generally smaller compared to natural images, and the dataset can be expanded using data augmentation to improve the generalisation ability of the model. In this paper, the samples in

the dataset are flipped left-right and vertically for data enhancement.

The expressions for left-right flip and vertical flip are as follows:

$$x'_{\max} = w - x_{\min} \ , \tag{12}$$

$$x'_{\min} = w - x_{\max} \ , \tag{13}$$

$$y'_{\min} = h - x_{\max} \ , \tag{14}$$

$$y'_{\max} = h - y_{\min} \ , \tag{15}$$

where w and h are the width and height of the image, $x_{\min}$ and $x_{\max}$ are the horizontal coordinates of the upper-left and lower-right corners of the labelled positions in the image, $x'_{\min}$ and $x'_{\max}$ are the horizontal coordinates of the upper-left and lower-right corners of the labelled position after the image is flipped, $y_{\min}$ and $y_{\max}$ are the vertical coordinates of the upper left and lower right corners of the labelled positions in the image, $y'_{\min}$ and $y'_{\max}$ are the vertical coordinates of the upper left and lower right corners of the image after the labelled positions are flipped.

4 Experiment and Result Analysis

4.1 Experiment Design

The PyTorch framework is used for code experiments in this paper; improved model is trained utilized stochastic gradient descent (SGD) as the optimizer The GPU model is NVIDIA GeForce RTX 3090, with 64G of RAM. The rest of the parameters are set in Table 1.

Table 1. Parameter selection

Parameter name	Parameter values
Epoch	20
Learning rate	0.01
Batch size	8
Momentum	0.9
Weight decay	0.0001

4.2 Dataset and Evaluation

The experiment used the public high-resolution remote sensing image object detection dataset HRRSD [14] to test the effectiveness of the modified approach. The HRRSD dataset contains a total of 13 item classes. Respectively Airplane (AE), Baseball Diamond (BD), Basketball Court (BC), Bridge (BE), Crossroad (CD), Ground Track Field (GTF), Harbor (HR), Parking Lot (PL), Ship (SP), Storage Tank (ST), T Junction (TJ), Tennis Court (TC), Vehicle (VE). The majority of which are portrayed as densely arranged small object in the images and are suitable for validating the proposed algorithm for detecting remotely sensed small objects. The trainval subset was chosen as the training set, the val subset as the validation set, and the test subset as the verifi-

cation set, with a total of 10818 training images, 5417 validation images, and 10943 test images.

The mAP (mean Average Precision) is widely used to object detection to measure the overall performance of a detector. The mAP is the average of all categories of AP (Average Precision). FPS (Frames Per Second) indicates how many images can be detected per second.

The definition of mAP in the context of object detection is:

$$mAP = (AP_1 + AP_2 + AP_3 + \cdots + AP_k)/k \ , \tag{16}$$

For AP this is usually defined as:

$$AP = \int_1^0 p(r)dr \ , \tag{17}$$

The loss plot of the improved algorithm is shown in Fig. 6. When training reaches 20 epoches, the loss stabilises and the learning rate also tends to be close to zero, the model starts to converge.

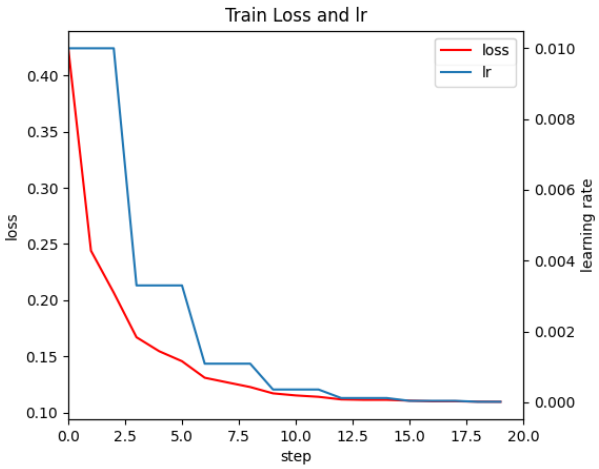

Fig. 6. The loss and lr curve of algorithm

4.3 Data Enhancement

Data augmentation methods have been diffusely used in various object detection models, in order to expand the dataset and enhance the number of training specimens to elevate the generalised and robustness of the model. The data set is simply flipped horizontally and vertically to account for the different sizes and orientations of the objects in the remote sensing dataset. In the cause of certify the influence of data enhancement on the experimental results, two methods of vertical and horizontal flipping are compared under the same conditions. The experimental results are presented in Table 2.

Table 2. Data enhancement comparison experiment

Methods	mAP@0.5/%	mAP@0.5:0.95/%	FPS
None	86.9	59.7	22.1
Vertical flip	88.5	60.9	21.2
Horizontal flip	88.9	61.1	21.0
All	89.8	62.3	20.6

The results in Table 2 show that both the horizontal and vertical flip methods can promote detection effect of the model to a certain region, the two data enhancement methods together provide a good improvement in accuracy, although the speed of detection is reduced a little, verifying the effectiveness of the data enhancement.

4.4 Attention Mechanism Selection

The purpose of introducing the attention mechanism is to reduce the redundancy of information brought about by multiple feature fusion and to prevent the remote sensing small object information from disappearing in the information conflict. In this experiment, three attention mechanisms, namely ECAM, ECA-Net and SENet, are selected to discuss the impacts of diverse attention mechanisms on detection results of the model under the same conditions of the laboratorial environment. The ultimate results are seen in Table 3.

Table 3. Results of the comparison of attentional mechanisms

Attention	mAP@0.5/%	mAP@0.5:0.95/%	FPS
ECAM	89.8	62.3	20.6
ECA-Net	89.4	61.2	20.2
SENet	89.1	61.1	20.3

It can be derived from Table 3, the ECAM attention mechanism has better accuracy values, and FPS is slightly better than ECA-Net and SENet, and under a comprehensive comparison, the more effective ECAM is selected in this paper.

4.5 Comparison of Feature Extraction Networks

The performance of feature extraction networks directly affects the accuracy of model detection and classification, In order to compare the impact of different feature extraction networks on the detection results of remote sensing small targets, VGG16 and ResNet50 were selected for comparative experiments in this experiment. The comparison results are shown in Table 4.

Table 4. Comparison of detection results of different feature extraction networks

Feature network	mAP@0.5/%	mAP@0.5:0.95/%
VGG16	81.5	56.6
ResNet50	82.7	58.4

From the results in Table 4, it can be seen that the detection effect of ResNet50 is superior to VGG16, reaching 82.7, which is 1.2% higher. ResNet50 can extract more feature information and improve the model's detection performance. Therefore, this article selects ResNet50 as the feature extraction network.

4.6 Experiment Result

To further validate the detection capability of the models in this paper, the comparison algorithms chosen were Faster RCNN, SSD, YOLOv5, CEFPN [15], YOLOX [16], Improved SSD [17], the results are presented in Table 5. Contrasted between the Faster-RCNN network, the improved algorithm model mAP improved by 8.3% compared to the previous one, and compared to the YOLOX network, mAP improved by 3.3%. The superiority, real-time, and robustness of the improved algorithm is demonstrated by comparing it with the traditional model algorithm as well as the latest algorithms.

So as to more directly reflect the impact of the improved model, the detection results of the 13 categories in the HRRSD dataset are shown in Table 6. Compared to the Faster-RCNN network, each category has a significant improvement, for example, the AP of Airplane is improved by 8%, the AP of Vehicle is improved by 13.1%, etc.; compared with YOLOX, the AP of Airplane Compared with YOLOX, the AP of Airplane improved by 7.9%, and the AP of Vehicle improved by 6.6%%, although detection accuracy of some categories was lower

than that of YOLOX, but the overall detection results were still better than it. Finally, compared with the latest Improved SSD algorithm, the AP of Basketball Court improved by 0.6%, the AP of Ship improved by 1.7%, the AP of Vehicle improved by 4.1%, etc. Among them, there are some categories whose detection results are lower than the latest improved SSD algorithm, and in general, the detection consequences of the improved model in this article are slightly better than it.

Table 5. Comparison of test results from varying models

Algorithm	Backbone	mAP@0.5/%	mAP@0.5:0.95/%
Faster-RCNN	VGG16	81.5	56.6
SSD	VGG16	80.8	48.9
YOLOv5	Darknet53	86.1	52.9
CEFPN	ResNet50	88.9	59.3
YOLOX	Darknet53	86.5	61.9
I-SSD	Darknet53	89.6	/
Ours	ResNet50	89.8	62.3

Table 6. Results of various types of detection by different models on the HRRSD dataset

Model	Faster-RCNN	SSD512	YOLOv5	CEFPN	YOLOX	I-SSD	Ours
AE	90.8	98.7	98.6	97.8	90.9	99.5	98.8
BD	86.9	83.0	88.1	89.1	85.0	96.2	90.9
BC	47.9	56.2	61.9	68.7	76.8	71.1	71.7
BE	85.5	83.3	85.4	91.0	90.1	91.9	92.0
CD	88.6	90.5	87.9	90.2	89.0	95.1	91.9
GTF	90.6	90.4	94.9	97.2	90.8	98.5	97.5
HR	89.4	85.0	92.3	95.4	98.7	96.4	95.7
SP	88.5	79.9	90.8	92.3	89.3	91.9	93.6
PL	63.3	43.5	63.6	70.0	68.7	58.5	71.0
ST	88.7	94.9	97.9	94.7	90.3	95.2	95.8
TJ	75.1	70.5	71.8	75.8	83.2	86.6	78.2
TC	80.7	88.4	95.6	92.0	90.5	93.8	94.0
VE	84.0	79.1	89.8	96.6	90.5	93.0	97.1
Mean AP	81.5	80.8	86.1	88.9	86.5	89.6	89.8

4.7 Ablation Experiment

Ablation experiments were carried out to prove the efficiency of two modules in the improved algorithm. We verified the model after adding MFEM alone, the model after adding ECAM alone, and the model after adding both modules in turn. As shown in Table 7, all experiments were tested on the HRRSD dataset. Faster RCNN was used as the baseline, the backbone of the model is a ResNet50 with and a four layers FPN, and the corresponding improved modules were added separately to assess the detection performance about the model by computing metrics.

Table 7. Results of various types of detection by different models on the HRRSD dataset

MFEM	ECAM	mAP@0.5/%	mAP@0.5:0.95/%	FPS
		86.4	56.6	22.7
√		88.1	61.0	21.7
	√	88.2	61.1	22.5
√	√	89.8	62.3	20.6

Compared to the detection results of the conventional model, the model after replacing the feature extraction network achieves a value of 86.4% detection accuracy; the addition of the feature enhancement module MFEM improved the detection performance of the model by 1.7%, indicating that the feature enhancement module enriches the information that can be taken from the feature map and further enhances the semantic information; The introduction of the attention mechanism ECAM improves the detection performance by 1.8%, suggesting that lead into the attention mechanism after fusion of feature maps can lessen the impact of information confusion, thus effectively promote the detection performance of the model; when the feature enhancement and attention mechanisms are used together, the model ameliorates the detection performance by 3.4%. The validity of improved model was further verified.

4.8 Images Visualization

This study completes a qualitative evaluation of two separate sets of situations including different categories to validate the effectiveness of the improved model more visually and clearly. Fig. 7 depicts the detection results of the Baseline algorithm, and the suggested method for small objects, where from left to right, the baseline detection results, the detection outcomes of this work are shown. It can be seen that when the remote sensing small objects are dense and affected by the backdrops both the baseline network have more or less problem of incorrect picking and missing detection. The improved model in this paper enhances the feature extraction information and at the same time reduces the effect of information redundancy brought about by multiple feature fusion, which significantly reduces the cases of missed detection and wrong pickup, thus improving the detection accuracy of the model for remote sensing small objects.

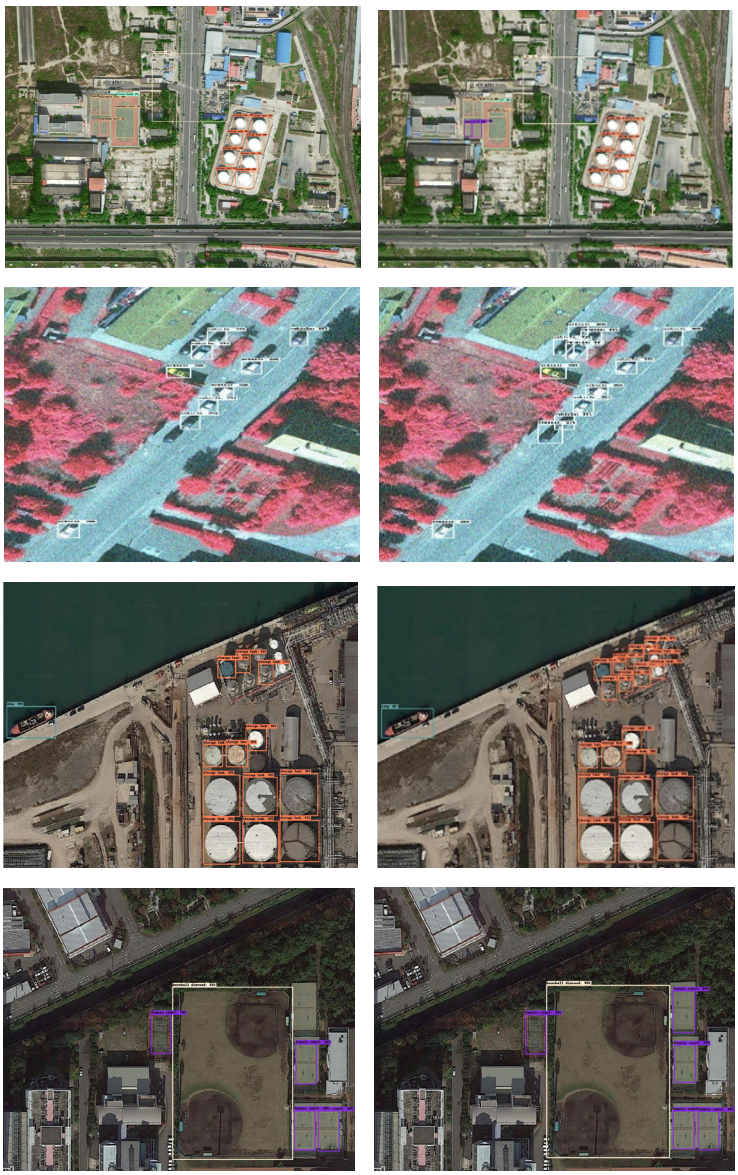

Fig. 7. Experimental visualisation results

5 Conclusion

Aiming at the problems that small objects in remote sensing images carry less information and it is more difficult to detect the objects, a small object detection method based on contextual information and attention is proposed. The feature pyramid is improved by adding a feature enhancement module to enhance the feature extraction capability of the model, and the attention mechanism is introduced to reduce the information redundancy caused by multiple fusion, and processing the data enhancement of remote sensing images to enhance the detection effect of the model on small objects. The experimental results show that the improved model has high detection accuracy and can meet the real-time detection.

Remote sensing small object detection is widely used in military, urban construction and other fields, this model mainly improves the feature extraction backbone network, strengthens the extraction of object information and reduces the impact of information redundancy, and does not improve the remote sensing small object detection in the direction of localisation difficulties, etc.; in the future work, research will be carried out on the detection frame of the existing model and the other features of the remote sensing small objects to improve the detection effect.

6 Acknowledgement

This research was supported by the Anhui Province Key R&D Program Special Project for International Science and Technology Cooperation NO. 202004b11020029.

References

[1] K. Tong, Y.Q. Wu, F. Zhou, Recent Advances in Small Object Detection Based on Deep Learning: A review, Image and Vision Computing 97(2020) 103910.

[2] S.-Q. Ren, K.-M. He, G.-S. Ross, J. San, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6)(2017) 1137-1149.

[3] H.-M. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proc. IEEE International Conference on Computer Vision (ICCV), 2017.

[4] G.-H. Yang, F. Wei, J.-T. Jin, X.-H. Li, G.-C. Gui, W.-J. Wang, Face Mask Recognition System with YOLOV5 Based on Image Recognition, in: Proc. International Conference on Computer and Communications (ICCC), 2020.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.-C. Berg, SSD: Single Shot Multibox Detector, in: Proc. European Conference on Computer Vision (ECCV), 2016.

[6] L. Shu, Q. Lu, H.F. Qin, J.P. Shi, J.Y. Jia, Path Aggregation Network for Instance Segmentation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[7] J.-S. Lim, M. Astrid, H.-J. Yoon, S.-L. Lee, Small Object Detection Using Context and Attention, in: Proc. International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021.

[8] Y.-N. Wang, X.-L. Wang, Remote Sensing Image Target Detection Model Based on Attention and Feature Fusion, Laser & Optoelectronics Progress 58(2)(2021) 0228003.

[9] S.-L. Jiang, W. Yao, M.-S. Wong, G. Li, Z.-H. Hong, T.-Y. Kuc, X. Tong, An Optimized Deep Neural Network Detecting Small and Narrow Rectangular Objects in Google Earth Images, in: Proc. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS) 13(2020) 1068-1081.

[10] M.-X. Tan, R.-M. Pang, Q.-V. Le, EfficientDet: Scalable and Efficient Object Detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[11] Y.-Y. Li, Q. Huang, X. Pei, L.-C. Jiao, R.-H. Shang, RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images, Remote Sensing 12(3)(2020) 389.

[12] J.-H. Yan, K. Zhang, T.-J. Shi, G.-Y. Zhu, Y. Liu, Y. Zhang, Multi-level Feature Fusion Based Dim Small Ground Target Detection in Remote Sensing Images, Chinese Journal of Scientific Instrument 43(3)(2022) 221-229.

[13] H. Gong, T.-K. Mu, Q.-X. Li, H.-S. Dai, C.-L. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, H. Li, X. Lang, Z. Li, B. Wang, Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images, Remote Sensing 14(12)(2022) 2861.

[14] Y.-L. Zhang, Y. Yuan, Y.-C. Feng, X.-Q. Lu, Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection, IEEE Transactions on Geoscience and Remote Sensing 57(8)(2019) 5535-5548.

[15] Y.-H. Luo, X. Cao, J.-T. Zhang, J.-J. Guo, H. Shen, T.-J. Wang, Q. Feng, CE-FPN: Enhancing Channel Information for Object Detection, Multimedia Tools and Applications 81(21)(2022) 30685-30704.
[16] Z. Ge, S.-T. Liu, F, Wang, Z.-M. Li, J. Sun, YOLOX: Exceeding YOLO Series in 2021. <https://arxiv.org/abs/2107.08430>, 2021.
[17] S. Wu, F. Zhou, Small Target Detection Based on Improved SSD Algorithm, Computer Engineering 49(7)(2023) 179-188.