Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

Remote Sensing Target Detection Algorithm based on CBAM-YOLOv5

Liang Bai [1], Xuewen Ding [1, 2], Limei Chang [1]

[1] College of Electronic Engineering, Tianjin Polytechnic Normal University, Tianjin 300222, China
[2] Tianjin Yunzhitong Technology Co., LTD., Tianjin 300350, China

Abstract: With the continuous improvement of traditional target detection algorithms, remote sensing targets have become a research hotspot. Aiming at the problems of low recognition accuracy caused by small target size and high background complexity in remote sensing images taken from overhead view, this paper proposes a remote sensing aircraft detection algorithm based on CBAM-YOLOv5. By introducing the lightweight convolutional attention module CBAM module into the YOLOv5 network, the feature extraction capability of the algorithm is improved to solve the problem that the small-size remote sensing target has little or even lost information on the feature map after multiple downsampling operations. The mAP of the proposed algorithm reaches 93.1%, which is 1% higher than that of the original algorithm, and the recognition of small-size remote sensing targets has been significantly improved.

Keywords: YOLOV5n; Target Detection; CBAM; Remote Sensing Image.

1. Introduction

Remote sensing images are generally captured by remote sensing satellite or UAV aerial photography, and are formed after a series of image processing. With the rapid development of remote sensing satellite technology and image processing technology, the number of highly informative and high-resolution remote sensing image data is increasing day by day, making remote sensing target detection become a research hotspot in the field of remote sensing, and has been widely used in military security, rescue and search and other aspects. However, remote sensing images have many problems, such as unique viewing Angle, small target size and high background complexity, which make the recognition process of remote sensing images more difficult than that of ordinary images.

With the rapid development of deep learning, object detection based on neural network has become a research hotspot. Object detection algorithms based on convolutional neural networks in neural networks can be divided into two categories: Use R-CNN [1], Fast R-CNN [2], Faster R-CNN [3], Mask The two-stage object detection algorithm based on candidate region represented by R-CNN [4] and the single-stage object detection algorithm represented by SSD [5] and YOLO [6] series directly generates its class probability and position coordinate value on the object. Because the two-stage detection algorithm forms a pre-selection box to predict the proposed area, the detection accuracy is relatively high. At the same time, due to the complex structure, the training time is relatively long. The single-stage detection algorithm directly detects the target, which has simple structure and short training time, but slightly low precision.

For the problem of small target size in remote sensing images, Zhang et al. [7] will expand the size of the feature map by upsampling in the candidate region of Faster R-CNN to improve the detection effect of small and medium-sized targets in remote sensing images. Zhang et al. [8] cascaded multilevel features in YOLOv3 network by using depth-separable attention to improve the feature expression capability of the network, thus improving the detection accuracy of small and medium-sized vehicles in remote sensing images. For the problem of complex remote sensing background, Song et al. [9] proposed to add learning parameters to each layer to achieve the adaptive fusion of deep and shallow features of adjacent and non-adjacent layers, and combined with the attention mechanism to suppress the problem of false detection and missing detection caused by remote sensing background. Liu et al. [10] proposed a module with dual attention mechanism to improve the extraction ability of target center and edge feature center to reduce the interference of complex background. With the continuous update and improvement of YOLO series algorithms, the detection accuracy has been greatly improved. Therefore, based on YOLOv5n algorithm, this paper optimizes and improves the algorithm to improve the detection effect of small and medium-sized targets in remote sensing images.

2. YOLOv5 Algorithm Introduction

The YOLOv5 algorithm model can be divided into four versions according to the model size and complexity. According to the model size and model depth, the ranking is S, M, L and X from small to small. Among them, "S" and "M" models are suitable for small and medium model training, while "L" and "X" are suitable for large model training. In version 6.0, an "N" model between "S" model and "M" model is introduced to balance the training speed and detection performance. While the "N" model is faster than the "S" model, the detection performance is also better than the "M" model. Therefore, this paper chooses YOLOv5n as the basis for further research.

The YOLOv5n structure can be divided into four parts: Input, Backbone, Neck, and Head detection. The specific network structure is shown in Figure 1. Input: Mainly performs Mosiac data enhancement for images. Mosiac mainly splice four images by means of random cropping, flipping, color gamut transformation, etc., which not only enriches the diversity of data, but also saves the consumption of GPU video memory. Backbone: It mainly extracts features from input images and forms multi-level feature maps for use by later networks. C3 module with residual structure is used

to deepen network depth and improve feature extraction capability. The specific structure is shown in Figure 2(a). Neck: Multi-scale feature fusion is realized through up-down sampling of feature maps transmitted by Backbone. Neck mainly performs multi-scale feature fusion without deepening the network, so residual structure in C3 structure is removed, as shown in Figure 2(b). Head: Obtains high-level feature information from Backbone and Neck to predict the location and category of the target.

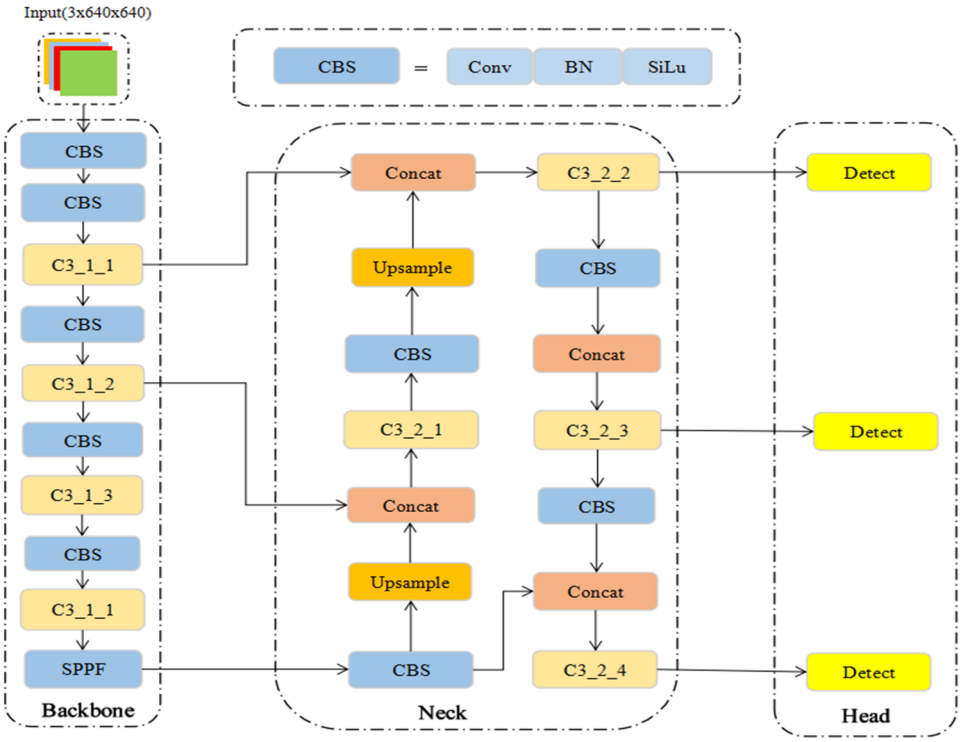

Figure 1. YOLOv5n Network Structure diagram

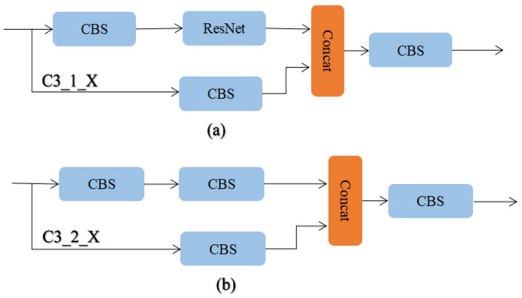

Figure 2. C3 Structure diagram

lightweight convolutional block Attention Module that includes the Channel Attention Module (CAM) and Spartial Attention Module (SAM)are two submodules. The input feature maps are extracted in channel dimension and space dimension respectively and the corresponding attention graphs are formed. Then the attention graphs are multiplied with the input feature maps to carry out adaptive feature optimization. The unique structure not only saves parameters and computation, but also can be freely integrated into the existing network structure. The specific structure is shown in Figure 3.

3. Attention Mechanism

Convolutional Block Attention Module (CBAM) [11] is a

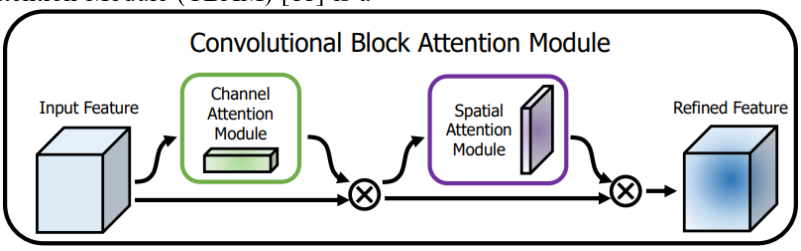

Figure 3. CBAM Structure diagram

The CAM submodule, as shown in Figure 4, compresses the input feature map in spatial dimension, then carries out maximum pooling and average pooling operations respectively, then passes it into the shared full connection layer to form a channel attention diagram, and finally multiplies the feature map to highlight the target feature information. The calculation formula is shown in (1). As shown in Figure 5, the SAM sub mode uses the channel attention map as the input feature map to compress the channel dimension, and then performs maximum pooling and average pooling in turn. Then the obtained feature map is splicing channels, and then 7x7 convolution operation is performed to form the spatial attention map, and finally the feature map is multiplied to highlight the location information of the target. The calculation formula is shown in (2).

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (1)$$

$$M_s(F) = (f([AvgPool(F'); MaxPool(F')]) \quad (2)$$

Where, F is the input feature map, $M_C(F)$ is the channel

attention mechanism, $M_S(F^{'})$ is the spatial attention mechanism, σ is the sigmoid function, MLP is the multi-layer perceptron, $AvgPool$ is the mean pooling, $MaxPool$ is the maximum pooling, f is the 7x7 Convolution operation.

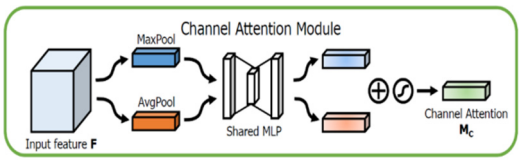

Figure 4. CAM Structure diagram

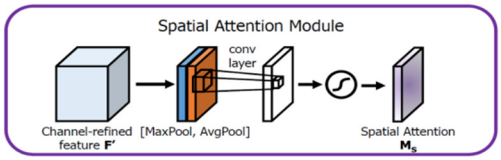

Figure 5. SAM Structure diagram

Due to the lack of feature information contained in small-size targets and the high background complexity in remote sensing images, the feature information of small-size remote sensing targets is not only small but also not obvious, which leads to the small feature information presented in the feature map formed by the small-size remote sensing targets after multiple downsampling. Therefore, this paper uses CBAM, an attention module that can generate attention feature maps in channel dimension and spatial dimension, to add to the original network structure to improve the information content of small-size targets on the feature maps, and then solve the impact of feature maps containing less target features on multi-scale feature fusion, classification and regression operations. The feature extraction task is mainly realized through the C3 module. Therefore, the convolutional attention CBAM structure is introduced into the C3 module to enhance the feature extraction capability and improve the problem of small target feature maps containing less information. The improved module is recorded as C3CBAM, and the specific structure is shown in Figure 6.

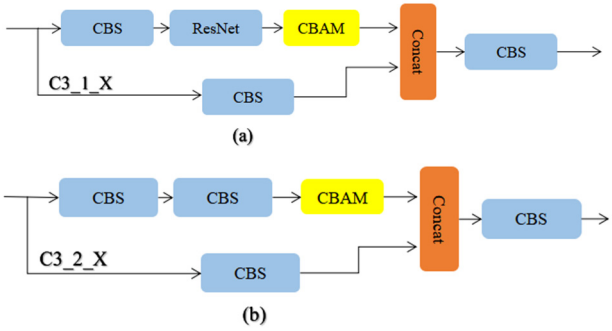

Figure 6. C3CBAM Structure diagram

4. Experimental Analysis

4.1. Introduction to Data Sets

The data used for model training in this paper is composed of RSOD data set and aircraft data from NWPU VHR-10 data set, with a total of 526 pieces, 420 pieces of training set, 106 pieces of verification set, and only one example aircraft. The length and width distribution of the training set is shown in Figure 7. It can be seen that the data points are mainly concentrated in the lower left corner, indicating that the target size is relatively small.

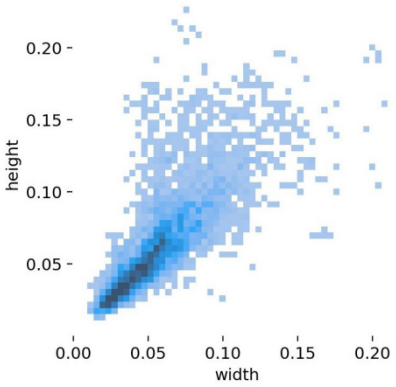

Figure 7. Length and Width Distribution of Training Set

4.2. Experimental Environment and Training Parameters

The experimental environment of this paper is shown in Table 1:

Table 1. Experimental environment

Name	Related configuration
CPU	Intel(R) Core (TM) i5-8300H CPU @ 2.30GHz 2.30 GHz
Internal memory	12GB
GPU	NIVIDIA GeForce GTX 1050
CUDA/CUDNN	11.0/8.1.5
Operating system	Windows 11 (64-bit)
Python/Pytorch	3.8/1.8.0

The training parameters of this algorithm are shown in Table 2:

Table 2. Training parameters

Name	Numerical value
Training Picture Size (imgsz)	640×640
batch-size	8
Training iterations (epochs)	200
Initial learning Rate (lr0)	0.01
Cycle learning Rate (lrf)	0.01
optimizer	SGD
momentum	0.937
Weight attenuation coefficient	0.0005

4.3. Evaluation Index

In this paper, the average precision mean (mAP), parameters (Params) and floating-point operations (FLOPs) are used to evaluate the performance of the model. mAP represents the performance measure of the target location and category predicted by the model, where mAP_0.5 represents the mean accuracy when the loss function threshold is 0.5. Params represents the number of parameters in the module training process, which is used to measure the complexity of the model. FLOPs represent the number of floating-point operations performed in a model network training period, which is used to measure the computational complexity of the model. GFLOPs represents 1 billion floating-point operations per second.

4.4. Experimental Results

The training results of CBAM-YOLOv5 algorithm on remote sensing aircraft data set are shown in Figure 8, the test results are shown in Figure 9, and the data results are shown in Table 3.

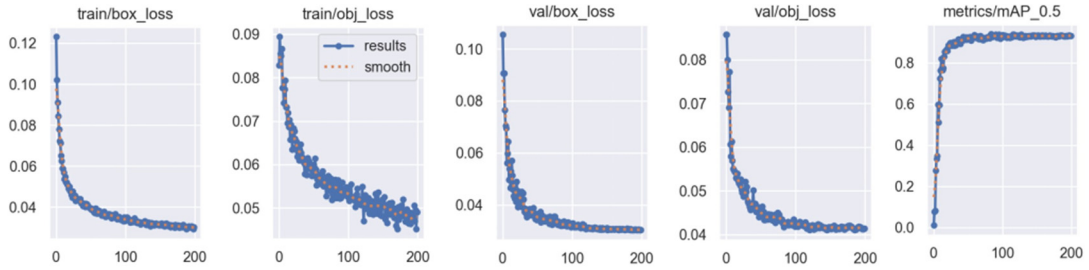

Figure 8. Training results of CBAM-YOLOv5

Figure 9. Test result of CBAM-YOLOv5

Table 3. Experimental Results

Model	mAP_0.5	GFLOPs	Params
YOLOv5	0.921	4.1	1760518
YOLOv5-C3CBAM (ours)	0.931	3.4	1485444

As can be seen from FIG. 8, after fewer training rounds, mAP_0.5 reached more than 80% and began to steadily increase, and box_loss and obj_loss was lower than 0.05 with fewer training rounds, and steadily decreased with the number of rounds. It can be seen from Figure 9 that the prediction probability of the proposed algorithm for remote sensing small-size targets is greater than 85%. As can be seen from Table 3, when lightweight convolutional attention CBAM is introduced into C3 module, the average accuracy of mAP_0.5 is increased by 1%, the number of parameters is reduced by 15.6%, and the amount of computation is also reduced.

5. Conclusion

In this paper, CBAM-YOLOv5 optimization algorithm is proposed to solve the difficult problem of extracting feature information of remote sensing small size targets. After experimental verification, the mAP_0.5 value of the proposed algorithm reaches 93.1%, and the parameter number is reduced by 15.6% compared with the original algorithm. The model not only improves the accuracy rate of remote sensing small-size target recognition, but also becomes lighter. However, there are still some shortcomings in the detection of small size objects in intensive remote sensing, and this problem will be studied based on the algorithm in this paper in the future.

References

[1] GIRSHICK R, DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Columbus: IEEE, 2014:580-587.

[2] GIRSHICK R, Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision.Santiago: IEEE, 2015: 1440-1448.

[3] REN S,HE K,GIRSHICK R,et al.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J] .IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

[4] HE K,GKIOXARI G,DOLLAR P,et al.Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision.2017:2961-2969.

[5] LIU W,ANGUELOV D,ERHAN D,et al.SSD: Single Shot MultiBox Detector [C]//Proceedings of 14th European Conference on Computer Vision.Amsterdam: Springer, 2016: 21-37.

[6] REDMON J,DIVVALA S,GIRSHICK R,et al.You only look once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Seattle:IEEE,2016:779-788.

[7] Zhang W, Wang S H, Thachan S, et al. Deconv R-CNN for small object detection on remote sensing images[C] // Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Los Alamitos: IEEE Computer Society Press, 2018: 2483-2486.

[8] Zhang Z Y, Liu Y P, Liu T C, et al. DAGN: a real-time UAV remote sensing image vehicle detection framework[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17(11): 1884-1888.

[9] Song K, Huang P M, Lin Z P, et al. An oriented anchor-free object detector including feature fusion and foreground enhancement for remote sensing images[J]. Remote Sensing Letters, 2021, 12(4): 397-407.

[10] Liu S, Zhang L, Lu H C, et al. Center-boundary dual attention for oriented object detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14.

[11] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.