Mathematical Biosciences and Engineering

MBE, 20(6): 10773–10789.
DOI: 10.3934/mbe.2023478
Received: 24 February 2023
Revised: 02 April 2023
Accepted: 07 April 2023
Published: 18 April 2023

http://www.aimspress.com/journal/MBE

Research article

An improved UAV target detection algorithm based on ASFF-YOLOv5s

Siyuan Shen, Xing Zhang, Wenjing Yan, Shuqian Xie, Bingjia Yu and Shizhi Wang*

Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing 210009, China

* **Correspondence:** Email: shizhiwang2009@seu.edu.cn; Tel: +862583272566; Fax: +862583324322.

Abstract: Object detection in drone-captured scenarios is a recent popular task. Due to the high flight altitude of unmanned aerial vehicle (UAV), the large variation of target scales, and the existence of dense occlusion of targets, in addition to the high requirements for real-time detection. To solve the above problems, we propose a real-time UAV small target detection algorithm based on improved ASFF-YOLOv5s. Based on the original YOLOv5s algorithm, the new shallow feature map is passed into the feature fusion network through multi-scale feature fusion to improve the extraction capability for small target features, and the Adaptively Spatial Feature Fusion (ASFF) is improved to improve the multi-scale information fusion capability. To obtain anchor frames for the VisDrone2021 dataset, we improve the K-means algorithm to obtain four different scales of anchor frames on each prediction layer. The Convolutional Block Attention Module (CBAM) is added in front of the backbone network and each prediction network layer to improve the capture capability of important features and suppress redundant features. Finally, to address the shortcomings of the original GIoU loss function, the SIoU loss function is used to accelerate the convergence of the model and improve accuracy. Extensive experiments conducted on the dataset VisDrone2021 show that the proposed model can detect a wide range of small targets in various challenging environments. At a detection rate of 70.4 FPS, the proposed model obtained a precision value of 32.55%, F1-score of 39.62%, and a mAP value of 38.03%, which improved 2.77, 3.98, and 5.1%, respectively, compared with the original algorithm, for the detection performance of small targets and to meet the task of real-time detection of UAV aerial images. The current work provides an effective method for real-time detection of small targets in UAV aerial photography in complex scenes, and can be extended to detect pedestrians, cars, etc. in urban security surveillance.

Keywords: UAVs; small target detection; YOLOv5s; ASFF

1. Introduction

In recent years, civil UAVs have been widely used in many fields, such as biological monitoring, agriculture and investigation. At the same time, the target detection technology in the UAV capture scene is also widely used in wildlife protection, urban monitoring, traffic monitoring and other aspects. However, due to the constant change of the flying height of the UAV, the scale of the object also changes greatly. When the flight height is high, the captured image contains a large number of small targets; when the scene of flight is complex, the image contains a variety of cluttered background information and the density of targets is high, which can easily produce mutual obscuration problems. In addition, the light intensity is also changing due to the flight time, making target recognition more difficult in night scenes. Also, achieve real-time recognition of aerial video streams, there are requirements for object detection speed.

Traditional target detection tasks use feature-based detection methods, which require targets with complex texture structures, large contrast between different classes and manual selection of target features. With the development of deep learning, convolutional neural networks have started to emerge with good results in tasks such as image classification, target detection and target segmentation.

For the problem of detecting small targets in aerial photography or remote sensing images, many scholars have conducted research. Huang et al. reduced the computational load of the network by introducing the shufflenetv2 feature extraction structure into the YOLOv5s network, and added the attention mechanism CBMA module to the network, which reduced the flops from 15.7 to 3.7 G, but the accuracy decreased significantly [1]. Xu et al. proposed a target detection algorithm based on improved YOLOv3 to solve the problems of large number, small size and low detection accuracy of vehicle targets in aerial photography. By improving the network structure, adding detection layers, and using the improved DIoUloss function in the regression loss function, we can more accurately identify road vehicle targets without reducing the speed, and reduce the error rate. However, there are still errors and omissions in some extreme cases [2]. To improve the accuracy of small target detection, Fang et al. improved the Faster R-CNN network by using the flexible context information fusion method, and tested it on the coco sub-data set, which shows that the improved algorithm has good performance on the accuracy and recall rate of small target detection. Experimental results show that the algorithm can achieve a balance between detection speed and detection accuracy, but the detection speed is slow and does not meet the real-time requirements [3]. Liu et al. proposed an improved feature fusion method for small target detection based on PANet and BiFPN (PB-FPN), which effectively improved the detection ability of the algorithm for small targets. By introducing the spatial pyramid pooling (SPP) in the backbone network into the feature fusion network and connecting it with the model prediction head, the performance of the algorithm has been effectively improved, but it has not well solved the problem that there is a large amount of occlusion of dense small targets in complex background, resulting in a large number of small targets missing detection [4]. Gong et al. added a fusion factor to the FPN to describe the coupling degree of adjacent layers to control the information passed from the deep layer to the shallow layer in order to make the FPN to adapt to tiny object detection, thus improving small target detection accuracy, but feature fusion in different layers remains a simple linear operation that cannot perform adaptive feature weight assignment according to the different scenes in which the features are located [5]. For small target detection as well as dense occlusion problems in agriculture, Arunabha M. Roy et al. have improved the YOLOv4 algorithm to achieve the detection of different classes of large-scale diseases in complex scenarios and also very small patch diseases, providing an effective method for early detection of different plant diseases in complex scenarios [6]. In addition, a Dense-YOLOv4 model is proposed for detecting different four different growth stages in mangoes, which can detect target objects in complex scenes with a high degree of occlusion and overlap [7].

In response to the needs of UAV target detection for real-time, model size and application scenarios, as well as the problems of poor performance, large computational volume and insufficient real-time detection of high-density occlusion targets in images of high-altitude shooting scenes by existing methods, this paper proposes a targeted approach based on the YOLOv5s algorithm with improvements in the algorithm network, loss function and other aspects, and proposes an improved ASFF- YOLOv5s algorithm for UAV aerial image target detection.

The main contributions of this paper are as follows: To improve the detection capability for fine targets, shallow feature maps are added to the feature fusion network to prevent the disappearance of small target features, and ASFF is improved to enable full fusion of multi-scale features. To improve the detection performance in complex backgrounds the attention mechanism module is fused into the backbone and prediction networks. Adopting IoU as the K-means++ anchor frame clustering indicator to re-cluster the anchor frame size of the VisDrone2021 dataset. Adopting SIoU as the bounding frame loss function to improve the convergence speed and accuracy.

It was found that with the above modifications, our model outperforms other state-of-the-art detection models with a model parameter count of 13×10^9 and other models with a parameter count of 30×10^9 or more, with a recognition speed of 73 FPS or more, much higher than other algorithms, and with higher recognition accuracy, especially in the detection of small targets.

2. YOLOv5s algorithm and improvements

The YOLO algorithm was proposed by Redmon et al. in 2016, and the main idea of the YOLO series [8–11] is to output both location and category information using the same network, and by 2020 the YOLO algorithm has evolved to the fifth generation. The YOLOv5 algorithm adds adaptive image scaling, Mosaic data enhancement, and adaptive anchor frame calculation on the input side compared to the original. YOLOv5 uses the CSPDarknet53 as the backbone network and the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) as the feature extraction network, which combines shallow and deep information to improve detection performance. The YOLOv5 series is divided into YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x according to the number of channels and model size, with the model size increasing in order.

The output of the original YOLOv5s algorithm contains three prediction layers for predicting large, medium and small size targets, while excessive down-sampling leads to missing position information of the targets, which is not conducive to the detection of small targets and is less effective for multi-size detection.

2.1. Algorithm improvements

In this paper, the YOLOv5s algorithm is improved, trained and tested on the VisDrone2021 UAV aerial photography image dataset. The improved network structure was shown in Figure 1. The complete network structure consists of three parts: a backbone for feature extraction, the neck for semantic representation of extracted features, and the head for prediction.

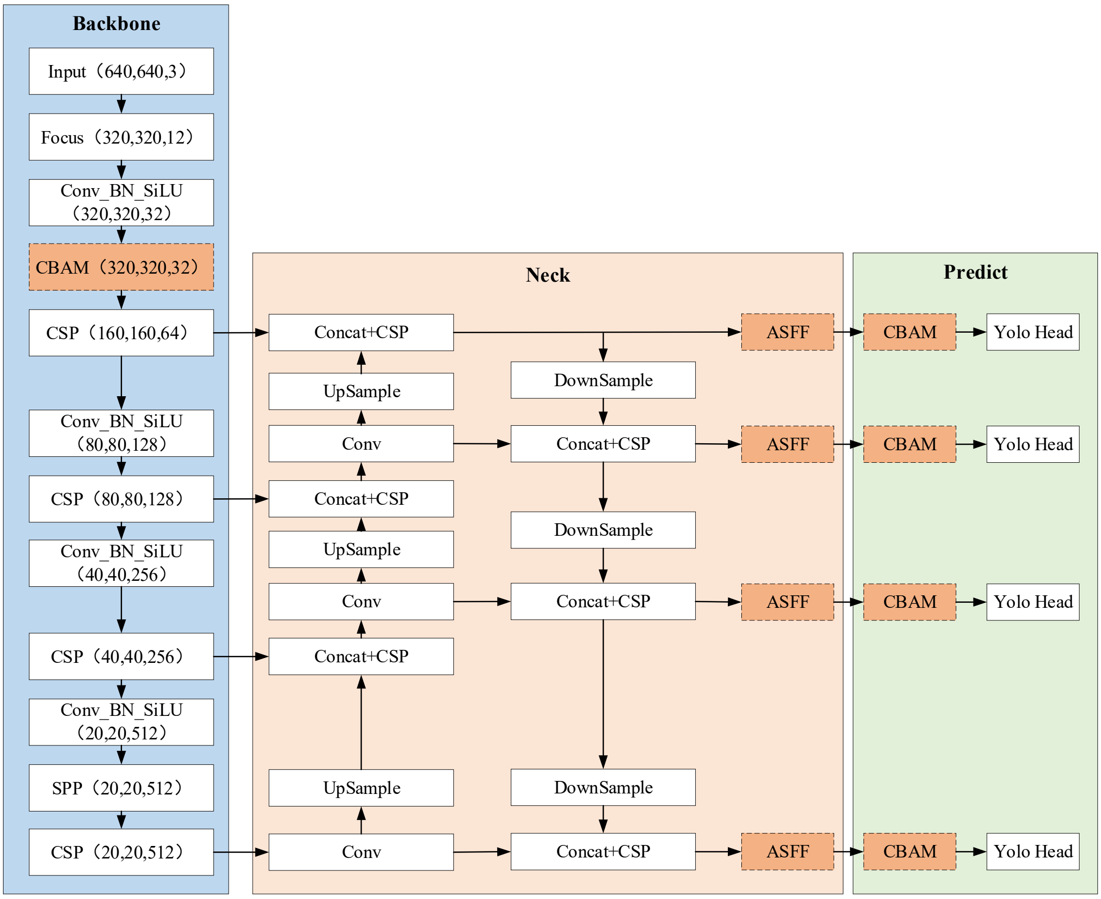
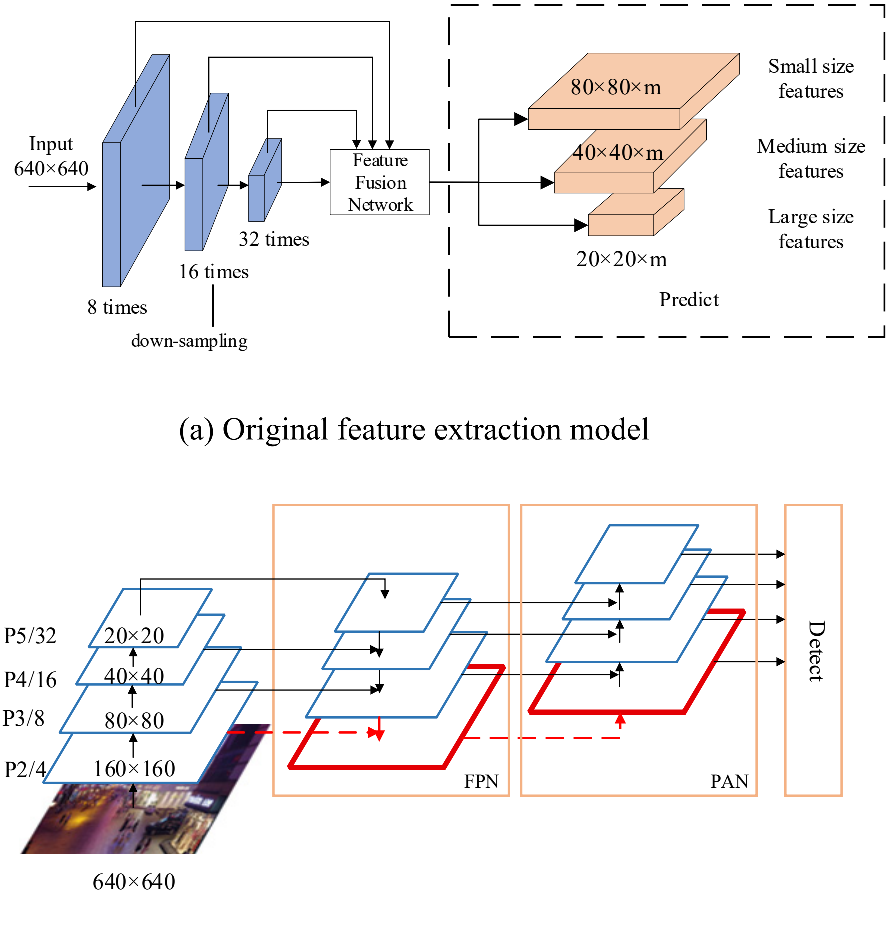


Figure 1. Improved YOLOv5s network structure.

First, we input 640 × 640 × 3 images into the network, and the input image needs to be preprocessed, including scaling, normalization and channel order adjustment. YOLOv5s uses RGB order, which requires scaling the pixel values between 0 and 1. Further, the anchor frame for the VisDrone2021 dataset was obtained by improving the k-mean algorithm before the model training. Then, the CSPDarknet53 network was used as the backbone network of YOLOv5s. The network is modified based on Darknet53, contains many layers of convolution and pooling, and improves the performance of the network by using techniques such as Cross-Stage Partial Connections (CSP) and multi-scale feature fusion. The use of CSP module can effectively reduce the amount of computation and the number of parameters, and improve the efficiency and accuracy of the network. In addition, we fuse the CBAM module before the CSP module so that important feature information can be input into the network. In the feature extraction process, to extract more semantic features of small targets, we input the shallow 320 × 320 × 32 feature maps obtained from CSPDarknet53 into the feature extraction network, connect them after convolution, then perform up-sampling, and down-sampling, and stack them with the remaining feature layers to enhance the feature fusion process. The final four feature maps are obtained and fed into the ASFF structure to fully combine the semantic information of the high-level features with the fine-grained features of the underlying features into the prediction network. In the prediction network, the four fused adaptive features are processed by the CBAM module to obtain four prediction detection heads of different scales and four bounding boxes of sizes 20 × 20, 40 × 40, 80 × 80, and 160 × 160. Further, And the NMS algorithm is used to filter the prediction frames and remove the prediction frames with high overlap. Finally, the bounding box regression fine-tunes the bounding boxes based on the detected prediction box location and size information to get more accurate target locations. In the loss function, we use SIoU as the bounding box loss function to further improve the convergence speed and accuracy.

2.2. Improved multi-scale feature fusion for ASFF

The feature fusion network was improved to address the problem of missed detection of small targets by the original network. In the original YOLOv5s algorithm, the input image was fed into the feature fusion network by down-sampling the feature map three times in the backbone network, as shown in Figure 2(a). Too much down-sampling will result in missing location information of the target, which was not conducive to the detection of small targets. Therefore, the down-sampled 4 times feature map was input to the FPN and combined with the deep semantic information and then input to the PAN to pass the target location information from the bottom up. This was shown in Figure 2(b). The added feature map was of high resolution and possesses richer location information, inputting more information about small targets from the backbone network into the feature extraction network, thus improving the detection of small targets.



(a) Original feature extraction model

(b) Improved feature fusion network

Figure 2. Comparison of multi-scale feature fusion networks.

In the PANet structure of the ASFF-YOLOv5 algorithm, after first augmenting the FPN structure with top-down semantic feature extraction, the ASFF algorithm [12] was introduced at each layer of the FPN structure for weighted fusion. The weight parameters were derived from the output of the convolutional feature layer, and the weight parameters were learned after gradient back-propagation so that the weighted fusion could be performed adaptively. The PANet structure of the proposed fused and improved ASFF algorithm was shown in Figure 3.

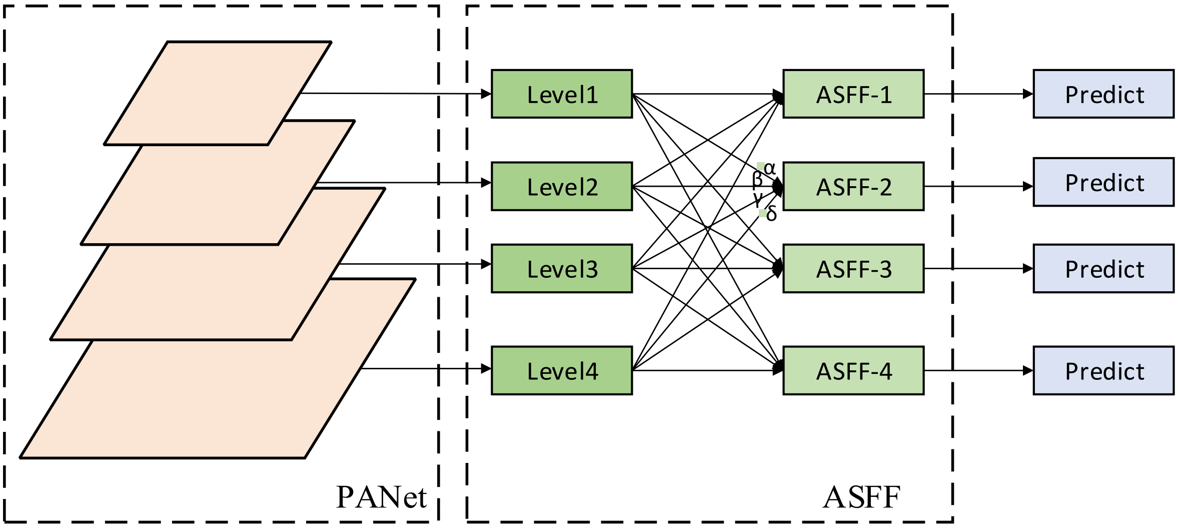


Figure 3. Improved multi-scale feature fusion with ASFF.

Using the ASFF-2 computational fusion as an example, the feature maps $Level1$, $Level2$, $Level3$ and $Level4$ were obtained from the PANet structure by fusing them with the ASFF algorithm to obtain ASFF-2. The pair $Leve\ l1$ feature map was convolved to obtain the same number of channels as the $Level2$ feature map and then up-sampled to obtain. For the $Level3$ and 4 feature maps, the number of channels and dimensions were adjusted by convolution and down-sampling operations to maintain the same number of channels as $Level2$ with the same number of channels and dimensionality to obtain. $Level2$ feature maps were adjusted by the number of channels after the convolution operation to obtain. After processing the four feature maps using the softmax function, the weight coefficients α, β, γ and δ were extracted from the YOLOv5 backbone network and then the ASFF fusion calculation was performed using the following equation.

$$y_{ij}^{l} = \alpha_{ij}^{l} X_{ij}^{1\rightarrow l} + \beta_{ij}^{l} X_{ij}^{2\rightarrow l} + \gamma_{ij}^{l} X_{ij}^{3\rightarrow l} + \delta_{ij}^{l} X_{ij}^{4\rightarrow l} \tag{1}$$

where was the new feature map obtained using the ASFF module and was the weight coefficients of the four feature maps, processed by softmax, as follows:

$$\alpha^l_{ij} = \frac{e^{\lambda^l_{\alpha_{ij}}}}{e^{\lambda^l_{\alpha_{ij}}} + e^{\lambda^l_{\beta_{ij}}} + e^{\lambda^l_{\gamma_{ij}}} + e^{\lambda^l_{\delta\ ij}}} \tag{2}$$

The adjustment of the feature fusion weight parameters by the improved ASFF algorithm fully realizes the multi-scale feature fusion of the model.

2.3. Fused attention mechanism

When UAV aerial images are taken, due to the interference of disordered geographical information, it often leads to the problem that the background information is mistakenly detected as the target object when detecting the performance of the object. Our research integrated the CBAM module [13] into the backbone network and prediction network, which was used to improve the feature extraction capability of the network, thereby improving the network performance, as shown in Figure 4.

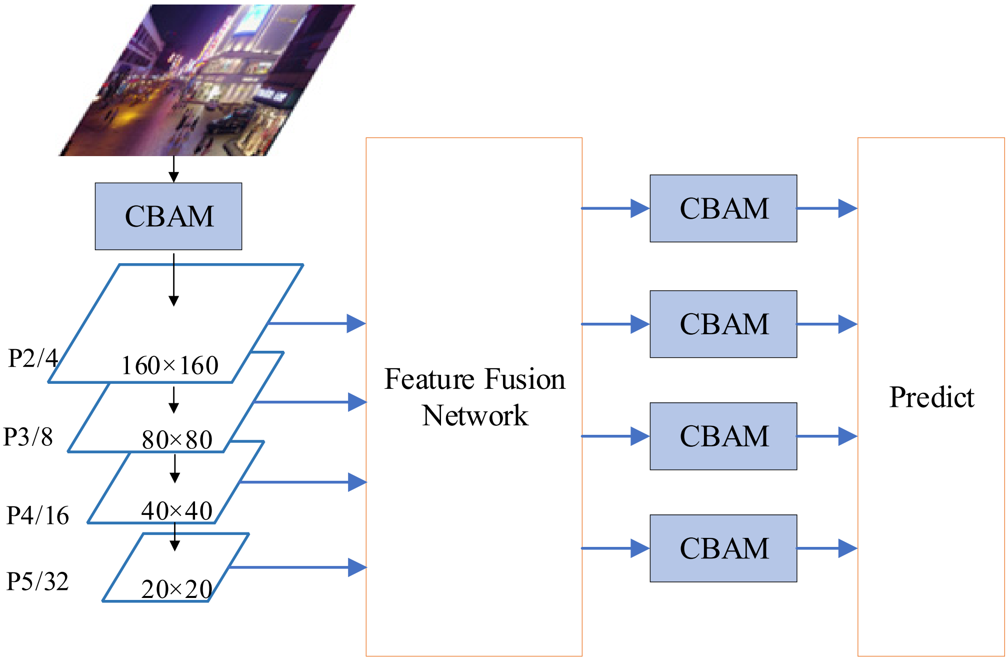


Figure 4. Network structure of the Convergent Attention Mechanism.

The CBAM module is a simple and efficient attention mechanism for improving the representation capability of the model. Its structure was shown in Figure 5, concluding two independent sub-modules: the channel attention module and the spatial attention module. In this paper, the CBAM module was integrated into the backbone network, and the output results were passed to the next layer and the feature extraction network after channel attention and spatial attention to improve the detection performance.

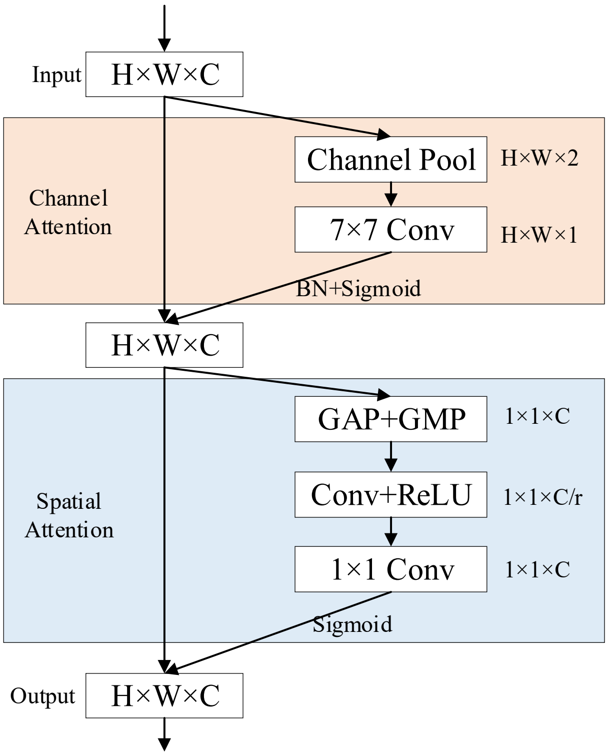


Figure 5. Structure of CBAM module.

2.4. *Improved K-means++ clustering of anchor frames*

YOLOv5s algorithm is a target detection algorithm based on anchor frame. The anchor frame parameters affect the final training effect of the algorithm to a certain extent. Because the initial anchor frame parameters are calculated by K-means clustering algorithm on the coco data set, which is not applicable to the UAV detection data set, this study used K-means++ clustering algorithm to calculate the anchor frame parameters, and re-seted the anchor frame parameters of the target data set. The original anchor frame clustering algorithm uses Euclidean distance as the metric, and large anchor frame clusters will produce more error than small ones when clustering. To address this problem, IoU was used as the metric, and the improved K-means++ algorithm was as follows.

Table 1. Improved K-means++ algorithm flow.

```
Algorithm: Improved K-means++ algorithm
Input: height and width of all sample boxes, number of clusters K, maximum number of iterations Max
Output: the height and width of the anchor boxes
K boxes randomly selected from the sample as the initial anchor box anchor.
while anchor is no longer changing or the maximum number of iterations is not reached do:
    Calculate the IoU of each sample with respect to the box;
    assign each box to the anchor with which it has the largest IoU;
    calculate the mean of the width and height of all boxes in each cluster, updating anchor;
return the height and width of the anchor boxes in cluster K
```

A total of K = 16 sets of anchor frames were selected for the experiment, with each of the four scale prediction layers containing four anchor frames, and the maximum number of iterations was set to 300 to obtain the new a priori frame scales, normalized as shown in Table 2.

Table 2. Anchor size.

Feature map	20 × 20	40 × 40	80 × 80	160 × 160
Anchor frame size	(34,71)	(32,23)	(12,18)	(3,6)
	(77,62)	(18,52)	(20,14)	(4,13)
	(55,107)	(27,41)	(11,34)	(10,10)
	(116,130)	(49,39)	(18,27)	(7,21)

2.5. Loss function

The loss function in the YOLOv5s algorithm includes bounding box loss, classification loss and confidence prediction loss [14], where GIoU was used as the loss function for bounding box loss [15] and the formula was calculated as follows.

$$GIoU = IoU - \frac{|C| - |AUB|}{|C|} \tag{3}$$

$$IoU = \frac{A \cap B}{A \cup B} \tag{4}$$

(A and B represent the prediction box and the real box, and C represents the smallest rectangular box that can surround the two boxes.)

When the prediction box is inside the real box, the GIoU loss cannot determine the position status of the prediction box, and at this time, the GIoU calculation value is the same as the IoU problem. In order to solve the above problem, subsequent studies have proposed CIoU [16], ICIoU, etc., while considering the distance, overlap area and aspect ratio between the prediction box and the real box. However, the mismatch between the two frames is not considered, resulting in slow convergence and low efficiency. To address the above problems, this paper used the SIoU loss function [17] as the bounding box loss function, which consists of four parts: angle cost, distance cost, shape cost and IoU cost.

Angle cost is defined as:

$$\Lambda = 1 - 2\sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{5}$$

$$x = \frac{c_h}{\sigma} = \sin(\alpha) \tag{6}$$

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2} \tag{7}$$

$$c_h = \max\left(b_{c_y}^{gt}, b_{c_y}\right) - \min\left(b_{c_y}^{gt}, b_{c_y}\right) \tag{8}$$

(where denote the horizontal and vertical coordinates of the centroids of the true and predicted boxes, respectively.)

Distance cost:

$$\Delta = \sum_{t=x,y} \left(1 - e^{-\gamma\rho_t}\right) \tag{9}$$

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2, \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2, \gamma = 2 - \Lambda \tag{10}$$

(where denotes the width and height of the smallest rectangular box enclosed by the real and predicted boxes, respectively.)

Shape cost:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\omega_t}\right)^\theta \tag{11}$$

$$\omega_w = \frac{\left|w - w^{gt}\right|}{\max\left(w, w^{gt}\right)}, \omega_h = \frac{\left|h - h^{gt}\right|}{\max\left(h, h^{gt}\right)} \tag{12}$$

(where denotes the width and height of the real and predicted boxes respectively.)

IoU cost:

$$\mathcal{L}_{IoU\mathrm{Cost}} = 1 - IoU \tag{13}$$

Final SIoU loss function:

$$L_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{14}$$

The SIoU loss function introduces directionality into the prediction frame. Considering the vector angle between the required regressions, it can make the model converge faster and more accurately in the training stage, and improve the model detection accuracy.

3. Experimental results sand analysis

3.1. Experimental dataset

This experiment used the VisDrone2021 dataset to evaluate the training model. The benchmark dataset consists of 288 video clips consisting of 261,908 frames and 10,209 stills with over 2.6 million annotated frames. The experiments were selected from 6471 training sets, 1610 test sets and 548 validation sets containing ten categories of pedestrians, cars, bicycles, buses, vans, trucks, people, motorbikes, awning tricycles and tricycles.

3.2. Experimental parameters and evaluation indexes

The training model parameters in this paper were set as follows: the SGD optimization algorithm was used with an initial learning rate of 0.01. Mish as the activation function; the training batches were 300 times, with 32 images passed in at a time in each training batch for training; the input image size was 640 × 640, and the last iteration model weights and the best performance model weights were stored. The environment for this experiment was shown in Table 3.

Table 3. Experimental configuration environment.

configuration items	Model
Programming Languages	Python
Deep learning frameworks	Pytorch
Operating system	Win10
CPU	Inter Core i9-9900
Running memory	32GB
GPU	NVIDIA GeForce RTX 2080Ti

In this experiment, mAP and F1-score were used as an evaluation indicator of model performance. mAP denotes the average of AP values for all categories calculated as follows.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{15}$$

$$AP = \int_0^1 P(R)\mathrm{d}(R) \tag{16}$$

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2PR}{P + R} \tag{19}$$

(The values of TP and FP were determined according to the set IoU threshold, usually 0.5, and the mAP value was calculated as the value of TP and FP was determined according to the set IoU threshold, usually 0.5.)

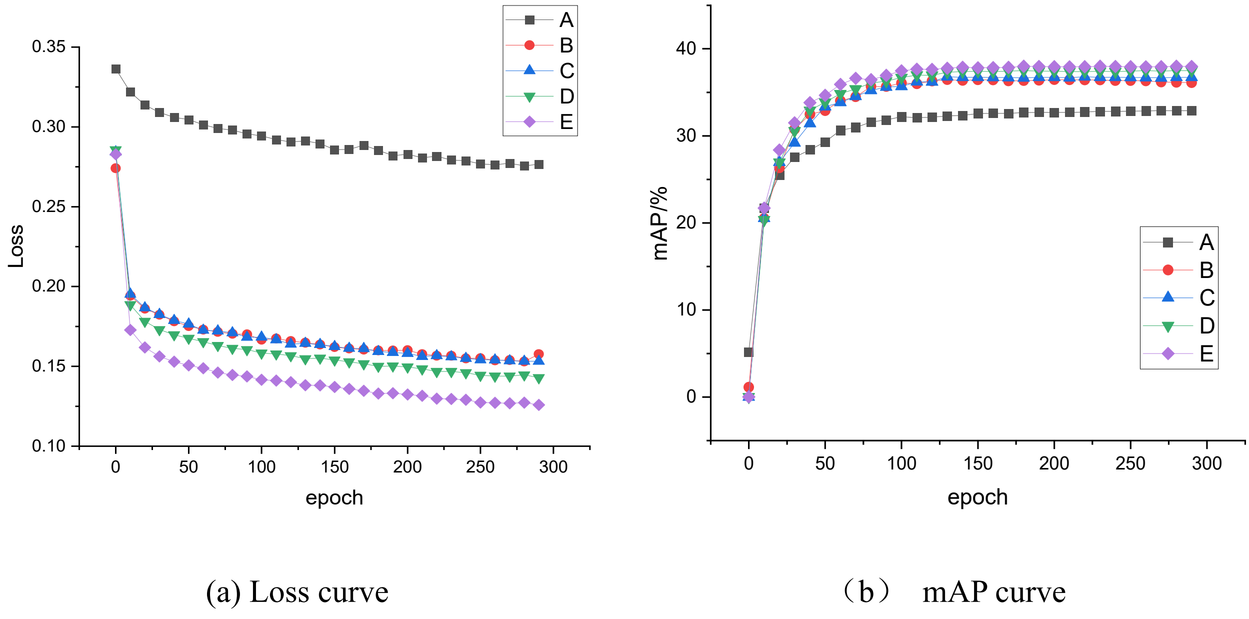
3.3. Ablation experiments

To verify the role of each module in improving the YOLOv5s algorithm, ablation experiments [18] were conducted to observe the impact of each module on the network performance. In this paper, five sets of comparison experiments were set up and tested on the test set after adding four sets of improved modules in turn, as shown in Table 4, with "√" indicating the use of the improved modules. The experimental results showed that the number of model parameters increased but the accuracy improved after adding each module in turn compared to the original YOLOv5s network. The improved model has a reduced FPS, but still meets the real-time requirements. A smaller reduction in detection speed in exchange for higher accuracy.

Table 4. Ablation experiments based on YOLOv5s.

Group	Improved multi-scale feature fusion for ASFF	Integration of attention mechanisms	Anchor frame optimization	Optimizing loss functions	P /%	R/%	F1-score /%	mAP/%	mAP 50:90/%	FPS
A					29.78	44.36	35.64	32.91	17.73	73.42
B	√				29.26	48.13	36.39	36.10	19.66	71.13
C	√	√			30.12	49.71	37.51	36.74	20.29	71.67
D	√	√	√		31.91	49.43	38.78	37.54	20.8	70.74
E	√	√	√	√	**32.55**	**50.61**	**39.62**	**38.02**	**20.96**	**70.36**

The accuracy and loss values in the training of each model were compared, as shown in Figure 6 (a),(b) for the loss curve and mAP curve of each model respectively, with the horizontal coordinates representing the training batches and the vertical coordinates representing the loss values and mAP respectively, and then the loss values decreased and the accuracy increased with the increase of training times after adding each module.



(a) Loss curve （b） mAP curve

Figure 6. Model mAP and loss curve.

3.4. Comparative experimental analysis

In order to verify the superiority of this paper's algorithm compared with other algorithms, this paper is compared with existing state-of-the art models, including Faster-RCNN [18], RetinaNet [19], Mask R-CNN [20], SSD, Cascade RCNN [21] and YOLOv3 [22], for comparison experiments, as shown in Figure 7, mainly for mAP50, mAP50: 95 and FPS, as shown in the table, this paper has a greater improvement in accuracy and FPS compared with other algorithms, and can identify aerial images faster and more accurately.

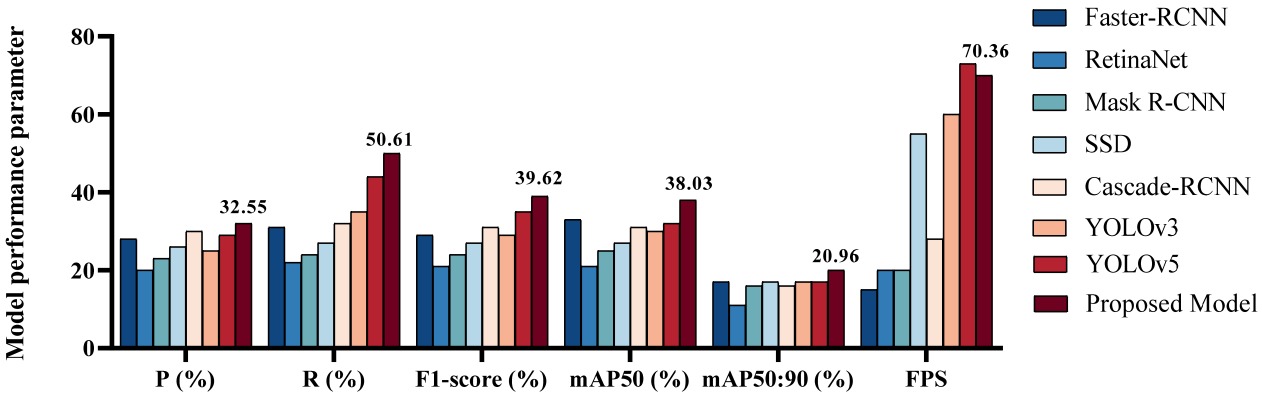
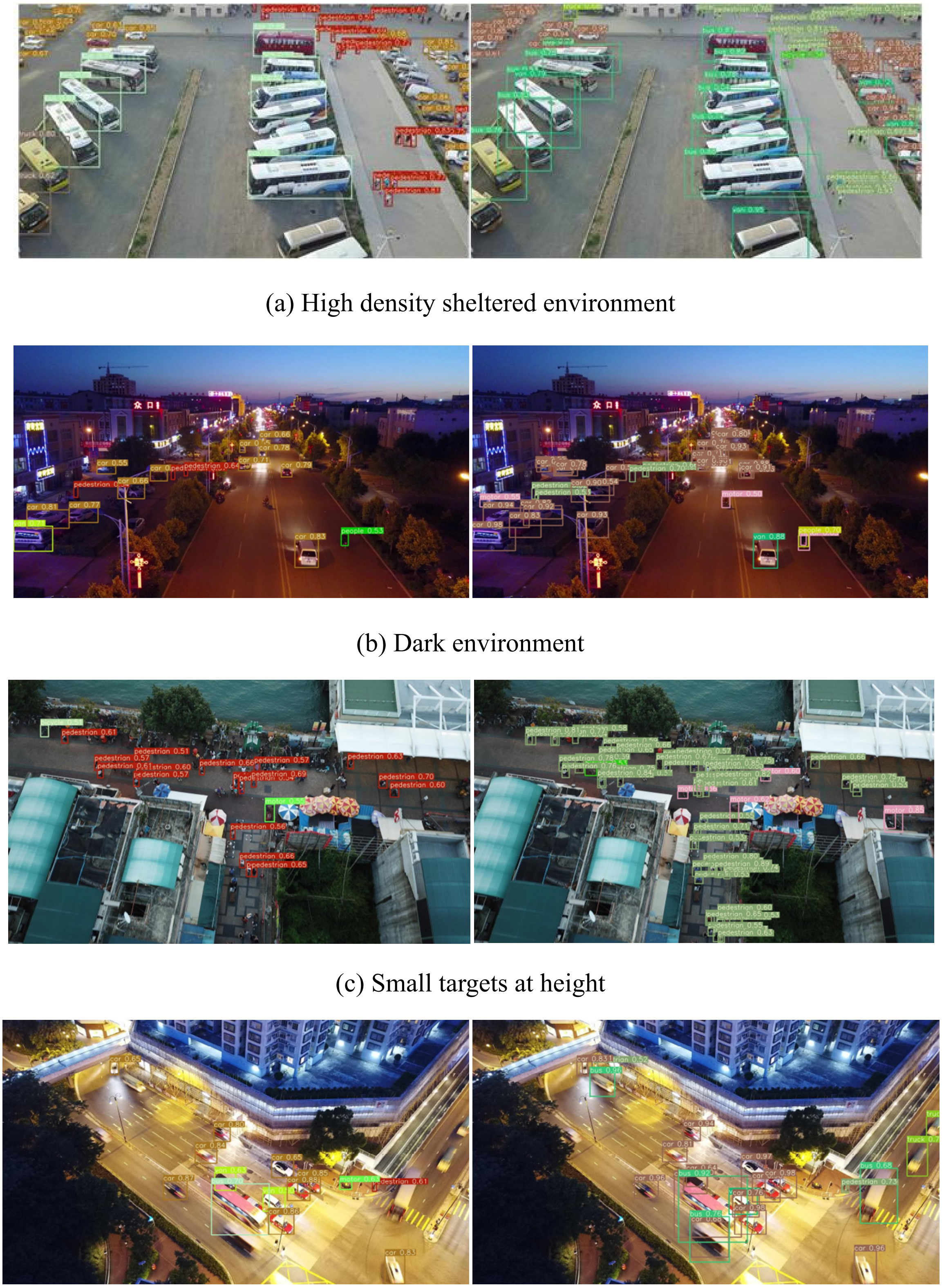


Figure 7. Comparison bar chart of precision, recall, F1-score, mAP and detection speed (in FPS) between proposed model (improved ASFF-YOLOv5s) and other state-of-the-art models: Faster R-CNN, RetinaNet, SSD, Mask R-CNN, Cascade R-CNN, YOLOv3 and YOLOv5.

3.5. Analysis of detection results



(a) High density sheltered environment

(b) Dark environment

(c) Small targets at height

(d) Motion blurred targets

Figure 8. Comparison of test results.

In order to verify the effectiveness of the improved algorithm, this paper used the VisDrone2021-DET test challenge dataset images on the original algorithm and the improved algorithm for visual comparison experiments. Several complex recognition scenarios were selected for detection, including high-density occlusion environment, dark environment, small targets at high altitude and blurred images in motion, as shown in Figure 8, the left side is the original algorithm, the right side is the improved algorithm.

According to the recognition results, the improved algorithm has partially improved the confidence level of object detection, and can accurately identify the obscured objects in high-density occlusion and dark environment, reduce the miss detection rate of small target objects at high flight altitude, and identify the objects captured in aerial images that are blurred due to fast movement. In summary, the improved algorithm has improved the detection performance of small target objects in complex environments, and the problem of wrong detection and leakage detection can be improved to a certain extent, which proves the effectiveness of the algorithm in this paper, and the detection time of each image is around 0.014 s, and the number of frames transmitted per second can reach 70, which meets the requirements of UAV for real-time detection of aerial images.

4. Conclusions

For the problem of missing detection of small targets in complex environments, such as high-density occlusion and background clutter in UAV aerial images, this paper proposed an improved small target detection algorithm of ASFF-YOLOv5s for the task of real-time detection of aerial images by UAVs. First, through multi-scale feature fusion, the new shallow feature map was introduced into the feature fusion network, and the ASFF was improved to improve the ability of extracting small target features and multi-scale information fusion. Second, CBAM module was added in front of the backbone network and prediction network to improve the ability to capture important features and suppress redundant features. Finally, the SIoU loss function was used to accelerate the convergence of the model and improve the accuracy. On the public UAV target dataset VisDrone2021, the algorithm proposed in this paper achieves the highest detection accuracy and speed compared to some existing state-of-the-art detection models, achieving precision, F1-score and mAP of 32.55, 39.62 and 38.03%, respectively, in detecting pedestrians, cars and other targets at a detection rate of 70.4 FPS. Compared with the original YOLOv5s algorithm, the improvement is 2.77, 3.98 and 5.1%, respectively. The current work provides an effective method for real-time detection of small targets in complex scenes. However, there are still certain problems, for example, changes in lighting and weather conditions can cause some interference in UAV target detection, making the detection results inaccurate. We can consider target detection by combining multimodal information (e.g., RGB images, infrared images, radar, etc.), which can improve the robustness of the model and thus enhance the reliability and accuracy of detection.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, Southeast University "Zhongying Young Scholars" Project and Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0162 and KYCX22_0301). The datasets generated during and/or analysed during the current study are available from the corresponding author

on reasonable request.

Conflict of interest

All authors disclosed no relevant relationships.

References

1. Y. Huang, H. Cui, J. Ma, Y. Hao, Research on an aerial object detection algorithm based on improved YOLOv5, in *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, (2022), 396–400. https://doi.org/10.1109/CVIDLICCEA56201.2022.9825196
2. M. Xu, X. Wang, S. Zhang, R. Wan, F. Zhao, Detection algorithm of aerial vehicle target based on improved YOLOv3, *J. Phys.*, **2284** (2022), 012022. https://doi.org/10.1088/1742-6596/2284/1/012022
3. P. Fang, Y. Shi, Small object detection using context information fusion in faster R-CNN, in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, (2018), 1537–1540. https://doi.org/ 10.1109/CompComm.2018.8780579
4. H. Liu, F. Sun, J. Gu, L. J. S. Deng, Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode, *Sensors*, **22** (2022), 5817. https://doi.org/10.3390/s22155817
5. Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, Z. Han, Effective fusion factor in FPN for tiny object detection, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2021), 1160–1168.
6. A. M. Roy, R. Bose, J. Bhaduri, A fast accurate fine-grain object detection model based on YOLOv4 deep neural network, *Neural Comput. Appl.*, **34** (2022), 1–27. https://doi.org/10.1007/s00521-021-06651-x
7. A. M. Roy, J. Bhaduri, Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4, *Comput. Electron. Agric.*, **193** (2022), 106694. https://doi.org/10.1016/j.compag.2022.106694
8. A. Bochkovskiy, C. Y. Wang, H. M. Liao, Yolov4: Optimal speed and accuracy of object detection, preprint, arXiv:2004.10934.
9. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 779–788.
10. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 7263–7271.
11. J. Redmon, A. Farhadi, Yolov3: An incremental improvement, preprint, arXiv:1804.02767
12. M. Qiu, L. Huang, B. H. Tang, ASFF-YOLOv5: multielement detection method for road traffic in UAV images based on multiscale feature fusion, *Remote Sens.*, **14** (2022), 3498. https://doi.org/10.3390/rs14143498
13. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19.

14. T. Han, Research on small object detection algorithm based on feature enhancement module, *J. Phys.*, **1757** (2021), 012032. https://doi.org/10.1088/1742-6596/1757/1/012032
15. H. Yu, L. Yun, Z. Chen, F. Cheng, C. Zhang, Neuroscience, A small object detection algorithm based on modulated deformable convolution and large kernel convolution, *Comput. Intell. Neurosci.*, **2023** (2023), 2506274. https://doi.org/ 10.1155/2023/2506274
16. Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU loss: Faster and better learning for bounding box regression, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 12993–13000. https://doi.org/ 10.1609/aaai.v34i07.6999
17. Z. Gevorgyan, SIoU loss: More powerful learning for bounding box regression, preprint, arXiv:2205.12740
18. S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, **28** (2015), 28.
19. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 2980–2988.
20. K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 2961–2969.
21. Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 6154–6162.
22. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, et al., Ssd: Single shot multibox detector, in *Computer Vision–ECCV 2016*, **9905** (2016), 21–37. https://doi.org/ 10.1007/978-3-319-46448-0_2