J. Meas. Sci. Instrum., 2024, 15(1): 128-139.
DOI: 10.62756/jmsi.1674-8042.2024013
http://jmsi.publish.founderss.cn/
jmsi@nuc.edu.cn

Road target detection algorithm based on improved YOLOv5 in UAV images

ZHANG Yi, MA Ronggui*, LIANG Chen

School of Information Engineering, Chang'an University, Xi'an 710064, China

***Corresponding author:** MA Ronggui (rgma@chd.edu.cn)

Received: May 23, 2023 **Revised:** June 28, 2023 **Accepted:** July 3, 2023

Abstract: Aiming at the problems such as low accuracy and poor robustness of target detection caused by missed detection of small road targets and occlusion between targets in UAV images, an improved road target detection algorithm based on YOLOv5 combining convolutional block attention module (CBAM), called YOLOv5s-FCC, was proposed. Firstly, a small target sensing layer was introduced to improve the multi-scale model, and a small target YOLO detection head was added to improve the feature extraction ability of the network for small road targets. Secondly, the CBAM fused space and channel information to enhance important information in the network after it was introduced into different locations of the Backbone network to obtain the best fusion location of CBAM. Finally, CIoU loss function was used to improve the speed and accuracy of the calculation required for predicting the bounding box of image. The experimental results showed that compared with YOLOv5 algorithm, YOLOV5-FCC algorithm can improve mAP50 and mAP50-95 by 2.0% and 4.2%, respectively. The effectiveness of YOLOv5-FCC algorithm was also verified on VisDrone dataset, and the results showed that the established system can realize automatic detection of road targets.

Key words: unmanned aerial vehicle (UAV); road target detection; YOLOv5; loss function; convolutional block attention module (CBAM)

0 Introduction

The intelligent maintenance of highways has developed into one of the primary study areas as the intelligent transportation sector has matured. Target detection technology is frequently used in the automatic detection and identification of road targets. Road target identification has been plagued by issues with low real-time and low accuracy because of the complexity of actual environments[1]. Currently, technical assistance for intelligent driving and intelligent highway maintenance provided by unmanned aerial vehicle (UAV) -based road target detection technology can realize automatic identification and detection of road key target information.

Most traditional target detection trely on feature extraction and classifier techniques such as histogram of oriented gradient (HOG) [2], scale invariant feature transform (SIFT) [3] and local binary pattern (LBP) [4], as well as using digital image processing techniques to achieve target detection and recognition. However, these techniques are vulnerable to the limitations of feature selection and may suffer from low accuracy and slow speed in complex environments. To address these issues, recent research has focused on the development of more advanced techniques, such as deep learning-based approaches, which have shown promising results in terms of detection accuracy and speed. Compared with traditional algorithms that require manual feature design, target detection algorithms based on deep learning can automatically learn image features, and the fusion of global and local features makes, it more flexible for information extraction at different scales and angles.

There are two types of deep learning-based target detection algorithms: two-stage algorithms and single-stage algorithms. Two-stage algorithms first send the image via a convolutional neural network (CNN) to generate a series of proposal boxes, and then filter them via the target area detection network. Representative algorithms include region with CNN feature (RCNN) [7], Fast RCNN[8], and Faster RCNN[9], characterized by high detection accuracy while low detection speed. Single-stage algorithms have quicker detection speed, such as "You Only Live Once" (YOLO) series[10-12], which can be implemented on mobile devices to attain real-time detection performance. To meet the requirements of detection efficiency and accuracy, deep learning has focused on the balance of

accuracy and speed.

Currently, target detection techniques based on YOLOv5 have been widely used in the field of intelligent transportation. Zhu et al.[13] proposed TPH-YOLOv5 that improves the original prediction head to transformer prediction head (TPH) on the basis of YOLOv5 and incorporates convolutional block attention model (CBAM), which can find the target attention region more accurately in the scenes with denser objects. Zhao et al.[14] proposed an improved vehicle target detection model based on YOLOv5s network. The attention mechanism squeeze-and-exictation (SE) modules at different positions of YOLOv5s are introduced to enhance important features of vehicles while suppress general features, which strengthens the detection network's ability to recognize vehicle targets. Zhang et al.[15] proposed a lightweight traffic sign detection algorithm based on YOLOv5 that combines CBAM with CA and optimizes the algorithm by model pruning. However, deep learning-based road target detection algorithms have certain drawbacks.

1) There are many small road targets in UAV images that are prone to missed and false detection by the existing algorithms.

2) The actual sizes of different road targets vary widely, which results in different target sizes in the images. However, it is difficult for the existing algorithms to adapt to multi-scale detection.

3) The presence of occlusion between road targets may lead to missed detection by the existing algorithms.

In our work, a modified YOLOv5 algorithm for road target detection with UAV images was proposed to detect six targets, including rpavement, guardrail, marker, crack, vehicle, and green belt. Firstly, a small target YOLO detection head was added, and the 22nd, 25th, 28th and 31st layers of the network were used as the output detection heads to improve the accuracy of small target detection. Secondly, the CBAM was introduced and the best integration location of the CBAM was explored at different locations of Backbone network. After that, CIoU loss function was used to improve the speed and accuracy of regression. Finally, a UAV image-based road target detection system was trained and verified on the self-built UAV image-based road target dataset, and the detection results were visualized.

1 YOLOv5

YOLOv5 is a classic single-stage target detection algorithm that is widely used nowadays. It has four different-scale network structures, namely YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. This study utilizes YOLOv5s, as shown in Fig.1.

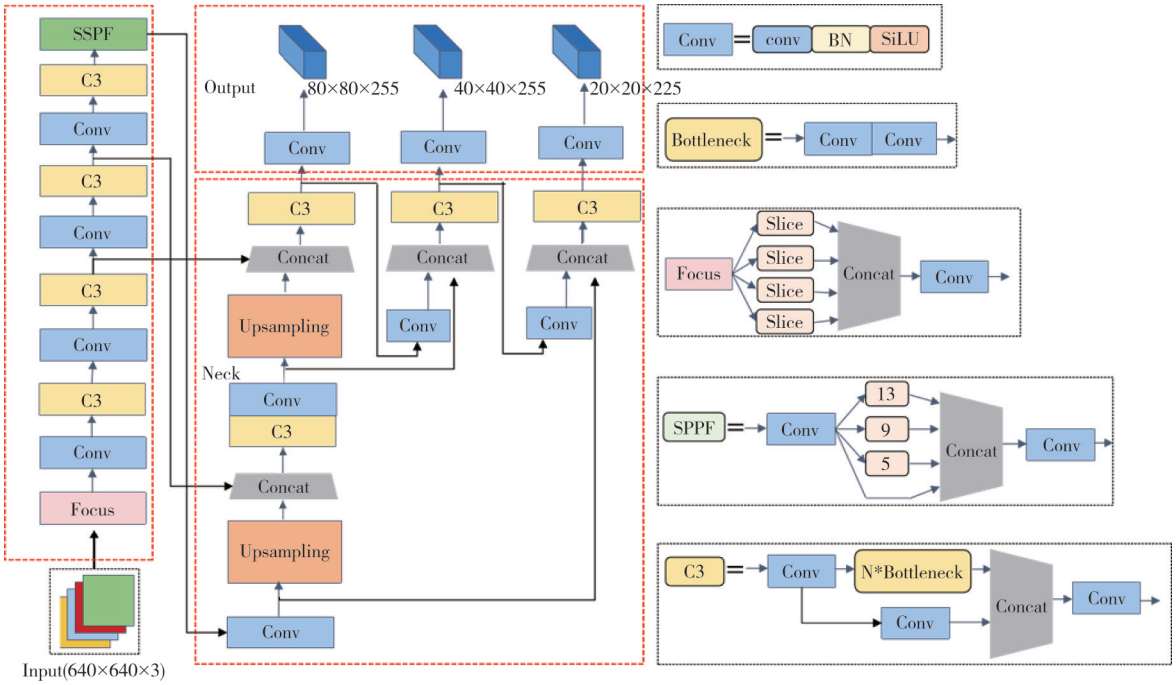


Fig. 1 YOLOv5s network structure

It consists of four parts: Input, Backbone, Neck and Output. The Input uses Mosaic data enhancement technology to rich the dataset by image blending, cropping and filling. It also uses adaptive anchor frame calculation technology to improve detection accuracy. The Backbone uses Focus module for fast down-sampling to reduce information loss. The C3 structure[16] contains three standard convolutional layers and the Bottleneck is used to

enhance the learning capability of the network. Furthermore, the spatial pyramid pooling-fast (SPPF) module is used to enhance the acceptance range of Backbone features, which can reduces the number of network layers based on SPP[17] so as to improve the algorithm's computation speed. The Neck uses FPN (feature pyramid network)[18]+PAN[19] (path aggregation network) to transfer the semantic information of image features from first deep to shallow and then from shallow to deep, thus enhancing the fusion ability of the network. In the outputr three detection heads are used to achieve prediction of large, medium and small targets.

2 Optimized YOLOv5

2.1 Small target detection layer

Four different-scale network structures of YOLOv5 have different depths and widths. For different-scale targets, there are three prediction heads with different sizes corresponding to the targets with different sizes. When a UAV takes a road image with a size of 640×640, the 80×80 detection layer can detect small targets in the image, the 40×40 detection layer can detect medium-sized targets, and the 20×20 detection layer can detect large targets. By using this multi-scale detection method, the impact of target scale changes can be effectively relieved.

Using UAV images as research objects, the changes of target scales in UAV images were particularly obvious, for example, the scale difference of road guardrail and crack in the image reached several times. Since the original YOLOv5 algorithm may miss small targets, an extra prediction head was added to alleviate the errors caused by the changes of target scales. Firstly, an anchor parameter suitable for small targets was added for the anchor box setting of target detection. Secondly, behind the 17th layer of the Backbone, an FPN structure was added to enhance the semantic information of features, and then a PAN structure was added to enhance the target positioning ability. The improved network can better enhance the target adaptability and the network structure is shown in Fig.2.

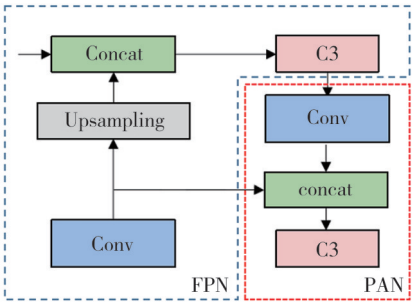


Fig. 2 Structural diagram of FPN+PAN

At the 20th layer of the improved network, the feature map obtained was fused with the 2nd layer feature map of the Backbone to obtain smaller targets. Finally, at the 31st layer of the improved network, a small target layer was added. The 22th, 25th, 28th and 31th layers of the network were the output detection heads, and a detection head with a size of 160×160 was added to detect small targets. The improved network structure is shown in Fig.3, where the blue dotted line represents the small target detection head.

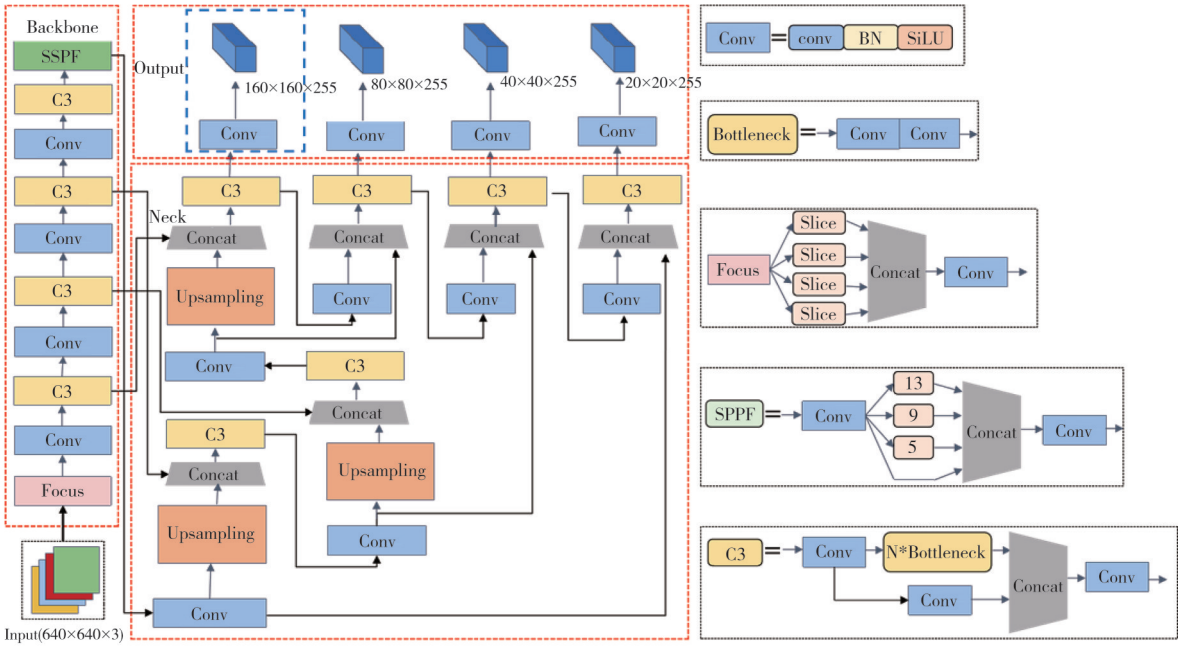


Fig. 3 Improved network by introducing a small target detection layer

2.2 Attention model

The principle of the attention model is to selectively focus on areas of greater interest and ignore other parts of the information. Introduction of attention model in YOLOv5 can improve the weights of road target area. In this way, the network can more effectively distinguish the road targets from environmental information and solve the problem of information loss of road targets due to deepening network layers. The mainstream attention models currently include SE-Net[20], efficient channel attention network (ECA-Net)[21], CBAM[22], etc. The SE-Net is not sensitive to target location information, resulting in poor performance in multi-target detection tasks. The ECA-Net generates channel attention through fast one-dimensional convolution, which is simple but lower accurate due to the use of local cross-channel interaction strategy without dimensionality reduction. The CBAM adopts average pooling and maximum pooling to aggregate features and integrates spatial and channel information.

In the highway images collected by UAVs, there are more targets and the targets vary drastically in size. To solve the problem of missed and false detection of small targets, it is necessary to make the network more focused on tiny targets such as cracks. In this study, the CBAM was adopted. It emphasizes meaningful features in two main dimensions: channel and spatial, and the channel and spatial attention modules are applied sequentially, as shown in Fig.4, and the specific computation process is as follows.

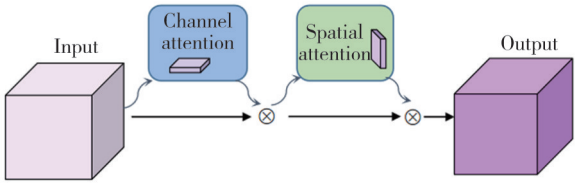


Fig. 4 Structure diagram of CBAM

1) After inputting the given intermediate feature map named $F \in \mathbf{R}^{C\times H\times W}$, the CBAM infers the channel attention module, ($M_{\mathrm{CH}} \in \mathbf{R}^{C\times 1\times 1}$), and then generates the channel attention map, which is multiplied by the input image to generate the intermediate variable F'. The calculation process is shown as

$$F' = M_{\mathrm{CH}}(F) \otimes F, \tag{1}$$

where $\otimes$ denotes weighted multiplication.

2) The two-dimensional spatial attention module ($M_{\mathrm{S}} \in \mathbf{R}^{1\times W\times H}$) is used to generate the spatial attention map from the feature map F', which is multiplied by the feature map F' after channel-adaptive refinement and then we obtain the final output refinement image F''. The calculation process is shown as

$$F'' = M_{\mathrm{S}}(F') \otimes F'. \tag{2}$$

Other scholars have experimentally demonstrated that CNNs with embedded CBAM have better performance in image classification and target detection tasks [23]. However, there is no clear research conclusion to show which specific position in the network can get the best effect. Since CBAM can augment important information in the network, it can be used for channel and spatial attention reconstruction of feature maps at different locations. In the Backbone, as the network layers become deeper, the width of the feature map becomes smaller, and some information may be lost. The Bottleneck in the C3 module serves to aggregate features at different levels, so that the CBAM can be placed behind or fused into the Bottleneck. Furthermore, the fourth C3 structure of the Backbone extracts the features and computes the final detection results, where attention model can be introduced to improve the network performance.

Two attention module fusion methods were designed. The first is to fuse the CBAM into the C3 structureof the Backbone, forming a C3_CBAM structure, as shown in Fig. 5. The network is named YOLOv5s-FC1. The second is to add the CBAM to the last layer of the Backbone. The network is named YOLOv5s-FC2. These two attention models are shown in Fig.6.

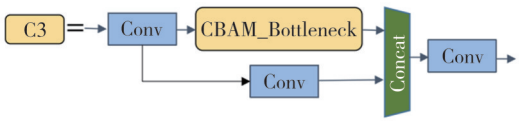


Fig. 5 Structural diagram of C3_CBAM

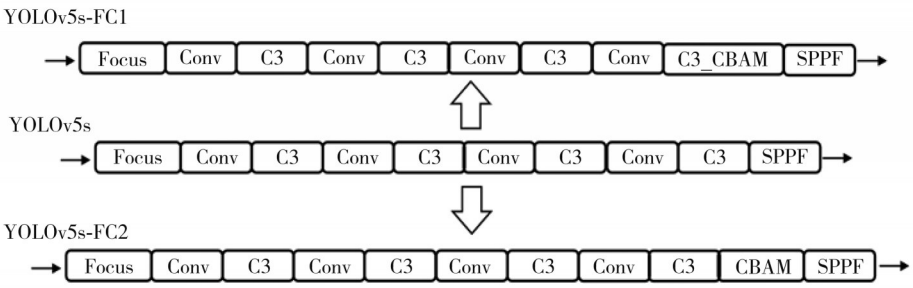


Fig. 6 Two kinds of YOLOv5s models embedded in CBAM

2.3 Improved loss function

To frame the detected target, it is necessary to predict the locations of the bounding box. By selecting the candidate box with the highest loss function value as the optimal prediction output, the overlapping candidate boxes are removed by using non-maximum suppression (NMS) method. Intersection over union (IoU) is adopted by YOLOv5 as the loss function.

IoU is defined as the ratio of intersection set to union set of target box A and prediction box B, as shown in Fig.7.

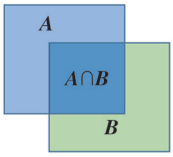


Fig. 7 IoU schematic diagram

IoU can be calculated by

$$IoU(A, B)=\frac{(A\cap B)}{(A\cup B)}. \tag{3}$$

There exist two special cases using IoU as the loss function. 1) If boxes A and B do not intersect, IoU is zero and gradient calculation cannot be performed. 2) IoU cannot accurately reflect how boxes A and B intersect as well as their overlapping degree. This means that the same IoU value does not mean the same overlapping degree.

In our work, GIoU was proposed to solve the above-mentioned problems. As shown in Fig. 8, GIoU introduces rectangle C, which is the smallest rectangle containing target box A and prediction box B.

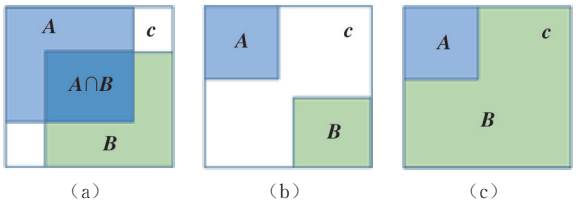


Fig. 8 GIoU schematic diagram

GIoU is calculated by

$$GIoU(A, B)=\frac{(A\cap B)}{(A\cup B)}-\frac{|C-(A\cup B)|}{|C|}. \tag{4}$$

In case that target box A and prediction box B do not intersect, as shown in Fig.8(b), the first term in Eq. (4) is zero, and the value of GIoU is equal to the second term in Eq. (4). Such a calculation method can effectively solve the problem that gradient calculation cannot be performed due to non-intersection of two boxes A and B. GIoU has been used in original YOLOv5 algorithm.

However, in case that prediction box B contains the target box A and C overlaps with box B, as shown in Fig. 8(c), if the second term in Eq. (4) equals zero, GIoU equals IoU and target box A is not optimized. Here, DIoU can effectively solve this problem because it has a faster convergence speed. On the basis of GIoU, DIoU introduces diagonal distance $\boldsymbol{c}$ of the smallest rectangular box C, and defines the center point $\boldsymbol{b}^{gt}$ of the target box A and the center point $\boldsymbol{b}$ of the prediction box B. It also introduces Euclidean distance $\boldsymbol{\rho}$ between point $\boldsymbol{c}$ and point $\boldsymbol{b}$, as shown in Fig.9.

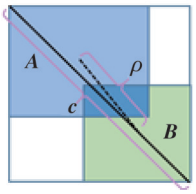


Fig. 9 DIoU schematic diagram

DIoU is calculated by

$$DIoU(A, B)=\frac{(A\cap B)}{(A\cup B)}-\frac{\boldsymbol{\rho}^2(\boldsymbol{b}, \boldsymbol{b}^{gt})}{\boldsymbol{c}^2}. \tag{5}$$

Furthermore, CIoU introduces aspect ratio influence factor $\alpha\upsilon$ on the basis of DIoU. At the same time, α is used as the coefficient to balance the ratio, and $\boldsymbol{\upsilon}$ is used to measure how the anchor frame and the target frame are proportionally consistent. In this way, the predicted box will be more consistent with the real box, and the calculation is performed as

$$CIoU(A, B)=\frac{(A\cap B)}{(A\cup B)}-\frac{\boldsymbol{\rho}^2(\boldsymbol{b}, \boldsymbol{b}^{gt})}{\boldsymbol{c}^2}-\boldsymbol{\alpha\upsilon}, \tag{6}$$

$$\boldsymbol{\upsilon}=\frac{4}{\pi^2}\left(\arctan^{-1}\frac{\boldsymbol{w}^{gt}}{\boldsymbol{h}^{gt}}-\arctan^{-1}\frac{\boldsymbol{w}}{\boldsymbol{h}}\right)^2, \tag{7}$$

$$\boldsymbol{\alpha}=\frac{\boldsymbol{\upsilon}}{(1-\frac{(A\cap B)}{(A\cup B)})+\boldsymbol{\upsilon}}, \tag{8}$$

where $\boldsymbol{w}^{gt}$ and $\boldsymbol{h}^{gt}$ denote the width and height of the target box, respectively; and w and h denote the width and height of the predicted box, respectively.

Compared with original IoU, the advantages of CIoU are as follows. First, CIoU can calculate the non-overlapping area of two boxes, which reflects the overlapping degree of target box and predicted box more clearly and unambiguously. Second, CIoU optimizes the calculation method of the distance between target boxes so as to improve convergence speed. Third, CIoU can more realistically and accurately describe the relationship between target box and prediction box.

For three different IoU loss calculation methods, namely GIoU, DIoU and CIoU, the improved

YOLOv5s was used to conduct multiple comparison tests on the self-built UAV road target dataset to obtain the best improvement scheme.

3 Experiment and analysis

3.1 Datasets

In this study, road target datasets was from the images collected by UAVs that was driven in July 2022 at an altitude of 40 m in clear, well-lit weather. The dataset was expanded to 2 640 images by panning, rotation, and brightness transformation operations. The detection targets were divided into six categories by LabelImg: pavement, marking, crack, vehicle, guardrail and green belt, as listed in Table 1. Some images in the dataset are shown in Fig.10.

Table 1 Dataset categories and quantity

Categories of road targets	Quantity
Pavement	3 648
Marking	14 483
Crack	5 692
Vehicle	1 868
Guardrail	3 872
Green belt	2 452

Fig. 10 Datasets display

Fig. 11 shows the label distribution of the self-built UAV road target dataset.

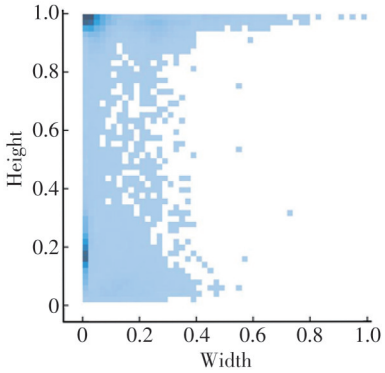


Fig. 11 Dataset label distribution

The horizontal coordinate represents the normalization width of the label box, and the vertical coordinate represents the normalization height. It can be seen that most of the road targets are distributed in the top left and bottom left. It indicates that most of the targets in the dataset are in slender and small size targets such as markers and cracks. Therefore, the migration learning based multi-scale YOLOv5 algorithm proposed in this study is targeted to solve the problem of missed and false detection of small targets.

The txt label format file was generated using LabelImg. The first column of the file contains the categories of all the targets, the other four columns contain the coordinates of the target center asd well as the width and height of the labeled rectangle. The target categories include pavement, markering, crack, vehicle, guardrail, and green belt corresponding to the integers $0, 1, \cdots, 5$, respectively, the center coordinates are normalized values. In this study, one-tenth of the images in the dataset were randomly selected to form the test set, one-tenth of the remaining images were selected as the validation set, and the other images were the training set, as listed in Table 2.

Table 2 Datasets partition

Dataset	Quantity
Training set	2 138
Validation set	238
Test set	264
Total	2 640

3.2 Experimental environment and model training

The experimental environment is as follows: OS: Windows 10; CPU: Intel(R) Core(TM) i5-11400 @ 2.60 GHz, 32 GB running memory; GPU: NVIDIA GeForce RTX 3060 with 16 GB video memory. Deep learning framework is PyTorch, and program language is Python3.7. The YOLOv5s network is used, and the training configuration basically adopts official recommended parameters. Epoch is set to be 300, input image size is 640×640, initial learning rate is set to be 0.01, momentum is 0.937, and weight decay coefficient is 0.0005, model training batch size is set to be 8. The accuracy metrics used include: mAP50, mAP50-95, computation(unit: Giga floating-point operations per second, GFLOPs), and network size(unit: MByte, MB), where mAP50 represents the average detection accuracy when the threshold of the loss function is equal to 0.5, and mAP50-95 represents the average detection accuracy within the threshold range of 0.5—0.95.

3.3 Experimental results

3.3.1 Small target detection layer

A small target detection head was added to the original YOLOv5s algorithm to construct an improved network named YOLOv5s-F(four layers). A comparison experiment was conducted to compare the improved and original algorithms on the self-built UAV road target

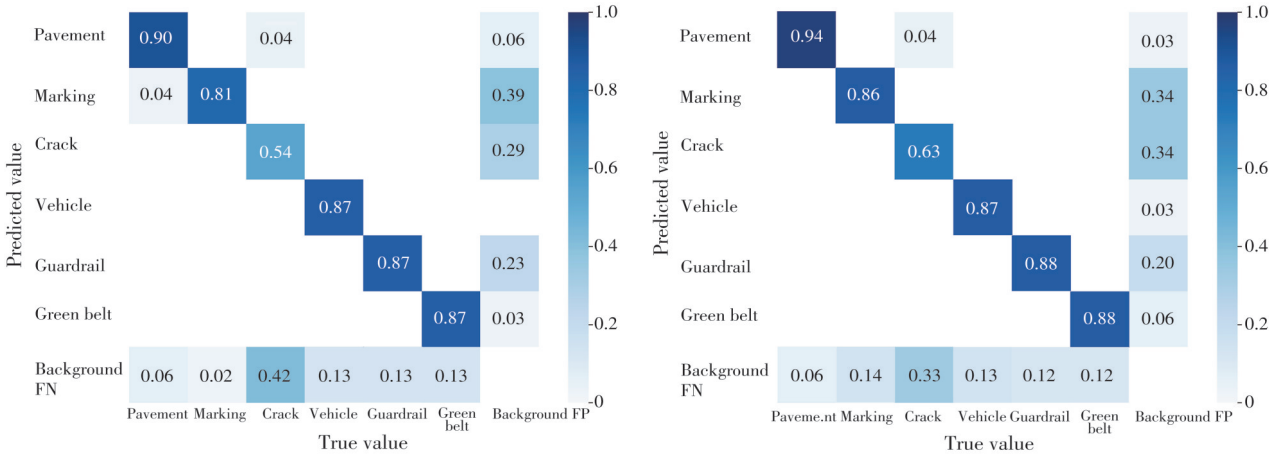
datasets, and the experimental results are shown in Table 3.

Table 3 Experiment results when introducing a small target detection layer

Algorithm	Network size/MB	Computation/GFLOPs	mAP50/%
YOLOv5s	14.7	20.1	86.0
YOLOv5s-F	15.4	27.0	86.9

It can be seen from Table 3 that the network size increases by 0.7 MB and the computation increases by 6.9 GFLOPs when introducing a small target detection layer. Since the increased values are within an acceptable range, the computation of real-time detection can be achieved. The mAP50 is improved by 0.9% when introducing a small target detection layer compared to the original algorithm. It shows that the improved network can significantly improve the target detection accuracy while increasing network size and computational effort.

To verify the effectiveness of the improved network to detect the small target, the detection accuracy comparison of seven categories of targets using YOLOv5s and YOLOv5s-F was performed, as shown in Fig.12.



(a) YOLOv5s (b) YOLOv5s-F

Fig. 12 Comparison of detection accuracy values

It can be seen that the detection accuracy values of five categories of targets are increased using YOLOv5s-F, especially, YOLOv5s-F has the highest improvement of 9% in detection accuracy for cracks compared with YOLOv5s. It is verified that YOLOv5s-F has a stronger ability to detect small targets.

3.3.2 Attention model

To verify the effectiveness of the improved network when introducing CBAM and to explore the optimal embedding location, two attention model embedding methods are compared and tested. YOLOv5s-FC1 introduces CBAM in the C3 structure of the Backbone while YOLOv5s-FC2 introduces CBAM in the 9th layer of the Backbone. The comparison was performed on the self-built UAV road target dataset, and the results are shown in Table 4.

Table 4 Experimental results of fused attention modules

Algorithms	Network size/MB	Computation/GFLOPs	mAP50/%
YOLOv5s-F	15.4	27.0	86.9
YOLOv5s-FC1	15.5	27.3	87.5
YOLOv5s-FC2	15.5	27.3	87.8

In Table 4, both CBAM embedding techniques can improve detection accuracy while slightly increasing network size and computation. The mAP50 of YOLOv5s-FC1 is improved by 0.6%, and the mAP50 of YOLOv5s-FC2 is improved by 0.9%, which means that YOLOv5s-FC2 can greatly increase detection accuracy while only slightly expanding the network size. Thus, the second attention module embedding method was adopted in this study. Based on this, we introduced three common attention modules, SE, ECA and CA, and conductsed comparative tests based on the second embedding method. The results are listed in Table 5.

Table 5 Experimental results of different attention modules

Algorithm	Network size/MB	Computation/GFLOPs	mAP50/%
YOLOv5s-F	15.4	26.9	86.9
YOLOv5s-F-SE	15.4	26.9	86.9
YOLOv5s-F-ECA	15.4	27.0	87.2
YOLOv5s-F-CA	15.5	27.3	87.4
YOLOv5s-F-CBAM	15.5	27.3	87.8

It can be seen from Table 5 that the introduction of different attention modules has basically no effect on network size and little effect on computation, but it brings a significant change in detection accuracy. The introduction of SE results in no change in computation and detection accuracy while the introduction of ECA and CA increases computation by 0.1 GFLOPs and 0.4 GFLOPs, respectively, with an improvement of mAP50 by 0.3% and 0.5%, respectively. The introduction of CBAM increases computation by 0.4 GFLOPs with an improvement of

mAP50 by 0.9%, which means that detection accuracy has been obviously improved. In summary, CBAM can greatly improve detection performance for UAV road targets compared with other common attention modules.

3.3.3 Loss function

Based on the improvement in Section 2.2, the final detection accuracy values of GIoU, DIoU and CIoU were compared horizontally. The results are shown in Table 6. It can be seen that the mAP50 of CIoU is 88.0%, which is higher than that of GIoU and DIoU, which verifies the effectiveness of CIoU.

Table 6 Comparative experimental results of loss functions

Loss function	mAP50/%
GIoU	87.8
DIoU	87.9
CIoU	88.0

3.3.4 Ablation experiments

Ablation experiments were conducted to compare YOLOv5s-F, YOLOv5s-FC and YOLOv5s-FCC with YOLOv5s. Batch size was set to be 300, and training was carried out on the labeled self-built UAV road target dataset. The results are listed in Table 7.

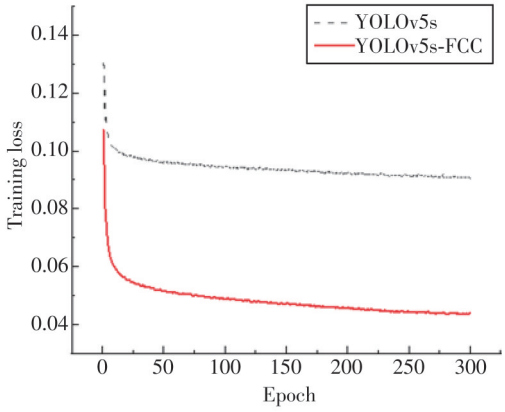
Table 7 Ablation experimental results

Algorithm	F	CBAM	CIoU	mAP50/%	mAP50-95/%	FPS/Hz
YOLOv5s				86.0	62.1	40.8
YOLOv5s-F	√			86.9	64.9	38.8
YOLOv5s-FC	√	√		87.8	66.1	37.4
YOLOv5s-FCC	√	√	√	88.0	66.3	37.5

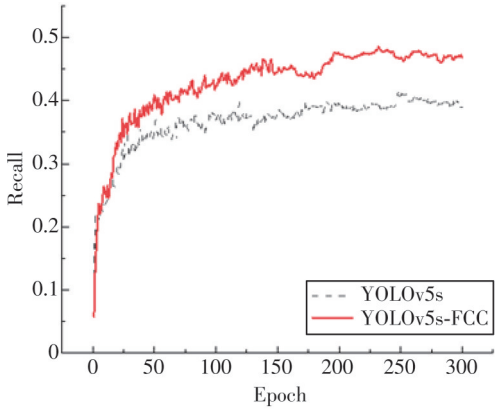
Table7 shows that compared with YOLOv5s, YOLOv5s-F improves on mAP50 and mAP50-95 by 0.9% and 2.8% respectively while decreases detection speed by 2.0 Hz. The YOLOv5s-FC improves on mAP50 and mAP50-95 by 1.8% and 4.0% respectively while decrease detection speed by 3.4 Hz. The highest detection accuracy is achieved by YOLOv5s-FCC, with improvements of mAP50 and mAP50-95 by 2.0% and 4.2% respectively and a decrease in detection speed by 3.3 Hz, which still meets the requirement of real-time detection and verifies the effectiveness and superiority of the improved algorithm.

3.3.5 Validation experiments on VisDrone

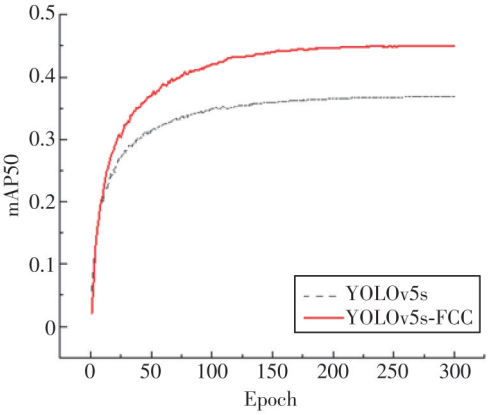
The VisDrone dataset is a UAV-view road target dataset containing 10 categories of targets: cars, people, pedestrians, trucks, vans, buses, motorcycles, bicycles, tricycles, and awning tricycles. To verify the superiority of YOLOv5s-FCC, the experiment was conducted on VisDrone dataset. The loss function, recall and mAP50 of two algorithms during the training process were compared, and the results are shown in Fig.13.



(a) Loss function



(b) Recall



(c) mAP50

Fig. 13 Comparison of three evaluation indicators of YOLOv5s and YOLOv5s-FCC during training process

In Fig.13(a), the loss function of YOLOv5s-FCC is finally converged to 0.04, with an improvement of 0.05 compared with that of YOLOv5s. In Fig.13(b) and (c), the recall and mAP50 values of YOLOv5s-FCC are significantly higher than that of YOLOv5s, which proves the effectiveness of YOLOv5s-FCC.

3.3.6 Comparative experiment of different algorithms

To verify the superiority of YOLOv5s-FCC, it was compared with the mainstream target detection algorithms: Faster RCNN, YOLOv3, YOLOv5s,

YOLOv7 and SE-YOLOv5s[14] under the same conditions on VisDrone dataset. The experimental results are shown in Table 8.

Table 8 Comparative experiment of different algorithms

Algorithm	mAP50/%	FPS/Hz
Faster RCNN	39.8	19
YOLOv3	30.6	7
YOLOv5s	36.9	41
YOLOv7	43.1	125
SE-YOLOv5s[14]	38.7	35
YOLOv5s-FCC	45.1	38

Compared with single-stage Faster RCNN, YOLOv3 and SE-YOLOv5s, the proposed YOLOv5s-FCC algorithm has a significant improvement on both mAP50 and FPS. Compared with YOLOv5s and YOLOv7, YOLOv5s-FCC has improvements on mAP50 by 8.2% and 2.0% respectively while a decrease in detection speed. The experimental results show that YOLOv5s-FCC has significant advantages in terms of detection accuracy in case of meeting real-time detection.

3.3.7 Experimental visualization

The detection results using YOLOv5s and YOLOv5s-FCC were visualized, as shown in Fig.14. In Fig.14(a) and (c), YOLOv5s encounters trouble detecting targets at markers 1 and 3 due to their small size, and it misses crack targets at marker 2 due to the targets close to each other. YOLOv5-FCC effectively resolves the missed detection problem at the markers, as shown in Fig.14(b) and (d), and the confidence level of detected targets is generally increased. By comparing the two algorithms, it is clear that YOLOv5s-FCC can effectively solve the problems of target occlusion as well as false and missed detection of small targets.

Fig. 14 Visualized results comparison of YOLOv5s and YOLOv5s-FCC. a,c: YOLOv5s; b,d: YOLOv5s-FCC

Fig. 15 shows the test results on VisDrone dataset. The road images from VisDrone dataset were selected and tested using YOLOv5s and YOLOv5s-FCC.

In Fig. 15(a), there are a large number of missed small targets such as pedestrians and bicycles by using YOLOv5s, especially in case of target occlusion. In Fig. 15(b), the above-mentioned problems are solved (mark in circles). The target missed detection rate is reduced and detection accuracy rate is improved. The experimental visualization shows that YOLOv5s-FCC also has better detection performance in dense road areas with small targets.

(a) YOLOv5s

(b) YOLOv5s-FCC

Fig. 15 Comparison of YOLOv5s and YOLOv5s-FCC

3.3.8 Road target detection system

The road target detection system was created. When it received the image or video to be detected, after training the weights by YOLOv5s-FCC, adjusting optimal IoU threshold, and so on, the detection system can display the category of detected targets, confidence level, and the number of matching targets, which verifies that the detection system can achieve the detection of single image and videos, especially real-time detection of camera images at 37 Hz. The system interface is shown in Fig.16.

Fig. 16 Detection system interface

Then, the detected targets were calibrated. By dividing the actual length of the pavement by the number of pixels occupied by the road target frame, the exact length and width of the detected road target can be determined. The detection system generates specific information about the road target in a text file, as shown in Fig.17. From this, inspectors can get a comprehensive picture of road conditions and vehicle status, which helps with road maintenance.

```
detect object name is crack
width 8.797 m
length 0.266 m
**************************************
detect object name is crack
width 0.323 m
length 0.893 m
**************************************
detect object name is marking
width 0.19 m
length 1.254 m
**************************************
detect object name is crack
width 1.558 m
length 0.285 m
**************************************
detect object name is crack
width 1.881 m
length 0.665 m
**************************************
detect object name is crack
width 0.266 m
length 1.1019999999999999 m
**************************************
```

Fig. 17 Detection system output

4 Conclusions

The YOLOv5s was optimized by adding a small target detection layer, introducing CBAM, and employing optimized CIoU loss function to form the YOLOv5s-FCC. Then, validation tests were performed on a self-built UAV road target datasets. The results showed that mAP50 and mAP50-95 of YOLOv5s-FCC algorithm were improved by 2.0% and 4.2% than that of YOLOv5s algorithm, respectively. YOLOv5-FCC was also verified on VisDrone dataset. The visualization results confirmed that the improved YOLOv5s algorithm has higher detection accuracy while increasing network footprint size and computational burden, which is a challenge to the next research.

Acknowledgement

This work was supported by Key Research and Development Project of China (No. 2021YFB1600104); National Natural Science Foundation of China; (No. 52002031); Scientific Research Project of Shaanxi Provincial Department of Transportation (No. 20-24K, 20-25X)

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] FAN Z H. Research and application of target detection based on YOLO. Chengdu: Sichuan University, 2021.

[2] DALAL N, TRIGGS B. Histograms of oriented gradients

for human detection//Computer Society Conference on Computer Vision and Pattern Recognition. June 20-25, 2005, San Diego, CA, USA. Piscataway, N. J.: IEEE, 2005: 886-893.

[3] GAO X, WU Y, YANG K, et al. Vehicle bottom anomaly detection algorithm based on SIFT. Optik: International Journal for Light and Electron Optics, 2015, 126(23): 3562-3566.

[4] HENG C K, YOKOMITSU S, MATSUMOTO Y, et al. Shrink boost for selecting multi-lbp histogram features in object detection// IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. Piscataway, N. J.: IEEE, 2012: 3250-3257.

[5] ARDIANTO S, CHEN C J, HANG H M. Real-time traffic sign recognition using color segmentation and SVM// International Conference on Systems, Signals and Image Processing, May 22-24, 2017, Poznan, Poland, USA. Piscataway, N. J.: IEEE, 2017: 1-5.

[6] ELLAHYANI A, ANSARI M E, JAAFARI I E. Traffic sign detection and recognition based on random forests. Applied Soft Computing, 2016, 46: 805-815.

[7] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation//Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. Piscataway, N. J.: IEEE, 2014: 580-587.

[8] GIRSHICK R. Fast R-CNN//IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. Piscataway, N. J.: IEEE, 2015: 1440-1448.

[9] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 2015, 28: 91-99.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection// Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. Piscataway, N. J.: IEEE, 2016: 779-788.

[11] REDMON J, FARHADI A. Yolov3: An incremental improvement. 2018-04-08[2023-01-04]. 2018: 1-6. https://arxiv. org/pdf/1804. 02767. pdf.

[12] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection. 2020-04-03[2023-01-04]. https://arxiv. org/abs/2004. 10934.

[13] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios// International Conference on Computer Vision. October 10-17, 2021, Montreal, Canada. Piscataway. N. J.: IEEE, 2021: 2778-2788.

[14] ZHAO L, WANG X, ZHANG Y. Vehicle object detection based on YOLOv5s fusion SENet. Journal of Graphics, 2022, 43(5): 776-782.

[15] ZHANG S, WANG H, RAN X K. A lightweight traffic sign detection method based on YOLOv5. Electronic Measurement Technology, 2022, 45(8): 129-135

[16] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN//Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. Piscataway, N. J.: IEEE, 2020: 390-391.

[17] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[18] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection//Conference on Computer Vision and pattern recognition, July 21-26, 2017, Honolulu, HI, USA. Piscataway, N. J.: IEEE, 2017: 2117-2125.

[19] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation//Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. Piscataway, N. J.: IEEE, 2018: 8759-8768.

[20] HU J, SHEN L, SUN G. Squeeze-and-excitation networks//Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. Piscataway, N. J.: IEEE, 2018: 7132-7141

[21] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks // Conference on Computer Vision and Pattern Recognition. June 13-19, 2020, Seattle, WA, USA. Piscataway, N. J.: IEEE, 2020: 11534-11542.

[22] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module//European Conference on Computer Vision. September 8-14, 2018, Munich, Germany. Berlin: Springer, 2018: 3-19.

[23] WANG P, HUANG H, WANG M. Complex road target detection algorithm based on improved YOLOv5. Computer Engineering and Applications. 2022, 58(17): 81-92.

改进YOLOv5的无人机影像道路目标检测算法

张　翼，马荣贵*，梁　辰

长安大学 信息工程学院，陕西 西安 710021

摘　要：针对无人机影像中道路小目标漏检和目标之间遮挡导致的目标检测精度低、鲁棒性差等问题，提出一种多尺度融合卷积注意力模块（Convolutional block attention module，CBAM）的YOLOv5道路目标检测算法，即YOLOv5s-FCC。首先，引入小目标感知层对模型进行多尺度改进，增加一个针对小目标的YOLO检测头以提高网络对道路中小目标的特征提取能力。其次，利用CBAM融合空间和通道信息以增强网络中的重要信息，通过将CBAM引入Backbone主干网络不同位置，以获得CBAM最佳融合位置。最后，采用CIoU作为损失函数，以提高边界框预测所需的计算速度和精度。在自建的无人机道路目标数据集上进行训练，结果表明，相较YOLOv5算法，YOLOv5-FCC算法可将mAP50和mAP50-95分别提高2.0%和4.2%。在VisDrone数据集上也验证了YOLOv5-FCC算法的有效性，并建立了基于无人机的道路目标检测系统，实现道路目标的自动检测。

关键词：无人机；道路目标检测；Y OLOv5；损失函数；卷积注意力模块

引用格式：ZHANG Yi, MA Ronggui, LIANG Chen. Road target detection algorithm based on improved YOLOv5 in UAV images. Journal of Measurement Science and Instrumentation, 2024, 15(1): 128-139.