IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 February 2024, accepted 13 March 2024, date of publication 19 March 2024, date of current version 26 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3378748

RESEARCH ARTICLE

MCA-YOLOv7: An Improved UAV Target Detection Algorithm Based on YOLOv7

ZHIYONG QIN, DIKE CHEN, AND HONGYUAN WANG
School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213000, China

Corresponding author: Hongyuan Wang (hywang@cczu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61976028, and in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant SJCX23_1487.

ABSTRACT Aiming at the problems of tiny targets, large target scale changes, and background information interference in target detection of UAV(Unmanned Aerial Vehicle) aerial images, a revised UAV target detection algorithm MCA-YOLOv7 based on YOLOv7 is proposed, and the algorithm advances from the following points: optimizing the FPN(Feature Pyramid Networks) structure to increase the small-target detection layer, and boosting the network's detection ability for small targets. To enhance the multi-scale feature extraction capability, the Efficient Multi-Scale Attention(EMA) is added. In order to reduce the complexity of the model and reduce the confusion of background information, the context aggregation block (CABlock) was introduced and improved, and an effective context aggregation block (ECABlock) was proposed. The loss function CIoU is enhanced and a new loss function FCIoU is proposed, which accelerates the convergence speed of the model, and obtains more accurate regression results. The experimental results demonstrate that the MCA-YOLOv7 model reduces the number of model parameters by 4.7 M and increases the average accuracy (mAP@0.5) by 2.9% when compared to YOLOv7 on the VisDrone2019 dataset. The new algorithm is more capable of handling situations involving UAV aerial photography.

INDEX TERMS UAV, object detection, YOLO, attention mechanism, context aggregation, loss function.

I. INTRODUCTION

With the rapid development of target detection and UAV technology, the application of UAV target detection has been extensively researched [1]. UAVs are widely used in scenarios such as oil pipeline inspection, electric power inspection, crop analysis, and disaster rescue due to their unique high-altitude perspective and efficient data acquisition capabilities [2].

Deep learning applications have developed very rapidly in recent years, such as steel surface defects detection [3], print surface defects detection [4], and pedestrian re-identification [5]. Deep learning-based target detection techniques have also achieved many significant results in UAV applications, but most of these algorithms are convolutional neural networks designed for natural scene images [6], and the optimization of the detection algorithms is still full of challenges for the UAV viewpoint images with a large percentage of small targets, large changes in the scale of the targets, dense targets, complex backgrounds, etc. [7]. These problems are visualized in Fig. 1, for example, cars have a huge change in scale due to different shooting distances, and pedestrians, motorcycles, etc. will appear to be very small targets when photographed from the air.

When it comes to the target detection task, single-stage target detectors have faster detection speeds but slightly lower accuracy compared to two-stage target detectors, which are not suited for real-time UAV identification due to their slower detection speeds. YOLO(You Only Look Once) series is the representative of single-stage detectors. In this paper, we propose an improved model MCA(Multi-Scale Context Aggregation)-YOLOv7 based on YOLOv7 [8] to address the issue of drone target detection. Firstly, the researchers added a small object detection layer to enhance the network's ability to detect small objects. Secondly, the EMA (Efficient Multi-Scale Attention) attention mechanism [9] is introduced into the backbone network to enhance the backbone's ability to

The associate editor coordinating the review of this manuscript and approving it for publication was Halil Ersin Soken.

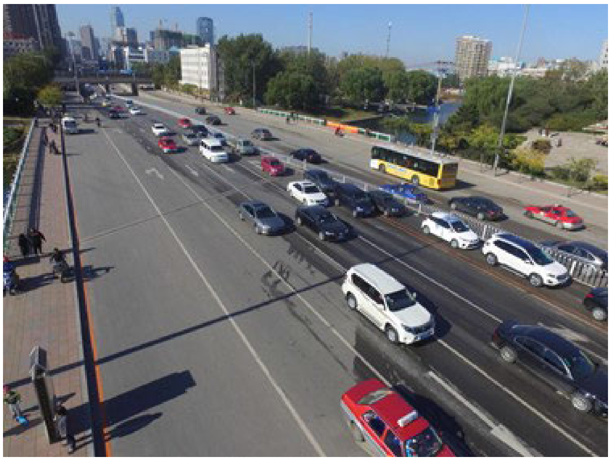

FIGURE 1. Characteristics of aerial images. Problems such as small targets and changes in scale can be seen in the figure.

extract multi-scale features. Afterward, in order to reduce background information interference and enhance feature fusion ability, we have improved based on CABlock(Context Aggregation Block) [10] and proposed an effective Context Aggregation Module (ECABlock). Then we replaced the RepConv module in YOLOv7 with it. Finally, the researchers utilized the idea of FocalL1 Loss [11] to optimize CIoU and proposed a new loss function FCIoU (Focal CIoU), which can optimize the convergence speed of the model and obtain more accurate regression results. Compared with YOLOv7, the improved MCA-YOLOv7 can better handle UAV-captured images. The main contributions of this paper are as follows:

(1) Added a detection layer to improve the model's detection accuracy for small targets.

(2) Integration of the EMA attention module into the model backbone helps the model to focus on target features at different scales.

(3) An effective context aggregation module, ECABlock, is proposed, responsible for aggregating spatial contexts by using a soft reweighting strategy to fuse local and global features while reducing the obfuscation of contextual information.

(4) Using the FocalL1 Loss idea to optimize the CIoU, the FCIoU is proposed to optimize the convergence speed of the model and improve the robustness of the model.

II. RELATED WORK

A. TARGET DETECTION

Deep learning based target detection methods can be classified into two types: the first is a two-stage detection algorithm, which first generates a series of candidate frames by convolutional neural network, and then completes localization and classification. As early as 2013, Girshick et al. [12] proposed R-CNN. R-CNN was a pioneer in the field of target detection after which Faster RCNN [13], Cascade RCNN [14], and so on emerged. Two-stage detection algorithms are usually more accurate but have slower detection speed, and can not meet the real-time requirements; the second is a single-stage detection algorithm, the use of regression ideas will be sent to the input image into the convolutional neural network, after the detection of the direct output to get the results. Representative algorithms are YOLO proposed by Redmon et al. [15] in 2015, RetinaNet proposed by Lin et al. [16] in 2017. After that YOLO series, DETR [17] series of algorithms have also made breakthroughs in camera. The single-stage algorithms lag behind the two-stage in terms of accuracy, but the detection speed is significantly improved.

At present, the application scenarios based on deep learning target detection are very wide. In 2022, Jiao et al. [18] proposed a wheel weld detection method based on the YOLOv4 algorithm, which improved the detection accuracy by optimizing the loss function and anchor frame. Liang et al. [19] proposed a traffic sign detection method for automatic driving scenes in 2022. This method combines ResNeSt and CoordAttention to improve sparse R-CNN and improve the extraction ability of important features. The target detection method for UAV also needs to enhance the ability of feature extraction. The difference is that this research also needs to solve the problem of small target and target scale change. Research on target detection algorithms based on UAV has also made some progress. In 2021 Zhu et al. [20] proposed the TPH-YOLOv5 algorithm, which effectively improves the network's detection performance for small targets by using a transformer prediction head and integrating a CBAM attention module. In 2021 Han et al. [21] proposed the ReDet, which improves remote sensing target detection by rotating the prediction frame and rotating the detector.

B. ATTENTION MECHANISM

The attention mechanism in deep learning is an approach that mimics human attention and is applied to neural networks to enable the model to selectively focus on the more important information in the input. Hu et al. [22] 2017 proposed SENet, which extracts informative features within the local receptive field by fusing spatial and channel-level information to improve the network's representativeness. Zhang and Yang [23] 2021 year proposed the SA attention mechanism, which utilizes feature grouping with channel substitution to improve network performance. Later CA (CoordAttention) [24] fused spatial and channel attention to achieve better results. Zhu et al. [25] proposed BRA (Bi-Level Routing Attention) in 2023, which can filter out the most irrelevant key-value pairs and remove redundant information. To increase the extraction capability of the backbone network for multi-scale features, this paper introduces the EMA attention mechanism in the backbone network part.

C. CONTEXTUAL AGGREGATION

The context aggregation module aims to improve the performance of dense prediction architectures by aggregating

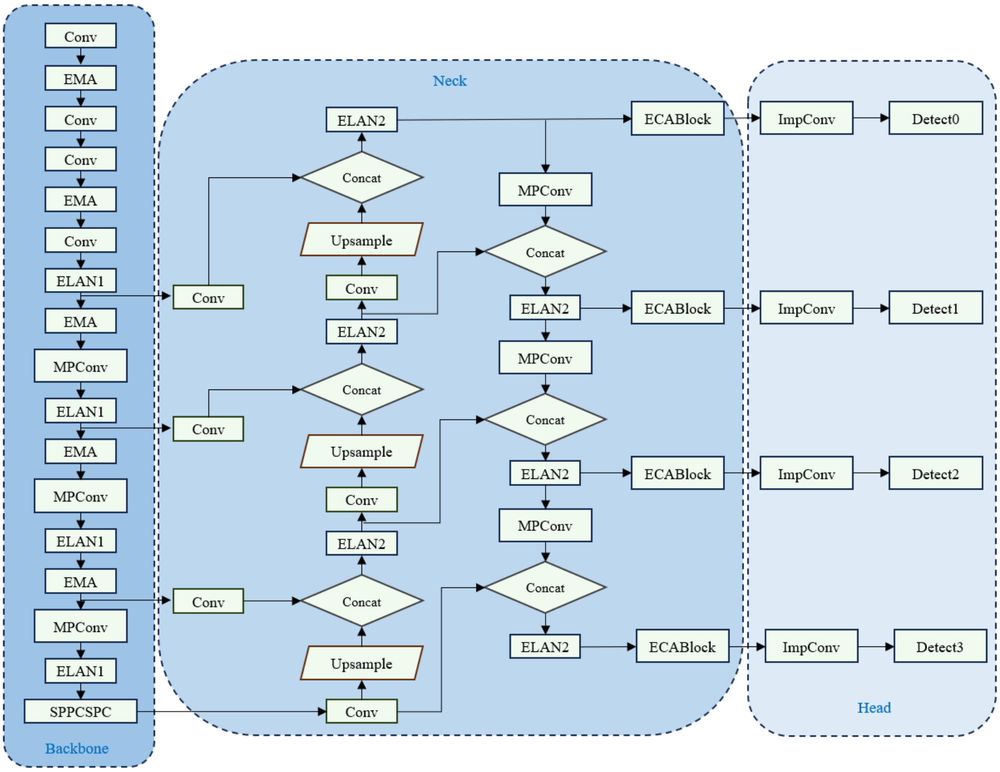

FIGURE 2. Overall structure of MCA-YOLOv7.

multi-scale contextual information. It reduces the confusion of contextual information by fusing local and global features. Yu and Koltun [26] proposed a convolutional network module based on dilated convolution in 2016, which avoids the loss of resolution by aggregating multiscale contextual information. Wang et al. [27] proposed a nonlocal neural network (NLNet) in 2018, which pioneered the aggregates of global spatial contexts by computing pixel-level pairwise correlations to achieve the effect of fusing global information but still suffers from high computational cost. These methods only consider the concept of context as remote spatial correlation, ignoring the global dependencies in the feature and instance domains. The ECABlock proposed in this paper can aggregate pixel-level spatial context in each block, which can improve the feature aggregation capability while reducing the number of parameters and computation.

III. METHODS

A. MCA-YOLOV7

The YOLOv7 target detection algorithm model is mainly divided into three parts: the Backbone, the Neck, and the Head [28]. The backbone part mainly adopts the ELAN module to control the shortest and longest gradient paths so that the network can learn more features to improve the model's robustness. The Neck section refers to the PAFPN structure to fuse feature maps of different scales separately. The Head section first adjusts the number of channels through RepConv and then uses 1 × 1 convolution to predict the results.

The overall structure of the improved MAC-YOLOv7 algorithm model in this paper is shown in Fig. 2. To address the problem that aerial targets contain many tiny targets, this paper adds a detection head that is introduced from the high-resolution feature maps at the lower level that contain more information about small objects, which improves the model's ability to detect small targets. Although the addition of the extra detection head increases the number of model parameters, the detection performance for small targets is also substantially improved. The four detection heads also work well when dealing with drastic changes in target scale. Attention mechanisms are added after the partial convolution and ELAN modules of the backbone to enhance multi-scale feature extraction. In this paper, RepConv is replaced by the proposed ECABlock, which improves the model performance by aggregating pixel-level contextual information while reducing the number of parameters and computation of the model. Finally, the loss function is optimized to accelerate convergence and improve model accuracy.

B. IMPROVED BACKBONE BASED ON EFFICIENT MULTI-SCALE ATTENTION

There are always a lot of background factors interfering with the images captured by UAVs. The height variation of drones can cause significant changes in the scale of the target. Using the EMA attention mechanism can help the neural network to pay more attention to useful information at different scales and resist confusing information. The structure of EMA attention is shown in Fig. 3. The feature map $X(X=[X_0, X_i, \ldots, X_{G-1}], X_i \in R^{C//G\times H\times W})$ will first be grouped into feature groups after being passed into the EMA module. The EMA will classify the X across the channel dimensionality direction into G sub-features, which will be used to learn different semantics. Next the feature subgraphs are passed into two parallel subnets, which use 1×1 kernel and 3×3 kernel respectively to be able to better capture multi-scale spatial structure information and enhance multi-scale feature extraction capability. The parallel placement of the two branches can avoid the sequential processing leading to the excessive depth of the model and realize the fast response of the module. Two lines are included in the 1×1 branch for 1D global average pooling along two spatial directions respectively, after which they are spliced together along the height direction, and the output is decomposed into vectors of two using a shared 1×1 convolution. The output is fitted to a 2D binomial distribution using a nonlinear Sigmoid function, after which the channel weights of each parallel branch are recalibrated. Another 3×3 branch captures local cross-channel interactions via 3×3 convolution to expand the feature space. This allows the EMA to adjust the importance of different channels while preserving spatial structure information. EMA also employs a cross-space learning approach that uses matrix dot product operations to fuse the output feature maps of two parallel subgrids, thus capturing pixel-level pairwise relationships and highlighting the global context of all pixels. Equation.1 is the formula for global average pooling in cross-space learning. z_c is the average value of the cth channel in the image, and x_c is the value of the cth channel denoting the (i, j) pixel location in the image.

$$z_c = \frac{1}{H \times W} \sum_{j}^{H} \sum_{j}^{W} x_c\,(i, j) \tag{1}$$

C. EFFECTIVE CONTEXT AGGREGATION BLOCK

CABlock is a context aggregation block with residual connections, which can be integrated into the network to reduce the number of parameters and computation, but the accuracy improvement is very limited. This article refers to the idea of parallel branching in EMA, which can avoid increasing the number of model layers and achieve fast response. The researchers placed a parallel branch of 1×1 convolution in CABlock. Although the sense field of 1×1 convolution is very small, it can introduce nonlinear features through the underlying nonlinear activation function to enhance the model representation, and only increase the number of parameters by a very small amount. This branch uses residual concatenation to stitch the input feature maps directly to the output after a single 1×1 convolution to retain more image information while mitigating network degradation that occurs as the network layers deepen. We refer to the improved module as ECABlock, and its structural diagram is shown in Fig.4.

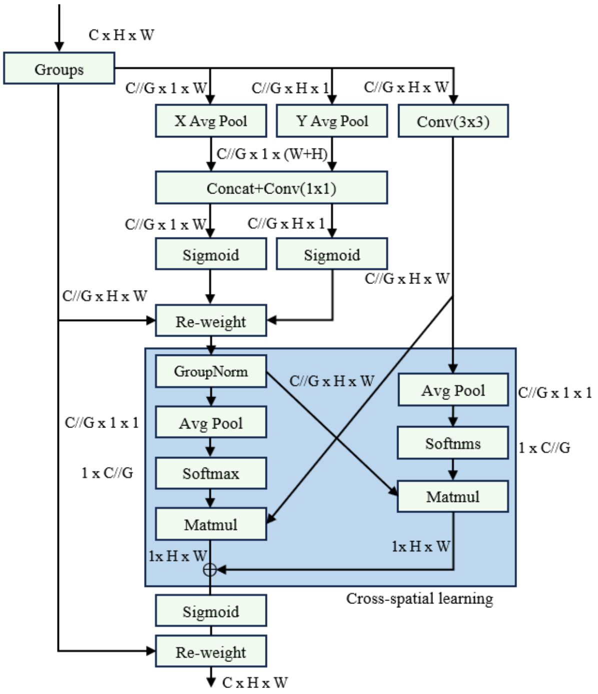

FIGURE 3. EMA Structure of Attention Mechanisms.

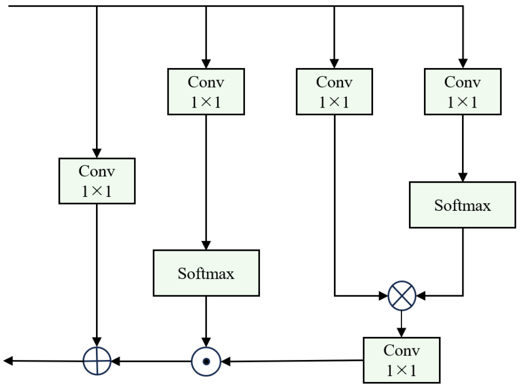

FIGURE 4. ECABlock Context Aggregation Module Architecture Diagram.

The ECABlock is responsible for aggregating spatial contexts by using soft-weighting strategies to fuse local and global features. ECABlock tends to aggregate objects from the same category or even global spatial contexts from similar or semantically related categories. This helps to cope with the challenge of scale variation in aerial images while reducing information confusion. The comparison experiments in Table 1 validate the effectiveness of ECABlock in this paper. Although the improved CABlock can significantly reduce the amount of computation and the number of parameters, the average accuracy improvement is only 0.1%.

The ECABlock proposed in this paper, on the other hand, achieves an accuracy improvement of 0.7%, while the number of parameters is reduced by 5.2M and the computational effort is reduced by 9.78%.

TABLE 1. Context Aggregation Module Before and After Improvement.

Method	mAP@0.5(%)	Parameters(M)	FLOPs(G)
YOLOv7(baseline)	48.3	36.53	103.3
YOLOv7+CABlock	48.4	30.99	92.6
YOLOv7+ECABlock	49.0	31.33	93.2

D. LOSS FUNCTION

The IOU loss function is commonly used in object detection, which represents the intersection union ratio of boxes A and B. The calculation formula for IoU is shown in Eq. 2. The localization loss calculation originally used in YOLOv7 is the CIoU loss function. The CIoU takes into account the three most important factors in the box-regression task: overlap area, aspect ratio, and centroid distance. The convergence of CIoU compared to the previous loss function speed and detection accuracy are significantly improved. Assuming that the predicted frame b and the real frame b^{gt} are given then the CIoU is computed as shown in Eqs. 3, 4 and 5. Where ρ represents the calculated distance, and c represents the diagonal length covering the overlapping part of two boxes, w and h are the width and height of the predicted box,w^{gt} and h^{gt} are the width and height of the realistic box, and α and v are the two coefficients measuring the aspect ratio.

$$L_{IoU} = 1 - \frac{\left|A \bigcap B\right|}{\left|A \bigcup B\right|} \tag{2}$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{3}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{4}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{5}$$

The VisDrone2019 dataset used in this paper has the problem of uneven training data, the data distribution is shown in Fig. 5, and this dataset is mostly small targets, which have very small true and predicted boxes. As long as the prediction box is slightly offset, the IoU value will undergo significant changes. This results in low-quality regression boxes, and the number of samples with smaller errors is much smaller than those with larger errors. The Focal L1 Loss proposed in Focal-EIoU sets a gradient value, and the value of the gradient should also be small where the error rate is small, and vice versa where the error rate is large, the gradient should be large, so that the low-quality samples can be suppressed very well. The gradient calculation and Focal L1 Loss calculation are shown in Eqs. 6 and 7, respectively. α, and β are the parameters that control the gradient value.

$$g(x) = \frac{\partial L_f}{\partial_x} = \begin{cases} -\alpha x \ln(\beta x), & 0 < x \le 1; 1/e \le \beta \le 1, \\ -\alpha \ln(\beta), & x > 1; 1/e \le \beta \le 1. \end{cases} \tag{6}$$

$$L_f(x) = \begin{cases} -\frac{\alpha x^2 (2\ln(\beta x) - 1)}{4}, & 0 < x \le 1; 1/e \le \beta \le 1, \\ -\alpha \ln(\beta) x + C, & x > 1; 1/e \le \beta \le 1. \end{cases} \tag{7}$$

This article reweights the CIoU loss function using IoU values to obtain the FCIoU loss function, which is calculated using formula $L_{FCIoU} = IoU^{\gamma} L_{CIoU}$. Where γ is a parameter controlling the degree of outlier suppression. The convergence speed and detection accuracy of the FCIoU exceeds that of the original CIoU, and the FCIoU has a very good relevance for the case of sample imbalance difficult samples are many.

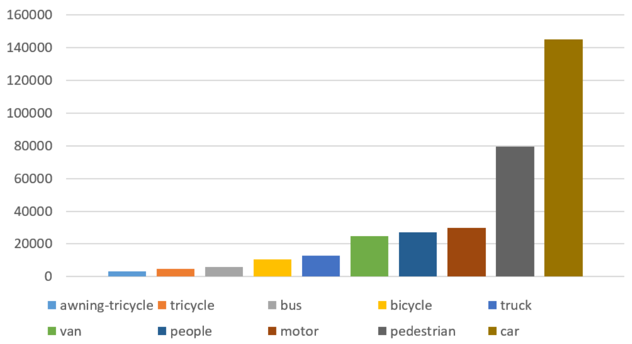

FIGURE 5. The figure shows the data distribution of the VisDrone2019 data set. The data volume for cars and pedestrians is far higher than that of other samples, and the data distribution is clearly highly imbalanced.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION INDICATORS

The dataset used in this paper is VisDrone2019 [29], which was collected by the AISKYEYE team of the Machine Learning and Data Mining Laboratory of Tianjin University. It contains 10,209 images, and the training set, validation set, and test set contain 6,471, 548, and 1,610 images, respectively. All of them are images collected using drones in different cities, different weather, and different periods. There are many small targets in the dataset, and most of the targets are densely distributed with large-scale variations. The dataset contains 10 categories: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. The comparison experiments in this paper are all conducted on the VisDrone2019 test set, and the ablation experiments conducted on the VisDrone2019 validation set were performed.

The evaluation metrics used in this paper are mean average precision mAP (mean Average Precision), parameter number

(Parameter), and computational volume (FLOPs). mAP is a comprehensive calculation of different categories of accuracy more comprehensive, parameter number, and computational volume that is mainly responsible for evaluating the model complexity. The details are described below.

(1)Average Accuracy Mean Value mAP: The mAP@0.5 used in this article refers to the mean average precision at an IoU threshold of 0.5, where average precision (*AP*) reflects the accuracy of object detection for a single class. The calculation formula for *AP* is shown in Eq. 8. Where *P* is Precision and *R* is Recall, their calculation methods are shown in Eqs. 9 and 10, respectively. Where *TP* denotes positive cases of correct prediction, *FP* denotes positive cases of incorrect prediction. *FN* is a negative case of incorrect prediction.

$$AP = \int_0^1 P(R)\,d(R) \tag{8}$$

$$P = \frac{TP}{(TP + FP)} \tag{9}$$

$$R = \frac{TP}{(TP + FN)} \tag{10}$$

The mAP is the average of the mean accuracies of all the categories, the higher the mAP value the better the model detection. mAP is calculated as shown in Eq. 11, where *n* is the number of categories.

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{11}$$

(2)Parameters: the number of Parameters is usually used to evaluate the algorithmic model space complexity, the larger the number of Parameters represents that the model can learn more knowledge, and at the same time requires more storage space and arithmetic power, and the opposite is true for the smaller number of Parameters.

(3)FLOPs(Floating Point Operations Per Second): FLOPs are the number of floating-point operations and are used to measure the time complexity of an algorithm.

B. EXPERIMENTAL ENVIRONMENT

The experimental environment of this paper is shown in Table 2. The hyperparameters of all experiments are kept unchanged using the default values of YOLOv7. The FPS values in the comparative experiment were obtained using GeForce RTX 3080 Laptop testing. The number of training rounds for the ablation experiments is 300 epochs, while the convergence speed of MCA-YOLOv7 becomes slower relative to YOLOv7 so it is set to 400 epochs.

C. COMPARISON EXPERIMENT

1) COMPARISON OF DIFFERENT ALGORITHMS

To verify the effectiveness of the algorithm proposed in this paper, this paper is compared with some mainstream target detection models and newer detection algorithms in recent years on the VisDrone2019 test set, and the results of the comparison experiments are shown in Table 3.

TABLE 2. Experimental environment configuration.

Experimental Environment	Versions
Programming Languages	Python3.8
Deep Learning Framework	Pytorch1.12.1
CUDA	10.2
OS	Ubuntu16.04
GPU	NVIDIA RTX2080Ti (11GB)

The experimental data in the table shows that, using the VisDrone data set, the mAP@0.5 of the MCA-YOLOv7 algorithm suggested in this paper is 42.7%. It significantly outperforms a few popular target identification algorithms, including YOLOv5, CenterNet [30], Cascade R-CNN, and Faster R-CNN. Although the average precision value of CDNet [31] is much lower than that of the algorithm presented in this article, it surpasses the algorithm of this article in the two similar categories of Tricycle and Awning-tricycle. Compared with the baseline model YOLOv7, the mAP@0.5 of MCA-YOLOv7 has increased by 1.8%, especially with a significant improvement in small object categories such as Pedestrian, People, and Motor. Compared with the latest detection algorithms of YOLOv8 and RT-DETR [32], the average precision value of MCA-YOLOv7 has increased by 5.9% and 4.3%, respectively. This study also compared the inference speed of some single-stage object detection algorithms. Although MCA-YOLOv7 leads in accuracy, the inference speed has decreased to some extent. The reason is that adding EMA will increase inference time, and how to improve the FPS of the model is a focus of future research.

2) COMPARISON OF DIFFERENT ATTENTION EFFECTS

To investigate the effect of different attentions in network models, this paper tests some of the attentional mechanisms in recent years with the baseline model YOLOv7.

As can be seen from Table 4, the excellent extraction ability of the EMA attention mechanism for multi-scale features results in a 0.7% increase in accuracy relative to the benchmark model. The accuracy of the EMA compared to the CA, SA, and BRA attention is improved by 0.6%, 0.3%, and 0.8%, respectively. The number of parameters and the amount of computation increase slightly relative to the baseline model, and the difference with the other attentions is very small.

To validate the impact of the channel grouping number G in EMA attention on detection performance, this study conducted comparative experiments using G=8, G=16, and G=32, and the results were evaluated on the VisDrone validation dataset. The experimental results are shown in Table 5. When G=8, the performance is the worst with a mAP@0.5 of only 41.2%. Similarly, with G set to 16, the mAP@0.5 remains at 41.2%. However, the best results are achieved when G is set to 32, with a mAP@0.5 of 41.6%.

TABLE 3. Comparison of experimental results of different algorithms.

Model	Pedestrian	People	Bicycle	Car	Van	Truck	Tri	Awn-Tri	Bus	Motor	mAP0.5(%)	FPS
Faster R-CNN	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2	21.8	-
CenterNet	22.6	20.6	14.6	59.7	24.0	21.3	20.1	17.4	37.9	23.7	26.2	-
Cascade R-CNN	22.2	14.8	7.6	54.6	31.5	21.6	14.8	8.6	34.9	21.4	23.2	-
YOLOv5-l	31.4	20.1	11.9	74.4	37.1	41.0	21.1	19.5	58.8	30.0	34.5	50
CDNet	35.6	19.2	13.8	55.8	42.1	38.2	**33.0**	**25.4**	49.5	29.3	34.2	-
YOLOv7	38.7	26.9	16.7	79.6	44.3	49.0	27.1	23.3	62.4	41.2	40.9	**53**
YOLOv8-l	31.2	17.5	14.3	75.0	42.8	47.6	22.5	20.3	62.2	34.6	36.8	40
RT-DETR-r50	38.2	27.6	13.2	77.8	38.1	43.8	25.4	20.6	57.8	41.4	38.4	28
MCA-YOLOv7	**42.0**	**27.8**	**18.1**	**81.6**	**47.1**	**52.0**	26.8	25.3	**63.2**	**43.1**	**42.7**	35

TABLE 4. Comparison of experimental results of different attention mechanisms.

Model	mAP@0.5(%)	Parameters (M)	FLOPs(G)
YOLOv7	40.9	36.53	103.3
YOLOv7+CA	41.0	36.57	103.4
YOLOv7+SA	41.3	36.53	103.3
YOLOv7+BRA	40.8	37.58	113.6
YOLOv7+EMA	41.6	36.54	107.0

TABLE 5. Experimental results of EMA when grouped with different characteristics.

Number of channel groups(G)	mAP@0.5(%)
8	41.2
16	41.2
32	41.6

3) COMPARISON OF THE EFFECT OF DIFFERENT LOSS FUNCTIONS

In this paper, the loss function used in MCA-YOLOv7 is FCIoU, to verify that the proposed FCIoU has a better effect on model convergence and improving the average accuracy, we conducted a comparison test of the average accuracy on the YOLOv7 benchmark model using CIoU, SIoU [33], MPDIoU [34], and Focal-EIoU, respectively, and the results of the experimental are shown in Table 6. It can be seen that the use of FCIoU gives the best results, the accuracy improvement is 0.2% compared to CIoU and Focal-EIoU, the use of MPDIoU resulted in a 0.1% improvement, while SIoU decreased by 0.1%.

TABLE 6. Experimental results with different loss functions.

Loss Function	mAP@0.5(%)
CIoU	40.9
SIoU	40.8
MPDIoU	41.0
Focal-EIoU	40.9
FCIoU	41.1

This paper also compares the convergence speed with MCA-YOLOv7 which uses CIoU. From Fig. 6, it can be seen that MCA-YOLOv7 using FCIoU converges much faster. It takes 400 rounds to converge using CIoU, while FCIoU converges after 360 rounds of training, reducing the training time by 40 rounds. The detection accuracy on the VisDrone test set is also slightly higher with FCIoU as shown in the results in Table 7.

TABLE 7. Comparison of CIoU and FCIoU experimental results.

Methods	mAP@0.5(%)	Epochs
+CIoU	42.6	400
+FCIoU	42.7	360

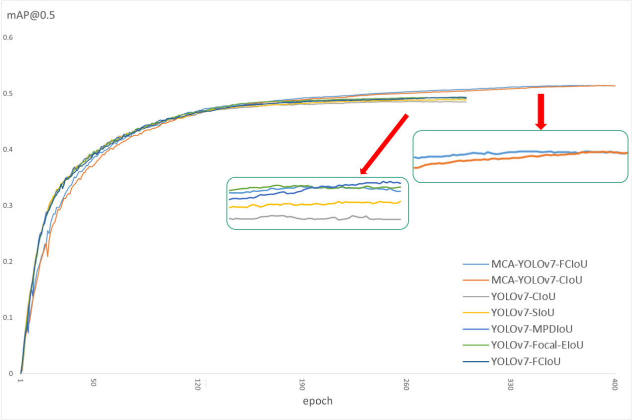

FIGURE 6. Convergence of different loss functions.

D. ABLATION EXPERIMENT

To verify the effectiveness of the introduced EMA attention module, the proposed ECABlock, and the proposed FCIoU, this article conducted an ablation experiment to compare the mAP@0.5 average accuracy value and the changes in model parameters and calculation amount after adding each module. The ablation experiment results were compared. Obtained from VisDrone validation set. The experimental results are shown in Table 8.

TABLE 8. Results of ablation experiments.

Method	mAP@0.5(%)	Parameters(M)	FLOPs(G)
YOLOv7(baseline)	48.3	36.53	103.3
YOLOv7+P2	50.5	37.08	117.1
YOLOv7+EMA	49.8	36.54	107.0
YOLOv7+ECABlock	49.0	31.33	93.2
YOLOv7+FCIoU	48.7	36.53	103.3
YOLOv7+P2+EMA	50.8	37.09	120.8
YOLOv7+P2+EMA +ECABlock	51.1	31.83	107.2
MCA-YOLOv7	51.2	31.83	107.2

After adding the p2 small target detection layer alone, the accuracy of the model is significantly improved, which is 2.2% higher than the baseline, indicating that the p2

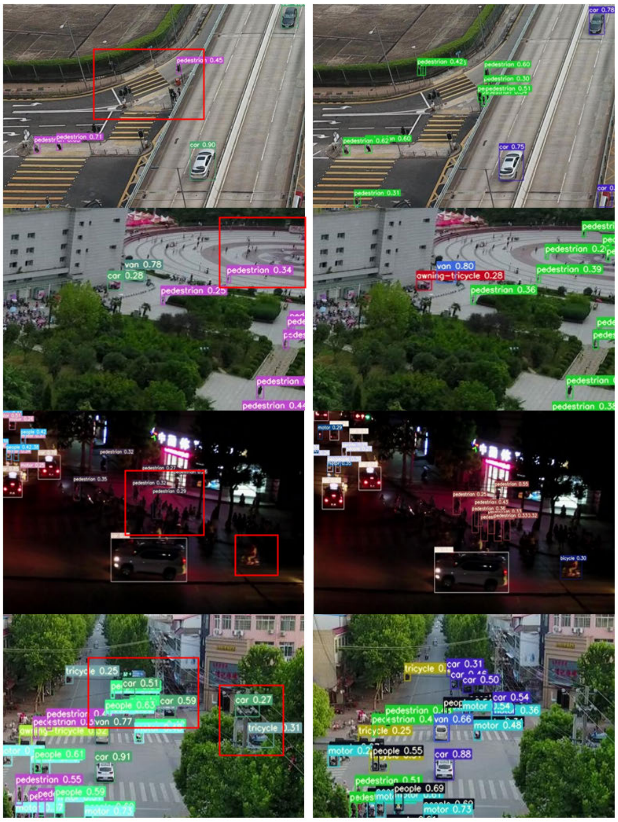

FIGURE 7. Comparison of YOLOv7 and MCA-YOLOv7 detection effect.The YOLOv7 detection result is on the left, while the MCA-YOLOv7 result is on the right. It is evident that, in comparison to the left, the right has a lower false and missed detection rate.

layer's detection performance for small targets has been greatly improved. However, adding a detection layer led to an increase in model complexity, with the number of parameters increasing by 0.55M and the FLOPs increasing by 13.36%. After adding EMA alone, mAP@0.5 increased by 1.5%, which proves that introducing EMA at different positions on the backbone enhances the model's attention to targets of different scales, and EMA serves as a lightweight attention module parameters, and FLOPs is only slightly improved. After adding the proposed ECABlock, the accuracy increased by 0.7%, the number of parameters was reduced by 5.2M and FLOPs were reduced by 9.78%, which improved the detection performance while reducing the complexity of the model. Using the improved FCIoU, the accuracy increased by 0.4%, which proves that suppressing low-quality samples can effectively improve model detection capabilities. The fusion of the p2 layer and EMA can improve accuracy, but the complexity of the model has also reached its maximum. After adding ECABlock, the progress has increased by 0.3%, and parameters and FLOPs have significantly decreased. After using FCIoU, there is no significant change in accuracy, but the convergence speed of the model has increased.

E. VISUAL ANALYSIS OF EXPERIMENTAL RESULTS

To more intuitively compare the improvement of the algorithm proposed in this article, Fig. 7 shows the comparison of the detection effect of the original YOLOv7 and the detection effect of MCA-YOLOv7. The left side is the YOLOv7 detection picture and the right side is MCA-YOLOv7. It can be seen from the pictures in the first and second rows that YOLOv7 is prone to missed detections for very small targets, and the algorithm proposed in this article is significantly better. The pictures in the third row are detection comparisons taken at night. It can be seen that the improved algorithm can better detect targets under poor lighting conditions. It can be seen from the fourth row of pictures that the original algorithm is prone to some false detections and missed detections, and the algorithm proposed in this article greatly improves this situation. In summary, compared with YOLOv7, MCA-YOLOv7 has improved detection capabilities when the target is small and the environment is complex.

V. CONCLUSION

Nowadays, UAV target detection still faces many challenges, and this paper proposes an improved detection algorithm MCA-YOLOv7 based on YOLO-v7 to address these challenges, which improves the model's performance on small target detection by adding a small target detection layer; enhances the multi-scale feature extraction capability of the backbone network by adding the EMA attention mechanism; enhances global feature fusion by incorporating the proposed ECABlock context aggregation module The proposed ECABlock context aggregation module is incorporated to enhance the global feature fusion, which reduces the number of parameters in the model while reducing the background information mixing; the proposed FCIoU is improved based on the ideas of CIoU and FocalL1 loss to accelerate the model convergence speed and improve the detection progress. The experimental results show that the average accuracy value of the proposed MCA-YOLOv7 algorithm on VisDrone2019 dataset exceeds that of YOLO-v7 by 2.9%, and the number of parameters is reduced by 4.7 M. However, the algorithm in this paper is still deficient, and in future work, it will continue to be studied to further reduce the complexity of the model and speed up the convergence speed of the model to reduce the training time while ensuring accuracy.

ACKNOWLEDGMENT

The authors sincerely thank all those who helped with this study.

REFERENCES

[1] J. Liu, Y. Lu, Y. Chen, Q. Zhao, Z. Qin, and Y. Fu, "Research on low-altitude UAV aerial photography target detection," in *Proc. Int. Conf. Comput. Netw., Electron. Autom. (ICCNEA)*, Sep. 2022, pp. 369–372.

[2] M. H. Sabour, P. Jafary, and S. Nematiyan, "Applications and classifications of unmanned aerial vehicles: A literature review with focus on multi-rotors," *Aeronaut. J.*, vol. 127, no. 1309, pp. 466–490, Mar. 2023.

[3] Q. Zhou, H. Wang, Y. Tang, and Y. Wang, "Defect detection method based on knowledge distillation," *IEEE Access*, vol. 11, pp. 35866–35873, 2023.

[4] X. Zihao, W. Hongyuan, Q. Pengyu, D. Weidong, Z. Ji, and C. Fuhua, "Printed surface defect detection model based on positive samples," *Comput., Mater. Continua*, vol. 72, no. 3, pp. 5925–5938, 2022.

[5] H. Chen, H. Wang, Z. Ding, and P. Li, "Dual attention network for unsupervised domain adaptive person re-identification," *IEEE Access*, vol. 11, pp. 88184–88192, 2023.

[6] J. Dai, L. Wu, and P. Wang, "Overview of UAV target detection algorithms based on deep learning," in *Proc. IEEE 2nd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA)*, vol. 2, Dec. 2021, pp. 736–745.

[7] F. Wang, H. Wang, Z. Qin, and J. Tang, "UAV target detection algorithm based on improved YOLOv8," *IEEE Access*, vol. 11, pp. 116534–116544, 2023.

[8] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[9] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[10] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, and C. W. Chen, "Learning to aggregate multi-scale context for instance segmentation in remote sensing images," 2021, *arXiv:2111.11057*.

[11] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–7.

[14] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[18] T. J. Liang, W. G. Pan, H. Bao, and F. Pan, "Vehicle wheel weld detection based on improved YOLO v4 algorithm," *Comput. Opt.*, vol. 46, no. 2, pp. 271–279, Apr. 2022.

[19] T. Liang, H. Bao, W. Pan, and F. Pan, "Traffic sign detection via improved sparse R-CNN for autonomous vehicles," *J. Adv. Transp.*, vol. 2022, pp. 1–16, Mar. 2022.

[20] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[21] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2785–2794.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[23] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2235–2239.

[24] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[25] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10323–10333.

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[27] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[28] Y. Wang, H. Wang, and Z. Xin, "Efficient detection model of steel strip surface defects based on YOLO-V7," *IEEE Access*, vol. 10, pp. 133936–133944, 2022.

[29] D. Du, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 213–226.

[30] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[31] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDNet: CNN-based cloud detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6195–6211, Aug. 2019.

[32] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "Detrs beat YOLOs on real-time object detection," 2023, *arXiv:2304.08069*.

[33] Z. Gevorgyan, "Siou loss: More powerful learning for bounding box regression," 2022, *arXiv:2205.12740*.

[34] M. Siliang and X. Yong, "MPDIoU: A loss for efficient and accurate bounding box regression," 2023, *arXiv:2307.07662*.

ZHIYONG QIN received the B.E. degree from Huaide College, Changzhou University, in 2022, where he is currently pursuing the M.E. degree. His research interests include computer vision and object detection.

DIKE CHEN received the Master of Science degree from The University of Manchester, in 2016. He is currently pursuing the Ph.D. degree in engineering with Changzhou University. His main research interests include computer vision and electric automatization.

HONGYUAN WANG received the Ph.D. degree in computer science from Nanjing University of Science and Technology. He is currently a Professor with Changzhou University. His research interests include pattern recognition, intelligence systems, and pedestrian trajectory discovery in intelligent video surveillance.