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

UAV Target Detection Algorithm with Improved YOLOv7

Fanrun Meng, Chen Liu, Zhiren Zhu, Liming Zhou

College of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China

Abstract: The wide application of UAV technology in various fields makes UAV target detection crucial. In this study, we propose an improved algorithm based on YOLOv7 to enhance the performance and robustness of UAV target detection. We utilize YOLOv7 as the infrastructure and introduce BiFPN (Bi-directional Feature Pyramid Network) to enhance the feature fusion, while adding the GAM attention mechanism to the model, which is trained and evaluated using the VisDrone2019 dataset. The experimental results of this study show that the improved model achieves an average accuracy mAP value of 45.6%, which is 2.7% higher than the traditional model, and is able to detect and localize UAV targets more accurately.

Keywords: Unmanned Aerial Vehicle (UAV); YOLOv7; BiFPN; GAM.

1. Introduction

UAV technology has been widely used in various fields such as military, civil and scientific research. UAV target detection has important practical significance as one of the cores of UAV applications. Accurately detecting and localizing UAV targets can be used in application scenarios such as surveillance, security, and rescue. However, UAV target detection tasks face many challenges, including complex backgrounds, multi-scale targets, and occlusion. Therefore, it is crucial to develop an efficient and accurate UAV target detection algorithm.

2. Overview of Target Detection Algorithm

2.1. Traditional Target Detection Algorithm

Traditional target detection algorithms use a sliding window strategy to select candidate regions on a given image, and then extract features (e.g., SIFT[1],HOG[2]) for these regions , and finally use a trained classifier (e.g., SVM[3], AdaBoost[4]) for classification. This method has the disadvantages of poorly targeted region selection strategy, high time complexity, and poor robustness. With the rapid development of artificial intelligence technology, it has been gradually replaced by deep learning methods of features.

2.2. YOLO Series Target Detection Algorithm

After entering the era of deep learning, target detection algorithms keep emerging like a spring, which can be roughly divided into two-stage algorithms represented by the R-CNN[5][6][7][8] series of algorithms and one-stage algorithms represented by the YOLO[9] series of algorithms. The method is divided into two steps, the first step is to generate a candidate area, and the second step is to divide the candidate area into a number of candidate areas, and then divide the candidate area into a number of candidate areas and correct the position of the candidate areas. The method has high accuracy and small probability of missed detection, but its calculation speed is slow and cannot meet the real-time requirements. The YOLO family of algorithms was developed to solve the above problems. The one-stage target detection approach, which treats the target detection problem as a regression problem, simultaneously predicts the location and class of targets through a single neural network model instead of selecting candidate regions one by one. This gives YOLO a good balance between speed and accuracy.

R-CNN is the first algorithm that utilizes CNN for target detection, which opens the era of target detection based on deep learning. YOLO is the first target detection algorithm, and many researchers have proposed a series of high-performance target detection algorithms under the influence of YOLO, and the R-CNN and YOLO algorithms have greatly promoted the development of target detection technology.

2.3. Theory Related to YOLOv7

YOLOv7 (You Only Look Once version 7)[10] is a deep learning model in the field of target detection, which is the latest version of the YOLO series. The main feature of the YOLO series of models is the ability to perform target detection in real time while maintaining high accuracy.YOLOv7 improves and optimizes on YOLOv5, aiming to improve the detection performance and speed. YOLOv7 is based on YOLOv5 and has been improved and optimized with the aim of increasing detection performance and speed. The model structure includes four parts: Input, Backbone, Head, and Neck.

Input: The main role is to perform a series of preprocessing operations on the input image, including Mosaic data enhancement, adaptive anchor frame calculation, and image scaling.

Backbone: The consists of several CBS convolutional modules, ELAN modules and MP-1 modules. the CBS module consists of a convolutional layer, a bulk-normalized BN layer and a SiLU activation function. the ELAN module consists of a number of convolutional modules, which are used to learn and converge more efficiently by controlling the shortest and longest gradient paths.

Head: It mainly includes SPPCSPC, ELAN-W, UPsmaple and MP-2. It performs feature processing on the output image of the backbone network, and adopts the Path Aggregation Feature Pyramid Network (PAFPN) structure for multi-scale feature fusion. The top-down structure is used to pass down the deep strong semantic features to enhance the features of the whole pyramid; finally, the bottom-up structure is used to pass up the shallow image structure, color, edge, position and other feature information, thus realizing the efficient fusion of features at different levels.

Neck: It uses the REP structure to adjust the number of

channels to the P3, P4, and P5 features output from the PAFPN structure. Finally, these features are fed into a 1×1 convolutional module for predicting the confidence, category and anchor frame information of the image and generating the final detection results.

3. Improvements to the YOLOv7

3.1. Introducing BiFPN

BiFPN (Bi-directional Feature Pyramid Network)[11] is a deep neural network architecture for target detection and semantic segmentation tasks, which is designed to extract multi-scale features in images and help the model to better understand and process targets at different scales. The design of BiFPN is inspired by network architectures such as FPN (Feature Pyramid Network) and network architectures such as PANet, which has higher efficiency and performance in processing multi-scale features. As shown in Figure 1.

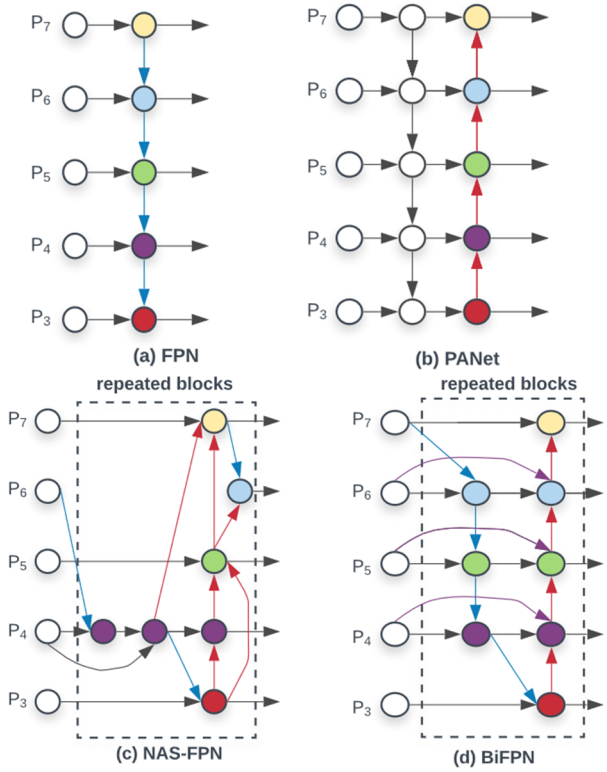

Figure 1. BiFPN structure

BiFPN consists of multiple repeating modules, each consisting of the following steps: a. Bottom-up connectivity: propagating information from high-resolution bottom features to low-resolution top-level features. b. Top-down connection: propagating information from low-resolution high-level features to high-resolution bottom-level features. c. Cross-layer connectivity: propagate information between different layers to make feature fusion more complete. d. Feature fusion: fuses feature from different connections to generate the final multi-scale feature pyramid.

In order to optimize the detection process, we try to embed BiFPN in the feature extraction network of YOLOv7, replacing the original backbone and neck networks to enhance the fusion of multi-scale features, thus improving the performance of target detection.

3.2. Attention

In the study of human vision, it has been found that humans selectively focus on certain visible information and ignore other information to rationally utilize the limited visual processing resources. At the same time, it has been found that selective encoding of input data can effectively improve the expression and generalization ability of neural networks. The attentional mechanism simulates that humans selectively focus on certain visible information and ignore other information to rationally utilize limited visual processing resources. At the same time, it also helps to solve the problem of the existence of a large amount of redundant information in the image or video, thus improving the neural network expressive ability and generalization ability.[12]

In order to better balance the model's lightweight and detection accuracy, this paper proposes to add the GAM (Global attention mechanism) attention mechanism to the YOLOv7.GAM redesigns the sub-module of CBAM, as shown in Figure 2, with two modules, the channel attention mechanism module and the spatial attention mechanism. The relevant information is extracted by selectively focusing on the desired parts of the channel and space, and important features are captured in the 3D channel, spatial width and spatial height to improve the recognition accuracy of the model.

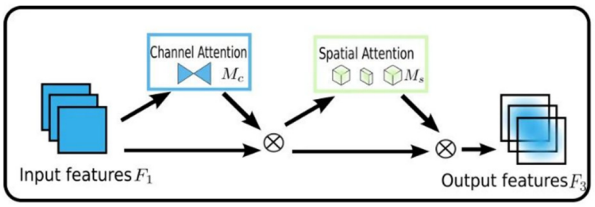

Figure 2. GAM structure

4. Experimental Design and Analysis of Results

4.1. Dataset

The VisDrone2019 dataset was collected by the AISKYEYE team at the Machine Learning and Data Mining Laboratory of Tianjin University. The benchmark dataset consists of 288 video clips consisting of 261,908 frames and 10,209 still images captured by a variety of UAV cameras covering a wide range of locations, environments (urban and rural), objects (pedestrians, vehicles, bicycles, etc.), and densities (sparse and crowded scenes). Some important attributes, including scene visibility, object class and occlusion, also provide better data utilization. This dataset is valuable to study for ships with small targets, overlapping targets, and complex backgrounds as a difficulty, resulting in a low mAP for detection. In this experiment, 1,550 randomly selected from the dataset and 464 images were selected to form the training and validation set for this experiment respectively.

4.2. Experimental Configuration

Table 1. Experimental configuration table

Name	Version&Model
CPU	Inter i7-11700
GPU	NVIDIA RTX A4000
Stystem	Windows10 64-bit
CUDA	11.7.1
Python	3.8
Pytorch	1.13.1

All the experiments were conducted in the same hardware environment, and the relevant environment configurations are

shown in Table 1.

4.3. Performance Indicators

Precision: the percentage of samples with true predictions that are correct.TP (True Positive) indicates the number of samples with correct predictions. FP (False Positive) is the number of samples with incorrect predictions. This is shown in Equation (1):

$$P = \frac{TP}{TP+FP} \tag{1}$$

Recall: Recall, also known as the check rate, indicates the proportion of predicted true positive samples to the total number of actual positive samples, which is used to reflect the leakage situation. FN (False Negative) denotesd the number of sample misses. Calculated as shown in Equation (2):

$$R = \frac{TP}{TP+FN} \tag{2}$$

Mean Average Precision (mAP): Precision and Recall are a pair of mutually constrained performance indicators, which have the limitation of single-point value and cannot fully evaluate the model performance, so mAP is introduced to equalize the results of the two calculations. Taking Precision as the vertical coordinate and Recall as the horizontal coordinate, the P-R curve can be obtained, and the area enclosed by the P-R curve and the coordinate axis is the value of AP, and mAP represents the average value of all the APs in the whole dataset, which is calculated as shown in Equation. (3) and (4).

$$AP = \int PdR \tag{3}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} APi \tag{4}$$

4.4. Analysis of Experimental Results

In order to verify the efficiency and adaptability of the improved model, this paper uses the original YOLOv7 model to train and validate this dataset with 300 iterations. A total of five performance metrics with parameters Precision, Recall, AP, mAP, etc. are selected for side-by-side comparison.

Compared with the original model, the improved model has some improvement in terms of recall, accuracy and precision, and the effect of improvement is obvious for the whole category of the dataset, and the comparison of the training results of the improved model and the traditional model for each category is shown in Tables 2 and 3.

Table 2. Compared YOLOv7 and YOLOv7-BiFPN

	YOLOv7-AP	YOLOv7-BiFPN-AP
pedestrian	0.571	0.599
people	0.494	0.511
bicycle	0.185	0.217
car	0.812	0.832
van	0.358	0.377
truck	0.374	0.403
tricycle	0.340	0.379
awning-tricycle	0.166	0.197
bus	0.416	0.445
motor	0.572	0.604

Table 3. Compared YOLOv7 and YOLOv7-BiFPN

Class	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv7	0.554	0.441	0.429	0.236
YOLOv7-BiFPN	0.567	0.454	0.456	0.251

From the experimental results, it is easy to see that the improved YOLOv7 has an increase of about 3% in the target AP value of each category, and the mAP@0.5 is also improved by 2.7%.

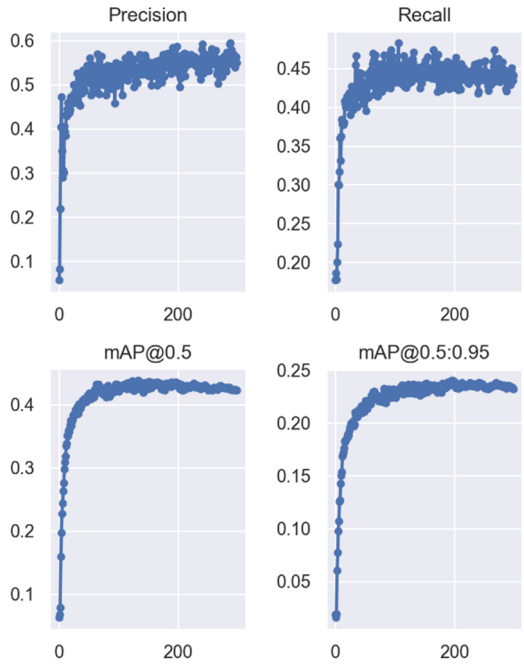

Figure 3. The experimental process of YOLOv7

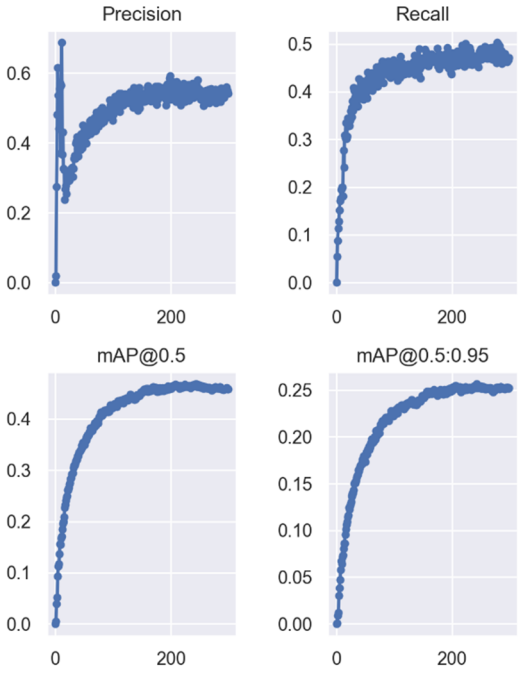

Figure 4. The experimental process of YOLOv7-BiFPN

Comparison of the training results Figure 3 and Figure 4 shows that the traditional Yolov7 model has the problems of unstable accuracy, unsatisfactory recall and low mAP value in the training process. The improved model is more stable in terms of training than the traditional model in terms of accuracy, the recall curve is smoother, and the mAP value is also improved. Therefore, our improved YOLOv7 algorithm is feasible.

5. Conclusion

Aiming at the challenges faced by UAV target detection with YOLOv7 algorithm's everything is fine except for one small defect, this study proposes an improved algorithm for UAV target detection based on YOLOv7, which improves the

accuracy of UAV target detection by introducing BiFPN and GAM attention mechanism. Our algorithm achieves satisfactory results on the VisDrone2019 dataset, indicating its broad potential for practical applications

Acknowledgments

We would like to express our gratitude to the authors of the YOLOv7 algorithm and the benchmark datasets used in this paper. We would also like to thank our colleagues for their helpful discussions and feedback on this research.

References

[1] A, Huiyu Zhou , Y. Y. B , and C. S. C . "Object tracking using SIFT features and mean shift." Computer Vision and Image Understanding 113. 3(2009):345-352.

[2] Yang Jinkun, et al."HOG and SVM algorithm based on vehicle model recognition." MIPPR 2019: PATTERN RECOGNITION AND COMPUTER VISION 11430.(2020).

[3] Joachims, Thorsten . "Making Large-Scale SVM Learning Practical." Technical Reports 8.3(1998):499-526.

[4] Viola, Paul , and M. J. Jones . "Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade." NIPS 2001.

[5] Chua, L. O. , and T. Roska . "The CNN paradigm." Circuits & Systems I Fundamental Theory & Applications IEEE Transactions on 40.3(1993):147-156.

[6] R. Girshick, "Fast r-cnn," Proceedings of the IEEE international conference on computer vision, vol. 12, pp. 1440-1448, 2015.

[7] S. Ren, K. He, Girshick R, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[8] K. He, G. Gkioxari, and P. Dollár, R. Girshick, "Mask r-cnn," Proceedings of the IEEE international conference on computer vision. vol. 10, pp. 2961-2969, 2017.

[9] Redmon, Joseph , and A. Farhadi. "YOLOv3: An Incremental Improvement." arXiv e-prints (2018).

[10] Sun, Y. X. , et al. "A CLASSIFICATION AND LOCATION OF SURFACE DEFECTS METHOD IN HOT ROLLED STEEL STRIPS BASED ON YOLOV7." Metalurgija (2023).

[11] Zhong, Lehai , et al. "Integration Between Cascade Region-Based Convolutional Neural Network and Bi-Directional Feature Pyramid Network for Live Object Tracking and Detection." Traitement du Signal: signal image parole (2021).

[12] Treisman, Anne M. , and G. Gelade . "A feature-integration theory of attention. " Cognitive Psychology 12. 1(1980):97-136.