Research on Airport Aircraft Target Detection Algorithm Based on YOLOv4

Yuqi Liao a, Ziyue Wang b and Xianfu Chen c
School of Microelectronics, University of Science and Technology of China, Hefei 230000, China

a liaoyq88@mail.ustc.edu.cn, b ziyuew@mail.ustc.edu.cn, c xfchen@ustc.edu.cn

Abstract. Aiming at the complex airport environment and the large scale transformation of aircraft, a target detection algorithm based on YOLOv4 is designed. Based on the YOLOv4 algorithm structure, a feature fusion scale is added to the original 3 detection scales, which has the ability to obtain smaller target information. Simultaneously, the residual network structure is simplified, and a certain amount of calculation is reduced. In order to optimize the loss, the regression loss function ClIoU is introduced, which accelerates the convergence of the network. The test results show that the algorithm of this article can reach a prediction accuracy of 93.31% for airplane target, a recall rate of 83.92%, a mean average accuracy of 91.02%, which is 2.49% higher than YOLOv4, which improves the recognition accuracy of the model in the airport environment.

Keywords: object detection, YOLOv4, deep learning.

1. Introduction
As a popular research direction of computer vision, object detection has been widely used in the fields of intelligent transportation, security monitoring, smart medical care, and autonomous driving. The layout of the airport scene is complicated, and there are many blind spots in the vision. Effective and reliable target detection technology is the basis for realizing the automatic control of airport video surveillance. With the development of computer technology, target detection algorithms based on deep learning have shown unique advantages in real-time and accuracy. Practical application scenarios require higher calculation speed and recognition accuracy of the algorithm, and the YOLO algorithm is a better one among the existing algorithms. However, for the airport environment, the scale of the aircraft changes greatly, and the detection of small targets is not accurate enough. Based on the YOLOv4 algorithm, this paper adjusts the network structure, simplifies the residual network structure, increases the feature fusion scale to obtain more small target feature information, and introduces the regression loss function ClIoU, and improves the accuracy of the recognition.

2. Algorithm Design
There are two main types of target detection methods based on deep learning: One is the two-stage R-CNN [1] series; the second is one-stage algorithm, including YOLO series and SSD, etc. YOLOv4 [2] is a target detection algorithm model with both speed and accuracy, and its structure is shown in Figure 1. Personal development only requires a single computing system (such as a single GPU) to run quickly.
In the airport application scenario, it is necessary to quickly and accurately identify the aircraft's take-off and landing, improve monitoring efficiency, reduce the workload of controllers, and assist command work. This article has done the following research work based on YOLOv4.

2.1. CSPdarknet53
CSPNet [3] (Cross Stage Partial) reduces the amount of data transfer and calculation in the network, without losing accuracy and light weight, reducing the calculation and memory consumption required by the CNN network. The input feature map is divided into two paths in the channel dimension, one of which is passed directly backwards, the other part is passed backwards through multiple residual blocks, and finally merged with the CSP end. Such cross-stage splitting and merging effectively reduces the possibility of gradient duplication, increases the diversity of gradients, and reduces the amount of data transfer and the amount of calculation in the network. In YOLOv4, each CPSX contains \( 32 \times X \) convolutional layers, so the entire backbone network backbone contains a total of 72 convolutional layers. With the purpose of reducing computational overhead, this article can compress the residual structure by reducing the convolutional layer, as shown in figure2 and figure3, and appropriately increase the running speed.

2.2. Feature Fusion Techniques
Feature fusion mainly refers to the direct feature fusion of different output layers. Traditional FPN only has a top-down one-way information flow. PANET [4] increases the bottom-up information flow, transfers the location information of shallow features to the deep network, and improves the ability of multi-scale feature layers to express location information. Due to the large scale transformation of
airport aircraft, the detection of small targets is more dependent on shallow features. This paper expands on the original 3 detection scales and adds a feature fusion scale of $104 \times 104$ to get more semantic information of small targets. The four scales are fused through PANET and input to the YOLO layer for category prediction and bounding box regression. Figure 4 shows the improved yolov4 structure diagram. The SPP [5] structure is placed on the last feature layer of CSPdarknet53 as an additional module, which is used to increase the receptive field and separate contextual features while maintaining the operating speed of the network.

2.3. Loss function.

IoU is an important function that expresses the performance of target detection algorithms and is not sensitive to the scale of the target object. Someone proposed to directly use IoU as the regression loss function, including CIoU which is considered more comprehensive. CIoU adds a penalty factor to DIoU[6] (Figure 5) to take the aspect ratio into consideration. CIoU takes into account the center point distance between bounding boxes, the ratio of overlapping areas, and the aspect ratio. It is more sensitive to scale transformation and solves the possible divergence problems of IoU and GIoU.

Among them, $d = \rho^2(B, B^g)$ represents the Euclidean distance between the center point of the predicted box and the ground truth box, $\beta$ represents the weight, $\nu$ is the parameter to measure the consistency of the aspect ratio, $c$ represents the diagonal distance of the smallest rectangular area that contains both the predicted box and the ground truth box, $w$ and $h$ represent the width and height of the predicted box, and $w^g$ and $h^g$ represent the width and height of the ground truth box.
\[ C_{IoU} = IoU - \frac{\rho^2(B, B^\prime)}{C^2 - \beta \nu} \]  
\[ \beta = \frac{\nu}{1 - IoU + \nu} \]  
\[ \nu = 4 \left( \frac{\arctan \frac{w^\prime}{h^\prime} - \arctan \frac{w}{h}}{\pi} \right)^2 \]  
\[ L_{C_{IoU}} = 1 - C_{IoU} = 1 - IoU + \frac{\rho^2(B, B^\prime)}{C^2} + \beta \nu \] 

2.4. Data Augmentation Techniques.
During training, data augmentation mainly improves the diversity of training samples, expands the data set, and improves the robustness and generalization ability of the model through random scale scaling, cropping, rotation and translation, and color gamut transformation on the training images. Mosaic Data augmentation [2] is an improved version of CutMix [7]. It mixes four pictures, flips, zooms, and transforms the color gamut of the four pictures respectively. Place the processed pictures in four positions, up, down, left, and right, and combine the pictures and frames. As shown in Figures 6 and 7, four pictures are randomly selected for Mosaic data augmentation.

![Fig. 6 The picture after flipping, zooming, and color threshold transformation.](image)

![Fig. 7 Picture after stitching.](image)

3. Network Training Results and Analysis

3.1. Experimental Environment and Training Parameters.
The hardware GPU used in this experiment is NVIDIA GeForce GTX 1660 Ti. The experiment process uses self-built data sets, inputs images and performs Mosaic data augmentation. Adjust the initial learning rate which is 0.001 by cosine annealing [8] to reduce it. The cosine function in cosine annealing undergoes a slow decline in the cosine value, and alternately accelerates the decline to reduce the cosine value, thereby reducing the learning rate. The reduction of the learning rate is to optimize the objective function through the gradient descent algorithm. When the global minimum of the loss function is reached, the model will reach the optimum. The improved network training loss function curve is shown in Figure 8. The improved network training loss function curve is shown in Figure 8.
3.2. Analysis of test results.
The target detection result is shown in Figure 9. The experiment uses the mean average precision as the evaluation standard to test the validity of the model. Among them, Precision is the proportion of the predicted positive samples and the number of samples with correct predictions to the predicted positive samples, and Recall is the proportion of the predicted positive samples and the correct predictions in the actual positive samples. AP is the area of the curve formed by the combination of different Precision and Recall points, and mAP is the average of AP of all classes.

The test results show that the improved algorithm can reach a prediction accuracy of 93.31% for airplane target, a recall rate of 83.92%, a mean average accuracy of 91.02%, which is 2.49% higher than YOLOv4, which effectively improves the recognition accuracy of the model in the airport environment.

4. Summary
Aiming at the airport environment, this paper simplifies the residual network structure based on YOLOv4, increases the feature fusion scale, and improves the detection accuracy. The experimental results on the self-built data set have certain application value. In the follow-up work, will continue to optimize the network structure to obtain higher accuracy and lower time cost. Subsequent consideration will be given to adopting better GPUs and research on tracking algorithms to improve the stability and reliability of recognition, so that it can be applied to real-time detection and recognition of airport targets.
References

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 580–587, 2014. 2.

[2] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[3] Wang, Chien-Yao, et al. "CSPNet: A new backbone that can enhance learning capability of CNN." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.

[4] Shu Liu, Lu Qi, Haifang Qin, et al. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8759–8768, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 37(9):1904–1916, 2015.

[6] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU Loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020.

[7] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, et al. CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 6023–6032, 2019.

[8] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.