Small object detection in aerial view based on improved YoloV3 neural network

Fuyan Lin
Faculty of Information Technology
Beijing University of Technology
Beijing, China
e-mail: 624778075@qq.com

Xin Zheng
Faculty of Information Technology
Beijing University of Technology
Beijing, China
e-mail: xz@bjut.edu.cn

Qiang Wu
Faculty of Information Technology
Beijing University of Technology
Beijing, China
e-mail: wuqiang@bjut.edu.cn

Abstract—**In application scenarios such as UAV inspection, deep learning-based object detection methods are increasingly used to improve the automation of line inspection. In the aerial view scene, the drone is usually fly at a high altitude from the ground, so the proportion of the object in the image is relatively small. When the YoloV3 network identifies small objects, the detection result would not be good because there is less information in the 8x downsampling feature map. In this paper, base on the LaSOT data set, the YoloV3 network has been modified by adjusting the values of anchors and establishing the 4x downsampling prediction layer to enhance the detection effect of small objects. Compared with the original YoloV3 network, the improved YoloV3 network has a certain improvement in convergence ability and detection accuracy compared to the original YoloV3 network.**

Keywords—object detection, YoloV3, small object identify, Kmeans clustering, Multi-scale feature fusion

I. Introduction

Object detection from aerial view can be used to check the condition of the target facility itself and find the risk factor around the facility to achieve the purpose of replacing manual line inspection. Because the flying of drones is not affected by terrain, etc., and it is more time-saving and labor-saving than manual inspection, it has developed rapidly in recent years.

At present, there are two main methods for object identification through neural networks: two-stage and one-stage [1]. The basic idea of the two-stage model is to divide the detection into two steps: generating candidate regions and classifying regression. The representative network structure mainly includes Fast R-CNN, Faster R-CNN, and R-FCN. The single-stage monitoring model does not require the stage of generating candidate regions, and directly classifies and returns the object's position. The representative network structure mainly includes YoloV1, YoloV2, YoloV3, SSD, and so on. In contrast, the two-stage detection model has higher detection accuracy but is slower; although the single-stage detection model has lower detection accuracy than the two-stage model, its detection speed has been greatly improved, which is more suitable for real-time scenarios use.

Among current neural network detection methods, the YoloV3 network has a high balance between detection speed and detection accuracy and has been widely used in various fields. Based on YOLOv2, YOLOv3 uses the ResNet network to deepen the feature extraction network layers and uses the Darknet53 network structure as the feature extraction network to obtain a detection effect comparable to ResNet101. The residual block structure and skip the connection mechanism of the ResNet network effectively solve The problem of gradient dispersion and gradient explosion caused by the deepening of the network. At the same time, the feature pyramid (FPN) [2]mechanism integrates multi-scale feature information to improve the detection accuracy. However, the excessively deep network structure can easily cause the loss of object's position information in shallow feature layers, Has a high rate of missed detection, is not conducive to the detection of small objects.

In the Darknet53 network structure, the shallow features have higher resolution and are richer in location information; the deep features have stronger semantic information, but the location information is relatively rough. As the network structure deepens, the shallow feature information is not fully utilized and resulting in the loss of many small object location information. Besides, too complicated and redundant network structure and too many parameters will lead to complex training, increased demand for data, and slow down the detection speed. This article proposes the following improvement methods for the above scenarios in aerial object detection:

1) For the data set that corresponding to the network application scenario, readjust the anchor value

2) For the aerial view detection scene, add a predict layer which using 4x downsampling layer as the main input

3) Remove predict layers that designed to detect large objects

II. YoloV3 network and improvement method

A. YoloV3 network

The YoloV3 object detection algorithm is based on the idea of regression. It executes both the object recognition and object localization step at the same time and returns the position of the bounding box and its category directly at the output layer. Its key features include using the Darknet53 classifier and making a multi-scale prediction based on different sampling layers. [2]

YOLOv3 divides the feature map into S×S grids according to the scale of the feature map. Each grid is used to predict B bounding boxes and is used to detect C categories of targets. Finally, the bounding boxes of each type of target are output and each Confidence of the bounding boxes. Confidence is determined by the probability of the detection target contained in each grid and the accuracy of the output bounding box.

978-1-7281-6521-9/20/$31.00 ©2020 IEEE

YOLOv3 sets each grid cell to predict three anchor boxes, each anchor box to predict three bounding boxes, and each bounding box to predict four values, which are x_t, y_t, w_t, and h_t respectively. There is an offset (x_c, y_c) in the upper left corner of the image, as shown in Figure 1, then correct it.

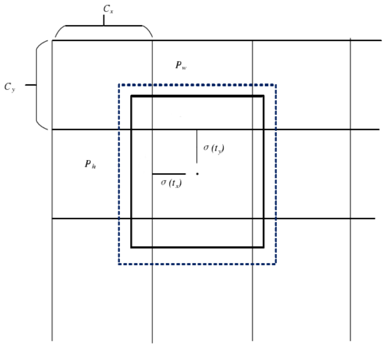


Fig. 1. Anchor in YoloV3

Among them, w_p and h_p represent the width and height of the grid corresponding to the anchor box, calculate the position of the bounding box, set the appropriate confidence score, filter out the low-scoring prediction boxes, and perform non-maximum suppression on the remaining prediction boxes to get the final prediction result.

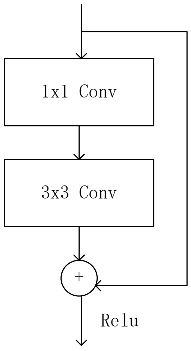
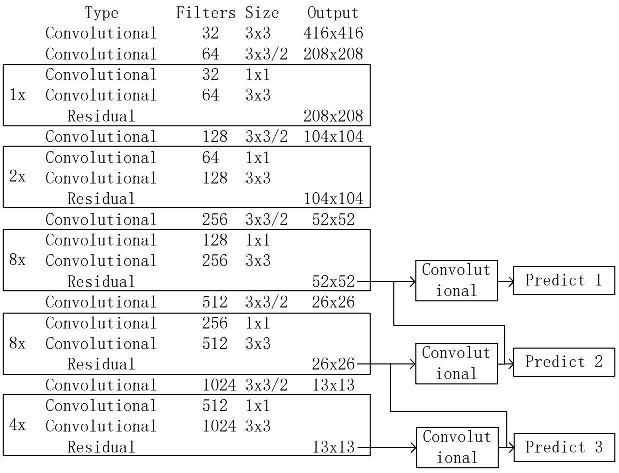


Fig. 2. Residual unit in Darknet53

Darknet53 contains 53 convolutional layers. It borrows the idea of the residual network[3] and uses a large number of residual units, as shown in Figure 2, in the network. The two convolutional layers and a residual block in the box form a residual unit. The number on the left indicates the number of times the residual unit is repeated in the Darknet53 network. In deep networks, this design can effectively ensure the effective convergence of the network and improve the effect of classification and detection.

Meanwhile, YoloV3 also borrows the idea of FPN and uses 3 predict layers of different scales to predict the object. The size of the feature maps of the 3 predict layers in pixels is 52, 26, and 13. It can be seen in Figure 3 that the input of each predict layer comes from the extraction layer of 32, 16, and 8 times downsampling characteristics. Then based on the preset bounding box size set by the predict layer, the pixel size of the bounding box in the actual image corresponding to each predict layer can be obtained. The differences are shown in the following table:



	Type	Filters	Size	Output
	Convolutional	32	3x3	416x416
	Convolutional	64	3x3/2	208x208
1x	Convolutional	32	1x1	
	Convolutional	64	3x3	
	Residual			208x208
	Convolutional	128	3x3/2	104x104
2x	Convolutional	64	1x1	
	Convolutional	128	3x3	
	Residual			104x104
	Convolutional	256	3x3/2	52x52
8x	Convolutional	128	1x1	
	Convolutional	256	3x3	
	Residual			52x52
	Convolutional	512	3x3/2	26x26
8x	Convolutional	256	1x1	
	Convolutional	512	3x3	
	Residual			26x26
	Convolutional	1024	3x3/2	13x13
4x	Convolutional	512	1x1	
	Convolutional	1024	3x3	
	Residual			13x13

Fig. 3. Residual unit in Darknet53

TABLE I. PREDICT LAYER AND PREDICTION BOUNDING BOX

Predict layer	Feature map size	Bounding box size
Predict layer 1	13x13	116x90;156x198;373x326
Predict layer 2	26x26	30x61;62x45;59x119
Predict layer 3	52x52	10x13;16x30;33x23

The original YoloV3 network was designed for the COCO dataset [4], and have achieved good performance results on the COCO dataset. However, compared with the application scenarios in this paper, the objects in the COCO dataset are mostly a variety of common things in daily scenes, the object sizes are different, and the proportion of the pictures is relatively large. Therefore, in the application scenario of this paper, to YoloV3 network is used to detect small objects, the YoloV3 network needs to be improved.

B. Small object detection in aerial view scenes

In the original YoloV3 network, three predict layers which in three different scales are used to detect objects. In general scenes, due to the uncertainty of object size and distance from the object, this method has a good detection effect on normal-sized objects. But when the object is too small, the detection effect on the object will be greatly reduced, and some objects may be miss inspected due to the downsampling operation has cut down lots of useful information of the object.

Regarding the condition that the objects in the LaSOT dataset are general in small pixel size, the YoloV3 network should be modified.

1) Clustering statistics of the target frame of the dataset

The Anchor Boxes idea used by YoloV3 is based on the design of Faster R-CNN using a fixed number of initial candidate boxes of fixed size to select the object for the box. The choice of Anchor Boxes will directly affect the speed and accuracy of object detection by the network and speed up the convergence speed during training. YoloV3 uses the kmeans method [5] to perform cluster analysis on the objects in the dataset to ensure that a limited number of candidate frames can match the actual size of the object.

In the original YoloV3, the average Intersection-Over-Union (Avg IOU) was used as the measurement index of the object cluster analysis and analyzed the COCO data set. The clustered Avg IOU objective function can be shown in Equation 1, where represents the sample, which is the object in ground truth; represents the center of the cluster; represents the number of samples in the cluster center; represents the total sample number of clusters; represents the number of clusters; represents the intersection ratio of the central mine of the cluster and the cluster box; represents the sequence number of the sample; represents the sequence number of the sample in the cluster center

As can be seen from the above figure, as the number of k clusters increases, the value of Avg IOU becomes more stable. It can be seen that when k = 3, the curve has an obvious inflection point, and then gradually flattens, so choosing the number of Anchor Boxes as three can effectively accelerate the convergence of the loss function and reduce the error caused by the candidate box. So according to the above analysis of data set, the number of the prediction boxes should be set to 3, and the width and height of the prediction boxes should be (27, 17), (34, 20), and (41, 25) due to the kmeans result.

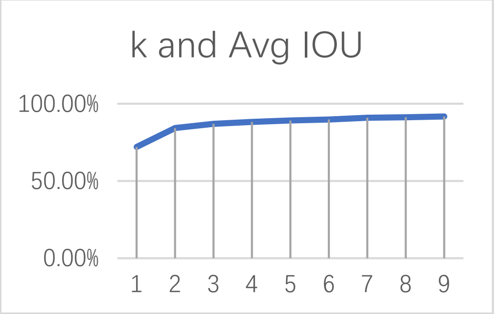


Fig. 4. Relationship lines between Kmeans clustering cluster number k and Avg IOU

Select the different , the number of clusters, to cluster the data set. When the relationship curve between and Avg IOU could be shown in Figure 4.

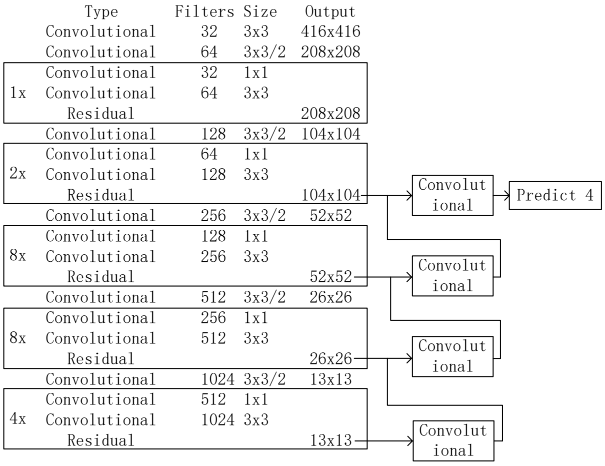
2) Improved YoloV3 network model

The original YoloV3 network uses the Darknet53 network structure as the backbone and uses three-scale feature maps to predict the object position. When detecting small objects, the original YoloV3 network mainly uses the information of the 8x downsampling layer to detect the object, which means that when the object's pixel is lower than 8x8, the network can hardly be predicted. Similarly, even if the object is just right Larger than 8x8 pixels, the network's ability to detect object is greatly reduced due to the lack of the object's information. Therefore, the information of the 4x downsampling layer can be used to detect the object, which could improve the detection effect on small objects. Meanwhile, based on the original YoloV3's method of fusion deeper layer's detection result residuals with feature layer to enhance the network performance, the improved YoloV3 first upsampling the eight times downsampling feature map by two times, and then fusion it with four times downsampling feature map to establish The input is a feature layer with a 4x downsampling feature map to detect small objects.

According to the clustering results of the object, if the object is detected using four predict layers, because the feature map sizes of the four predict layers are 104, 52, 26, and 13, respectively. The deep predict layer cannot effectively predict small objects, so we only keep the predict layer 4, which been added in the improved YoloV3 model, to detect the small object.

At this point, to improve the network's detection effect on small objects in the scene, the original YoloV3 network's output detection layers on large scales were deleted, and add a four times downsampling feature layer to detect small objects. The improved YoloV3 network model is shown in Figure 5.

$$f = \operatorname{argmax} \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_k} I_{IOU}(B,C)}{n} \tag{1}$$



	Type	Filters	Size	Output
	Convolutional	32	3x3	416x416
	Convolutional	64	3x3/2	208x208
1x	Convolutional	32	1x1	
	Convolutional	64	3x3	
	Residual			208x208
	Convolutional	128	3x3/2	104x104
2x	Convolutional	64	1x1	
	Convolutional	128	3x3	
	Residual			104x104
	Convolutional	256	3x3/2	52x52
8x	Convolutional	128	1x1	
	Convolutional	256	3x3	
	Residual			52x52
	Convolutional	512	3x3/2	26x26
8x	Convolutional	256	1x1	
	Convolutional	512	3x3	
	Residual			26x26
	Convolutional	1024	3x3/2	13x13
4x	Convolutional	512	1x1	
	Convolutional	1024	3x3	
	Residual			13x13

Fig. 5. Improved YoloV3 network

III. Experiment and Result Analysis

This article is about the improvement of the YoloV3 network in specific scenarios. Therefore, the test mainly compares the improved YoloV3 network with the original YoloV3 network on the object detection performance under the data set.

The input size of the YoloV3 network model image is 416x416, so the original test image size needs to be resized to the size of the input image. Also, it is necessary to ensure that the target does not deform during the resize process. Since the pixel ratio of the original image is not 1: 1, and the pixel ratio of the object in the image should be changed, so the inconsistent portion of the image is complemented by increasing the black border and resize it to 416x416.

Experimental conditions: Windows 10 operating system, i7-7700 CPU, 24GB memory, Nvidia GeForce GTX 1060 GPU.

A. Network training

The Large-scale Single Object Tracking dataset (LaSOT) dataset used in this paper provides high-quality datasets for single-type targets [6]. The vehicle images in aerial view from the LaSOT has been selected as the dataset used for training and testing in this paper.

On the training of the original YoloV3 network and the improved YoloV3 network, the training parameters were set to start with a learning rate of 0.001 and a decay rate of 0.0005. At the same time, image rotation and enhancement of contrast are used to enhance and expand the images in the data set.

The loss curve of the improved YoloV3 network during training is shown in Figure 6.

After about 400 epochs, the parameters of loss is stable, so it can be determined that the two types of YoloV3 network training result at this time is ideal. The weights of the model that trained above can be used for network testing.

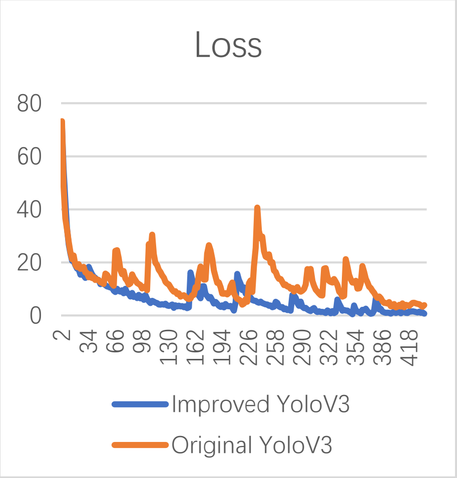


Fig. 6. Two kinds of YoloV3 Loss curve for network training

B. Network test results and analysis

The network's mAP value is used as an indicator to measure the performance of the network in the data set, and the final mAP values of the original YoloV3 and improved YoloV3 networks in the data set are listed, as shown in Table II.

TABLE II. MAP RESULT

Network Model	**mAP**
Original YoloV3	85.82%
Improved YoloV3	91.68%

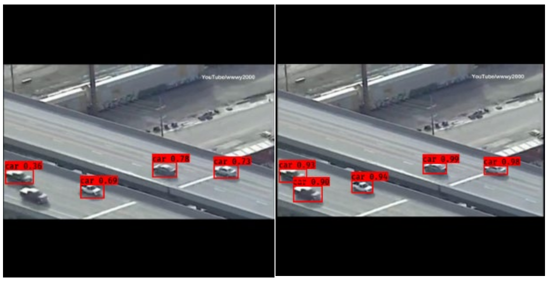


Fig. 7. Compare original YoloV3（left）and improvedYoloV3（right）

From the experimental data, it can be seen that the mAP of the original YoloV3 network model trained on the data set is 85.82%, and the improvement of the YoloV3 network model has increased the mAP of the network to 91.68%, which is 5.86% higher than the original network.

According to the above experimental results, it turns out that:

(1) The Anchor Boxes of the original YoloV3 model were obtained based on the COCO data set. When the YoloV3 model working with the new data set, recalculating the Anchor value can effectively improve the convergence speed and mAP value for a specific data set.

(2) For small objects in the data set, by selecting the feature map of the shallow network, more object information can be retained, which can effectively improve the network's performance for small object objects in the data set.

IV. CONCLUSION

In this paper, the YoloV3 algorithm is applied to the aerial view object detection scene. For small object detection problems, using two kinds of method: k-means clustering to optimize the Anchor size and modify the YoloV3 model base on the object characteristics. The detection accuracy of the network. According to the experimental data, compared with the original YoloV3 network, the improved YoloV3 model's mAP is improved by 5.86%, and model convergence is faster. Subsequent research work will be conducted on different data and experiments to explore the generalization ability of the model further while focusing on image enhancement and data enhancement research to improve the robustness of the model.

ACKNOWLEDGMENT

First of all, I would like to thank my supervisor Qiang Wu for his guiding opinions and recommendations on the research direction of my paper and giving careful guidance to me when I encountered difficulties and doubts.

Second, I would also like to thank my classmate for their strong support in the preparation of my paper and the inspiration provided to me in relevant aspects, which enabled my dissertation to be completed faster and better.

Finally, due to the 2019-nCoV epidemic, this paper had to be completed at home, so I would like to thank my mom for her support.

REFERENCES

[1] DaLei Kou, YiChuan Quan, ZhongWei Zhang. Research on target detection framework based on deep learning [J]. Computer engineering and Application, 2019, 55(11):25-34.

[2] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[3] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified Real-Time Object Detection[J]. 2015.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[J]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015:770-778.

[5] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context [J]. 2014.

[6] ARTHUR D, VASSILVITSKII S. k－means++: the advantages of careful seeding [C] Eighteenth Acm-siam Symposium on Discrete Algorithms. 2007.

[7] H. Fan*, L. Lin*, F. Yang*, P. Chu*, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling. LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking, CVPR, 2019.