International Journal of Science and Engineering Applications
Volume 11-Issue 01, 22 - 26, 2022, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1101.1003

A Vehicle and Pedestrian Detection Method Based on Improved YOLOv4-Tiny

Hui Xiang
College of Electromechanical Engineering
Qingdao University of Science and Technology
Qingdao, China

Junyan Han
College of Electromechanical Engineering
Qingdao University of Science and Technology
Qingdao, China

Hanqing Wang
College of Electromechanical Engineering
Qingdao University of Science and Technology
Qingdao, China

Hao Li
College of Electromechanical Engineering
Qingdao University of Science and Technology
Qingdao, China

Shangqing Li
College of Electromechanical Engineering
Qingdao University of Science and Technology
Qingdao, China

Xiaoyuan Wang
College of Electromechanical Engineering
Qingdao University of Science and Technology
Qingdao, China

Abstract: Aiming at the problems of low detection accuracy and poor recognition effect of small-scale targets in traditional vehicle and pedestrian detection methods, a vehicle and pedestrian detection method based on improved YOLOv4-Tiny is proposed. On the basis of YOLOv4-Tiny, the 8-fold down sampling feature layer was added for feature fusion, the PANet structure was used to perform bidirectional fusion for the deep and shallow features from the output feature layer of backbone network, and the detection head for small targets was added. The results show that the mean average precision of the improved method has reached 85.93%, and the detection performance is similar to that of YOLOv4. Compared with the YOLOv4-Tiny, the mean average precision of the improved method is increased by 24.45%, and the detection speed reaches 67.83FPS, which means that the detection effect is significantly improved and can meet the real-time requirements.

Keywords: computer vision; deep learning; YOLOv4-Tiny; vehicle detection; pedestrian detection

1. INTRODUCTION

In recent years, with the continuous improvement of hardware computing performance, the target detection method based on convolutional neural networks (CNN) has become the mainstream because of its strong learning ability. The traditional vision-based target detection methods construct detectors by manually extracting target image features or by machines selecting different target image features. In contrast, the CNN-based method can autonomously and centrally calculate the multi-layer features of the target image, and as the network depth increases, higher-dimensional features will be learned and have better feature expression capabilities for the target. For advanced driving assistance system, the improvement of target detection method will contribute to improve the perception ability of the external environment of the vehicle, so as to make real-time response to the driving road environment.

At present, the target detection methods based on CNN are mainly divided into two categories. One is a two-stage target detection algorithm and the detection process is divided into two steps: generating candidate regions and extracting image features of candidate regions for classification and regression. The typical algorithms are: R-CNN [1], Fast-RCNN [2], Faster R-CNN [3], R-FCN [4] and Mask R-CNN [5]. Limited by the complex network structure, the real-time performance of the two-stage method is not enough for practical application. The other is one-stage target detection method. The target location and category are classified and regressed by a single network, which takes into account the detection speed and accuracy. The typical algorithms include SSD [6], DSSD [7], YOLO [8], YOLOv2 [9], YOLOv3 [10] and YOLOv4 [11].

In order to realize the timely and accurate detection of vehicles and pedestrians in front under the complex driving road environment, YOLOv4-Tiny [11] is used as the basis to improves the network structure and the detection accuracy while ensuring the detection speed, so as to meet the actual needs of the driving environment.

2. INTRODUCTION TO YOLOV4-TINY

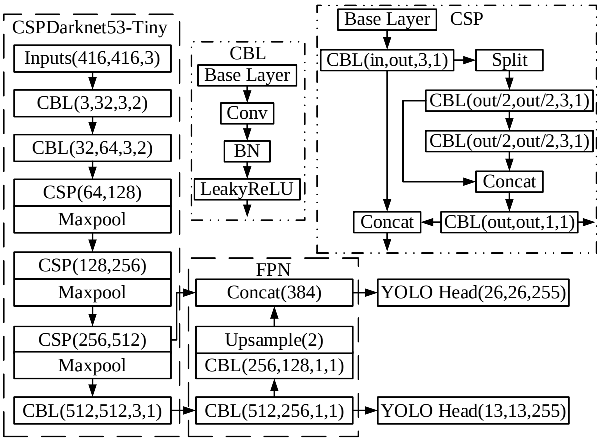


Figure 1. Network structure of YOLOv4-Tiny

In April last year, YOLOV4 was proposed on the basis of YOLOv3, and its detection speed and detection rate have been improved. YOLOv4-Tiny is a lightweight version of YOLOv4. Compared with YOLOv4, the network structure of YOLOv4-Tiny is simpler, the detection speed is faster, and it performs well in the target detection algorithm. The network structure of YOLOv4-Tiny is shown in Figure 1.

As we can see in Figure 1, CSPDarknet53-Tiny is used as the backbone for YOLOv4-Tiny, which mainly includes Convolution Block (CBL), Cross Stage Partial Connections (CSP) and down sampling module. CBL is composed of ordinary convolution (Conv), batch normalization layer (BN), and activation function (Leaky ReLU). In CSP, after the input feature is convolved with a convolution kernel size of 3×3 and a step size of 1, the convolution result X is obtained. Then the features obtained by result X passing through CBL twice and CBL once are stacked in the channel dimension, and the result after stacking is named as Y. Finally, the result Y is convolved with 1×1 convolution to increase the nonlinear expression ability of the network, and the result after convolution is named as Z. The result Z can be selected as the input of the feature fusion part. The result X and Z are stacked in the channel dimension as the output of the CSP. Max pooling is chosen as the down sampling module. Feature pyramid network (FPN) is chosen to be the feature fusion part of YOLOv4-Tiny, which can fuse the 13×13 feature layer output by the backbone network with the 26×26 feature layer to improve feature extraction capabilities. Finally, Yolo head is used to perform target regression and classification on the feature layer coming from FPN.

3. IMPROVEMENT OF YOLOV4-TINY

The simple network structure of YOLOv4-Tiny is not enough to learn sufficient vehicle and pedestrian characteristics. Based on the original YOLOv4-Tiny, a real-time lightweight vehicle detection model is designed, which can reduce the false and missed detection in vehicle and pedestrian detection and improve the detection performance.

With the continuous deepening of the backbone network level, the resolution of the feature layer continues to decrease, and the semantic dimension continues to deepen. The shallow features have a higher resolution and contain more regional texture details, which are suitable for determining the location of the target. The deep features contain higher-dimensional regional semantic information and are suitable for target classification. In the feature fusion part, the original YOLOv4-Tiny only selected two different scale features of 13×13×512 and 26×26×256 in the backbone as the input of FPN, so that the network contains both high-dimensional semantic information and low-dimensional texture details. With this fusion strategy, feature information of different scales is less fused, the model is not sensitive to targets of different sizes, and the detection accuracy of the model is low. For this reason, on the basis of the original backbone network output, the 52×52×128 shallow feature layer output by the second CSP convolutional layer in CSPDarknet53-Tiny, which is the 8-fold down sampling feature layer of the input image, is added as the input of the feature fusion part.

In addition, through the feature fusion strategy of bottom-up and top-down, the Path Aggregation Network (PANet) is used to perform bidirectional fusion for the deep and shallow features from the output feature layer of backbone network. Therefore, three scale feature maps of 13×13×512, 26×26×256 and 52×52×128 are output to form the final feature expression. Finally, the feature maps of three scales are respectively used to output the prediction results through three convolutional layers with the same number of channels. And the number of channels in the convolutional layer is set to (5+3)×3=24. Above all, based on YOLOv4-Tiny, the improved network called YOLOv4-Tiny-DL3 has been constructed, and its specific network structure is shown in Figure 2.

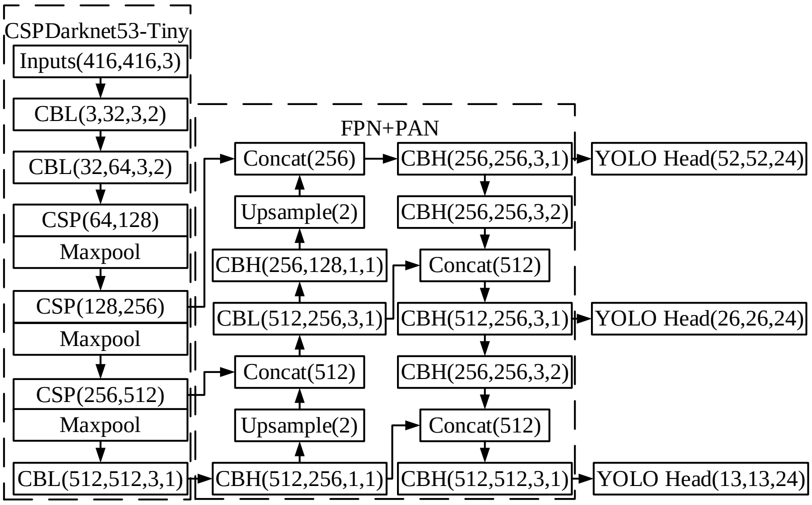


Figure 2. Network structure of YOLOv4-Tiny-DL3

4. EXPERIMENT AND ANALYSIS

4.1 Dataset

In the experiment, the KITTY dataset is selected as the vehicle and pedestrian detection dataset. The KITTY training set has 7481 pictures. Since the KITTY test set has no labeled information, the KITTY training set is redivided as the training set and validation set of this article according to the ratio of 9:1. The original dataset annotation file includes nine category labels. In this paper, the original “Truck”, “Van”, “Tram” and “Car” tag categories are merged into “Car”; the original “Pedestrian” and “Person_sitting” tag categories are merged into Person; the original “Cyclist” tag category is retained, and the other two tag categories are eliminated. Next, the annotation file is transformed from the original KITTY dataset format to the VOC format. As a result, the annotation files required for the dataset in this article are obtained. The number of categories in the divided training set and test set is shown in the figure below.

Table 1. Number of categories in datasets

	Car Number	Person Number	Cyclist Number
Training set	30079	4251	1481
Verification set	3182	458	416

4.2 Experimental platform

The hardware platform selected in this paper is as follows: the system processor is Intel(R)Core(TM)i7-10870H CPU @2.20GHz, the memory is 16G, the operating system is ubuntu18.04, the deep learning framework is Pytorch1.7, the model parallel computing framework is CUDA 11.1, and the model acceleration library is CUDNN 8.0.5.39.

4.3 Experimental setup

During training, the input image size of each model is 416×416, the input batch size is 8, and the training epoch is 300. The COCO pretraining model was used for the first 50 epoch, and the model parameters are adjusted by freezing the backbone network. The initial learning rate is set to 0.001; the cosine annealing learning method is used later to gradually reduce the learning rate from 0.001 to 0.0001. The prior boxes used during training are: (8, 11), (13, 25), (28,1 9); (25, 51), (52, 28), (49, 38); (97, 75), (131, 166), (314, 274). During training, the loss function used in the paper is the same as YOLOv4.

4.4 Experimental result

Model detection performance is evaluated based on Precision (P), Recall (R), mean Average Precision (mAP), and FPS. FPS is used to measure the detection efficiency, which represents the number of images that the model can process per second. The mAP value is defined as the mean value of the average precision (Average Precision, AP) of each class, and the AP value corresponds to the area under a certain type of P-R curve. The calculation formulas are shown as follows.

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

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

$$AP = \int_0^1 PdR \tag{3}$$

Among them, TP means that the prediction box matches the label box correctly; FP means that the background is predicted to be an object; FN means that the object is predicted to be the background.

Table 2. Model performance comparison table

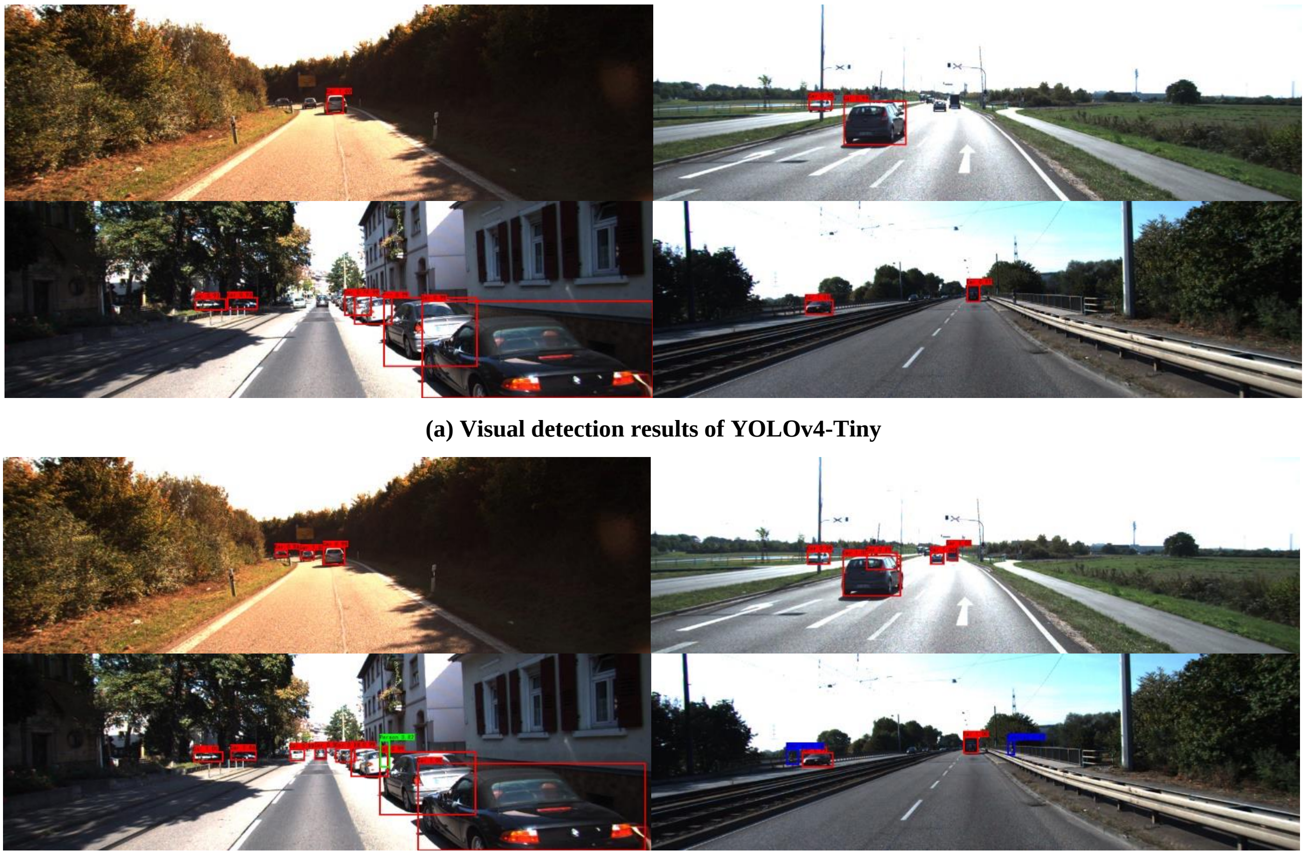
Models	AP/%			mAP/%	Parameters/M	BFLOPs	FPS
	Car	Person	Cyclist				
YOLOv4-Tiny	86.79	57.22	63.15	69.05	6.06	3.47	79.75
YOLOv4-Tiny-DL3	95.88	75.64	86.58	85.93	12.67	8.93	67.83
YOLOv4	95.80	76.78	88.35	86.98	63.95	29.89	30.13

Table 2 shows the experimental results of YOLOv4-Tiny, YOLOv4 and YOLOv4-Tiny-DL3 proposed in this paper. The experimental results are all obtained from training set and testing set selected from the KITTI dataset. It can be seen from the table that the YOLOv4-Tiny-DL3 model is improved by the YOLOv4-Tiny model, and its detection effect is better after the improvement; the AP value of Car has increased by 9.09, the AP value of Person has increased by 18.42, the AP value of Cyclist has increased by 23.43 and the overall mAP value has increased from 69.05 to 85.93. It can be confirmed that the fused multi-scale feature map contains more useful information owing to the 8-fold down-sampling feature layer is added for feature fusion in the YOLOv4-Tiny-DL3 model. In addition, the added small target detection head and PAN feature fusion strategy are beneficial for small target detection. The improved model has a small increase in network parameters. Although the FPS has dropped to a certain extent, the overall FPS can still reach 67.83, which has high real-time performance.

What’s more, compared with the YOLOv4 model, the YOLOv4-Tiny-DL3 model proposed in this paper achieves similar detection performance and more than twice the detection speed with a lower number of model parameters and computational requirements.

4.5 Qualitative analysis of results

In order to analyze the experimental results of this paper more intuitively, some results of YOLOv4-Tiny and YOLOv4-Tiny-DL3 on the KITTI dataset are respectively visualized in Figure 3(a) and Figure 3(b) when the input image resolution is 416×416. In the figure, the red detection box refers to the “Car” category; the green detection box refers to the “Person” category; the blue detection box refers to the “Cyclist” category. Compared with YOLOv4-Tiny, YOLOv4-Tiny-DL3 can detect farther "Car" targets. For "Person" targets and "Cyclist" targets with a small proportion of pixels in the image, YOLOv4-Tiny did not detect it, while YOLOv4-Tiny-DL3 can detect it. In summary, YOLOv4-Tiny-DL3 can detect smaller targets, reduce the missed detection rate of



(a) Visual detection results of YOLOv4-Tiny

(b) Visual detection results of YOLOv4-Tiny-DL3

Figure. 3 Comparison of visual detection results between YOLOv4-Tiny and YOLOv4-Tiny-DL3 models

small targets, and have better detection accuracy for small targets.

5. CONCLUSION

A vehicle and pedestrian detection method based on improved YOLOv4-Tiny was proposed. Owing to the original YOLOv4-Tiny network only contains two prediction scales of 13 × 13 and 26 × 26, the model has a good recognition rate for large and medium objects in the image, except for small targets objects. Therefore, the 8-fold down sampling feature layer from the backbone network is used for prediction, and then the feature fusion strategy of PANet is added to fuse the features of three different scales output by the backbone network, which is beneficial to promote the bidirectional fusion of deep features and shallow features. In addition, the detection head for small targets is added to the network to make it more sensitive to small targets, so as to overcome the shortcomings of the original network's lack of small target detection capabilities. Through the improvement of the above method, the overall detection performance of YOLOv4-Tiny-DL3 is improved significantly. Compared with the previous network model, the final mAP value of the YOLOv4-Tiny-DL3 has increased from 69.05 to 85.93 and the FPS of the model has reached 67.83, which has high real-time performance.

6. ACKNOWLEDGMENTS

This study was funded by the Qingdao Top Talent Program of Entrepreneurship and Innovation (Grant No.19-3-2-11-zhc), the Natural Science Foundation of Shandong Province (Grant No. ZR2020MF082), the Foundation of Shandong Intelligent Green Manufacturing Technology and Equipment Collaborative Innovation Center (Grant No. IGSD-2020-012), and the National Key Research and Development Project (Grant No.2018YFB1601500).

7. REFERENCES

[1] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[2] Girshick R. Fast R-CNN[C]. 15th IEEE International Conference on Computer Vision, ICCV 2015, December 11, 2015 - December 18, 2015, 2015: 1440-1448.

[3] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] Dai J, Li Y, He K, et al. R-FCN: Object detection via region-based fully convolutional networks[C]. 30th

Annual Conference on Neural Information Processing Systems, NIPS 2016, December 5, 2016 - December 10, 2016, 2016: 379-387.

[5] Tian J, Yuan J, Liu H. Road Marking Detection Based on Mask R-CNN Instance Segmentation Model[C]. 2020 International Conference on Computer Vision, Image and Deep Learning, CVIDL 2020, July 10, 2020 - July 12, 2020, 2020: 246-249.

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. 14th European Conference on Computer Vision, ECCV 2016, October 8, 2016 - October 16, 2016, 2016: 21-37.

[7] Fu C-Y, Liu W, Ranga A, et al. DSSD : Deconvolutional Single Shot Detector[J]. arXiv e-prints, 2017: arXiv:1701.06659.

[8] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 779-788.

[9] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6517-6525.

[10] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018: arXiv:1804.02767.

[11] Bochkovskiy A, Wang C-Y, Liao H-Y. YOLOv4: Optimal Speed and Accuracy of Object Detection[M]. 2020.