Vehicle and pedestrian detection method based on improved YOLOv4-tiny*

LI Jing[1], XU Zhengjun[2], and XU Liang[2]**

1. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China

2. School of Electrical Engineering and Automation, Tianjin University of Technology, Tianjin 300384, China

(Received 2 May 2023; Revised 13 June 2023)

Aiming at the problem of low detection accuracy of vehicle and pedestrian detection models, this paper proposes an improved you only look once v4 (YOLOv4)-tiny vehicle and pedestrian target detection algorithm. Convolutional block attention module (CBAM) is introduced into cross stage partial Darknet-53 (CSPDarknet53)-tiny module to enhance feature extraction capabilities. In addition, the cross stage partial dense block layer (CSP-DBL) module is used to replace the original simple convolutional module superposition, which compensates for the high-resolution characteristic information and further improves the detection accuracy of the network. Finally, the test results on the BDD100K traffic dataset show that the mean average precision (mAP) value of the final network of the proposed method is 88.74%, and the detection speed reaches 63 frames per second (FPS), which improves the detection accuracy of the network and meets the real-time detection speed.

Document code: A **Article ID:** 1673-1905(2023)10-0623-6

DOI https://doi.org/10.1007/s11801-023-3078-x

The core technology of automatic driving system includes three parts, environment sensing positioning, decision planning and executive control. Environment sensing positioning is the identification and positioning of pavement information, such as pedestrians, vehicles and traffic signs in road scenes, which is a key link in autonomous driving technology. In the process of vehicle driving, rapid and accurate detection of pedestrians and vehicles is of great significance for rational decision-making and safe driving.

In recent years, with the rapid development of deep learning, more and more excellent target detection algorithms have been proposed and applied to the field of automatic driving. Currently, target detection algorithms can be mainly divided into two categories, two-stage target detection and single-stage target detection. Two-stage target detection algorithms include classical algorithms, such as faster R-convolutional neural network (R-CNN)[1,2], R-fully convolution network (R-FCN)[3] and mask R-CNN[4]. The single-stage target detection algorithms include single shot detection (SSD)[5] and you only look once (YOLO) series[6]. In the two-stage object detection algorithm, feature extraction and region proposal are carried out separately, and they adopt different methods. In the single-stage target detection algorithm, all the locations are detected, which has the characteristics of simplicity and efficiency. Both methods have been used successfully in practice. Among them, the two-stage target detection algorithm has higher detection accuracy, but the calculation process of this algorithm is complicated, and the number of parameters and calculation amount of the network are large, which directly leads to the slow detection speed of the network and cannot meet the real-time detection requirements of pedestrians and vehicles in the driving process. The one-stage target detection network model is small, the end-to-end detection is realized, and the detection speed is fast. The deployment of target detection network on vehicle-mounted mobile devices has high requirements on the size and real-time performance of the detection network model, so the single-order target detection algorithm has more advantages in the field of automatic driving.

LI et al[7] proposed a multi-scale vehicle and pedestrian detection algorithm based on YOLOv3 network embedded attention mechanism. The network used spatial pyramid pool module to complete the fusion and stitching operation of multi-scale feature information, and finally the average accuracy was improved by 2.2%. MENG et al[8] proposed a vehicle pedestrian detection method based on SSD network embedded attention mechanism, and used Res Next50 network to replace visual geometry group (VGG) network, the backbone network of the original SSD algorithm. The improved method can effectively improve the detection accuracy of vehicles and pedestrians in traffic scenes. The improved mean average precision (mAP) reaches 79.6% and the

* This work has been supported by the National Natural Science Foundation of China (Nos.61975151 and 61308120).

** E-mail: lxu@email.tjut.edu.cn

detection speed reaches 44 frames per second (FPS). GUO et al[9] proposed an improved YOLOv4 network for target detection of pedestrians and vehicles in road infrared scenes. K-means algorithm was mainly used to regain prior box values, and SENet module was added to enhance the feature description ability. The above studies are aimed at some improved methods proposed by the one-stage target detection algorithm for vehicle and pedestrian detection. The algorithm performance has been improved to varying degrees, but the detection accuracy still needs to be further improved.

Aiming at the problems of missing detection and misdetection in the moving target detection experiments of pedestrians and vehicles in road scenes, this paper uses YOLOv4-tiny[10] as the basic model of the experiment, integrates dual attention module to enhance the feature extraction capability of the network, and uses bidirectional feature pyramid module to enhance the feature fusion capability of the network.

The specific network structure of YOLOv4-tiny is shown in Fig.1.

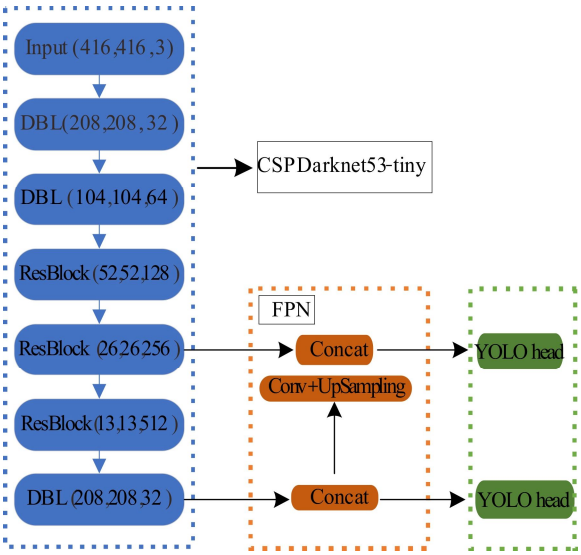

Fig.1 YOLOv4-tiny network structure

The backbone feature extraction network of YOLOv4-tiny is cross stage partial Darknet-53 (CSPDarknet53)-tiny[11]. This module has a total of 15 convolutional layers, which are composed of Darknet Conv2 Leaky-ReLU module and Res block module. Res block is to carry out residual nesting combination of four dense block layer (DBL) units, and then carry out maximum pooling processing, so as to expand the number of gradient paths and reduce the number of parameters and parameters of the model.

The feature fusion module of YOLOv4-tiny network is the feature pyramid network (FPN)[12]. FPN is composed of a bottom-up path, a top-down sum-up path and the middle part of transverse connection. The top-down process and horizontal connection are used to solve the problem of low level of semantic information of low-level features. While maintaining the advantage of low-level features for small target detection, the detection accuracy is improved.

After the feature fusion operation is completed, the network sends the fused feature information to the YOLO head module for classification prediction. CIOU loss and non-maximum suppression (DIOU NMS) functions are introduced into the YOLOv4 series head module to improve the accuracy of head prediction results.

In principle, the attention mechanism can be divided into three modules, spatial attention, channel attention and spatial channel mixed attention.

The principle of channel attention mechanism (CAM) is a model established based on the key of information and the correlation between different information channels, and weights the feature layer with key information, pays attention to meaningful feature channels, and suppresses unimportant information channels.

The essence of spatial attention module (SAM) is to emphasize the spatial feature information of key parts by analyzing the position relationship between pixels, and supplementing the output information of the channel attention module.

Due to the shortcomings of a single channel attention module and the spatial attention module, the spatial and channel information in the channel domain is ignored. The hybrid attention module convolutional block attention module (CBAM)[13] composed of the sequential arrangement of the channel attention module and the spatial attention module makes up for the shortcomings of the single attention module, and the attention weight parameter is introduced sequentially through the two dimensions of space and channel, and then multiplied with the input feature map to make adaptive adjustments to the feature information. The specific structure is shown in Fig.2, and the channel attention module is carried out first. After channel attention and spatial attention processing, the obtained feature output contains both spatial feature information and channel feature information, which enhances the expression ability of the network model for local key.

The attention module is a plug-and-play module, which can theoretically be placed behind any feature layer, that is, it can be placed in the backbone network or the feature fusion network, and this paper applies the attention mechanism to strengthen the feature extraction network.

The block diagram of the CBAM module is shown in Fig.2.

By incorporating the CBAM attention mechanism into CNNs, the model can better capture the important features in the input data, which can lead to better performance in a variety of computer vision tasks, such as image classification, object detection, and segmentation.

The CBAM mechanism can also help to reduce the number of parameters in the model by allowing the network to focus on the most important features, thereby improving the efficiency of the model.

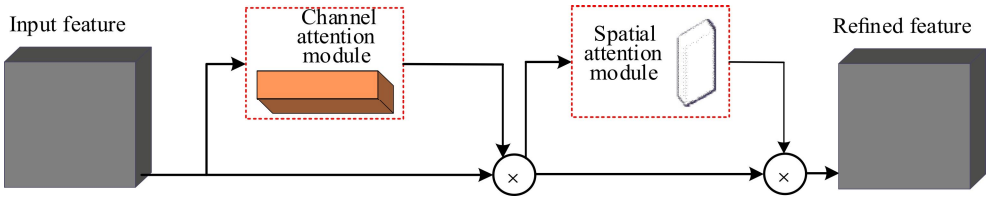

Fig.2 CBAM module structure

In deep neural networks, the lower layers serve as feature extraction modules that retain more low-level image information while having relatively less high-level semantic information. This provides more prediction basis for detecting small and occluded objects. In contrast, the higher layers serve as feature extraction modules that have more high-level semantic information but less low-level information after multiple convolutions[14]. To complement the limitations of both low-level and high-level features, this paper has made further improvements to feature fusion in the neck section and enhanced the details of feature maps.

Cross stage partial network (CSPNet)[15] aims to achieve richer gradient combination while reducing computational complexity. The main idea is to add paths in the network to divert the gradient flow during computation, thereby demonstrating the associated differences in selecting connection and transformation paths. The implementation method is to segment the feature map and fuse it through cross-stage strategies. As an improvement solution, it is often combined with network structures such as ResNet, DenseNet, and EfficientNet to achieve network lightweighting and reduce computational burden while enhancing the learning ability and maintaining accuracy of the CNN. It eliminates computational bottlenecks and reduces memory consumption.

Inspired by the CSP network, the CSP2-DBL module is designed as shown in Fig.3, using DBL as the computational unit in the CSP module to supplement the detail information, and reducing the computational complexity compared to the original CSP residual module. The five DBL superimposed processing on the 32×32 feature map in YOLOv4 is canceled and replaced by CSP2-DBL to achieve cross-stage feature fusion, complement the feature details, reuse features, and only pay a small computational speed cost.

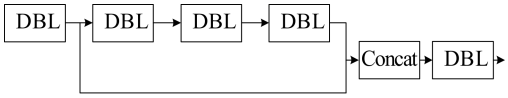

Fig.3 CSP2-DBL module structure

The fused feature P4 is directly upsampled and then fused with the CA-processed P3 feature map. The fused P3 is involved in the first prediction of the head section, with a size of 64×64, preserving more detailed information and mainly used for detecting small objects. The downsampled P3 is fused again with the feature map processed by CSP2-DBL to obtain a new P4, which is involved in the second prediction with a size of 32×32. After participating in the prediction, P4 is downsampled again and fused with P5 to obtain a new P5, which is involved in the third prediction. By complementing the detail features, it is beneficial for better detection of small and occluded objects in the first and second predictions.

Finally, the improved network structure diagram of YOLOv4-tiny is shown in Fig.4. The processing process of the network is as follows. First, the CSPDarknet53-tiny module extracts the features of the input data, obtains three feature maps of different sizes, and sends them to the CBAM module for processing to improve the grasp of the input information. Then, the obtained feature information is sent to the improved feature fusion module of CSP-DBL for feature fusion, and the feature details are compensated, so as to realize feature reuse and improve the detection accuracy of the network.

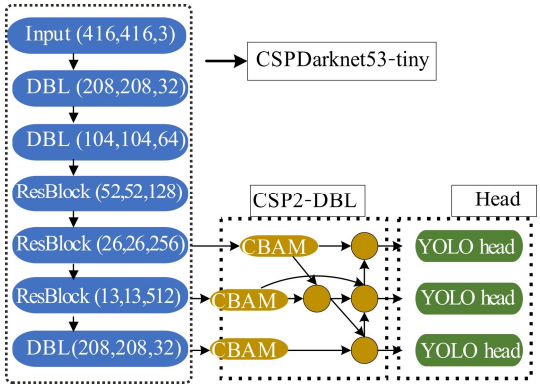

Fig.4 Improved YOLOv4-tiny network model

The processor of the computer used to complete this experiment is Intel(R) Core (TM) i5-4590 CPU @ 3.60 GHz, the operating system is 64-bit Windows10 Professional Edition, and the memory size is 8 GB. The running environment is based on Python3.8 and uses cuda10.1 and cudnn7.8.6, GPU-accelerated.

The dataset used in the experiment is BD100K, which is widely used in model training in the field of computer vision. The dataset consists of 100 000 video files collected from different parts of the United States. As

shown in the figure above, the database covers different weather conditions. It includes sunny, cloudy and rainy days as well as different times of day and night. It is the largest open source video dataset at present. 6 452 images were selected for training in this experiment. Fig.5 is part of the dataset selected in this experiment.

Fig.5 BDD100k dataset

After datasets are prepared, simulation tests are carried out. The training process is shown below.

(1) First, randomly divide the dataset into training set, verification set and test set, set the initial parameters in the network, set the number of iterations to 100 times, set the batch processing size to 640×640, import the corresponding yaml file, modify the training set and dataset path, and then start the training.

(2) The network preprocesses the dataset, such as Mosaic data enhancement, adaptive anchor frame calculation, etc.

(3) Load the network model, the network starts iterative learning, carries out object location classification and feature extraction on the dataset, at the beginning of each iteration, the network will judge whether the current iteration number is the last one, if not, the current mAP value will be calculated, if the updated model has better performance, the new model with better performance will replace the old model, and the best model will be obtained after the training.

(4) After the training, the training results are derived.

The evaluation indexes of this paper are average precision (AP) and mAP to measure the detection accuracy of the algorithm, and the frames of images detected per second (FPS) is used to measure the detection speed of the algorithm.

The calculation of AP and mAP requires specific values of accuracy rate and recall rate. The calculation formulas of accuracy rate and recall rate are shown as follows

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

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

In this paper, for the target of pedestrian and vehicle detection, person and car, true positive (*TP*) represents the part considered positive by the classifier and is indeed positive. False positive (*FP*) represents the part considered positive but negative by the classifier. False negative (*FN*) represents the part of the classifier that is considered negative but positive. In the instruction coordinate system with accuracy rate *P* as the vertical coordinate and recall rate *R* as the horizontal coordinate, a *P*-*R* curve of a class of targets is drawn, and the area of the area enclosed by the curve and the horizontal and vertical axes is calculated. The area value obtained is AP of this class of targets, and mAP is the sum and then average of the AP of all categories. The calculation formula is as follows

$$AP=\int_0^1 P(R)\mathrm{d}R, \tag{3}$$

$$mAP=\frac{\sum_{i=1}^{n} AP_i}{n}. \tag{4}$$

To validate the improvement in accuracy and speed of network recognition by introducing the CBAM attention mechanism module, CSP2-DBL module, and network pruning, we conducted ablation experiments on the BDD100K dataset. As shown in Tab.1, YOLOv4-tiny-1 refers to only introducing the CBAM attention mechanism, YOLOv4-tiny-2 refers to a network that only introduces CSP2-DBL for feature fusion, and improved YOLOv4-tiny refers to a network that simultaneously introduces all methods.

As can be seen from the data in Tab.1, the mAP is 87.09 % when using the YOLOv4-tiny network. When the CBAM attention mechanism was introduced, the detection accuracy of the model for people and car was improved by 0.9% and 2.67%, respectively, and the overall mAP was improved by 0.65%. The introduction CSP2-DB L fuses features to enhance detailed

information, enhances attention to small targets and occlusion targets, and improves the accuracy of people by 0.33%, and the model speed is improved due to CSP2-DBL's shunt of gradient streams. When lightweight design of models is individual, convolution operations are reduced. The number of channels for feature tensors is reduced, and the detection speed is increased by about 10%. The improved YOLOv4-tiny-1 network, which introduces CBAM module and CSP2-DBL module at the same time, increases mAP by 1.65% and reduces the speed by 4 FPS, about 6%, but still meets the requirements of real-time detection. Experiments show that the innovation proposed in this paper can effectively improve the detection accuracy of small targets and occlusion targets.

Fig.6 shows the detection results of the original YOLOv4 tiny network, YOLOv4 network, and the improved network, respectively.

Tab.1 BDD100K dataset test results

Model	Class	Precious (%)	Recall (%)	AP (%)	mAP (%)	FPS
YOLOv4-tiny	car	90.86	79.43	87.86	87.09	65
	People	87.65	78.89	86.32		
YOLOv4-tiny-1	car	91.76	80.06	88.31	87.74	71
	People	90.35	74.86	87.69		
YOLOv4-tiny-2	car	90.15	78.69	87.16	86.42	73
	People	87.98	76.59	85.67		
Improved YOLOv4-tiny	car	91.69	80.16	89.54	88.74	63
	People	88.16	78.90	87.93		

YOLOv4 Improved YOLOv4-tiny YOLOv4-tiny

Fig.6 Network detection results

By observing the best detection effect of YOLOv4, pedestrians and obscured vehicles in the distance can be detected. The YOLOv4-tiny detection effect is the worst, and the missed detection problem is serious. The improved network improves the missed detection problem of YOLOv4-tiny. It can be concluded that the improved network detection effect is improved compared with the original network.

In this paper, a pedestrian vehicle detection algorithm based on the improved YOLOv4-tiny network is proposed. First, in order to ensure the authority of the dataset, the public BDD100K dataset is adopted, and then the YOLOv4-tiny network is improved. The CBAM attention mechanism module was added, and the feature fusion module of the network was improved using the CSP-DBL module. The final experiments show that the detection accuracy of the improved YOLOv4-tiny network is improved, and the average accuracy reaches 89.93%. The detection speed reaches 63 FPS, which meets the requirements of real-time performance and has excellent performance in pedestrian vehicle inspection. Subsequent research will continue to explore how to increase detection speed without sacrificing accuracy for better deployment on edge computing devices.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] GIRSHICK R. Fast R-CNN[C]//Proceeding of the IEEE International Conference on Computer Vision, December 11-18, 2015, Santiago, Chile. New York：IEEE, 2015：1440-1448.

[2] REN S Q, HE K M, 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.

[3] DAI J F, LI Y, HE K M, et al. R-FCN：object detection via region-based fully convolution networks[C]// Proceeding of the 30th Annual Conference on Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. Neural Information Processing Systems Foundation, 2016：379-387.

[4] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceeding of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York：IEEE, 2017：2980-2988.

[5] FANG L P, HE H J, ZHOU G M. Research overview of object detection methods[J]. Computer engineering and applications, 2018, 54(13)：11-18. (in Chinese)

[6] SHAO Y H, ZHANG D, CHU H Y, et al. A review of YOLO object detection based on deep learning[J]. Journal of electronics & information technology, 2022, 44(10)：3697-3708. (in Chinese)

[7] LI J Y, YANG J, KONG B, et al. Multi-scale vehicle-pedestrian detection algorithm based on attention mechanism[J]. Optics and precision engineering, 2021, 29(06)：1448-1458. (in Chinese)

[8] MENG L X. Research on vehicle-pedestrian detection method based on deep learning[D]. Taiyuan：North University of China, 2021. (in Chinese)

[9] GUO Z J, LI J Y, QI H J, et al. Detection algorithm for infrared pedestrian and vehicle based on the improved YOLOv4[J]. Laser & infrared, 2023, 53(4)：607-614. (in Chinese)

[10] ZHOU H P, WANG J, SUN K L. Pedestrian detection algorithm based on improved YOLOv4-tiny[J]. Radio communications technology, 2021, 47(4)：474-480. (in Chinese)

[11] YI X, SONG Y H, ZHANG Y L. Enhanced darknet53 combine MLFPN based real-time defect detection in steel surface[J]. Chinese Conference on Pattern Recognition and Computer Vision (PRCV), October 16-18, 2020, Nanjing, China. Berlin, Heidelberg：Springer Science and Business Media Deutschland GmbH, 2020：303-314.

[12] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, USA. New York：IEEE, 2017：936-944.

[13] WOO S, PARK J, LEE J Y, et al. CBAM：convolutional block attention module[C]//15th European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin, Heidelberg：Springer Verlag, 2018：3-19.

[14] TAN M M, PANG R M, LE Q V. EfficientDet：scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 14-19, 2020, Virtual. New York：IEEE, 2020：10778-10787.

[15] CHIEN Y W, HONG Y M L, YEH I H, et al. CSPNet：a new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 14-19, 2020, Virtual. New York：IEEE, 2020：1571-1580.