ITC 2/51

Information Technology and Control
Vol. 51 / No. 2 / 2022
pp. 294-312
DOI 10.5755/j01.itc.51.2.30667

Lightweight Deeplearning Method for Multi-vehicle Object Recognition

Received 2022/02/06 | Accepted after revision 2022/03/12

http://dx.doi.org/10.5755/j01.itc.51.2.30667

HOW TO CITE: Li, X., Yun, X., Zhao, Z., Zhang, K., Wang, X. (2022). Lightweight Deeplearning Method for Multi-vehicle Object Recognition. *Information Technology and Control,* 51(2), 294-312. http://dx.doi.org/10.5755/j01.itc.51.2.30667

Lightweight Deeplearning Method for Multi-vehicle Object Recognition

Xun Li

School of Electronics and Information, Xi'an Polytechnic University, 58 Shangu Road, Lintong District, Xi'an City, Shaanxi Province; Phone: +86 13898231612; e-mail: lixun@xpu.edu.cn

Xi'an Polytechnic University Branch of Shaanxi Artificial Intelligence Joint Laboratory, 19 Jinhua Road Xincheng District, Xi'an City, Shaanxi Province; Phone: +15191418916; e-mail: lixun@xpu.edu.cn

Xin Yun

School of Electronics and Information, Xi'an Polytechnic University, 58 Shangu Road, Lintong District, Xi'an City, Shaanxi Province; Phone: +86 18392164708; e-mail: 1601611194@qq.com

Zhengfan Zhao

Reliability Data Center, The Fifth Institute of Electronics, Ministry of Industry and Information Technology, Guangzhou City, Guangdong Province; Phone: +86 18610171729; e-mail: 108983231@qq.com

Kaibin Zhang

School of Electronics and Information, Xi'an Polytechnic University, 58 Shangu Road, Lintong District, Xi'an City, Shaanxi Province; Phone: +15191418916; e-mail: xihua_0169@163.com

Xi'an Polytechnic University Branch of Shaanxi Artificial Intelligence Joint Laboratory, 19 Jinhua Road Xincheng District, Xi'an City, Shaanxi Province; Phone: +15191418916; e-mail: xihua_0169@163.com

Xiaohua Wang

School of Electronics and Information, Xi'an Polytechnic University, 58 ShanguRoad, Lintong District, xi'an City, Shaanxi Province; Phone: +18946004933; e-mail: wangxiaohua@xpu.edu.cn

Xi'an Polytechnic University Branch of Shaanxi Artificial Intelligence Joint Laboratory, 19 Jinhua Road Xincheng District, Xi'an City, Shaanxi Province; Phone: +86 13759969743; e-mail: wangxiaohua@xpu.edu.cn

Corresponding author: lixun@xpu.edu.cn

The recognition method based on deep learning has a large amount of calculation for the changes of different traffic densities in the actual traffic environment. In this paper, an integrated recognition method YOLOv4-L is proposed for reducing computational complexity based on the YOLOv4. The characteristics of multi-lane traffic flow with different flow densities were analyzed for statistical data sets, and k-means++ clustering algorithm was used to optimize the prior frame parameters to improve the matching degree between the prior frame. GhostNet was used to replace CSPDarknet53 of original network structure of YOLOv4 as the feature extraction network. The depthwise separable convolution module was introduced to replace the original 3×3 common convolution in feature extraction network, reduce model parameters and improve detection speed. The network model is further improved both with accuracy and robustness with the help of comprehensive method of Mosaic data enhancement, learning rate cosine annealing and label smoothing. Experimental results show that, Recognition speed is greatly improved at the expense of minimal recognition accuracy reduction: the recognition speed improvement value is 47.81%, 49.15% , 56.06% in detection speed (FPS), respectively in free flow, synchronous flow and blocked flow, the reduction value of accuracy is 2.21%, 0.67%,, 0.05% mAP, respectively.

KEYWORDS: Multi-object recognition, YOLOv4, GhostNet, Depthwise separable convolution.

1. Introduction

With the acceleration of urbanization in China, vehicle ownership is increasing and the pressure on road traffic is growing. This has led to increasingly serious problems such as traffic congestion, blockage and traffic accidents. Municipalities and traffic management departments have been increasing the efficiency of vehicle use of roads by increasing road infrastructure construction, expanding roads and erecting viaducts in order to alleviate the many problems caused by traffic congestion. Obviously, the monitoring of multi-lane quasi-freeways is almost devoid of human involvement, and in order to ensure traffic safety in the city and effectively combat violations, camera-based surveillance systems have been established in all sections and important locations of the city [15], and the background of such surveillance systems requires a large number of personnel to perform behavioral recognition from the collected image information. In multi-lane traffic, when the speed and number of vehicles increase dramatically, it is difficult to capture various abnormal behaviors in vehicles and traffic flow accurately in real time based on manual recognition methods. Therefore, accurately detecting vehicles from the background through intelligent traffic management has become a popular research topic nowadays. To grasp the number and distribution characteristics of vehicles in a large area as much as possible is a prerequisite for optimizing intelligent traffic, and only by accurately detecting vehicles from the background can traffic flow statistics, vehicle identification and tracking be performed.

By observing the characteristics of traffic flow, it is found that the individuals participating in traffic are not independent of each other, and there are mutual influences and mutual constraints between vehicles and vehicles, and between vehicles and road infrastructure, thus forming a complex traffic system. When the scale of the city becomes larger and larger, the expansion of the number of lanes causes various contingency, randomness and uncertainty in the system. At the same time, in the urban traffic environment, the travel rules of vehicles, the temporal and spatial characteristics of vehicle concentration during morning and evening peak hours, and the synchronization of traffic flow and density Increasing, many problems occur simultaneously in a multi-lane environment, where the interaction of individual motor vehicles in the traffic flow is particularly obvious. For example, the number of vehicles in the same visual frame is high, the speed of the vehicles is high, the number of vehicles of the same color and type is large, and the vehicle target in the same visual frame changes lanes frequently, which will affect the detection of the same vehicle in the adjacent frames before and after. It can be seen that there are still many complex problems in the modern traffic environment. Therefore, how to establish a suitable vehicle object detection dataset to describe the complex relationship between vehicles, how to correctly detect more vehicle objects in the case of real-time detection, and how to construct a lightweight and fast vehicle detection accuracy under the premise The detection model is the key research content of intelligent transportation system to identify vehicle targets in different traffic flows.

In view of the important position of vehicle detection in the intelligent transportation system, the research on more robust, accurate and efficient vehicle detection methods undoubtedly has important academic value and broad application prospects. Vehicle detection method is a branch of target detection method. From the perspective of the whole development process, the earliest detection method is the VJ (Viola-Jones) detector proposed by Paul Viola and Michael J Jones in 2001. Based on the AdaBoost algorithm, Haar-like wavelet features and integral graphs are added to make the obtained features more targeted to the face, and multiple AdaBoost strong classifiers are cascaded at the same time to put more resources in the target window. above. The HOG algorithm was proposed by Dalal et al. at the 2005 CVPR conference. Its main innovation is that in the detection process, the local shape of the object can be described by the distribution of light intensity gradients or edge directions, which is often used in pedestrian detection. In 2010, Felzenszwalb et al. proposed the DPM (Deformable Parts Model) method, which is an excellent target detection method and won three consecutive championships in the VOC (Visual Object Class) competition in 2007-2009. Its idea is similar to HOG, but the traditional HOG feature only uses one-to-one feature representation, while DPM divides the model into a root model and a partial model. The root model is similar to the traditional HOG feature and is used for object location positioning, and part of the model is used for further confirmation. Since the DPM method requires more computation than HOG, its detection effect is much better than the traditional HOG method. In 2012, the convolutional neural network achieved great success in the large-scale visual recognition challenge ImageNet competition, and people subsequently applied it to various application fields such as speech recognition, image classification, and face recognition, and achieved far more than The performance of traditional methods, researchers have introduced deep learning into the field of vehicle recognition, and breakthrough results have been emerging in recent years.

Based on the deep learning method of detection and recognition, especially for the high accuracy of image samples, we will naturally consider the introduction of artificial intelligence technology into traffic monitoring. As the core technology of intelligent monitoring system, vehicle recognition is to detect the position of vehicle objects in the image and identify the type of objects In practical application [1], vehicle recognition is easily affected by factors such as background clutter [8], illumination conditions and partial occlusion, resulting in the reduction of object detection accuracy [26]. However, recent research results show that the detection accuracy of multi-lane vehicles is well solved based on deep learning, the real-time problems caused by massive computing needs. Fast and accurate vehicle object detection in traffic scene has always been the research content in the cross field of image processing and traffic engineering.

With the continuous development of deep learning, convolutional neural networks have excellent performance on vehicle target detection tasks, but current research focuses on how to build deeper networks for the purpose of improving detection accuracy. This has led to overly large network models, and most of the best-performing networks can only run on high-performance graphics processors(GPUs). In order to apply deep learning models more widely on embedded platforms, such as intelligent surveillance systems and unmanned systems, building lightweight networks can effectively reduce hardware costs and improve the operational efficiency of the networks.

Such networks are difficult to be applied in embedded platforms due to computational and memory limitations. Therefore, for the task of vehicle target detection in real-time embedded scenarios of intelligent surveillance systems, this paper optimizes the YOLOv4 network based on the current excellent lightweight convolutional network GhostNet and constructs the lightweight vehicle detection network YOLOv4-L. The YOLOv4-L network reduces the model parameters by optimizing the traditional convolutional operation and using depth-separable convolution to improve operational efficiency.

2. Related Work

Compared with traditional detection algorithms for target detection, there are defects such as poor robustness, large calculation amount, and weak applicability [11], the detection algorithm based on deep learning [31] has the advantages of high accuracy. Deep learning techniques are widely used in various fields, such as computer vision, machine vision, and speech recognition, among which computer vision is one of the most popular fields in which promising results to have been obtained in image classification tasks [32, 33]. Object detection methods based on deep learning are mainly divided into two categories. One is the two-stage algorithm represented by R-CNN [6], Fast R-CNN [7] and Faster R-CNN [18]. Although this method has high detection accuracy and positioning accuracy, it has some defects such as cumbersome training steps and poor real-time performance. One is the single-stage algorithm represented by SSD [12] and YOLO series algorithms [19, 20, 21]. This method cancels the candidate region generation mechanism and performs classification and regression prediction directly by convolutional operations, so as to generate the category probability and coordinate information of the object.

In the research of object detection by deep learning method, vehicle is a kind of special object, and many research results have been presented in recent years. Yang Wei et al. [29] proposed an improved Faster R-CNN vehicle real-time detection algorithm. The algorithm adopts a multi-scale strategy in model training, which improves the generalization ability of the model. However, the vehicle detection effect is not effective in complex environments such as dense vehicles and severe occlusion. For the Faster R-CNN algorithm, the vehicle detection effect is not good in complex traffic environments such as dense vehicles and serious occlusions. Nguyen et al. [17] proposed an improved algorithm based on Faster R-CNN for fast vehicle detection. The algorithm adopts MobileNet architecture to build the basic network of Faster R-CNN framework, and introduces context-aware pool to improve the detection accuracy of network model for occluded vehicles and small target vehicles. The network model effectively improves the detection accuracy of occluded vehicles and small object vehicles. However, further improvements are needed in real-time detection. Li Xun et al. [13] proposed a new target detection network YOLO-vocRV based on YOLOv2, which transforms the detection problem into a binary classification problem and improves the detection accuracy of the model. The proposed YOLO-vocRV method is suitable for multi-target detection of different traffic densities, and the average accuracy of the YOLO-vocRV model can reach more than 90% for different traffic densities. However, since the vehicle images in the dataset used for the experiments are collected under good visible light conditions. Therefore, the vehicle detection under night light or low light conditions is poor. Shi Binbin [22] et al. proposed a YOLOv2-voc_mul, a multi-objective recognition and classification method for vehicles based on the YOLOv2 algorithm, for solving the problems of low detection rate, poor robustness, and unsatisfactory impact on real road environment of classical multi-objective classification methods. The model improves the network structure of YOLOv2_voc based on the YOLOv2 algorithm according to the real road conditions, and obtains the classification network structure YOLOv2-voc_mul for sensitive vehicle changes using the ImageNet data after multiple adjustments. The improved algorithm reduces the model framework and parameter computation and improves the accuracy. However, the method is not accurate enough for the detection of tiny targets at a distance, and the detection of tiny objects at a distance may still be missed when the method is used to detect objects. Zhu Maotao et al. [30] proposed the YOLO-TridentNet algorithm, which uses the TridentNet algorithm weight sharing based on YOLOv3 to improve the detection accuracy of small targets in more distant vehicles. However, the model contains three branch networks with different expansion rates, the structure is complex and the parameters are large, which cannot meet the needs of real-time detection. Choi et al. [4] improved YOLOv3 by using Gaussian modeling bounding box coordinates combined with improved loss function, which has a good balance between detection accuracy and detection speed, but it cannot satisfy real-time detection for larger size input images.

Although the above methods and their improvements have been improved in detection accuracy, they have complex network structures and a larger number of network parameters. They require powerful GPU computing power to achieve the real-time object detection. To solve this problem, many researchers have proposed lightweight target detection methods. These methods have comparatively simpler network structures and fewer parameters. As a result, the demand for computational resources and memory is low and the detection speed is fast. Weiping [23] proposed using MobileNetv2 [24] to replace the backbone network of YOLOv3. The model size was reduced to 26MB, which was 90% lower than that of YOLOv3, but sacrificed 10.42% of the mAP value. Chen [25] et al. improved the YOLOv3-tiny algorithm by using model pruning combined with parameter quantification, which greatly increased the detection speed and met the demand for real-time vehicle detection. However, parameter quantification has a greater impact on detection accuracy and still needs to be further improved. Qinghe [34] proposed a holistic pruning method named Drop path to reduce model parameters of 2D deep convolutional neural networks, utilizing redundancy inter parameters per layer under PAC-Bayesian framework. This results in smaller memory footprint and computational requirements for real-time image processing, making deep CNN easier to be deployed on mobile systems or embedded devices, and effectively accelerating the inference of the network. In 2020, Bochkovskiy et al. [3] combined mainstream optimization skills and more complex network architecture on the basis of YOLOv3 to design YOLOv4, which supports fast and accurate training and detection on a single graphics card. However, due to the large amount of model parameters and large volume, it is difficult to meet the needs of high detection speed and high detection accuracy at the same time. Wu et al. [27] combined with the channel pruning algorithm on the basis of YOLOv4, greatly reduced the model size and the number of parameters, and improved the model efficiency, but the optimization effect is unstable for different data sets. Hu et al. [9] proposed adding dense modules to the backbone network of YOLOv4 to reduce the network depth. Although the network parameters are reduced, due to the connection of each module and multiple modules in the network, the network structure becomes complex and the amount of calculation is increased.

In summary, the improvement ideas of the YOLO series of vehicle recognition algorithms focus on improving the detection speed of the algorithm while increasing the detection accuracy. It reduces the number of model parameters by improving the feature extraction network, model pruning, etc., and improving the feature fusion strategy, convolution method, and combining multiple algorithms to reduce accuracy loss. Although the improved vehicle detection algorithm has improved inference speed and detection accuracy, there is still much room for improvement in the detection performance of the multi-vehicle target model recognition based on the improved vehicle detection algorithm of the YOLO series at this stage. Aiming at the shortcomings of the YOLOv4 model in vehicle multi-target recognition, such as large model size, high computational complexity, low operating efficiency, etc., an optimized backbone feature extraction network and PANet and detection head part of the common convolution block are proposed. YOLOv4-L multi-vehicle target recognition model. Use the improved GhostNet to replace the original CSPDarknet53 feature extraction network, which reduces the amount of model parameters and improves the detection accuracy. It is proposed to introduce a deep separable convolution block to replace the ordinary convolution used in the network, which further reduces the model parameters The K-means++ clustering algorithm is used to optimize the prior frame parameters to improve the accuracy of model detection. Finally, the Mosaic data enhancement, cosine annealing and label smoothing methods are combined to enhance the convergence effect of the model and improve the model The generalization ability. Finally, the detection methods in this paper are tested under different road traffic conditions by comparing with YOLOv3, YOLOv3-tiny, and YOLOv4 model.

3. Lightweight Multi-object Recognition Model YOLOv4-L

YOLOv4 is an improved object detection algorithm based on the YOLOv3 model. The algorithm is mainly composed of three parts: the feature extraction network Backbone, the neck feature enhancement network Neck, and the head detection network Head.

3.1. Object Detection Algorithm Based on YOLOv4 Model

According to the previous researches [30, 25, 9, 16], when YOLOv4 is used for vehicle detection and recognition, The specific implementation process of yolov4 for vehicle object detection is as follows:

1 After the input actual road vehicle pictures pass by the feature extraction network, the feature maps of three scales of 52×52, 26×26, and 13×13 are out-

put, respectively. The feature maps of different scales contain semantic information of different dimensions.

2. In the feature fusion part, the 13×13 size feature map enters the SPP (Spatial Pyramid Pooling) structure. The SPP stacks and convolves the obtained new feature map and the feature map before entering the network and then outputs it to the feature fusion Network PANet.
3. The PANet upsamples the 13×13 feature map twice, and then stacks the results of the first upsampling and the second upsampling with the feature maps of the 26×26 and 52×52 scales and convolves them. Then a series of similar down-sampling and stacked convolutions are performed from top to bottom to fully integrate the features of three different scale feature maps.
4. Finally, the feature map after feature fusion is output to three YOLO detection heads of 52×52, 26×26, and 13×13. Each detection head contains 3 sets of adjustment parameters of candidate frames, and the adjustment parameters of each set of candidate frames include 1 confidence parameter, 4 parameters for adjusting length, width and coordinate offset, and 80 category parameters. With these adjustment parameters, the YOLOv4 algorithm will adjust the coordinates and width and height of the candidate frame to generate the final prediction frame.

3.2. Improvement of Multi-object Detection and Recognition Model

The improved lightweight multi-vehicle target recognition model YOLOv4-L mainly includes the following three parts: GhostNet feature extraction network, feature enhancement extraction module composed of SPP and PANet and YOLO-Head. The overall network structure is shown in Figure 1, and the Submodules of YOLOV4-L is described in Figure 2.

As shown in Figure 1, (a) module is the end-to-side residual module Ghost BottleNecks (see Section 3.2.1). (b) module is a CBL module that combines batch normalization and LeakyRelu activation function. (c) module is SPP module, It is composed of multiple pooling layers, and the feature maps of the same size are spliced at the output end (see Section 3.2.2).

Figure 1

Network structure of YOLOv4-L

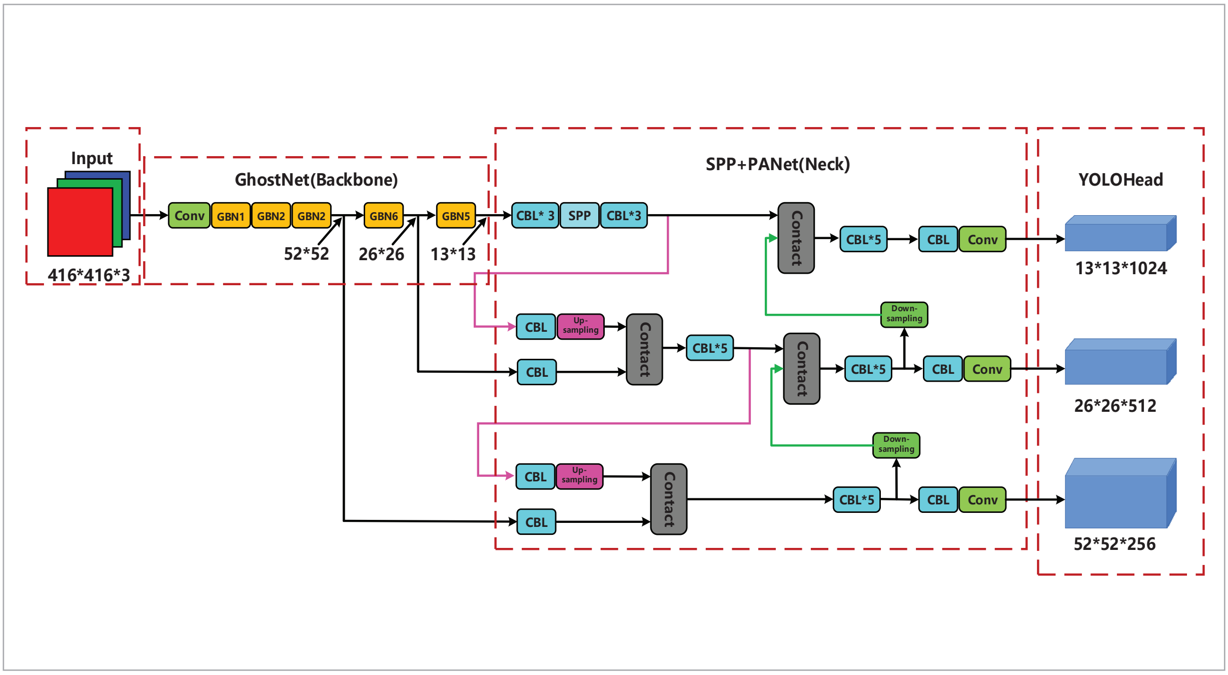
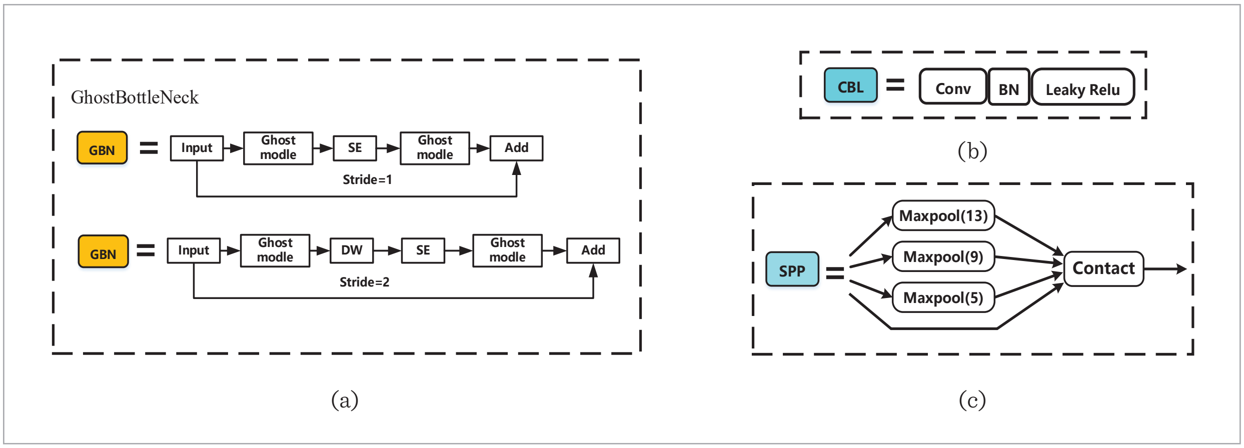


Figure 2

Submodules of YOLOV4-L



(a) (b) (c)

3.2.1. GhostNet Feature Extraction Network

For the detection of vehicles, the vehicle detection speed will slow down since a large number of operational parameters in the CSPDarknet53 feature extraction network, Overfitting is easily occurred when the data set sample size is small. Many of the network feature layers are similar, and the redundant part in the feature layer may be an important part. Eliminating the same features and retaining important feature information is the first problem to be solved by model YOLOv4-L.

The Ghost module in GhostNet [10] can generate the same features with fewer parameters. Therefore, the redundant information is reserved in Ghost, and the feature information is obtained with lower computational cost.As this part, GhostNet, which removes the average pooling layer and the full connection layer, is used as the feature extraction network of YOLOv4-L. The function of this part is to extract the preliminary features of the input vehicle image data sets under different traffic environments. After preliminary feature extraction, it can be extracted when G-bneck is the sixth, twelfth, and sixteenth layers, and three effective feature layers can be obtained. They are 52×52 large target feature layer, 26×26 medium target feature layer, and 13×13 small target feature layer.

GhostNet is based on the end-to-side residual module (Ghost bottlenecks, G-bneck), G-bneck is mainly composed of two Ghost modules. The first Ghost module is mainly used to increase the number of channels and expansion layer, second Ghost module is mainly

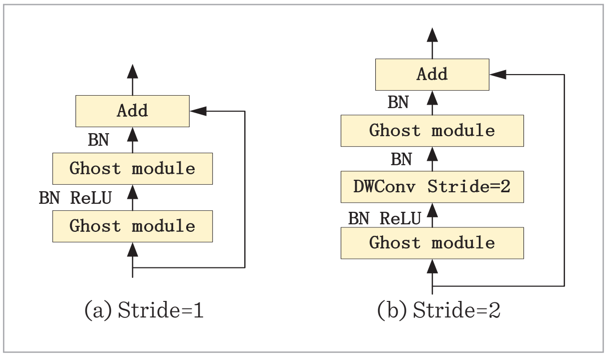
Table 1

GhostNet network structure

Input	$416^2\times3$	$208^2\times16$		$104^2\times24$		$52^2\times40$		$26^2\times80$			$26^2\times112$		$13^2\times160$				$1^2\times960$	$1^2\times1280$
Operator	Conv2d 3×3	G-bneck												Conv2d 1×1		AvgPool 7×7	Conv2d 1×1	FC
exp size	–	16	48	72		120	240	200	184	480	672		960	–				
out size	16		24		40		80			112	112	160	160	960		–	1280	1000
SE	–	–	–	–	√		–	–	–	√			–	√	–			
Stride	2	1	2	1	2	1		2	1			2	1			–	1	–

used to reduce the number of channels and connect the input and output of the two Ghost modules. The specific structure is shown in Figure 3.

Figure 3
G-bneck structure of different stride



(a) Stride=1 (b) Stride=2

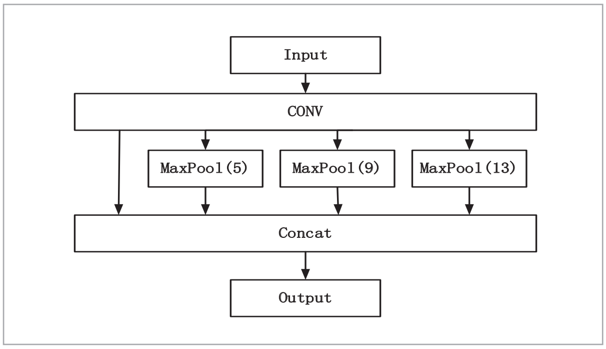
The Ghost module is used as the main building block. The first layer adopts the standard convolution process. The G-bneck operation increases the number of channels, and G-bneck is grouped into different stages according to feature maps of different sizes. At the same time, the squeeze-and-excitation module is introduced to make the extracted features more objects and the features are more fully utilized through the attention mechanism. Conv2d represents a convolutional layer, AvgPool represents an average pooling layer, and FC represents a fully connected layer.

3.2.2. Feature Enhancement Extraction Module

For avoiding the over-fitting problem mentioned in Section 3.2.1, a feature enhancement extraction module composed of SPP and PANet was introduced into YOLOV4-L. The purpose is to perform feature fusion on the three preliminary feature layers extracted by the backbone feature extraction network, so as to obtain three more generalized optimized feature layers. In this part, replace the 3×3 ordinary convolution in the tertiary convolution block and the fifth convolution block in the PANet network with the depth separable convolution to further reduce the network parameters.

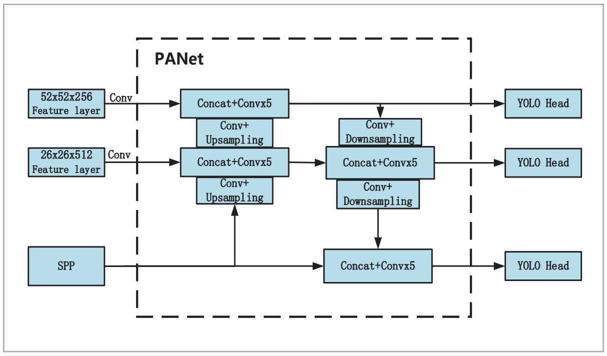
The SPP module obtains the receptive field information of the local region and the near global receptive field of the feature map by using the maxpool layer of cores with different sizes, and performs feature fusion. This operation of fusing receptive fields of different scales can effectively enrich the expression ability of feature map, enhance the acceptance range of output features of feature extraction network, separate important context information, effectively expand receptive fields and reduce over fitting. The structure of the SPP module is shown in Figure 4.

Figure 4
SPP structure diagram



In YOLOv4-L, The specific implementation process of the SPP module is as follows: After the input feature map is subjected to convolution operation, it is sent to the Max Pooling layer of 5×5, 9×9, and 13×13 for maximum pooling processing, and then the vectors of different scales and the vectors of the original feature map are concatenated into multi-scale vectors. The number of output channels becomes 4 times the original number of channels, and the size of the feature map remains unchanged. The final feature layer is input to the subsequent PANet feature extraction pyramid.

Figure 5
SPP structure diagram



PANet is a further improvement of the Feature Pyramid Network (FPN) [28]. On the basis of FPN, a Bottom-up Path Augmentation structure is added to avoid the problem of shallow information loss in the transmission process and improves the accuracy of network prediction. Figure 5 shows the PANet network structure with improved input feature layer size. It combines up-sampling, down-sampling and features, so that the results of up-sampling and down-sampling and the results of the corresponding effective feature layer convolution are simultaneously Concat, and After the multi-level information is integrated, the prediction is made, and the underlying information is effectively used, and finally 3 effective feature layers of YOLO Head are obtained. PANet provides three feature layers, respectively, the sizes are 52×52×256, 26×26×512, 13×13×1024, respectively corresponding to the middle, middle and lower prediction frames.

3.2.3. Feature Enhancement Extraction Module

YOLO-Head uses the multi-scale features obtained from the PANet structure to perform regression and classification prediction.

The YOLO-Head output matrix scales are 52×52×255, 26×26×255, 13×13×255, respectively. The third dimension is 255 because it can be split into 3×(80+5). 80 is the number of sample classifications in the COCO data set, and 5 represents the X-axis offset, Y-axis offset, height *H* and width *W*, confidence and classification results of the prediction frame. When the prediction result is obtained, the prediction box can be drawn directly on the original map. The predicted three output scale matrices also need result decoding. The original image will be decomposed into 76×76, 38×38, and 19×19 matrix. The center of the prediction frame can be obtained by adding its corresponding X-axis offset and Y-axis offset to each matrix point, and then the length and width of the prediction frame can be calculated by using the a priori frame, prediction height *H* and width *W*, and finally the prediction frame can be drawn on the original image.

3.3. Parameters Reducing by Depthwise Separable Convolution

The parameter quantity directly determines the size of the network model and also affects the memory consumption during inference. The depthwise separable convolution [5] replaces the standard convolution with fewer parameters and calculations. In the process of feature extraction, the size of convolution kernel is usually 3×3 size. Therefore, the amount of calculation and parameters of depth separable convolution are about 1/9 of that of conventional convolution. There are useful changes in YOLOv4-L:

1. Replace the 3×3 ordinary convolution in the tertiary convolution block and the fifth convolution block in the PANet network with a deep separable convolution.
2. Replace the 3×3 ordinary convolution in the prediction network YOLO-Head with a depthwise separable convolution.
3. Replace the 3×3 ordinary convolution in two down-sampling with depth separable convolution. The changes in the parameters of the network model after replacement are shown in Table 2. YOLOv4-Ghostnet is a network model that only replaces the feature extraction network. YOLOv4-L is based on the YOLOv4-Ghostnet model after performing the above deep separable convolution replacement.

Table 2
Comparison of parameters in different network structures

Model	Parameters	Model size/MB
YOLOv4	64040001	244.29
YOLOv4-Ghostnet	39689409	151.40
YOLOv4-L	11428545	43.60

After using the lightweight feature extraction network GhostNet to replace CSPDarknet53 in YOLOv4, the parameters of the network model YOLOv4-Ghostnet are reduced by 38% compared with YOLOv4. After replacing the ordinary convolution with depth separable convolution, the parameter amount of the YOLOv4-L network model are decreased by 29%. The parameter quantity of the YOLOv4-L model is only 1/6 of YOLOv4.

4. Experimental Results and Analysis

In the multi-target recognition experiment, a workstation is equipped with an Intel i7-6800 CPU, one NVIDIA GeForce Titan X 1080TI 11GB GPU and eight 32GB memories.

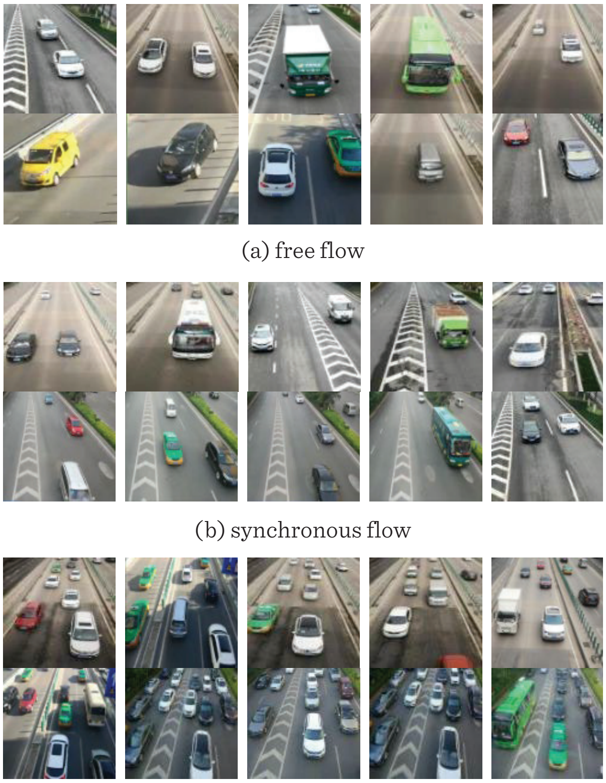
To ensure fairness, for the four algorithm models compared in the experiment, YOLOv3, YOLOv3-tiny, YOLOv4 models and YOLOv4-L were trained and tested independently for different algorithm models under the same initialization conditions on the workstation, using the same number of training sample sets and the same ratio of training set to test set allocation, both being 8:2. Ensure that the CPU share does not fluctuate by more than 1% before training and uniformly use the Pytorch 1.7 deep learning framework. The initial learning rate was 0.001, the batch size was set to 32, and the training was unfreezed after 50 iterations. In order to prevent the program from reporting an error due to insufficient display memory or the program termination during operation, the unfreezed Batch size was set to 16,, the learning rate was set to 0.0001, and 70 iterations were performed, for a total of 120 iterations. The momentum size is 0.9 and the smooth_Label was set to 0.01.

4.1. Data Sample

The experiment constructed three road scene object detection data sets in different traffic environments according to the illumination and traffic density, which are free flow, synchronized flow and blocked flow [14] The data sets are obtained by actual shooting, and the shooting periods are 6:00-7:00 (number of vehicles < 300 vehicles/hour, free flow), 7:30-8:30 (number of vehicles between 900 and 1300 vehicles/hour, blocking flow), and 9:00-11:00 (number of vehicles between 300 and 900 vehicles/hour, synchronous flow), the samples obtained are shown in Figure 6.

The data set contains 4 vehicle types: car, van, bus, and truck. The images are in RGB format, totaling 6000. The data in each time period are randomly divided into training set and verification set in the ratio of 80% and 20%. Before training, the input image size was adjusted to 416×416, and the width, height and center point coordinates of the bounding box of the labeled information were normalized following the PASCAL VOC data set format to reduce the influence of abnormal samples on the data.

Figure 6
Data samples of different flow



(a) free flow

(b) synchronous flow

(c) blocked flow

4.2. Selection of Anchor Box

For object detection, the choice of anchor box will directly affect the final performance of the model. The appropriate anchor box is determined by the object size of the data set. The size of the anchor box in YOLOv4 is obtained from the COCO data set by clustering. However, considering that the COCO data set contains 80 types of objects, the aspect ratio of the detection targets varies greatly and is not applicable to the road vehicle detection dataset in this paper. In the task of road vehicle detection, the aspect ratio of the object vehicle is a relatively fixed value, and the original anchor box size of the model cannot be used directly. The anchor box in the training set used need to be clustered and redistributed. In this paper, the k-means++ clustering [2] method is chosen to calculate the clusters, and the basic steps of the algorithm are as follows:

1. A randomly selected point from the data set as the initial cluster center.
2. Calculate the shortest distance $D(x)$ between each sample and the existing cluster center.
3. Calculate the probability $D(x)^2/\sum_{x\in X} D(x)^2$ that each point becomes the next cluster center, and select the next cluster center according to the roulette rule.

4 Repeat steps (2)-(3) to select all clustering centers.

5 Using the traditional k-means algorithm to cluster the selected clustering centers.

The YOLOv4-L algorithm detects vehicle objects by 3 feature detection scales, assigns 3 anchor boxes to each feature detection scale. Each cell in each feature detection scale is predicted by 3 anchor boxes for each of the 3 bounding boxes, so 9 anchor boxes are required. The k-means++ clustering algorithm is used to cluster the length and width of the label box in the vehicle data set. Based on the clustering results, the updated anchor boxes sizes are: (30, 20), (42, 29), (44, 49), (56, 38), (71, 57), (98, 91)), (137,140), (301,252) and (374,347), respectively. Compared with the original anchor boxes of YOLOv4, the size of the anchor boxes obtained by clustering based on the vehicle data set is more consistent with the aspect ratio of the road vehicle in the training set. Using the updated anchor boxes to train the vehicle data set can make the localization more accurate.

4.3. Ablation Experiment

The ablation experiments, respectively, compared the YOLOv4 recurring network with the three modified lightweight networks proposed in this paper to prove the performance improvement brought by the network design in this paper. The design of the ablation experiment is shown in Table 3. Model 1 is the original YOLOv4 object detection model. Model 2 is the initial optimized model using GhostNet as the backbone feature extraction network. Model 3 represents a light-weight network model based on model 2 using depth-separable convolution to replace the general convolution in the enhanced feature extraction network and the YOLO-Head prediction network. Model 4 represents a further optimized network model based on the network structure of model 3 using kmeans++ clustering to reselect the anchor box, which is the YOLOv4-L lightweight vehicle detection model finally improved in this paper.

Table 3

Models of Ablation experiments

Model	GhostNet	Introduce depth separable convolution	kmeans++
model 1	×	×	×
model 2	√	×	×
model 3	√	√	×
model 4	√	√	√

Table 4

Different flow ablation experiment results

Model	AP%(car)	AP%(truck)	AP%(bus)	AP%(van)	mAP%	FPS
Free flow						
model 1	99.14	94.79	94.78	96.43	96.28	36.37
model 2	98.85	87.13	91.64	95.35	93.24	42.60
model 3	98.99	84.69	96.56	91.16	92.85	52.73
model 4	99.03	90.63	96.99	92.13	94.70	53.76
Synchronous flow						
model 1	98.97	89.76	97.95	99.76	96.11	35.42
model 2	98.91	89.48	96.57	95.15	95.03	42.38
model 3	98.08	90.27	94.41	92.89	93.91	52.66
model 4	98.10	88.14	96.32	96.18	94.69	52.83
Blocking flow						
model 1	99.26	92.58	97.59	94.98	96.10	33.34
model 2	99.15	91.60	97.74	89.28	94.44	41.35
model 3	99.13	89.38	98.12	87.97	93.64	51.62
model 4	99.07	90.86	96.55	93.71	95.05	52.03

The experimental results of the ablation experiment in different traffic flow environments are shown in Table 4, the performance of the four models was evaluated in terms of average accuracy (AP), mean average accuracy (mAP) and frames per second (FPS) recognizable by the models under three different traffic flow environments, respectively.

Comparing the performance of model 1 and model 2 in terms of FPS, it can be seen that under three different traffic flow environments, the FPS of model 2 improves by 17.13%, 19.65% and 24.01%, respectively, indicating that the improvement in detection speed is obvious and can meet the real-time requirements well. Observing the mAP, it can be seen that the mAP accuracy of model 2 has been reduced, decreasing by 3.04%, 1.08% and 1.66% under the free flow, synchronous flow and blocked flow conditions, respectively. The small sacrifice in accuracy brings a great improvement in detection speed, which is acceptable in the application scenario of road vehicle detection and inspection. Theoretically, the backbone feature extraction network CSPDarkNet53 in YOLOv4 contains 53 layers of convolutional networks, and the model memory consumption is up to 244 MB, which is a huge amount of computation. After replacing it with GhostNet, the model occupies 151.40 MB, so the detection rate is improved. Due to the introduction of the Squeezing-and-Excitation attention mechanisms in GhostNet, the accuracy of the replaced model is not affected too much. In conclusion, model 2 sacrifices a small amount of detection accuracy, but greatly improves the detection speed.

Comparing the performance of model 2 and model 3 in terms of FPS, it can be seen that model 3 improves FPS by 23.78%, 24.26% and 24.84% in three different traffic flow environments, and FPS reaches more than 52, which meets the requirement of real-time performance; observing the mAP, it can be seen that the mAP of model 2 decreases from 94.44% to 93.64% in the obstructed flow traffic environment, which decreases Theoretically, the use of depth-separable convolution instead of general convolution reduces the memory consumption of the model from 152 MB to 44 MB, so the detection efficiency is further improved. However, the probability of missed or false detection in the prediction frame of model 2 has increased, resulting in a small decrease in the accuracy of model 2 in all three traffic flow environments.

Comparing Model 3 and Model 4, it can be seen that after recalculating the anchor dimensions, the FPS values of Model 4 in the three traffic flow environments are almost the same as those of Model 3, indicating that the detection speed is not affected. By observing the mAP, the overall average detection accuracy in model 4 is improved by 1.85%, 0.78%, and 1.41%, respectively. Theoretically, this is because the anchor frame size is more suitable for the vehicle dataset constructed in this paper after optimizing the anchor frame size. Therefore, compared with model 3, the detection accuracy of model 4 for all four vehicle types was improved to some extent.

In summary, based on both network model size and detection accuracy, Model 4 meets the performance requirements for real-time detection of vehicle detection tasks while taking into account detection accuracy, and has the best overall performance among all optimized models.

4.1. Analysis of Model Verification Results

At present, there are a variety of current deep learning models. In order to further verify the efficiency of the model in this paper, it is trained with the popular YOLOv3, YOLOv3-tiny and YOLOv4 models in the self-built data set of this paper, respectively. In the experiment, mAP (mean average precision) and FPS are used as indicators to measure the detection accuracy. The results of detecting different traffic density are intuitively compared.

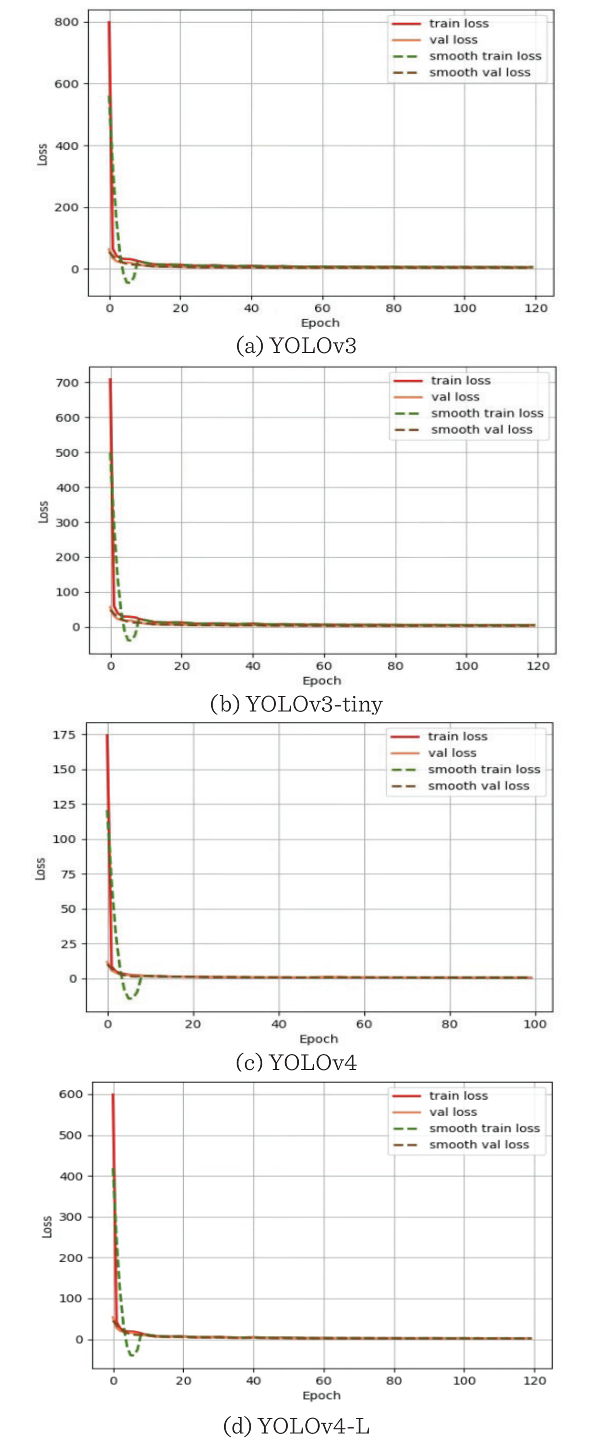
4.4.1 Loss Curve Analysis

In Figure 7, the subgraph (a), (b), (c), and (d) are the loss curves of the YOLOv3, YOLOv3-tiny, YOLOv4 and YOLOv4-L under the blocking flow. The four curves in the figure show the trend of the training set loss value, the validation set loss value and the training set loss value and validation set loss value after the label smoothing operation during the training process, respectively.

At the initial learning rate, The maximum loss value of YOLOv3 has reached more than 800 at the beginning of the training, which is the highest among all models. The maximum loss values of the other models are below 700. The validation set losses are all in the range of 10-50. Among the training set loss curves after the smoothing operation, YOLOv4 has only 118 loss values at the beginning of training, which is the

Figure 7

Loss curve of blocking flow



(a) YOLOv3

(b) YOLOv3-tiny

(c) YOLOv4

(d) YOLOv4-L

lowest among all the models, and the remaining three models are all between 400 and 600. In terms of convergence speed, all four models converge to close to 0 after about 20 epochs. As can be seen in Figure 7, the loss curves of the remaining three models do not differ much except for the YOLOv4 model. During the experiment, we found that the difference convergence speed in the early stage has little effect on the recognition effect in the later stage.

As the results of the loss curve show, firstly, the combination of the more lightweight Ghostnet feature extraction network and the introduction of depth-separable convolution instead of normal convolution in the network can reduce the number of parameters and inference time of the model more effectively. Secondly, the use of learning rate cosine annealing decay in the training process of the network can effectively jump out of the local optimal solution, which enables the network model to converge in fewer epochs.

4.4.2. Analysis of Experimental Results Under Different Flow Environments

1 *Analysis of detection results in free flow traffic environment*

The detection results of free flow show that the four models perform well in different vehicle categories, and the AP values all reach more than 84%. This is due to the fact that vehicles in the same frame in a free-flowing traffic environment have fewer numbers, less density, and do not block each other, so detection is less difficult and it is easier to obtain better recognition results.

Compared with YOLOv3 and YOLOv3-tiny, YOLOv4-L has significantly higher detection speed and detection accuracy than the above models. Compared with the

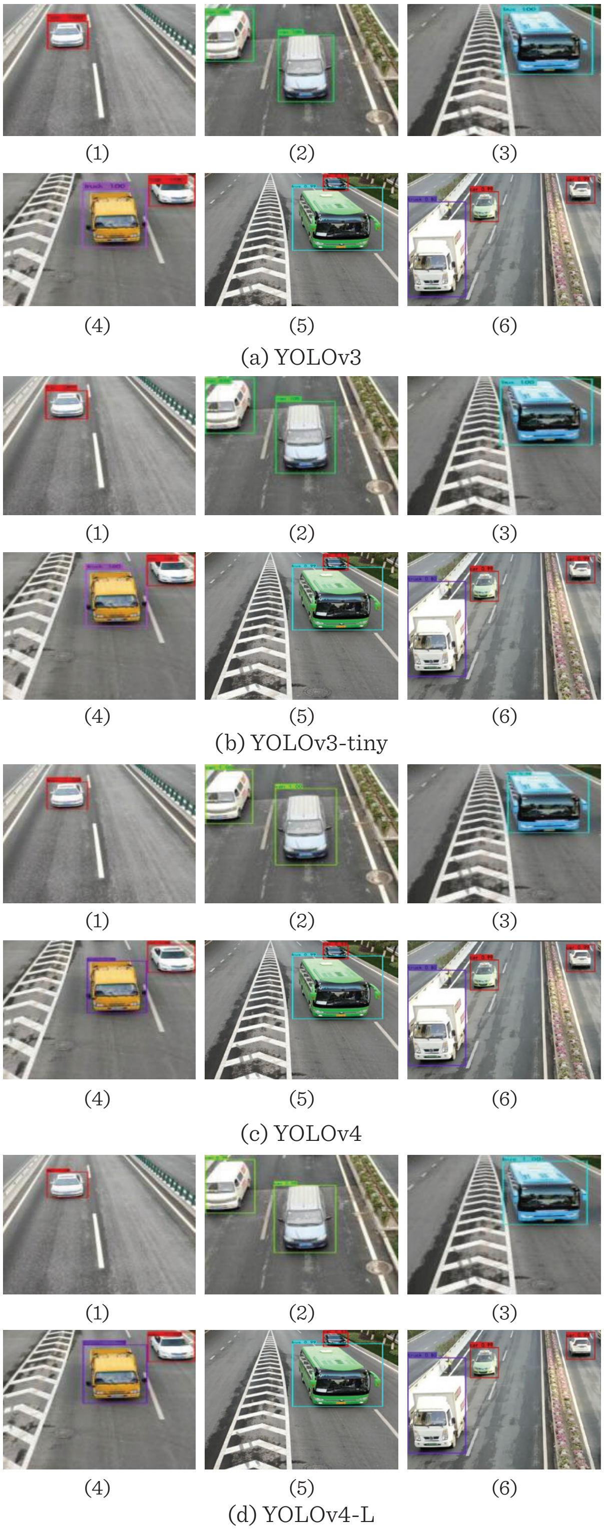
Table 5

Detection results of different model under free folw

evaluating indicator	YOLOv3	YOLOv3-tiny	YOLOv4	YOLOv4-L
AP% (car)	97.80	94.40	99.14	99.03
AP% (truck)	86.83	84.29	94.79	90.63
AP% (bus)	90.04	88.71	94.78	96.99
AP% (van)	92.61	87.76	96.43	92.13
mAP/%	91.82	88.79	96.28	94.70
FPS	32.05	42.92	36.37	53.76

Figure 8

Test results for training samples with free flow



(a) YOLOv3

(b) YOLOv3-tiny

(c) YOLOv4

(d) YOLOv4-L

YOLOv4 model, the mAP of the YOLOv4-L model lost 1.58%, but the FPS increased by 17.39%, which lost a smaller detection accuracy and achieved a faster detection speed.

The detection results of each model under the condition of free flow training samples can be seen from Figure 8(a)-(d): because the vehicle density in the free flow traffic environment is small, the number of vehicles in each frame is small, the vehicle spacing is large, it is not easy for vehicles to block each other, and the detection difficulty is relatively small, the four models can effectively detect all vehicles in the video frame, The identification of four different models is also more accurate.

2 *Analysis of detection results in synchronous traffic environment*

From Table 6 four models are more accurate in identifying cars, and the AP values are all above 93%, while the truck classification efficiency is significantly lower than other vehicle classifications. We speculate that this phenomenon is due to the diversity of truck categories and the relatively small number of truck data samples. Compared with the YOLOv3 and YOLOv3-tiny models, the YOLOv4-L model is superior to the above two models in terms of detection accuracy and detection speed. The accuracy increased by 4.57% and 6.95%, respectively, and speed increased by 66.40% and 26.18%, respectively. compared with the YOLOv4 model, the YOLOv4-L model has increased by 49.15% increase in speed, and the average accuracy of all categories is only 1.42% lower than that of YOLOv4.

From Figure 9, in the subgraph (a), the YOLOv3 model does not detect the tiny objects in (2) and (4) in the distance. In subgraph (b) and (c), the YOLOv3-tiny

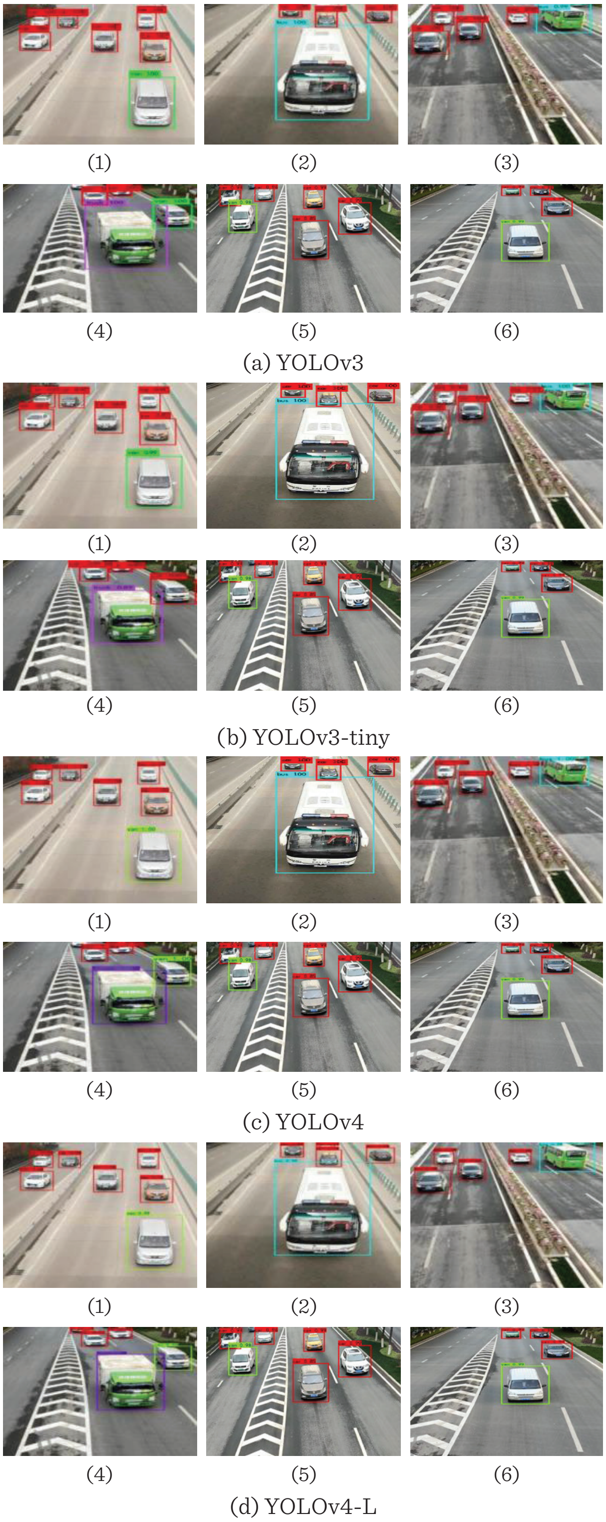
Table 6

Detection results of different model under Synchronous flow

evaluating indicator	YOLOv3	YOLOv3-tiny	YOLOv4	YOLOv4-L
AP% (car)	96.97	93.58	98.97	98.10
AP% (truck)	86.06	84.30	89.76	88.14
AP% (bus)	90.95	87.64	97.95	96.32
AP% (van)	86.50	85.73	97.76	96.18
mAP/%	90.12	87.74	96.11	94.69
FPS	31.75	41.87	35.42	52.83

Figure 9

Test results for training samples with synchronous flow



(1) (2) (3)

(4) (5) (6)

(a) YOLOv3

(1) (2) (3)

(4) (5) (6)

(b) YOLOv3-tiny

(1) (2) (3)

(4) (5) (6)

(c) YOLOv4

(1) (2) (3)

(4) (5) (6)

(d) YOLOv4-L

model and YOLOv4 model have the same missed detection as the YOLOv3 model, and there is also a false detection in subgraph (b), which recognized “van” as “car”. In subgraph (d), it eliminates the miss detection phenomenon and detects the “van” objects in the distance correctly. It can be seen that our proposed model can also achieve better results in the detection of vehicle objects in a synchronous traffic environment.

3 *Analysis of detection results in congested traffic environment*

Due to the high density of vehicle flow and low vehicle speed in the traffic environment of blocking flow, the vehicle spacing is small, and it is easy for larger vehicles such as buses or trucks are likely to block smaller vehicles such as cars or vans, thus affecting the vehicle detection effect. From Table 7, the detection accuracy of the four models trained with blocked flow samples is slightly lower than that of free flow and synchronous flow. However, under the obstructed flow traffic environment, the traffic flow is larger, and the number of vehicles of each category can be collected more, which can effectively solve the situation that the number of truck samples is insufficient, so the recognition rate for trucks is improved, and the recognition rates of YOLOv4 and YOLOv4-L models for trucks reach more than 90%.

Compared with the YOLOv3 and YOLOv3-tiny models, the YOLOv4-L model has increased accuracy by 5.32% and 7.50%, and speed increased by 71.15% and 28.41%,table respectively. Compared with the YOLOv4 model, the YOLOv4-L model has achieved 56.06% improvement in speed, the average accuracy rate of all categories is only 1.05% lower than that of YOLOv4.

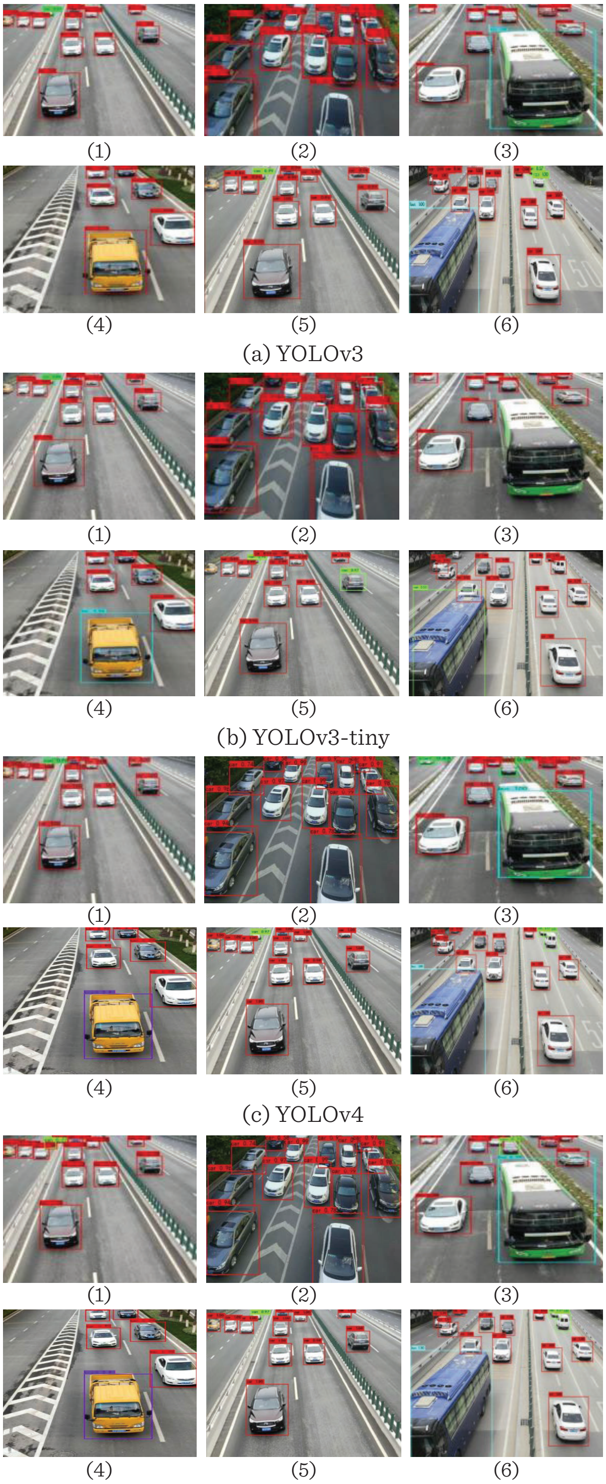
Table 7

Detection results of different model under blocking flow

evaluating indicator	YOLOv3	YOLOv3-tiny	YOLOv4	YOLOv4-L
AP% (car)	96.99	94.13	99.26	99.07
AP% (truck)	85.92	84.24	92.58	90.86
AP% (bus)	90.02	89.73	97.59	96.55
AP% (van)	86.33	84.76	94.98	93.71
mAP/%	89.82	88.22	96.10	95.05
FPS	30.40	40.52	33.34	52.03

Figure 10

Test results for training samples with blocking flow



(a) YOLOv3

(b) YOLOv3-tiny

(c) YOLOv4

(d) YOLOv4-L

From Figure 10, in subgraph (a), there are heavy re-inspection phenomenon in the lower left and upper right corners of Figure (2), and there is a false detection in (4), which identifies “truck” as “car”, and a missed detection in the upper left corner of (5). In subfigure (b), the YOLOv3-tiny model shows the same re-detection phenomenon as the YOLOv3 model, and there is a missed detection of the large proximal target in (3), and a false detection in (4), which identifies “truck” as “bus” in (5), a missed detection in the upper left corner and a misdetection of “car” as “van” in (6), a missed detection in the lower left corner and a misdetection of “bus” as “van” in the lower left corner in (7). In the lower left corner of (6) and “bus” is mistakenly detected as “van” in the lower left corner. In subgraph (c) and (d), the YOLOv4 model and the YOLOv4-L model eliminate the reinspection and false detection in the above-mentioned YOLOv3 and YOLOv3-tiny models, and the recognition results of the four vehicle types are YOLOv4-L It has similar test results with YOLOv4, but the model size and test speed of YOLOv4-L are better than YOLOv4.

5. Conclusion

The YOLOv4-L aim at providing a multi-moving target recognition method, which has advantages of accuracy and real-time. In this paper we have the following contributions as: (1) A lightweight YOLOv4-L multi-vehicle target recognition model with optimized backbone feature extraction network and PANet and detection head part of ordinary convolutional blocks is proposed to address the disadvantages of large size, high computational complexity and low operational efficiency of YOLOv4 target recognition model. An improved GhostNet is used to replace the original CSPDarknet53 feature extraction network, which reduces the number of parameters of the model and improves the detection accuracy. The introduction of depth-separable convolution blocks is proposed to replace the ordinary convolution used in the network, which further reduces the number of parameters of the model and decreases the computational complexity, and the K-means++ clustering algorithm is used to optimize the prior frame parameters and improve the model detection accuracy. Finally, Mosaic data enhancement, cosine annealing and label smoothing are combined to enhance the convergence effect of the model and improve the generalization ability of the model. (2) A lightweight improvement model YOLOv4-L of model YOLOv4 is proposed. The YOLOv4-L model is validated separately for traffic density features in different traffic environments, and the experimental results show that the average accuracy reaches more than 90% and the FPS reaches more than 52 in all three different traffic densities, which indicates that the YOLOv4-L multi-vehicle target recognition model has good applicability in three major traffic scenarios. (3) Comparing the classical models YOLOv3, YOLOv3-tiny and YOLOv4, the training set with different traffic densities and the test set with different traffic densities are tested and analyzed. The detection accuracy and detection speed of the improved model are better than those of YOLOv3 and YOLOv3-tiny models, and the mean value of the lost detection accuracy is significantly smaller than that of the YOLOv4 model YOLOv4-L under the condition of obtaining higher detection speed, which is suitable for multi-target detection with different traffic densities and can obtain better car recognition under different lighting conditions. As shown in the experimental results, the detection accuracy and detection speed of the improved model are better than those of YOLOv3 and YOLOv3-tiny models under different traffic flow densities. Compared with YOLOv3, the minimum improvement of YOLOv4-L in mAP and FPS values are 2.88% and 21.08, respectively. compared with YOLOv3-tiny, the minimum improvement of YOLOv4-L in mAP and FPS values are 5.91% and 10.84, respectively. Compared with the YOLOv4 model, the average detection accuracy loss of YOLOv4-L is smaller when higher detection speeds are obtained, with 1.58%, 1.42%, and 1.05% accuracy loss for three different traffic flow environments, respectively.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61971339); Shaanxi Natural Science basic research project (No. 2022JM407); Key Research and Development Program of Shaanxi (No.2022GY-276); Training Program for University Students' Innovation and Entrepreneurship (No.S202110709142); The Youth Innovation Team of Shaanxi Universities; Innovation Capability Support Program of Shaanxi(Program No.2021TD-29

References

1. Abdul, R., Ahim, K., Salam, R. A. Traffic Surveillance: A Review of Vision Based Vehicle Detection, Recognition and Tracking. International Journal of Applied Engineering Research, 2016, 11(1), 713-726.
2. Arthur, D., Vassilvitskii, S. K-Means++: the Advantages of Careful Seeding. Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms. Stroudsburg, PA: Association for Computational Linguistics, 2007, 1027-1035.
3. Bochkovskiy A, Wang C-Y, Liao H-Y M. Yolov4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934[2020-4-23].
4. Choi, J., Chun, D., Kim, H., et al. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving[C]. International Conference on Computer Vision (ICCV), 2019, 502-511. https://doi.org/10.1109/ICCV.2019.00059
5. Chollet, F. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 1251-1258.
6. Girshick, R., Donahue, J., Darrell, T., et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81
7. Girshick, R. Fast R-CNN. Proceeding of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169
8. Hua, X., Wang, X. Q., Wang D, et al. Multi-objective Detection of Traffic Scenes Based on Improved SSD. Journal of the Optical, 2018, 38(12), 221-231. https://doi.org/10.3788/AOS201838.1215003
9. Hu, X., Liu, Y., Zhao, Z., et al. Real-time Detection of Uneaten Feed Pellets in Underwater Images for Aquaculture Using an Improved YOLO-V4 Network. Computers and Electronics in Agriculture, 2021, 185: 106135. https://doi.org/10.1016/j.compag.2021.106135

10. Han, K., Wang, Y. H., Tian, Q., et al. GhostNet: More Features from Cheap Operations. http:// arXiv:1911.11907v2, 13, Mar, 2020. https://doi.org/10.1109/CVPR42600.2020.00165

11. Li, X., Xu, X., Li, J. Small Target Detection in Remote Sensing Images for Aviation Flight Safety. Aviation Weapons, 2020, 27(3), 54-61.

12. Liu, W., Anguelov, D., Erhan, D., et al. SSD: Single Shot Multibox Detector. Proceeding of the 2016 European Conference on Computer Vision, LNCS 9905. Cham: Springer, 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

13. Li, X., Liu, Y., Zhao, Z. F., et al. A Deep Learning Approach of Vehicle Multitarget Detection from Traffic Video. Journal of Advanced Transportation, 2018, 1-11. https://doi.org/10.1155/2018/7075814

14. Li. X., Zhao, Z. F., Liu, L., et al. An Optimization Model of Multi-intersection Signal Control for Trunk Road Under Collaborative Information. Journal of Control Science and Engineering, 2017, 2017, 1-11. https://doi.org/10.1155/2017/2846987

15. Ma, Y. J., Cheng, S. S., Ma, Y. T., et al. Convolutional Neural Network and Its Application in Intelligent Transportation System. Journal of Traffic and Transportation Engineering, 2021, 21(04), 48-71.

16. Ma, L. P., Yun, X., et al. Road vehicle Multi-Target Detection Method Based on Improved Yolov3 Model. Journal of Xi'an University of Engineering, 2021, 35(05), 64-73.

17. Nguyen, H. Improving Faster R-CNN Framework for Fast Vehicle Detection. Mathematical Problems in Engineering, 2019, 2019, 1-11. https://doi.org/10.1155/2019/3808064

18. Ren, S., He, K., Girshick, R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

19. Redmon, J., Divvala, S., Girshick, R., et al. You Only Look once: Unified, Real-time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

20. Redmon, J., Farhadi, A. YOLO9000: Better, Faster, Stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 7263-7271. https://doi.org/10.1109/CVPR.2017.690

21. Redmon, J., Farhadi, A. Yolov3: An Incremental Improvement. arXiv preprintarXiv: 180402767, 2018.

22. Shi, B., Li, X., Nie, T., et al. Multi-object Recognition Method Based on Improved YOLOv2 Model. Information Technology and Control, 2021, 50(1), 13-27. https://doi.org/10.5755/j01.itc.50.1.25094

23. Shao, W. P., Wang, X., Cao, Z. R., et al. Design of Lightweight Convolutional Neural Network Based on MobileNet and YOLOv3[J]. Journal of Computer Applications, 2020(S01), 8-13.

24. Sandler, M., Howard, A., Zhu, M., et al. Mobilenetv2: Invert ed Residuals and Linear Bottlenecks. CVPR 2018: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

25. Schen, S., Lin, W. Embedded System Real-Time Vehicle Detection Based on Improved YOLO Network. IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), 2019, 1400-1403. https://doi.org/10.1109/IMCEC46724.2019.8984055

26. Wang, P., Wu, J., Wang, H. Y., et al. Low-light-level Image enhancement Algorithm Based on Integrated Networks. Multimedia Systems, 2020(10), 1-11. https://doi.org/10.1007/s00530-020-00671-8

27. Wu, D., L, S., Jiang, M., et al. Using Channel Pruning-based YOLOv4 Deep Learning Algorithm for the Real-Time and Accurate Detection of Apple Flowers in Natural Environments. Computers and Electronics in Agriculture (0168-1699), 2020, 178, 105742. https://doi.org/10.1016/j.compag.2020.105742

28. Wang, W., Xie, E., Song, X., et al. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. ICC V 2019: Proceedings of the IEEE/CVF International Conference on Computer Vision.Piscataway: IEEE, 2019, 8440-8449. https://doi.org/10.1109/ICCV.2019.00853

29. Yang, W., Wang, H. Y., Zhang, J., et al. An Improved Vehicle Real-Time Detection Algorithm Based on Faster-RCNN. Journal of Nanjing University (Natural Science), 2019, 55(2), 231-237.

30. Zhu, M. T., Xing, H., Fang, R. H. Vehicle Detection Based on YOLO-TridentNet. Journal of Chongqing University of Technology (Natural Science), 2020, 34(11), 1-8.

31. Zheng, Q. H., Zhao, P. H., Zhang, D. L., et al. MR-DCAE: Manifold Regularization-based Deep Convolutional Autoencoder for Unauthorized Broadcasting Identification. International Journal of Intelligent Systems, 2021, 36(12). https://doi.org/10.1002/int.22586

32. Zheng, Q. H., Yang, M. Q., Yang, J., et al. Improvement of Generalization Ability of Deep CNN via Implicit Regularization in Two-Stage Training Process. IEEE Access, 2018, 6, 15844-15869. https://doi.org/10.1109/ACCESS.2018.2810849
33. Zheng, Q. H., Yang, M. Q., Tian, X. Y., et al. A Full Stage Data Augmentation Method in Deep Convolutional Neural Network for Natural Image Classification. Discrete Dynamics in Nature and Society, 2020, 2020. https://doi.org/10.1155/2020/4706576
34. Zheng, Q. H., Tian, X. Y., Yang, M. Q., et al. PAC-Bayesian Framework Based Drop-Path Method for 2D Discriminative Convolutional Network Pruning. Multidimensional Systems and Signal Processing, 2019, 31. https://doi.org/10.1007/s11045-019-00686-z