Research on small target detection technology on assisted driving road based on Yolov3

Qing Zhou [1, a], Gongquan Tan [2, b]

[1]Sichuan University of Light Chemical Technology, School of Automation and Information Engineering, Yibin City, China

[2]Sichuan University of Light Chemical Technology, Sichuan Provincial Key Laboratory of Artificial Intelligence, Yibin City, China

[a]10714327552@qq.com, [b] tgq77@126.com

Abstract. Aiming at the huge difference in the size of the bounding box of different types of targets on the road in natural traffic scenes. The existing original target detection algorithm yolov3 cannot well balance the detection accuracy of large and small targets, so the target detection algorithm based on yolov3 is redesigned. Firstly, the detection module is improved and designed, and a new feature output module for small targets is added to obtain a new road target detection method yolov3 with four detection scales_ T. At the same time, focal loss is added to the loss function to reduce the result error caused by the imbalance of positive and negative samples [1]. The results show that the improved yolov3_ The average detection accuracy of T algorithm on the data set mixed with bdd100k and Kitti is 0.465, which is 0.048 higher than that of the original yolov3, especially for small targets.

Keywords: Rroad target detection; BDD100K; KITTI dataset; Improved YOLOv3 model.

1. Introduction

In recent years, with the continuous increase in market demand, the number of motor vehicles in the world is also increasing year by year. With this growth is also a variety of large and small traffic accidents. At the same time, due to people's increasingly high requirements for intelligent life, many researchers have begun to vigorously study autonomous vehicles. At present, some domestic and foreign auto companies such as Google, BMW and Tesla are also gradually involved in the field of automatic driving [2]. Target detection is a very important research direction in the field of automatic driving. The main detection targets are traffic signs, lanes, obstacles, etc.; moving targets such as vehicles, pedestrians, non-motor vehicles, etc. However, its detection has the following difficulties: 1) In the process of detection, the detection effect of small targets is not good; 2) In detection, the detection speed and accuracy cannot be met at the same time. Therefore, whether the above two problems can be solved directly affects the safety performance of vehicles in automatic driving.

Based on the yolov3 network model, aiming at the different size of the target caused by the distance in the automatic driving scene, in order to balance the detection accuracy of large and small targets, this paper maximizes the target object feature extraction ability of the convolution neural network and improves the detection performance of the network in the actual target detection, In this paper, focal loss is introduced into the loss function to solve the errors caused by the unbalanced distribution of positive and negative samples and the imbalance between simple samples and difficult samples. Finally, based on the yolov3 algorithm, the detection module is improved, and a multi-scale road target detection method with four detection scales is designed. The experimental results show that the improved yolov3_ Compared with the original yolov3 algorithm, t algorithm has a certain improvement in the detection accuracy of small target objects.

2. YOLOv3 main structure

Yolov3 network is an improved network based on yolov1 and yolov2. It is an end-to-end detection algorithm. The network structure is divided into backbone network darknet-53 and detection network [3]. Firstly, backbone network darknet53 is a convolutional neural network with 53 layers. Compared

with networks resnet-152 and resnet-101, darknet-53 not only improves the classification accuracy, And its computing speed is much better than resnet-152 and resnet-101. The residual block is added to make the network structure deeper and have stronger feature extraction ability; The residual block structure is shown in Figure 1. Each residual component consists of two convolution layers and a quick link. The superimposed characteristic map is used as a new input to the next layer. The main body is composed of multiple residual modules [4], which reduces the risk of gradient explosion and strengthens learning ability.

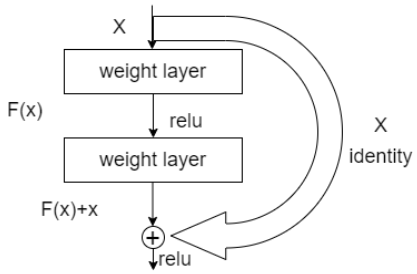

Figure 1. Residual block structure

Secondly, the Feature Pyramid (FPN) method is adopted to extract multiple feature maps of different scales for detection respectively, which improves the detection ability of the algorithm for targets of different sizes, and outputs 13 × 13, 26 × 26 and 52 × A total of 52 features of 3 scales [5] Into the detection network. Next, the detection network will have three scales of feature regression,predicts multiple prediction frames, and uses non-maximum suppression to retain the optimal candidate frame. The YOLOv3 network structure is shown in Figure 2:

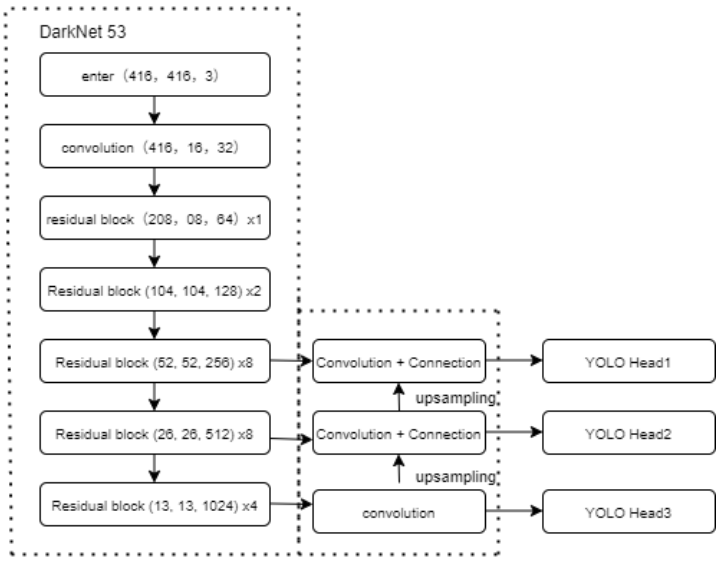

Figure 2. YOLOv3 network structure

3. Related work

3.1 Data set

The training data set used in this article is the BDD100K data set released by the AI Laboratory of the University of California, Berkeley. Bdd100k is by far the largest and most diverse public driving data set, and it is one of the most common datasets in the field of autonomous driving [6]. The BDD100K dataset contains 100,000 high-definition videos, each about 40s, with a resolution of 1280 × 720 and a frame rate of 30 frames. The key frames are sampled in the 10s of each video, and 100,000 images are obtained and annotated. Among them, 70,000 labeled images are divided into training set, and 10,000 labeled images are used as validation set. There are 10 categories of GT box labels in the

dataset, namely: Bus, Light, Sign, Person, Bike, Truck, Motor, Car, Train, Rider.There are about 3 different calibration types, as shown in the figure3.

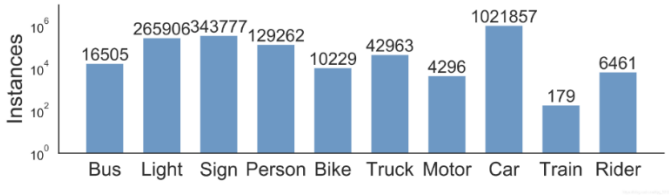

Figure 3. Number of objects in the BDD dataset

The categories of bdd100k dataset are unbalanced. There are 1021857 instances of car in the dataset, while there are only 179 instances of target train. The distribution of dataset categories is shown in Figure 3. In the case of uneven distribution between training sets, the feature extraction ability of the network will be enhanced for a large number of instance targets such as car in neural network training, but the feature extraction ability of the network will be reduced for a small number of instance targets. At the same time, the purpose of road target detection model is to accurately detect common targets in natural driving scenes. The number of pedestrians and cyclists in the data set used in this experiment is much less than that of motor vehicles, so it will lead to over fitting problem to a certain extent, and the generalization ability of the final trained detection model is poor [7]. To solve this problem, this paper adopts the method of mixed data set training in the experiment. Therefore, select 10000 labeled data sets from bdd100k and remove the labels of train, rider and Moro to form seven categories (traffic _light, traffic _sign, car, truck, bus, person and bike). At the same time, Download Kitti pedestrian data set, which contains 600 pictures of the training set and 2000 pedestrian samples. Finally, a total of 12600 images from the two data sets are used to assist the training of multi category target detection models in driving scenes at the same time. 10000 of them are divided into training sets and 2600 into verification sets.

Figure 4. Dataset example

3.2 Improved YOLOv3 detection model

According to the good effect of convolutional neural network in detecting small targets on large-scale feature maps, on the basis of the original three detection scales of YOLOv3, the feature fusion part and the detection layer of YOLO are redesigned, and a detection layer is added. The large-scale feature output map of the small target, the improved overall structure is shown in Figure 5. Compared with the original YOLOv3 network structure, in order to adapt to the scale change of the required feature map caused by the new detection layer, the feature fusion network is redesigned. , a new feature fusion module is added on the basis of the original three shallow feature fusion modules. The first three detection maps 13×13, 26×26, 52×52 are the same as the original YOLOv3 architecture, and the added detection map with a scale of 104×104 is to upsample the output 52×52 of the 108th layer of the network once, Increase the resolution to 104 × 104, and then add a feature fusion layer to splicing the feature map of the 11th layer to the channel of the 110th layer output feature map, and initially generate three different scale anchor boxes on the feature map obtained after feature fusion.and then alternately use 3×3 and 1×1 convolution operations to map to obtain tensor data under

104×104 [8]; the improved YOLOv3 network structure outputs 4-scale detection maps after feature fusion is completed, which is better than YOLOv3. An increase of 104 × 104 is added, and the added 1-scale detection map provides important features for the detection of small targets. The improved network considers the detection of small targets in the long-range more fully than YOLOv3, and does not affect the detection of large targets.

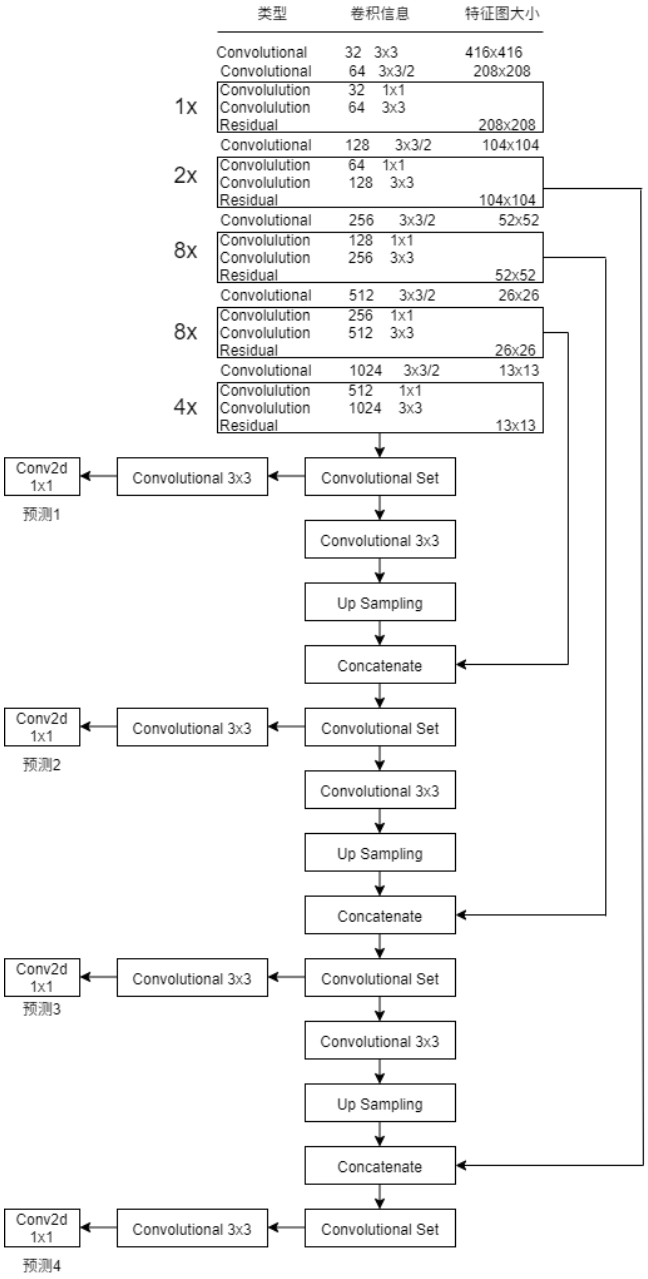

Figure 5. Improved YOLOv3 network structure

3.3 Loss function design

Loss function is a standard to measure the error between the predicted value and the real value. The speed of loss function learning in the network has a great impact on the final model detection effect.For the problem of moving target detection in the road studied in this paper, the network needs to measure three aspects during the training process: the coordinates of the target, the confidence score and the category. The loss function of the YOLOv3 network model is shown in formula (1):

$$
\begin{aligned}
Loss = {} & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} (2 - \hat{\omega}_i^j \times \hat{h}_i^j) \\
& \left[\begin{array}{l} (t_{w_i^j} - \hat{t}_{w_i^j})^2 + (t_{h_i^j} - \hat{t}_{h_i^j})^2 + \\ \left(\sigma(t_{x_i^j}) - \sigma(\hat{t}_{x_i^j})\right)^2 + \left(\sigma(t_{y_i^j}) - \sigma(\hat{t}_{y_i^j})\right)^2 \end{array} \right] \\
& - \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\hat{C}_i^j \cdot \lg\left(C_i^j\right) + \left(1 - \hat{C}_i^j\right) \cdot \lg\left(1 - C_i^j\right) \right] - \\
& \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} \left[\hat{C}_i^j \cdot \lg\left(C_i^j\right) + \left(1 - \hat{C}_i^j\right) \cdot \lg\left(1 - C_i^j\right) \right] \\
& - \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \sum_{c \in class}^{n} \left[\hat{P}_{c_i^j} \cdot \lg\left(P_{c_i^j}\right) + \left(1 - \hat{P}_{c_i^j}\right) \cdot \lg\left(1 - P_{c_i^j}\right) \right]
\end{aligned} \tag{1}
$$

In the formula, λ_{coord}, λ_{noobj} respectively represent the coordinate loss weight and the confidence loss weight that does not contain the target [9], S is the horizontal or vertical number of grid division, B is the number of Bounding Boxes that the grid has, I_{ij}^{obj} , I_{ij}^{noobj} Indicates whether the th mesh th Bounding Box is responsible for the detection of an object. $\left(t_{x_i^j},\ t_{y_i^j},\ t_{w_i^j},\ t_{h_i^j},\ C_i^j, P_{C_i^j}\right)$ Indicates the coordinates, width and height offset of the network output relative to the Anchor Box, and the prediction box confidence and category probability prediction. $\left(\hat{t}_{x_i^j},\ \hat{t}_{y_i^j},\ \hat{t}_{w_i^j},\ \hat{t}_{h_i^j},\ \hat{C}_i^j, \hat{P}_{C_i^j}\right)$ Represents the real target box coordinates, width and height, confidence and class probability prediction.

Ignore_ Thread is used in the yolov3 model and the use of small confidence in the sample box without target can solve some problems of imbalance between positive and negative samples to a certain extent [10]. But even with ignore_ Thread will still have the imbalance problem of positive and negative samples. After the introduction of focal loss, the model can better solve the imbalance problem of positive and negative samples, and because the detection effect is better when detecting difficult samples, it can improve the detection effect of the model.

Typical cross-entropy loss is widely used in current image classification and detection CNN networks, as shown in Equation (2) Among them, p $\in$ [0,1], represents the model output class probability, y is the class label, and its value is 0 or 1.

$$
CE(p, y) = \begin{cases} -\log_a p, y = 1 \\ -\log_a (1 - p), y = 0 \end{cases} \tag{2}
$$

Due to the imbalance between positive and negative samples in the data set, it can be corrected by using a coefficient inversely proportional to the target existence probability in the cross entropy loss. In this way, the weight coefficient of a small number of positive samples is larger, and its contribution to the model will also increase. The weight coefficient of a large number of negative samples is smaller, and its contribution to the model will be relatively weakened. Therefore, the model will learn more useful information. Add weight coefficient α The cross entropy loss after is as follows:

$$
CE(p, y) = \begin{cases} -\log_a p, y = 1 \\ -(1 - \alpha) \log_a (1 - p), y = 0 \end{cases} \tag{3}
$$

In addition, some categories in the sample are relatively clear, while others are more difficult to distinguish. Based on the loss of cross entropy, focal loss automatically reduces the loss of simple samples by adding a dynamic scaling factor, which helps the model focus on some samples that are more difficult to train.In the calculation of Focal loss, a new hyperparameter is introduced, and the calculation of Focal loss is shown in formula (4):

$$FL(p,y)=\begin{cases}-(1-p)^{\gamma}\log_a p, y=1\\ -p^{\gamma}\log_a(1-p), y=0\end{cases} \tag{4}$$

Combining Focal loss with weight α[12], the final calculation formula of Focal loss is:

$$FL(p,y)=\begin{cases}-\alpha(1-p)^{\gamma}\log_a p, y=1\\ -(1-\alpha)p^{\gamma}\log_a(1-p), y=0\end{cases} \tag{5}$$

The use of Ignore_thread in the YOLOv3 model and the use of smaller confidence levels for sample frames without targets can solve the problem of positive and negative sample imbalance to a certain extent. However, the problem of positive and negative sample imbalance still exists. After the introduction of Focal loss, the model can better solve the problem of positive and negative sample imbalance, and because the detection effect of difficult samples is better, the model detection effect can be improved. After introducing Focal loss, the loss function of the YOLOv3 model is shown in Equation (6):

$$\begin{aligned}&-\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{obj}\left[\begin{array}{l}\alpha(1-C_i^j)^{\gamma}\hat{C}_i^j\cdot\lg\left(C_i^j\right)+\\ (1-\alpha)(C_i^j)^{\gamma}\left(1-\hat{C}_i^j\right)\cdot\lg\left(1-C_i^j\right)\end{array}\right]\\ &-\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{noobj}\left[\begin{array}{l}\alpha(1-C_i^j)^{\gamma}\hat{C}_i^j\cdot\lg\left(C_i^j\right)+\\ (1-\alpha)(C_i^j)^{\gamma}\left(1-\hat{C}_i^j\right)\cdot\lg\left(1-C_i^j\right)\end{array}\right]\end{aligned} \tag{6}$$

4. Experiment and result analysis

4.1 Experimental environment and evaluation index

The experimental environment configuration of this paper is shown in Table 1.

Table 1. Lab Environment Configuration

device name	Device Information
CPU	Intel Xeon E5-2620 v3
GPU	NvidiaGe Force GTX TITAN X
operating system	Ubuntu 16.04.2
Python version	3.7.9
CUDA version	10.0
CUDNN version	7.6.5
Pytorch version	1.7.1

The network parameter configuration during the experiment is as follows: the momentum is 0.9, the weight attenuation is 0.0005, the number of iterations is set to 50000, the learning rate uses the step-by-step strategy [11], the initial value is set to 0.001, the number of changes is 30000 and 40000, and the ratio is 0.15 and 0.1.

In the scene of assisted driving, the vehicle has high requirements for the speed and accuracy of target detection. In the detection process, the corresponding category targets are detected into other categories, or the detection speed is too slow, resulting in the vehicle's too late to respond, which will cause very serious consequences. Therefore, on the issue of evaluating the detection effect, the average accuracy rate mean map and the number of frames detected per second FPs are used as the evaluation indexes of the road target detection model in the process of automatic driving. Precision and recall rates are defined as follows:

$$\Pr ecision(classes)=\frac{TP}{TP+FP} \tag{7}$$

$$\mathrm{Re}\,call(classes)=\frac{TP}{TP+FN} \tag{8}$$

$$mAP(\%) = \frac{\sum AP}{NC} \times 100\% \tag{9}$$

Take the bus category in the road target studied in this paper as an example. Firstly, TP refers to the number of buses recognized as buses by the detection model, FP refers to the number of pedestrians or cyclists recognized as buses, and FN refers to the number of pedestrians or cyclists recognized as buses.

AP Represents the sum of the average accuracy of single category pictures [12]. NC is the total number of categories.

4.2 Analysis of results

The optimal weight of the unmodified yolov3 road target detection algorithm trained on the mixed data set of BDD and Kitti is compared with the improved yolov3_ According to the optimal weight [12] trained by T, 7 target types and the overall detection average accuracy are tested on the data set, as shown in Table 2.

Table 2. Model Detection Accuracy Comparison

class name	mAP@0.5	
	YOLOv3	***YOLOv3_T***
Car	0.524	0.613
Person	0.481	0.542
Bus	0.497	0.509
Traffic_light	0.419	0.480
Traffic_sign	0.402	0.456
Bike	0.431	0.471
Truck	0.503	0.522
mean	0.465	0.513

It can be seen from Table 2: Compared with the maximum mAP@0.5 value obtained by the road target detection algorithm based on YOLOv3 in the test set, the maximum mAP@0.5 value of the road target detection algorithm based on YOLOv3_T is 0.513 on the test set, and the detection accuracy is increased. 0.048. From the analysis of the changes in mAP@ 0.5 of each category before and after the improvement of the detection model, it can be seen that the detection accuracy of YOLOv3_T for small targets has been greatly improved. The typical small targets in [13], achieved 0.54 and 0.061 accuracy improvements, respectively. At the same time, YOLOv3_T also achieved better performance for detection targets such as cars (Car) and pedestrians (Person). Accuracy improvement of 0.064. For targets with more difficult side detection such as Bus and Truck, the detection accuracy of YOLOv3_T is also improved by 0.012 and 0.019 respectively, which means that YOLOv3_T can fully improve the detection accuracy of small targets while taking into account common Detection accuracy of large objects [14].

On the BDD100K test set with annotation information, two pictures in typical traffic scenarios are selected, and the optimal weights of the two networks are used for detection, and the detection results are compared with the actual annotation information. The comparison of the two detection network effects in the same scene is shown in Figure 6.

(a) YOLOv3 detection results (b) YOLOv3_T detection results

Figure 6. Comparison of two detection network effects in the same scene

As can be seen from Figure 6: (a) The YOLOv3 detection network failed to detect the Traffic_sign traffic sign in the upper right corner of the picture, (b) The YOLOv3_T detection network detection result was 0.60, and (a) The interior of the car was detected as The car class, which shows that the original YOLOv3 network is inaccurate for small target object detection, has been improved in YOLOv3_T. In addition, for large targets and close-range targets, for example, in (a), the detection accuracy of bus (Bus) is improved from 0.74 to 0.86, and the accuracy of car (car) is improved from 0.89 to 0.94. It shows that YOLOv3_T can fully improve the detection accuracy of small targets while taking into account the detection accuracy of common large targets.

Table 3 shows the comparison of the detection performance between the YOLOv3_T detection network and the current mainstream detection algorithms. The test data set used is the mixed test set of BDD100K and KITTI, with a total of 10,000 pictures.

Table 3. Performance comparison between YOLOv3_T and each detection algorithm on the mixed test set

network structure	mAP@0.5	Speed/fps
Fast R-CNN	0.457	10.7
Faster R-CNN	0.466	18.5
SSD	0.456	29.8
YOLOv2	0.423	24.5
YOLOv3	0.495	37.3
YOLOv3_T	0.513	35.1

It can be seen from the values in Table 3 that the average accuracy of the two-stage target detection algorithms fast r-cnn and fast r-cnn are 45.7% and 48.6% respectively, but the detection speed is low, which is mainly due to the fact that the candidate area generation network (RPN) generates the candidate box containing the object to be tested during the detection process of the two-stage target detection algorithm, which will increase the calculation time of the network model. Using single-stage target detection algorithms yolov2 and yolov3 to detect targets [15], although the detection speed has been greatly improved, the average accuracy has decreased. Compared with the two-stage target detection algorithms fast r-cnn and fast r-cnn, the single-stage target detection algorithm SSD has a great improvement in average accuracy and detection speed, but it has a certain gap with the original yolov3 algorithm in average accuracy and detection speed [16]. This paper improves yolov3_T algorithm and the original yolov3 algorithm have further improved the average accuracy of target detection, reaching 51.3%.

5. Concluding remarks

This paper improves the loss function of YOLOv3 and adds a detection layer. The improved YOLOv3 with increased feature scale was applied to the detection and recognition of road targets, and achieved good results. The experimental results show that the detection model obtained by the above method can ensure high detection accuracy for moving objects of different sizes in different traffic scenarios. However, due to the addition of a detection layer to this model, the detection speed is reduced. For automatic driving scenarios, both detection accuracy and detection speed need to be included in the evaluation criteria. In the future, solving the above problems will be the focus of research.

Acknowledgments

The Project of Sichuan provincial science and Technology Department (Grant No 2020YFG0178).

References

[1] Chen Fei, Liu Yunpeng, Li Siyuan. A Review of Traffic Sign Detection and Recognition Methods in Complex Environments [J/OL]. Computer Engineering and Applications, 2021, 57(16):65- 73 [2021-06-22]. http:/ /kns.cnki.net/kcms/detail/11.2127.TP. 20210527.1455.011.html. CHEN Fei, LIU Yunpeng, LI Siyuan. Survey of traffic sign detection and recognition methods in complex environment [J/OL]. Computer Engineering and Applications, 2021, 57(16):65-73 [2021-06-22].

[2] Dong Tiantian, Cao Haixiao, Kan Xi, et al. Research on multi-target recognition method in traffic scene under complex weather [J]. Information and Communication, 2020(11):72-74. DONG Tiantian, CAO Haixiao, KAN Xi, et al. Multi target recognition method of traffic scene in complex weather[J]. Information & Communications, 2020(11):72-74.

[3] Zhao Kun, Liu Li, Meng Yu, et al. Traffic sign detection and recognition under low light conditions [J]. Journal of Engineering Science, 2020, 42(8):1074-1084. ZHAO Kun, LIU Li, MENG Yu, et al . Traffic signs detection and recognition under low-illumination conditions[J]. Chinese Journal of Engineering, 2020, 42(8):1074-1084.

[4] MOGELMOSE A, TRIVEDI M M, MOESLUND T B. Vision based traffic sign detection and analysis for intelligent driver assistance systems: perspectives and survey[J]. IEEE transactions on intelligent transportation systems, 2012, 13(4): 1484-1497.

[5] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition[J]. Neural networks, 2012, 32:323- 332.

[6] Zhang Zhongwen, Gao Yu, Wang Jing, et al. A deep learning traffic sign recognition system based on YOLOv3 [J]. Building Electric, 2020, 39(7):64-68. ZHANG Zhongwen, GAO Yu, WANG Jing, et al. Deep learning traffic sign recognition system based on YOLOv3[J]. Building Electricity, 2020, 39(7):64-68.

[7] Zhang Zhongwen, Gao Yu, Wang Jing, et al. A deep learning traffic sign recognition system based on YOLOv3 [J]. Building Electric, 2020, 39(7):64-68. ZHANG Zhongwen, GAO Yu, WANG Jing, et al. Deep learning traffic sign recognition system based on YOLOv3[J]. Building Electricity, 2020, 39(7):64-68.

[8] DU Xinlei. Research on traffic sign recognition based on improved YOLOv3 network in natural environment[D]. Dalian: Dalian Maritime University, 2020. DU Xinlei. Research on traffic sign recognition based on improved YOLOv3 network in natural environment[D]. Da lian: Dalian Maritime University, 2020

[9] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 2117-2125.

[10] EEE conference on computer vision and pattern recognition. 2018. p. 7132-7141.

[11] Li Y T, Kuo P, Guo J N. Automatic industry PCB board DIP process defect detection with deep ensemble method[C]// 2020 IEEE 29th International Symposium on Industrial Electronics (ISIE). Delft, Netherlands: IEEE, 2020: 453-459.

[12] Redmon J, Farhadi A. YOLOv3: An incremental improvement[C]//IEEE conference on Computer Vision and Pattern Recognition, 2018, arXiv: 1804.0276.

[13] Zhang Zhen, Li Haofang, Li Mengzhou. Research on YOLO algorithm in abnormal images of security inspection [J]. Computer Engineering and Applications, 2020, 56(21): 187-193.

[14] Mao Q C, Sun H M, Liu Y B, et al. Mini-YOLOv3: real-time object detector for embedded applications[J]. Ieee Access, 2019, 7: 133529-133538.

[15] Yi Z, Yongliang S, Jun Z. An improved tiny-yolov3 pedestrian detection algorithm[J]. Optik, 2019, 183: 17-23.

[16] Huang Y Q, Zheng J C, Sun S D, et al. Optimized YOLOv3 algorithm and its application in traffic flow detections[J]. Applied Sciences, 2020, 10(9): 3079.