IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Research on Small Target Detection in Driving Scenarios Based on Improved Yolo Network

Qiwei Xu[1], Runzi Lin[1], Han Yue[2], Hong Huang[1], Yun Yang[1], ZhigangYao[1]

[1] State Key Laboratory of Power Transmission Equipment& System Security and New Technology, Chongqing University, Chongqing 400044, China; linrunzi@cqu.edu.cn (R.L.)

[2] Northeast Branch of State Grid Corporation of China, Shenyang 110180, China

Corresponding author: Qiwei Xu (xuqw@cqu.edu.cn).

This work was supported by the Chongqing Science and Technology Commission of China under Project No. cstc2018jcyjA3148； graduate scientific research and innovation foundation of Chongqing, China under Grant No. CYS1807；graduate research and innovation foundation of Chongqing, China under Grant No. CYB18009.

ABSTRACT The obtainment of road condition information during driving is extremely important for a driver. However, drivers usually cannot notice multiple information at the same time, which definitely increases certain safety risks. Considering this problem, this paper designs a road information collection plus alarm system based on artificial intelligence to monitor road information. The underlying core algorithm of this system adopts the YOLO v3 network with the best comprehensive detection performance in the end-to-end network. We use this network's advantage of fast detection speed to optimize on its original basis, and propose to "copy" part of the backbone network to build an auxiliary network, which enhances its feature extraction capability. Further, we apply the attention mechanism to the feature information fusion of the auxiliary network and the backbone network, suppress the invalid information channel, and improve the network processing efficiency. Besides, the training part of the network is optimized, and the mAP (mean Average Precision) is improved by setting the scale that meets the target to be detected. Through the test, the average test accuracy of the optimized network model reaches 84.76%, and the real-time detection speed on the 2080Ti reaches 41FPS. Compared with the previous network, the detection accuracy increases by 5.43% after optimization.

INDEX TERMS Convolutional neural network, residual network, target Detection, YOLO v3

I. INTRODUCTION

In recent years, road condition information recognition technology plays an important role in advanced driving assistance and automatic driving of unmanned vehicles as an important part of intelligent driving system. It is the key and foundation for the research and application of intelligent driving systems [1]. The road condition information intelligent identification system collects road condition pictures, identifies the interested area such as lane lines and traffic signs, and transmits the information to the driver in an appropriate manner. It can be used not only to assist the normal driver to drive the vehicle, but also to effectively reduce the occurrence of traffic accidents. It also increases the possibility of driving on the road for people with visual impairments, providing the necessary conditions for the realization of a comprehensive unmanned vehicle system and the establishment of a smart city system [2].

People has attached greater importance to road traffic safety these days. In the field of intelligent transportation, scholars from various countries have achieved some outstanding research results. Literature [3] proposes VG-RAM wavelet neural network to identify the benchmark of German traffic signs. Literature [4] proposes an intelligent traffic identification method combines with digital map, and integrates GPS with vehicle odometer to develop a matching algorithm based on the particle filter, which improves the recognition performance at night. However,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

necessary to keep the vehicle running at a low speed to ensure the accuracy of recognition. Literature [5] proposes a feedforward back propagation neural network technology which extends supervision to realize traffic identification. This method not only gives an appropriate combination of the number of hidden layers and the number of input nodes, but also trains the network to the best state in a short period of time. However, this literature does not further explain the recognition accuracy of the method. Literature [6] adopts the geodesic transformation (GDT) method to generate the distance map based on superpixel, and the classification performance is significantly improved by incorporating the shape feature. However, the authors only tested on one data set and did not test the real-time property of the identification. In the computer vision field, target detection of images is the earliest application direction. From the early traditional detection algorithms to today's deep learning algorithms, the categories of detection are becoming more and more diverse. From the initial pedestrian detection to the complex workpiece defect detection, the detection speed and accuracy are getting higher and higher. As for the combination of deep learning and computer vision technology, after Alex Krizhevsky won the championship [7] with the AlexNet neural network model in the 2012 ILSVRC competition, many researchers have invested in it, and many excellent network models have also been designed. Literature [8] applies the YOLO algorithm to the detection of pedestrians, which improves the positioning accuracy. However, the detection method only uses the static picture information of the pedestrian as the detection basis, so the detection method has a great limitation on the detection of dynamic pedestrian, and the detection result on the pedestrian data set is not so ideal. Literature [9] proposes a scene image vehicle target discovery method based on fast region convolutional neural network (Fast RCNN). The shortcoming of this method is that its construction requires a large number of valid samples, and the preliminary sample candidate region extraction process is time consuming. Literature [10] proposes a region-based convolutional neural network (RCNN), based on the extracted candidate regions, the convolutional neural network is used to extract the interested region in the image that may contain vehicle features to detect and identify the target. However, the calculation is relatively complicated, and it is difficult to meet the need of real-time detection. Applying deep learning and computer vision to assisted driving systems and automatic driving systems, and combine them with multi-source data and liner model [11-13], to conduct reasonable path planning through data processing [14,15], is undoubtedly the main development direction of the future car.

When performing target detection in driving scenarios, the background is complex, the targets to be measured are densely distributed or overlapping, and the viewing distance of the camera is not fixed, resulting in different sizes of the targets. Especially small targets, such as distant vehicles, pedestrians, and traffic signs, have the characteristics of fewer pixels, low resolution, and inconspicuous features, resulting in the detection of small targets in images with low detection rates and high false alarm rates. However, the performance of YOLO network for small target detection is not ideal. All of these problems bring great challenges to small target detection in driving scenarios. For these reasons, this paper detects the road condition information based on the optimized Yolo V3 network. In addition to common pedestrian and vehicle detection, it adds the identification of traffic signs for small targets and identification lines. The feature extraction auxiliary network is obtained by "copying" the backbone network, and the feature extraction network of Yolo is optimized to improve the performance of the whole feature extraction network. Attention mechanism is adopted for the feature information fusion between the auxiliary network and the backbone network. It focuses on the processing of the effective feature channels, suppresses the invalid information channels, and improves the network processing efficiency. And optimized the network training part, the performance and function of the entire detection system meet the practical conditions after testing.

II. YOLO V3 NETWORK STRUCTURE ANALYSIS

The YOLO series algorithms [16, 17] is a typical end-to-end network structure. This type of network structure is more concise compared with the two-stage network of R-CNN series algorithm [18-20] that firstly generates candidate recommended regions and then performs detection plus judgment. Integrated the candidate area mechanism and detection into the same network, making the detection speed faster than the R-CNN series.

A. TESTING PROCESS

The network structure of YOLO v3 replaces the RPN (Region Proposal Network) network in the R-CNN network by using predefined candidate areas. The detection process of YOLO v3 is shown in Figure 1. It divides the feature map into a $S \times S$ grids, and each grid then generates B bounding boxes that predicts the target [21]. Finally, $S \times S \times B$ prediction bounding boxes are generated on the feature map, which covers the entire area of the feature map and directly perform border regression on the generated forecast bounding box.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

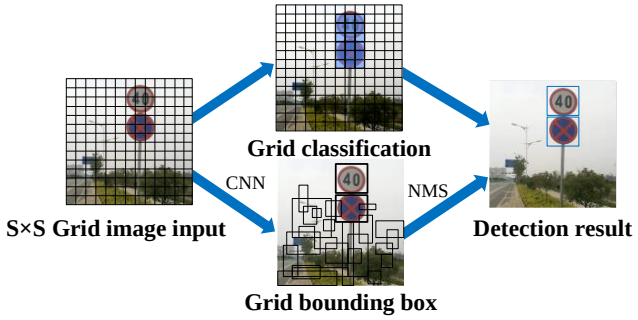


FIGURE 1. Testing process of YOLO v3

In order to prevent the prediction bounding box from being redundant, it is necessary to perform a confidence calculation for each prediction bounding box, and then set a threshold for the confidence. The bounding boxes above the threshold are reserved for regression, and the converse ones which are lower than the threshold will be directly deleted. The confidence of each bounding box is composed of two parts: the predicted target category probability and the coincident degree between the predicted bounding box and the actual frame. The formula is:

$$C = \Pr(class_i \mid object) * \Pr(object) * \mathrm{IOU}_{pred}^{truth} \quad (1)$$

Where $\Pr(class_i \mid object)$ is the probability of the objects' classification in the bounding box;

$\Pr(Object)$ is a parameter to determine whether there is a center point of the object to be detected in the divided grid. If so, it is 1. If not, it is 0.

$\mathrm{IOU}_{pred}^{truth}$ is the ratio of the intersection area to the union area between the prediction frame and the actual frame.

Setting a threshold for the prediction bounding box can eliminate most of the useless bounding boxes, but a single object may hold multiple bounding boxes at the same time to predict the object, resulting in redundant prediction bounding boxes on the feature map. Use non-polar non-max suppression (NMS) algorithm combines multiple prediction bounding boxes by judging the area coincident degree, and removes the ratio of the intersection area to the actual area between the prediction frame and the actual frame which is larger and the prediction bounding box whose confidence score is lower. Retains the prediction bounding box with a higher confidence score as the target detection box [22]. The loss function of the predicted bounding box consists of four parts, and the formula is:

$$loss = loss_1 + loss_2 + loss_3 + loss_4 \quad (2)$$

Where $loss_1$ is the loss of predicted central coordinate, $loss_2$ is the loss of width and height of the prediction bounding box, $loss_3$ is the loss of the predicted category, and $loss_4$ is the loss of the predicted confidence. The calculation formula for each part is as shown in Equation (3).

$$\begin{cases} loss_1 = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \text{Д}_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \\ loss_2 = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \text{Д}_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \\ loss_3 = \sum_{i=0}^{S^2} \text{Д}_{i}^{obj} \sum_{c \in class} (p_i(c) - \hat{p}_i(c))^2 \\ loss_4 = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \text{Д}_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \text{Д}_{ij}^{obj} (C_i - \hat{C}_i)^2 \end{cases} \quad (3)$$

Where (x, y) is the position of prediction bounding box. $(\hat{x}, \hat{y})$ is the actual position obtained from the training data. w_i and h_i are the width and height of the predicted bounding box respectively. λ_{coord} is to control the prediction position loss of the prediction box. λ_{noobj} is to control the no target loss in a single grid, Д_{ij}^{obj} is to determine whether the j^{th} (j=0,…,B) prediction bounding box in the cell i (i=0,…,S^2) is responsible for this target. If there is a target in the cell i, then the prediction value in the j^{th} bounding box is effectively for the prediction. Д_{i}^{obj} is to determine if the targeted center in the cell i. When there is an object in a cell, $\text{Д}_{i}^{obj}=1$. Otherwise, $\text{Д}_{i}^{obj}=0$. C_i is the confidence score. $\hat{C}_i$ is the intersection part of the predicted bounding box and the actual box.

B. NETWORK STRUCTURE

As the network continues to deepen, problems such as gradient disappearance and gradient explosion that occur during the training process can be solved by introducing a residual network [23]. Combining the features before entering the residual module with the features output by the residual module can extract deeper feature information. YOLO v3 adopts the new network structure Darknet-53. Darknet-53 is mainly composed of 53 convolutional layers, and contains a large number of 3×3, 1×1 convolution kernels. Compared with the network structure of v1 and v2, YOLO v3 draws on the residual network to design the shortcut connection modules. The module structure is shown in Figure 2. Shown as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

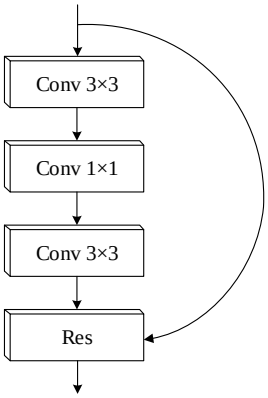


FIGURE 2. Residual structure network diagram

The participation of the shortcut connection module not only solves the problem of gradient disappearance caused by too many network layers, but also makes the entire network's total layer number reaches 106 layers, which is better for feature extraction. At the same time, YOLO v3 uses a multi-scale detection mechanism to detect the feature maps of 13×13, 26×26 and 52×52, respectively, and enhances the ability to extract small targets. Its network structure is shown in Figure 3:

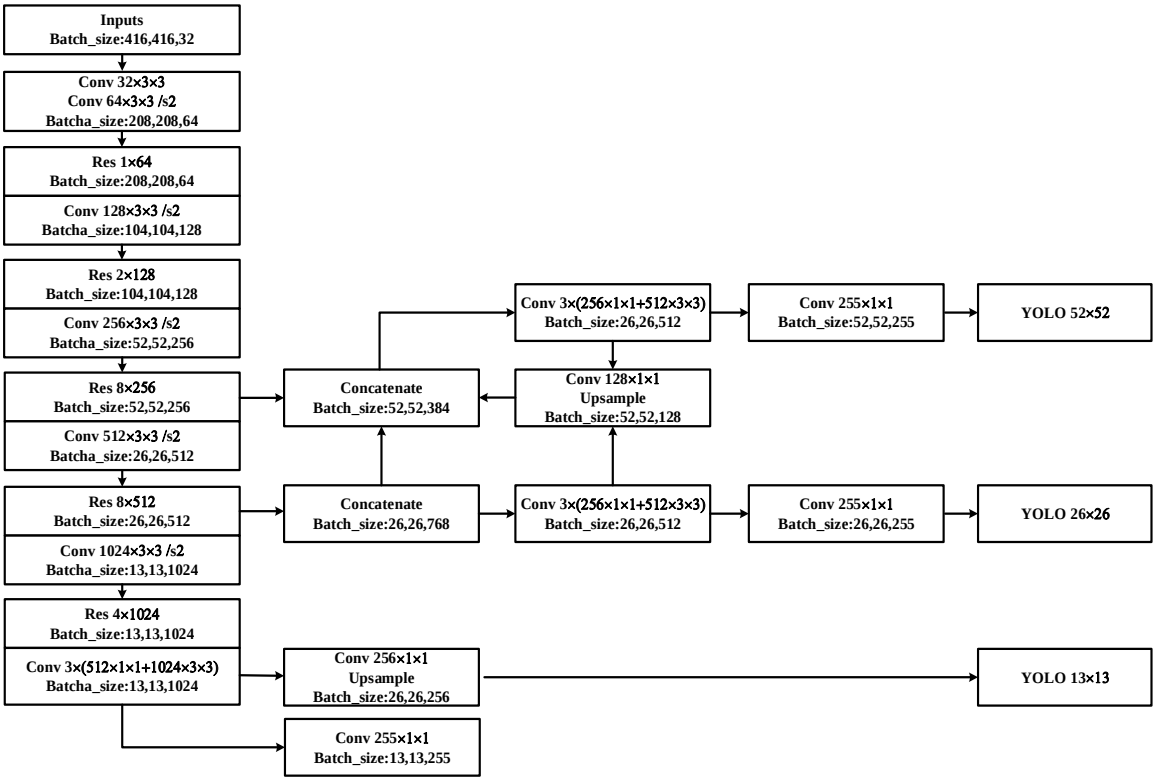


FIGURE 3. YOLO v3 network structure

YOLO v3 uses 3 different scale feature maps to predict the detection results. Regarding a resolution input image, the size of the basic scale feature map is 1/32 of the original resolution, and the remaining 2 scales are 1/16 and 1/8 respectively. For instance, a 416×416 training image, whose size of the basic scale feature map is 13×13×N, obtains a 26×26×N feature map by upsample. Fusion it with the output of the previous convolution layer to obtain the second scale feature map 26×26×M; Then, based on the second scale feature map, use the same method to obtain the third scale feature map 52×52×W. Finally, predict the 3-d tensor coding by the bounding box, goal score, and category prediction on each scale feature map. The detection frame has 4 parameters, the target evaluation holds 1 parameter, and the number of category is 80. Each scale feature map cell predicts 3 sets of information mentioned above, that is, 3× (4+1+80) = 255-dimensional information. The output tensor dimensions of the final three scales are y_1 = 13×13×255, y_2 = 26×26×255, and y_3 = 52×52×255.

In YOLO v3, the complexity of the model is further improved, and the multi-scale fusion method is used for prediction. Perform the position and category prediction on the multi-scale feature map could improve the accuracy of target detection. While introducing the residual network, a better feature extraction network Darknet-53 is proposed. Based on the above improvements, YOLO v3 model has a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

good effect on the accuracy and speed compared with the v1 and v2 models.

III. RESEARCH ABOUT IMPROVING YOLO v3

This paper's design is aimed at the detection of road information. The categories to be detected include: vehicles, pedestrians, traffic signs, and traffic sign lines. Due to the complex environment, the targets to be tested are also diverse, especially for the small targets such as traffic signs. In view of the above situation, this paper optimizes the YOLO's feature extraction network in two parts. The first part is to obtain the feature extraction auxiliary network by "replicating" the backbone network, which improves the performance of the entire feature extraction network. In the second part, attention mechanism is adopted for the feature information fusion of auxiliary network and backbone network, focusing on the effective feature channel, restraining the invalid information channel and improving the network processing efficiency.

A. FEATURE EXTRACTION AUXILIARY NETWORK STRUCTURE

The feature extraction network of YOLO v3 adopts Darknet-53, which deepens the sampling depth in the form of residual structure. The internal structure of the residual module is relatively simple, which makes the whole network simple and easy to use, but the feature extraction ability cannot be optimized. The emergence of the residual module enables the depth of neural networks to be further deepened. At present, it is a common network structure design method to rely on more network layers to extract the target features and enrich the semantic information of the convolution layer to improve the detection accuracy. However, for the recognition of small targets, as a result of the small size of the small target in the image's pixel area, its area will be further reduced after multiple downsampling operations in the neural network. The representativeness of features will also decline, which is not easy to be learned by the network layer. Therefore, it is not feasible to improve the accuracy of recognition by deepening the network structure.

In this paper, the main way to optimize the backbone network is to widen the overall network through "copying" the residual module, and fine-tune the residual module structure obtained by "replication". The optimized network structure is shown in Figure 4:

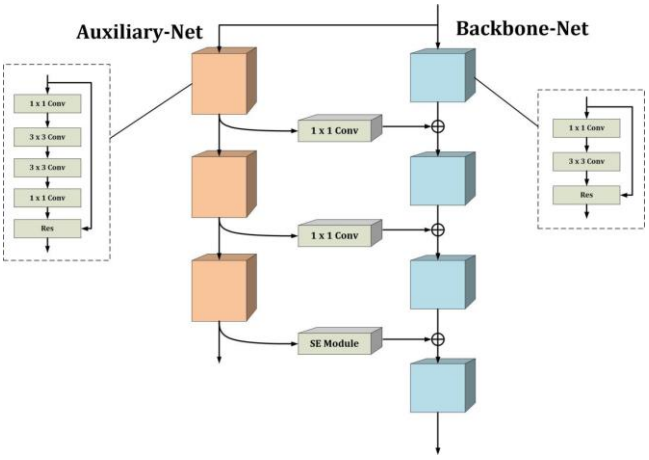


FIGURE 4. Optimized YOLO v3 network structure

Compared with the original network with a single structure, this paper adds a feature extraction auxiliary network with a smaller scale than the backbone network and composed of multiple residual blocks in the bypass of backbone networks. The residual module of the auxiliary network is improved compared with the residual module of YOLO, the original residual module uses 3×3 convolution kernel to carry out feature extraction, while the residual module in the auxiliary network uses two consecutive 3×3 convolution kernel to obtain a 5×5 receptive field, and then the extracted features will be merged into the backbone network. The receptive field used in the auxiliary network is 5x5 size. Using a large receptive field such as 5x5 to perform global feature extraction on the feature map can obtain the area features of the target. The auxiliary network transmits the acquired location features to the backbone network. The backbone network uses the 3x3 local receptive field in combination with the target location information provided by the auxiliary network to learn the target detailed features more accurately. Auxiliary network makes the whole network structure more closely related to high-level and low-level semantic features to a certain extent, which greatly improves the detection performance of the network. The original network of the YOLO v3 has a deep depth. If add the auxiliary network to the entire backbone network, more computing will be introduced, resulting in a slow running speed. Based on the above reasons, this paper only adds the auxiliary network to the feature extraction layer of three corresponding scale detection of Yolo.

B. ATTENTION MECHANISM

There are two different ways to connect the secondary network with the backbone network. The first way is that the output of the auxiliary module is firstly integrated by the 1×1 convolution kernel, and then transfers into the backbone network. The second way is to add attention mechanism between the connections of the two networks for the deep auxiliary network. When the network reaches a certain depth, its semantic information also becomes more

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

advanced. The auxiliary module can focus on processing and transmitting the effective features, and channel suppresses the invalid features. The implementation of the attention mechanism between the two networks uses the SE module (Squeeze-and-Excitation Module), which is simple in construction and easy to deploy. The connection structure is shown in Figure 5.

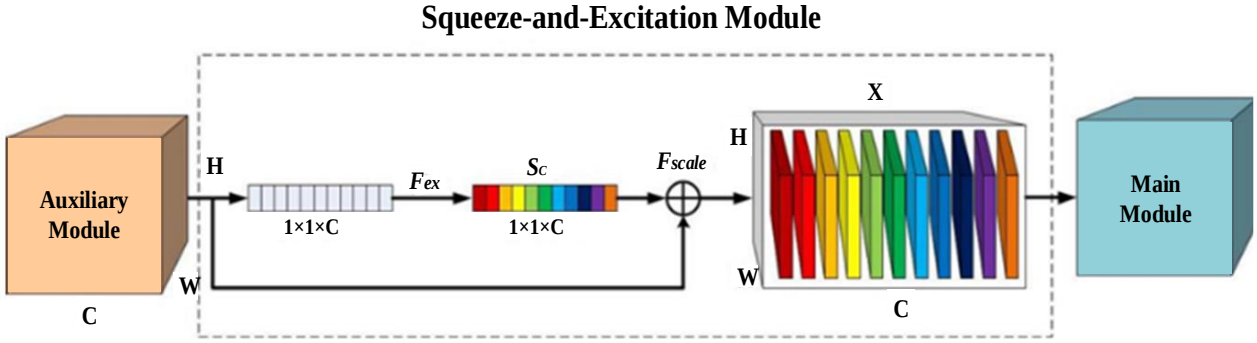


FIGURE 5. Connection structure diagram

The purpose of adding the SE module is to recalibrate the output features of the auxiliary module. The workflow can be roughly divided into squeeze and excitation. Firstly, the feature map is compressed, and its two-dimensional feature channel is changed into one-dimension by average pooling. At this time, the feature map size is converted to 1×1×C. The purpose of transforming two-dimensional feature map into a one-dimensional feature map by pooling is to better display the distribution of eigenvalues of each channel in this layer.

After the compression of the feature map is completed, the one-dimensional feature map will be excited. Its calculation formula is:

$$S_C = F_{ex}(Z,W) = \sigma(g(Z,W)) = \sigma(W_2\delta(W_1 Z)) \quad (4)$$

Where S_C is the characteristic map after excitation; δ function is ReLU; σ is sigmoid activation function; Z represents the input, one-dimensional convolution layer, after compression; W_1Z represents the full connection operation, and the dimension of W_1 is the fully connected layer of $\frac{C}{r} * C$; r is a scaling factor, and its function is mainly to scale the number of channels to reduce the amount of parameters. The scale of the feature map represented by Z is 1×1×C, and the scale of the feature map becomes $1*1*\frac{C}{r}$ through W_1; W_2 is a fully connected layer as well, and its dimension is $C*\frac{C}{r}$. The feature map output by W_1Z will output a feature map of 1×1×C scale through W_2, and finally it will be activated to get the feature map S_C via sigmoid function.

S_C is the core of the connection module and is used to indicate the channel weight of the auxiliary module output. S_C is activated via the sigmoid function, and its value ranges from 0-1, indicating the importance of each channel. Then, using the result of multiplying S_C and U_C to get the last redirected feature graph. The calculation formula is:

$$X_C = F_{scale}(U_C, S_C) = S_C \cdot U_C \quad (5)$$

By multiplying the different channel values with different weights, the attention on critical channel domains can be increased. Finally, the auxiliary residual module is redirected through the SE module and then get into the backbone network to complete feature fusion.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. PRODUCTION OF DATA SET

The data set used in this experiment is from Baidu's open source autopilot dataset ApolloScape. ApolloScape covers many complex road conditions and is relatively close to the realistic traffic scenes. This paper has selected four categories of data, including vehicles, pedestrians, left and right steering ground markings and forbidden traffic signs. 2000 samples of each class are selected as training sets, and another 500 samples are marked as test sets.

Then, perform the K-means mean calculation on the processed tag data set, and reset the size of anchors. The new nine candidate window sizes are calculated as follows: [23, 68], [39, 126], [51, 191], [93, 173], [71, 274], [105, 318], [153, 244], [172, 355], [291, 358]. After that, find the optimal parameters through multiple adjustments, and operate iterative training of the model for 45,000 times.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

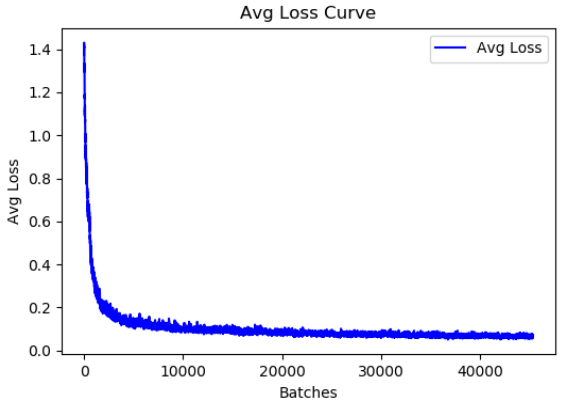


FIGURE 6. Training loss function graph

During the training process, the dynamic process of training can be visually observed by drawing the loss curve. Figure 6 shows the corresponding average loss curve during the training of the method model in this paper. The abscissa indicates the number of training iterations and the ordinate indicates the Loss value during training.

B. Model testing and lateral comparison

Use the test set to test the trained model. The test indicators mainly include precision, recall, mAP (mean Average Precision), AMC (Average Minimum Confidence), and missed detection rate (Loss). Among these indicators, the average minimum confidence is used to evaluate the classification performance of a multi-class target algorithm for a certain target. The miss detection rate is used to test the performance of the algorithm related to identify the target frame. The indicators are evaluated using the following formulas:

$$\text{Precision}=\frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{7}$$

Where TP represents a positive sample that is correctly classified in the algorithm, FP represents a positive sample of the misclassified, and FN represents a negative sample of the misclassified.

$$\text{mAP}=\frac{\sum AP}{N_C} \tag{8}$$

$$AP=\frac{\sum \text{Precision}}{N} \tag{9}$$

Where N_C indicates the target type, and N indicates the number of pictures.

$$AMC=\frac{\sum MC}{N} \tag{10}$$

Where MC represents the lowest confidence of a certain type of target that is correctly identified by the system in a picture.

$$\text{Loss}=\frac{\sum N_M}{N_B} \tag{11}$$

Where N_B represents the total number of targets in the picture, and N_M represents the number of undetected targets.

2500 samples are used to test and compare the optimized Yolo V3 model with the original network model. The comparison results are shown in Table 1:

TABLE I
COMPARISON BETWEEN OPTIMIZED MODEL AND ORIGINAL NETWORK MODEL TEST RESULT

	Precision	Recall	mAP	Loss
Our-YOLO	83%	85%	84.76%	14.56%
YOLO v3	79%	82%	79.33%	18.71%

The comparison of the Table 1 shows that the optimized Our-YOLO network has improved detection accuracy and recall rate compared to the original network. The improvement of this part is mainly based on the improvement of the detection ability aiming at small targets. The improved accuracy of each type can be referred to the Table 2.

The AMC indicator pairs for each category are shown in Table 2:

TABLE II
AMC INDICATORS COMPARISON

	Vehicle	Pedestrian	Left and right steering line	No-stop traffic sign
Our-YOLO	93.16%	88.54%	87.93%	78.50%
YOLO v3	92.95%	87.69%	85.44%	64.12%

It can be seen from the comparison of the above table that although the optimized network does not significantly improve the detection performance in terms of large targets, the detection accuracy of small targets such as traffic signs in complex environments is significantly improved.

The above test results are plotted as an accuracy-recall rate curve (PR curve). The graph is shown in Figure 7. Compare the improved network with the original network visually, and it is obvious that the overall detection performance of the Our-YOLO network can also be seen from the curve. The detection performance is better than the previous YOLO v3 network.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

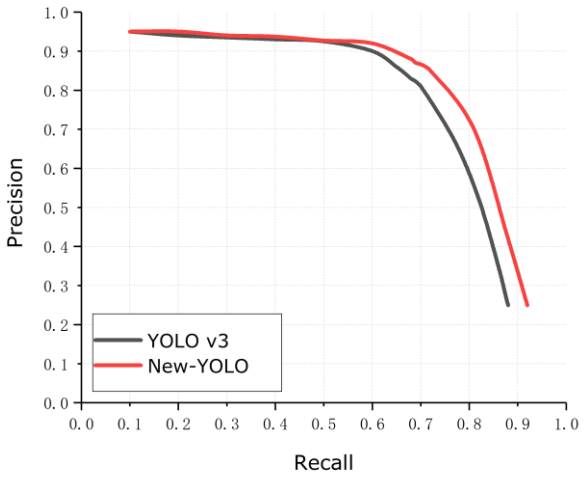


FIGURE 7. PR curve

Since the application scenario of this algorithm is to detect real-time road conditions on vehicles in the future, the detection rate was also tested. The test environment was tested under NVIDIA 2060 and 2080Ti graphics cards. The comparison of the test results is shown in Table 3:

TABLE III
DETECTION RATE TEST COMPARISON

	2060	2080Ti
Our-YOLO	25FPS	41FPS
YOLO v3	27FPS	44FPS

Although adding the auxiliary network will increase the computation of the whole network to a certain extent, the test results show that the detection speed of the whole network conforms to the actual application scenarios.

The Our-YOLO network was tested and compared with typical networks in one-stage and two-stage. The road condition information dataset used in the previous experiment was still used to compare the average accuracy and processing time. The comparison results are shown in Table 4.

TABLE IV
NETWORK TEST COMPARISON

	mAP	Frame rate
Our-YOLO	84.76	41FPS
PFP-Net	80.54	42FPS
SNIPER-Net	85.84	37FPS

Both PFP-Net and SINPER-Net are networks that have performed well in recent years. PFP-Net belongs to one-stage type network. From the above test results, Our-YOLO has higher accuracy than PFP-Net and frames. And the rate is almost the same. While SNIPER-Net is a two-stage type network, Our-YOLO is only slightly lower than the mAP by 1.08%, but the detection speed is faster. That is to say, Our-YOLO has a good performance in detection accuracy and detection speed.

(a) Original images (b) YOLO v3 (c) Our-YOLO

FIGURE 8. Comparison of test results

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

Performed the real-world tests on the models, and compared the recognition results of the pictures taken by the vehicle camera on the two models. The test results of the two groups are shown in Figure 8.

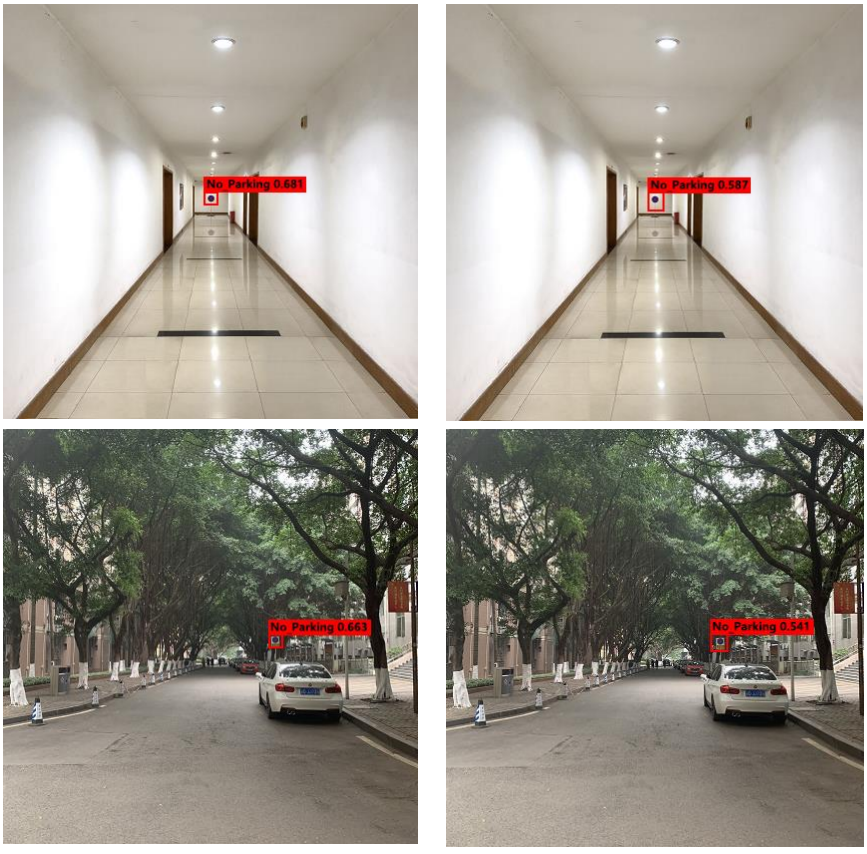
It can be seen from the comparison of the two sets of actual measurements that the Our-YOLO network after adding the auxiliary network improves the effectiveness and accuracy of small target detection than the original YOLO v3 network. For instance, the original YOLO v3 cannot detect the no-stop sign in Figure 8, but the improved network can effectively detect the sign, and even in a complex environment far away, the new network still can detect and identify accurately.

Longer distance targets usually show the characteristics of small targets. In addition to the regular model index test, this paper adds a single target test of the distance index. The no-parking traffic sign is set as the target sample for this experiment. Three sets of distances are given: 10m, 20m, and 30m. Each distance consists of 100 samples to form a test set, and the samples includes both indoor and outdoor scenarios. The results of the comparative tests are shown in the following Table 5:

TABLE V
SINGLE TARGET DISTANCE TEST COMPARISON

	10m-AP	20m-AP	30m-AP
Our-YOLO	87.53%	82.91%	68.76%
YOLO v3	87.46%	67.12%	55.12%

It can be seen from the comparison data in the table above, at a distance of 10m, the accuracy of YOLO v3 is similar to optimized one. Starting from the distance of 20m, it can be reflected that the detection performance of YOLO v3 for small targets has begun to decline, and Our-YOLO can still maintain its accuracy. Further, the gap between YOLO v3 and Our-YOLO is more obvious in the 30m test. Although the accuracy of Our-YOLO has decreased, the accuracy decline is not too obvious compared to the decrease rate of YOLO v3 accuracy. The indoor and outdoor within distance of 30m sample test results are shown in Figure 9:



(a) Our-YOLO (b) YOLO v3

FIGURE 9. Comparison of indoor and outdoor single target sample test at 30m

As can be seen from the above comparison chart, in addition to the accuracy of Our-YOLO is better than YOLO v3, the degree of fitting between the predicted bounding box position and the target is also more accurate than YOLO v3, which proves that the added auxiliary network is helpful for the location feature learning and determination.

IV. CONCLUSION

This paper mainly introduces the detection of road information based on our-yolo network model optimized by Yolo V3 network. The main work of this paper are the following two points:

1) Our-yolo network model, based on the original Yolo V3 model, uses a dual feature extraction network structure. The backbone network at the scale of 13 × 13, 26 × 26, 52 × 52 is equipped with different feature extraction auxiliary network of receptive field.

2) Attention mechanism is adopted for the feature information fusion of the auxiliary network and the backbone network. It focuses on the processing of the effective feature channels, suppresses the invalid information channels, and improves the network processing efficiency.

Compared with YOLO v3, the detection capability of the network after optimization is greatly improved. And the detection accuracy is improved by 5.43% compared with that before the optimization. Due to the addition of auxiliary network, the amount of calculation will also increase. The impact is to increase the detection time. But after the video stream test, although the detection speed is slower than that of Yolo v3, it also meets the conditions of practical application.

The main work of the next step can be divided into two steps. One is to optimize the structure of the network to improve the detection speed. the other is to use edge computing, so that the model can be used in embedded devices.

REFERENCES

[1] A.Y. Cao, "Research on Traffic Sign Recognition Algorithm," M.S. thesis, EIE. Dept., BJTU. Univ., BJ, CHN, 2017.

[2] C. Chen, C. Liao, X. Xie, et al. Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes. Personal and Ubiquitous Computing, 2019, 23(1): 53-66.

[3] M. Berger, A. Forechi, A.F.D. Souza, et al. " Traffic Sign Recognition with WiSARD and VG-RAM Weightless Neural Networks," Journal of Network and Innovative Computing, vol.87, no.1, pp.87 -98, Jan, 2013.

[4] A. U. Peker, O. Tosun, H.L. Akin, et al. " Fusion of map matching and traffic sign recognition," in 2014 IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI,USA,2014,pp.867-872.

[5] S.M. Karis, N.M. Ali, J. Safei, "Hidden Nodes of Neural Network: Useful Application in Traffic Sign Recognition," in ICSIMA., Kuala Lumpur, Malaysia, 2014, pp.1-4.

[6] K. Fu, I. Y. G. Gu, et al. "Geodesic distance transform-based salient region segmentation for automatic traffic sign recognition," in IV., Gothenburg, Sweden, 2016, pp.948-953.

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012, pp.1097-1105.

[8] Z. Gao, J. N. Chen, Z.J. Li, "Pedestrian detection method based on YOLO network," Comput Eng., vol.44, no.5, pp.215-219, May.2018.

[9] L. H. Li, Z. M. Lun, J. Lian, "Road vehicle detection method based on convolutional neural network," Journal of Jilin University (Engineering and Technology Edition), vol.47, no.2, pp.384-391, Mar.2017.

[10] Y. Tang, C. Z. Zhang, R.S.GU, et al. "Vehicle detection and recognition for intelligent traffic surveillance system," Multimedia Tools and Applications, vol.76, no.4, pp. 5817-5832, Mar.2015. DOI. 10.1007/s11042-015-2520-x

[11] S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, Z. Yu, et al. ROD-Revenue: Seeking Strategies Analysis and Revenue Prediction in Ride-on-demand Service Using Multi-source Urban Data. IEEE Transactions on Mobile Computing. https://ieeexplore.ieee.org/document/8733999

[12] S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, et al., A simple but quantifiable approach to dynamic price prediction in ride-on-demand services leveraging multi-source urban data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 2, Issue 3, pp 112:1-112:24

[13] S. Guo, Y. Liu, K. Xu, D. M. Chiu, Understanding ride-on-demand service: Demand and dynamic pricing. 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp 509-514

[14] C. Chen, Y. Ding, X. Xie, S. Zhang, Z. Wang, and L. Feng. TrajCompressor: An Online Map-matching-based Trajectory Compression Framework Leveraging Vehicle Heading Direction and Change. IEEE Transactions on Intelligent Transportation Systems, to appear, 2019.

[15] C. Chen, Daqing Zhang, Xiaojuan Ma, Bin Guo, Leye Wang, Yasha Wang and Edwin Sha. CrowdDeliver: Planning City-wide Package Delivery Paths Leveraging the Crowds of Taxis. In: IEEE Transactions on Intelligent Transportation Systems (T-ITS), 18(6): 1478-1496, 2017.

[16] J. Redmon, A. Farhadi, "YOLO9000: Better, Faster, Stronger," in CVRP., Honolulu, HI, USA, 2017, pp.6517-6525.

[17] J. Redmon, S. Divvala, R. Girshick, et al. "You Only Look Once: Unified, Real-Time Object Detection," in CVPR., Las Vegas, NV, USA, 2016, pp.779-788.

[18] R. Girshick, J. Donahue, T. Darrelland, et al. "Rich feature hierarchies for object detection and semantic segmentation," in CVPR., Columbus, OH, USA, 2014, pp.580-587.

[19] R. Girshick, "Fast R-CNN," in ICCV., Santiago, Chile, 2015, pp.1440-1448.

[20] S. Ren, R. Girshick, K. He, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Transactions on Pattern Analysis & Machine Intelligence, vol.39, no.6, pp.1137-1149, June, 2017.

[21] M. Rajchl et al., "DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks," in IEEE Transactions on Medical Imaging, vol. 36, no. 2, pp. 674-683, Feb. 2017.

[22] K. He, X. Zhang, S. Ren, et al. "Deep Residual Learning for Image Recognition," in CVPR., Las Vegas, NV, USA, 2016, pp.770-778

[23] N. Bodla, B. Singh, R. Chellappa, et al. "Soft-NMS--Improving Object Detection With One Line of Code," in ICCV., Venice, Italy, 2017, pp.5562-5570.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2966328, IEEE Access

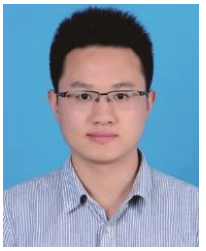
Qiwei Xu was born in Heilongjiang, China, in 1983. He received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2006, 2008, and 2013, respectively, all in electrical engineering. He is currently an associate professor with Chongqing University, Chongqing, China. His current research interests include design and control of special electric machines, electric drive system of electric vehicles, and control and simulation of hybrid electric vehicles.

Runzi Lin received the bachelor's degree in automation from Northeastern Electric Power University, China, 2015. She is currently pursuing a master's degree in electrical engineering at Chongqing University, China. Her main research interests are in the field of pattern recognition and image processing.

Han Yue was born in Heilongjiang, China, in 1984. She received the B.S. and M.S. degrees from Harbin Institute of Technology, Harbin, China, in 2006 and Xi'an Jiaotong University in 2009, respectively, all in electrical engineering. She is currently a senior engineer in Northeast Branch of State Grid, Shenyang, China. Her current research interests include operation and control of large power grid and new energy consumption.

Hong Huang has been studying for a master degree at Chongqing University since 2018, His research interest covers target detection and computer vision.

Yun Yang was born in Sichuan, China, in 1993. He received the B.S. degree in Vehicle engineering from Nanchang University, Nanchang, China in 2017. He is currently working toward the M.S. degree in Vehicle engineering at Chongqing University. His research interests including the electromagnetic optimization design of motor and fault diagnosis such as a PMSM inter-turn short fault.



Zhigang Yao (S'18) was born in Chongqing city, China, in 1991. He received the B.S. degree in Electrical Engineering from Chongqing University, Chongqing, China, in 2014.He is currently working towards his Ph.D. degree in Electrical Engineering from Chongqing University, Chongqing, China. His current research interests include interleaving techniques, high power density bidirectional power converters, gird-connected inverter, and renewable energy systems.