Traffic Sign Detection in Complex Scenarios based on YOLOV7

Yicong Wu[1, *], Shuhan Wang[2]

[1]Tianjin University, Tianjin, 300072, China

[2]Hubei University of Technology, Wuhan, Hubei province, 430068, China

*Corresponding author: 3021202292@tju.edu.cn

Abstract. In recent years, thanks to the development of the automobile industry and the advancement of industrial intelligence, driving assistance systems have been developing rapidly under the requirements of national policies and the demands of the consumer market. Traffic sign detection module, as an indispensable part of the driving assistance field, is in urgent need of new target detection algorithms in this field to solve the problems of poor detection accuracy, low computing speed and poor adaptability of the existing algorithms to the vehicle hardware. To solve the application problem, in this paper, we propose a traffic target detection algorithm based on YOLOV7. The model is trained using the CCTSDB dataset and tested in complex scenarios such as different lighting conditions and weather conditions. The mAP of YOLOV7 is stable at over 82.46% and the mAP value of YOLOV7-tiny is stable at over 78.57%, demonstrating the effectiveness of our method that it has the ability to detect and recognize our traffic signs in practical scenarios. In addition, this paper conducts a comparative study on the commonalities and differences of the above two models and draws the corresponding conclusions. In summary, YOLOV7 series can be effectively used in traffic sign detection work.

Keywords: Object detection; traffic sign detection; YOLOV7; deep learning.

1. Introduction

In recent years, with the development of artificial intelligence technology, driver assistance systems have been widely used in the field of civilian automobile control and have been greatly emphasized by the country. According to the "Guidelines for the Construction of the National Internet of Vehicles Industry Standard System" issued by the Ministry of Industry and Information Technology and the National Standardization Management Committee, by 2025, the system will form a standard system for intelligent connected vehicles that can support the combination of driving assistance and self-driving general functions [1]. Adding to the fact that the rise of new energy vehicles has greatly stimulated the consumption desire of the consumer market, many countries have implemented policies that favor the development of new energy vehicles equipped with intelligent driver assistance systems, which has created a wave of driver assistance systems worldwided and expanded demand for the production sector. In this context, the research and upgrading of driver assistance systems have received extensive attention. Traffic sign detection, as an indispensable part of driver assistance system, urgently needs to introduce advanced algorithms to meet the development requirements of the country and society.

Much effort has been devoted to realizing the traffic sign detection, which applies deep learning networks to traffic sign recognition in the field of autonomous driving. However, they struggle to do their job efficiently due to Poor suitability and outdated versions of deep learning networks, and also suffer from difficulty adapting to in-vehicle hardware environments. As a single-stage target detection network, YOLO has many unique advantages in the field of traffic sign detection and recognition. The low hardware requirements, good adaptability and fast response capability almost perfectly match the requirements of autonomous driving for target detection systems. Applying YOLO to traffic sign detection can effectively optimize the experience of using driver assistance systems and improve safety. For L2-level driver assistance systems, the vehicle can provide traffic sign information back to the driver through the dashboard, HUD system or in-vehicle voice announcement after detecting traffic signs. In this way, it helps drivers to get more comprehensive traffic information,

avoid missing important traffic signs caused by distraction or environmental factors, improve driving perception, and reduce the possibility of car accidents. For L3-level and above driver assistance systems, the traffic sign detection system can not only provide more complete road information to the vehicle, but also detect whether the car is in normal driving mode based on the status of the road traffic signs, further avoiding accidents and improving traffic safety.

As a newcomer to the deep learning network target detection task, YOLOV7 has a major advantage over previous convolutional neural networks in terms of speed and accuracy in the range of 5 FPS to 160 FPS and excels in terms of the degree of hardware suitability. Considering the limitations of hardware facilities due to the special characteristics of the vehicle environment and the higher requirements on the sensitivity of the target detection system response put forward by the driver assistance function under complex road conditions, YOLOV7-tiny, as the lightweight version of YOLOV7 series, effectively makes up for the shortcomings of YOLOV7 in this situation.

Based on the above considerations, this paper innovatively proposes to apply YOLOV7 and YOLOV7-tiny in the field of traffic sign detection. This paper carries out targeted debugging and specific analysis, so that it can be better suited to the domestic road environment, which will facilitate the development of driver assistance technology in the application field. In order to achieve the realistic scenarios of Chinese roads, the Chinese Traffic Sign Detection Dataset (CCTSDB) is adopted to train the model. We further explore the performance under different scenarios, such as various weather environments and lighting conditions, respectively. The test results were analyzed and studied to compare the test results of different versions of YOLOV7 series. In this paper, side-by-side comparisons are made to produce refined

2. Method

2.1. Revisiting the development of YOLO

YOLO (you only look once) is a single-stage object detection network [2], which is influenced by the regression idea and the running steps are very simple and concise. The detection target is used as input data, and the output data such as the position of the bounding box, the category to which the target belongs, and the degree of confidence can be derived after only one regression operation. It reduces the runtime while reducing the redundancy of the deep learning system and has good recognition accuracy and precision. Research on the application of the first-generation YOLO convolutional neural network in industry, transportation and even the military had emerged. With the YOLO convolutional neural network's fast computing speed, high recognition accuracy and low resource consumption, coupled with its strong adaptive ability, researchers had developed it into an image inspection system that can be used for highway vehicle inspection, weld qualification inspection, wind turbine blade defect detection, and many other industrial aspects. However, as the range of applications continues to expand, problems in the use of the first-generation YOLO were gradually exposed. For example, the authors of YOLO greatly reduced the volume of the algorithmic model and simplified the computational process at the beginning of the design in order to ensure its good adaptability, which led to the first generation of YOLO in the competition with other target detection models, both in terms of detection accuracy and accuracy had a significant disadvantage. On top of that, YOLO suffered from problems such as low recall rates, further narrowing its viability. These shortcomings greatly limited the application and development of YOLO convolutional neural networks in areas that require high accuracy.

In order to solve such problems, YOLOV2 [3] replaced the backbone network of YOLO used for feature extraction, improved the loss function, and introduced an anchor frame mechanism in the feature image to predict the target position in advance [4]. After testing, the YOLOV2 model showed significant improvement in metrics such as mAP. For YOLOV3 model [5], improvements are made such as adding feature pyramid networks, which allows it to output image information at three scales simultaneously and further balances the requirements of the convolutional neural network model in terms of operational accuracy and speed. In 2020, YOLOV4 [6] further introduced the ASPP [7]

(Atrous Spatial pyramid pooling) invented by DeepLab[8] and used a new activation function Mish, and added the Mosaic data enhancement, resulting in a significant increase in detection accuracy for YOLOV4 compared to YOLOV3. After minor modifications to YOLOV4, the YOLOV5 was launched by UltralyticsLLC[9]. YOLOV5 continued to make some improvements and local optimizations in terms of computational efficiency. In 2022, the YOLOV6 augmented the idea of RepVGG, focused on optimizing the adaptability of YOLO series models to hardware, and reduced the computation time, which was another important development of target detection algorithms. The development of successive generations of YOLO is shown in Figure 1.

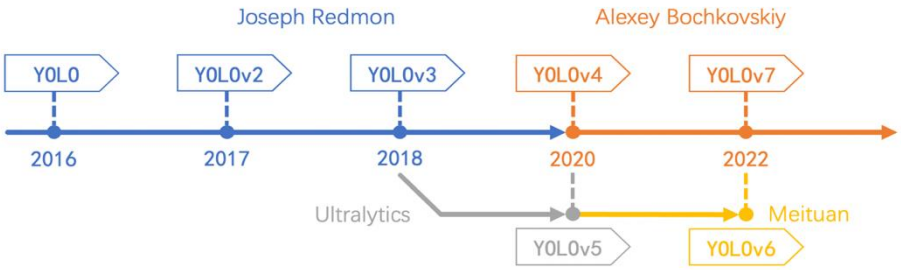


Fig. 1 The development of YOLO

2.2. Details of YOLOV7

In 2022, Wang et al. proposed a new generation of convolutional neural networks for target detection, known as YOLOV7[10]. YOLOV7 series is a new generation of target detection convolutional neural network optimized by adding four different tricks based on YOLOV4, YOLOV5, and YOLOV6, which performs only one regression operation.

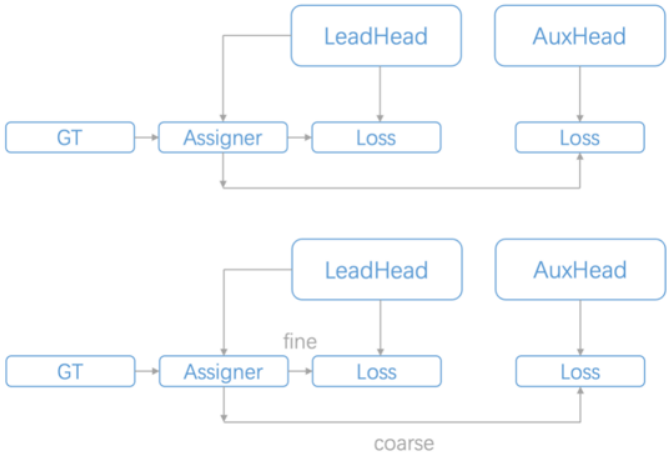


Fig. 2 The schema of 4 tricks in YOLOV7

As shown in Figure 2, the first trick is to use an efficient clustering network structure, E-ELAN, which is a novel structure that can dramatically improve the learning ability of the network model without destroying the original gradient paths. The second trick is model scaling based on the concatenate model, which performs scaling operations on depth, width, or module scale to optimize detection accuracy. The third trick is to add RepConv structures for convolutional reparameterization operations, and to analyze the problem of pairwise use of reparameterization modules through gradient propagation paths. The fourth strategy is to incorporate an auxiliary training module that generates various assigners from the prediction results of the Leadhead to help the Auxheads and Leadheads learn.

As shown in Figure 3, the basic structure of YOLOV7 is mainly composed of three parts: feature extraction network, enhanced feature extraction layer and YOLOHead[11]. The input side follows the Mosaic data enhancement method proposed by YOLOV4[12]. The left block diagram portion of Figure 3 shows the main feature extraction network, which completes the initial information

extraction of the traffic sign image by multiple CBS structures, ELAN structures and MP-1 structures. The upper right part of the block diagram in Figure 3 shows the basic structure of the enhanced feature extraction layer, where the initial image information obtained by the main feature extraction network will be fused in successive convolutional and up-sampling operations to fuse the feature information from the different feature layers and to enhance the extracted effective information. YOLOHead is the core structure of YOLOV7, whose role is to perform the regression operation on the extracted information, analyze and derive the information such as confidence degree and bounding box, and use it as the basis to complete the subsequent work of the target detection system[11].

Combining the advantages of YOLOV7 over other target detection models in the above aspects, YOLOV7 is selected as the basic algorithmic model in this paper to be applied to traffic target detection in driver assistance systems. Considering that YOLOV7-tiny, a derivative of YOLOV7 series, is further reduced in size and has lower requirements for the hardware environment of the vehicle-mounted device, this paper also applies YOLOV7-tiny to the traffic sign detection work. After conducting refinement experiments on both of YOLOV7 and YOLOV7-tiny in different segmentation scenarios, this paper also conducts a side-by-side comparison of the two convolutional neural networks based on the experimental data obtained. Based on the commonalities and differences demonstrated by the data, the target detection capabilities of YOLOV7 and YOLOV7-tiny are investigated under different weather conditions and different lighting conditions.

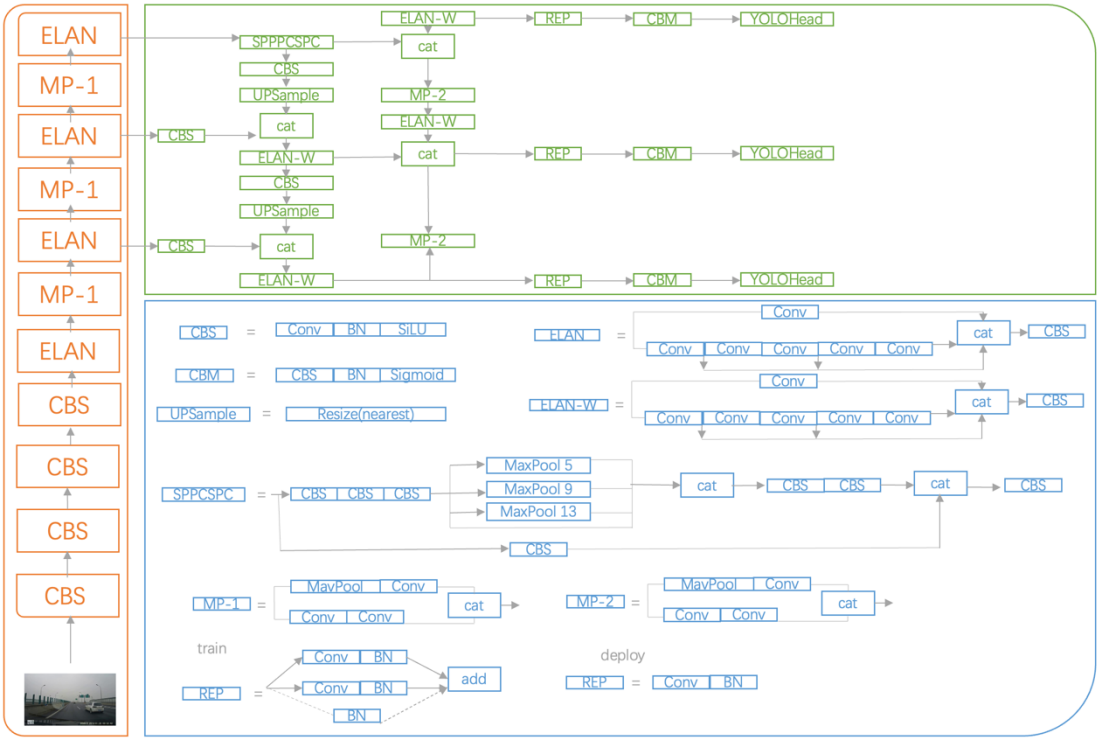


Fig. 3 Structure of YOLOV7

2.3. Original data set

This paper proceeds to the actual road scenarios in different regions of China, taking into full consideration the special characteristics of China's highway traffic. Therefore, this paper uses the China Traffic Sign Detection Dataset (CCTSDB). In this paper, CCTSDB dataset are categorized according to different weather and different lighting conditions, and images containing out different meteorological conditions such as rainy, cloudy, and sunny are selected. The data are then imported into VOLOV7 and YOLOV7-tiny convolutional neural networks, respectively, for model adaptation and improvement. After a period of sufficient training, the VOLOv7 series is initially equipped with the ability to recognize and detect traffic signs in photo data. The paper goes on to analyze the detection results and attempts to test the trained convolutional neural network against the available

information. This project imitates the height and distance of on-board recorders and on-board cameras, and collects photos under different traffic road conditions on site to fully guarantee the authenticity of the test atlas. The images are classified. Experiments are conducted to recognize the photographs using a convolutional neural network to check the performance of the trained model. Comparison of VOLOV7 and VOLOV7-tiny's detection capabilities in different environments, specifically in terms of recognition speed and recognition accuracy. Analyze to draw conclusions.

2.4. Model construction

The main process of model construction in this paper is to import the CCTSDB into the YOLOV7 model and its derivative version YOLOV7-tiny model respectively. During the import process, this project team found that the original YOLOV7 and YOLOV7-tiny had an error in compatible datasets and were not able to directly utilize the existing datasets for training. As a result, in this paper, the original model training code module is modified before training the dataset so that it can complete the training on the dataset in the current hardware environment. In this paper, with the modified model, YOLOV7 and YOLOV7-tiny are allowed to complete the training containing the same number of images under the same hardware environment using only the Chinese traffic signs dataset. After that, the training results are saved and subsequent experiments are performed using the trained model.

3. Experiment

3.1. Hardware environment

This project uses a Windows 10 system, based on NVIDIA GeForce RTX 3060 Laptop GPU, AMD Ryzen 5000Series 7 hardware conditions, at room temperature of 26 degrees Celsius and sufficient heat dissipation conditions to complete the operation test. The experiment was conducted using Anaconda to build a virtual environment with the programming environment set to Python 3.8.17 and using the PyTorch 2.0.1 framework.

3.2. Experimental setting

In this paper, the same hyperparameters are set for YOLOV7 and YOLOV7-tin models, the IoU threshold used for non-maximum suppression values is set to 0.3, and the prediction box is set to remain when the confidence level is greater than 0.5 to ensure training efficiency.

At the beginning of model training, this paper introduces the general pre-training weights of the YOLOV7 series, and uses freezing training for 16 epochs of pre-training in order to detect the feasibility of model training of YOLOV7 series in road sign detection while the pre-training weights used in this paper are also obtained. Adaptive dynamic learning rate is used during model training. In the case of the existing pre-trained weights that have been preliminarily trained, compared with the same common Adam optimizer, the advantage of the SGD optimizer is that it is more conducive to the stable convergence of the model. Therefore, the SGD optimizer is used in this paper, and the maximum learning rate of the model is set to 1e-2 and the minimum learning rate is 1e-4. During training, the batch_size parameters of the YOLOV7 model are set to 8 and epochs to 16. At the same time, set the batch_size parameter of the YOLOV7-tiny model to 12 and epochs to 16. The training process uses the Cuda compute architecture to assist training. In view of the problem of insufficient computing power of home computers found in early attempts, this paper chooses to adopt the mixed precision training mode in the experiment, which achieves a significant reduction in the memory occupied during the training process. The data set used in this article is a VOC format data set, and the ratio of training and test sets is 9:1. Since the input image format is not uniform, this paper standardizes the input of the image and obtains 640×640 standard drawings to be inspected. At the same time, this paper sets the mosaic or mixup data enhancement to be detected with a 50% chance as a pretreatment in order to achieve the effect of enhancing the training effect of the model and avoiding overfitting problems. In addition, in this data set, the objects to be detected are divided into three categories for training: mandatory, prohibitory, and warning.

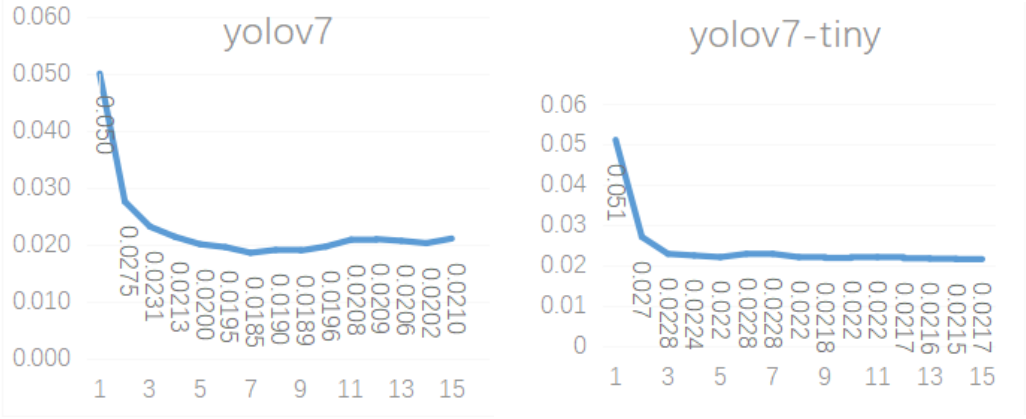
3.3. Evaluation metric

In order to verify the effectiveness of model training, this paper uses the value of mAP (Mean Average Precision) as the basis to track the training effectiveness of YOLOV7 network model and YOLOV7-tiny network model. The following is the formula for calculating mAP, where AP is the average accuracy:

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{1}$$

3.4. Model performance analysis

After the tested model, the convergence was determined, and the val loss values of YOLOV7 and YOLOV7-tiny were shown in Figure 4.



(a) Val loss of YOLOV7 (b) Val loss of YOLOV7-tiny

Fig. 4 Val loss of YOLOV7 and YOLOV7-tiny models

The mAP parameter calculation results are shown in Figure 5, where the mAP calculation value of YOLOV7 is 82.46%, and the mAP calculation value of YOLOV7-tiny is 78.57%.

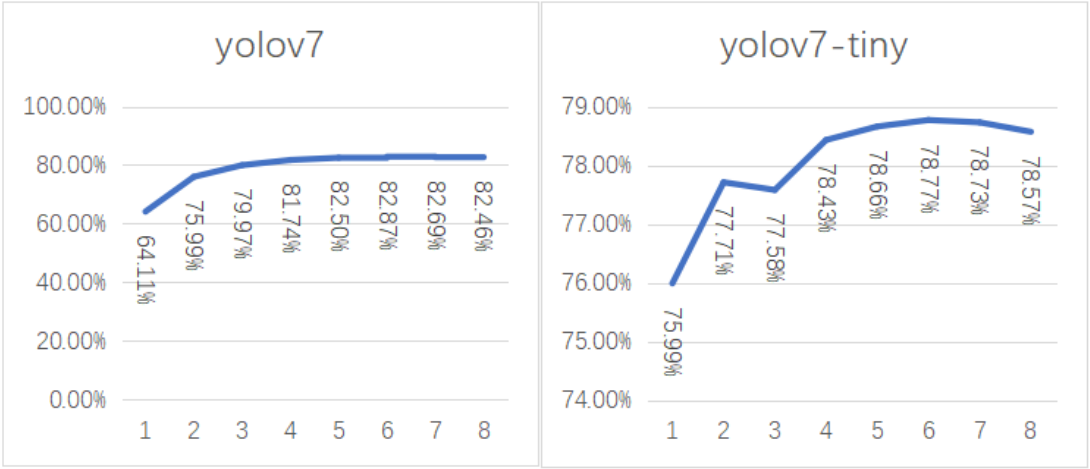


Fig. 5 The mAP calculation values of YOLOV7 and YOLOV7-tiny

3.5. Model performance for different scenes

In this paper, the YOLOV7 model and the YOLOV7-tiny model were used to sample the test set.Among them, the weather elements are divided into three types: sunny days, rainy and snowy days, and nights.Based on the division of the original data set, the target to be detected is further divided into three categories: small (xs, s), medium (m, l), and large (xl). If all the road signs in a scene are correctly identified and there are no misidentifications, the object to be detected is considered successful.The successful recognition rates of YOLOV7 and YOLOV7-tiny are rounded to the test traffic signs, and the results are shown in Table 1 and Table 2.

Table 1. Recognition success rate of YOLOV7

YOLOV7	sunny	rainy and snowy	night	average
large	82%	78%	72%	77%
medium	85%	86%	75%	82%
small	65%	62%	40%	55%
average	77%	75%	62%	

Table 2. Recognition success rate of YOLOV7-tiny

YOLOV7-tiny	sunny	rainy and snowy	night	average
large	80%	74%	68%	74%
medium	78%	82%	65%	75%
small	55%	52%	30%	45%
average	71%	69%	54%	

From the comparison of Table 1 and Table 2, it can be seen that YOLOV7 and YOLOV7-tin can both detect medium and large types of road signs under good lighting conditions. However, the information loss is caused by overexposure of traffic signs in too dark environments or to be detected, and the detection success rate of small traffic signs, especially very small (xs) traffic signs, is insufficient.

Looking at the horizontal comparison results, it is not difficult to find that the overall success recognition rate of YOLOV7 is higher than that of its simplified YOLOV7-tiny version, especially for the detection of smaller traffic signs, the recognition success rate of YOLOV7-tin is significantly lower than that of YOLOV7. The main reasons for the influence are shown in the example in Figure 6, YOLOV7-tiny, as a small model, reduces the accuracy of detection and increases the probability of missed and false positives while lightweighting. It is worth noting that although YOLOV7-tiny sacrifices some of the accuracy and robustness of detection, it is not much different from YOLOV7 in the success recognition rate of medium and large category traffic sign detection.

Fig. 6 Visualization of detection results of YOLOV7 and YOLOV7-tiny models

4. Discussion

In this paper, the detection capabilities of the YOLOV7 model and YOLOV7-tiny model in the field of road traffic sign detection were quantitatively studied with the help of China Traffic Sign Detection Dataset (CCTSDB). At the same time, the research data were compared and analyzed from the two dimensions of different weather conditions, various lighting conditions and the size of traffic signs to be measured. Compared with the two models, this paper argues that the two models have their own advantages, but further modifications and enhancements are still needed to apply them to the actual road sign detection application.

Among them, the large model YOLOV7 has good robustness and accuracy, and can detect and judge medium and large targets well in various weather environments. However, there are still

deficiencies in the accuracy of small target detection, the size of the YOLOV7 model itself and the computing power occupation. In view of these problems, this paper will further prune the YOLOV7 model in future research to make it easier to adapt to the computing power of vehicle-mounted embedded systems. At the same time, try to increase the attention mechanism so that the model can better detect small targets.

The small model YOLOV7-tiny has the advantage of being lightweight, and its detection speed and adaptability to the in-vehicle embedded system are stronger than the YOLOV7. In contrast, YOLOV7-tiny has disadvantages compared with YOLOV7 in generalization ability, and is prone to misjudgment and missed judgment. In view of this problem, this paper will further improve the neural network of the YOLOV7-tiny model in future research, and treat it dropout to increase the generalization ability of the model.

In addition, this paper argues that in the field of convolutional neural network traffic sign detection, the original image input of the vehicle camera also has a great influence on the detection of the model. The degree of distortion and color difference of the input image directly affects the accuracy of the model judgment. In future research, this paper will try to add the preprocessing module of the input graph in combination with the specific hardware configuration to increase the feasibility and accuracy of the model in real-world applications. At the same time, this paper notes that although the China Traffic Sign Detection Dataset (CCTSDB) has a large amount of picture data, there are only three types of classification that are not detailed, which also has a certain impact on the accuracy and practical application of model training. In future research, this article will try to further refine the classification of existing datasets to ensure the accuracy and practicality of model training.

5. Conclusion

In summary, after model tests and practical experiments, the YOLOV7 and YOLIv7-tin models applied in the field of traffic sign target detection in this paper reflect good detection accuracy, pass the subdivision scene test, and can complete the detection task of traffic signs in actual scenes.

References

[1] Bailly K, Dubuisson S. Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests[J]. IEEE Transactions on Affective Computing, 2017, 10(2): 167-181.

[2] Liu D, Ouyang X, Xu S, et al. SAANet: Siamese action-units attention network for improving dynamic facial expression recognition[J]. Neurocomputing, 2020, 413: 145-157.

[3] Li S, Deng W. Deep facial expression recognition: A survey[J]. IEEE transactions on affective computing, 2020, 13(3): 1195-1215.

[4] Li J, Jin K, Zhou D, et al. Attention mechanism-based CNN for facial expression recognition[J]. Neurocomputing, 2020, 411: 340-350.

[5] M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

[6] Man Tang, Cheng Li, and Huimin Yu, "Face recognition using SVM-based committee machine," in Proc. 2008 19th International Conference on Pattern Recognition (ICPR), 2008, pp. 1-4.

[7] Y. Gao and K.H. Leung, "Facial Expression Recognition Using Line Edge Map," IJCSI International Journal of Computer Science Issues, vol. 8, no. 3, pp. 295-301, 2011.

[8] S. Zafeiriou, I. Kotsia, and I. Pitas, "Illumination normalization of y-faces in the wild," IEEE Transactions on Information Forensics and Security, vol. 6, no. 1, pp. 107-122, 2011.

[9] J. Yang, Q. Wu, and F. Zhang, "Illumination normalization for facial expression recognition based on adaptive gamma correction," Signal, Image and Video Processing, vol. 10, no. 8, pp. 1591-1598, 2016.

[10] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in Proc. 2006 7th International Conference on Automatic Face and Gesture Recognition (FGR06), 2006, pp. 211-216.

[11] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," Computer Vision and Image Understanding, vol. 138, pp. 1-24, 2015.

[12] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.