Journal of Signal Processing Systems (2021) 93:899–911
https://doi.org/10.1007/s11265-020-01614-2

An Efficient Small Traffic Sign Detection Method Based on YOLOv3

Jixiang Wan [1,2] · Wei Ding [1,2] · Hanlin Zhu [1,2] · Ming Xia [1,2] · Zunkai Huang [1] · Li Tian [1] · Yongxin Zhu [1] · Hui Wang [1]

Received: 31 July 2019 / Revised: 2 November 2020 / Accepted: 3 November 2020 / Published online: 20 November 2020

Abstract
In recent years, target detection framework based on deep learning has made brilliant achievements. However, real-life traffic sign detection remains a great challenge for most of the state-of-the-art object detection methods. The existing deep learning models are inadequate to effectively extract the features of small traffic signs from large images in real-world conditions. In this paper, we address the small traffic sign detection challenge by proposing a novel small traffic sign detection method based on a highly efficient end-to-end deep network model. The proposed method features fast speed and high precision as it attaches three key insights to the established You Only Look Once (YOLOv3) architecture and other correlated algorithms. Besides, network pruning is appropriately introduced to minimize network redundancy and model size while keeping a competitive detection accuracy. Furtherly, four scale prediction branches are also adopted to significantly enrich the feature maps of multi-scales prediction. In our method, we adjust the loss function to balance the contribution of error source to the total loss. The effectiveness, and robustness of the network is further proved with experiments on Tsinghua-Tencent 100 K traffic sign dataset. The experimental results indicate that our proposed method has achieved better accuracy than that of the original YOLOv3 model. Compared with the schemes in relevant literatures our proposed method not only emerges performance superiors in detection recall and accuracy, but also achieves 1.9–2.7x improvement in detection speed.

Keywords Computer vision · Convolutional neural networks · YOLO · Traffic sign detection

1 Introduction

Traffic sign is meant to be one of the most critical elements in transport systems because it provides instructive or warning messages like road conditions and real-time traffic conditions for vehicles and pedestrians. Complying with traffic sign lawfully can greatly prevent traffic accidents and reduce congestion. For human beings, identifying the traffic sign is an easy task. However, for self-driving cars, locating and classifying the traffic sign accurately and quickly remains an incredible challenge. Therefore, traffic sign detection in autonomous vehicles has been catching the attention from the computer vision community incessantly for several decades [1–3]. Generally, traditional visual approaches for object detection, which usually use manual features including color, texture, and geometric to extract the regions of interest in an image, are difficult to achieve desirable results in the field of traffic sign detection [4–7]. With the vigorous development of artificial intelligence and computer vision technologies, deep learning appears to be one of the most efficient solutions for complex detection tasks, such as traffic sign detection, which has high demands in the aspects of detecting accuracy and response speed in multi-objects detection [8, 9].

Convolutional Neural Networks (CNNs) have been proved to be capable of achieving superior performance in image classification and object detection. The development of deep learning has brought new directions to target recognition. As a result, various excellent algorithms for object detection have been reported successively. The representative methods can be generally divided into two categories, respectively proposal-based methods and proposal-free methods. The proposal-based methods includes Region-based Convolutional Neural Network (R-CNN) series works [10–13], and the proposal-free methods mainly contains You Only Look Once (YOLO) model and Single Shot Multibox Detector (SSD), which indeed work well on Pascal Visual

✉ Zunkai Huang
huangzk@sari.ac.cn

✉ Yongxin Zhu
zhuyongxin@sari.ac.cn

✉ Hui Wang
wanghui@sari.ac.cn

[1] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, People's Republic of China

[2] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

Object Classes (VOC) [14] and ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [15]. In conventional target recognition system, the objects to be detected are conspicuous and typically occupy a large fraction of the whole image size. Unfortunately, for the traffic sign detection, size of the objects captured in real driving circumstances is much smaller. Since inadequate pixel features can be extracted, the small but significant objects are usually neglected by the ordinary models. To solve this problem, the intuitive solution is to design a more complex networks with a large number of candidate boxes or capture images that contain clearer objects with higher resolution. But this will additionally cause other troubles such as exponentially increase of calculation complexity and uncontrollability during training. Some constructive convolutional neural networks [16–18] aimed at detecting small traffic signs have been proposed, but failed to achieve satisfactory results in terms of detection accuracy and detection speed.

In addition, detecting small object from a relativity large image is not the only challenge. Under real-world conditions, the captured pictures are usually filled with complicated backgrounds, such as skies, buildings, roads, trees, pedestrians, vehicles and streetscape, rather than with clean and monotonous background. Numerous interference factors in the environment, such as advertising symbols and other indications, usually feature indistinguishable color saturation and contrast from traffic signs. Besides, under realistic traffic conditions, the detection environment is badly interfered by lighting, blocking, shadowing, and even bending, tilting or color fading. Furthermore, the counterfeit traffic signs from surface reflecting are also misleading elements for detection models. All of these difficulties make the traffic sign detection in the outdoor environment still an open problem.

In order to deal with above-mentioned challenging issues, we propose an efficient algorithm based on the state-of-art YOLOv3 model for real-life traffic sign detection. We optimize the feature extraction network to reduce the redundant residual block while keeping a competitive detection accuracy. In this way, we successfully decrease the amount of network parameters and effectively speed up the calculation process meanwhile. Moreover, the fourth scale prediction branch is attached to shallower network, and smaller and denser anchors are also introduced. Therefore, more finely grained information can be utilized in the extra feature map. This naturally enriches the feature maps of multi-scales prediction and improve the detection accuracy. Additionally, we intensify the penalties for category prediction to balance the weights of location error and classification error.

Briefly, the main contributions of this paper are described as follows:

1) We propose an optimized model and corresponding algorithm to efficiently classify and detect small objects like traffic signs in high-resolution images.
2) Network pruning is elaborately utilized to effectively minimize the network redundancy and model size while keeping a competitive detection accuracy.
3) Four-scale prediction branch is introduced to further enrich the feature maps of multi-scales prediction, which facilitates the utilization of more fine-grained information in additional feature maps.
4) The loss function optimization is performed by intensifying the penalties for category prediction to balance the contribution of each component error to the total loss.

Our proposed scheme performs well on the Tsinghua–Tencent 100 k benchmark with 92% recall and 94% accuracy. The experimental results demonstrate that our proposed model offers major advantages over previous works [19, 20], especially in terms of detection recall and accuracy. Moreover, the detection speed is roughly increased by a factor of 1.9–2.7.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related studies about object detection algorithms based on CNNs for traffic sign detection and recognition. The detailed description of our algorithm and model architecture are thoroughly described in Section 3. Section 4 discusses the overall experimental results of our proposed traffic sign detection algorithm. Finally, certain conclusions and further remarks are summarily presented in Section 5.

2 Related Work

2.1 Object Detection Algorithms Based on CNNs

Object detection is a computer vision task to exactly locate the bounding boxes of objects in a given picture and mark the category of the objects. As deep learning methods are widely applied to computer vision over the past few years, most of the state-of-the-art object detection algorithms, such as R-CNN series works [10–13], OverFeat, SSD [21], and YOLO series works [22–24], have used CNNs and achieved impressive success in various fields of object detection and classification.

Faster Region-based Convolutional Neural Network (Faster R-CNN) and Region-based Fully Convolutional Networks (R-FCN) have a made a quantum leap forward among the series networks of R-CNN based on regional proposals, which are known as the two-stage method and have certain advantages in terms of accuracy of object detection base on the deep learning method. Both of them optimize the detection speed and precision by employing region proposal network (RPN) instead of selective search algorithms. R-FCN proposed a concept of the position-sensitive score map which introduced location information of the target into the ROI pooling. However, the two-stage approaches are

computationally inefficient, and the computation process is resource-intensive .

Furthermore, Sermanet et.al. team proposes an end-to-end network OverFeat and has taken the crown of ILSVRC2013. They synthesize identification, localization and detection tasks into a CNN-only based framework, rather than recognition after the region is proposed. The SSD algorithm introduces multi-scale feature maps for detection by adding convolutional feature layers to the end of the truncated base network. The detection speed of SSD is considerably faster than faster-RCNN with the same detection accuracy. Joseph J. et al. sequentially proposed YOLO series works [22–24]. YOLO model converted the target classification and localization tasks into regression problems, and the detection speed of it has increased dramatically. With new multi-scale predictions techniques, the state-of-art model, named YOLOv3, not only provides high detection accuracy and speed, but also greatly improves the performance of small object detection.

2.2 Traffic Sign Detection and Recognition

Since the traffic signs play an important role in guiding driving behavior on the road, traffic sign detection has been considered essential in automated driving. Traditional visual object detection methods failed to adapt to the complex environment, multiple and small target sample detection, and real-time response requirements in real traffic scenarios. Benefiting from the rapid development of vehicle networks and intelligent transportation systems, extensive research on real-life traffic sign detection has been conducted in recent years.

Zhu et al. [19] at Tsinghua University create a large-scale Chinese traffic sign benchmark named Tsinghua-Tencent 100 K, which covers more realistic scenes, traffic sign categories, and image instances. Furthermore, they build on the work of Huval et al. [25] by improving OverFeat framework and introducing three streams after the final branch to simultaneously detect and classify traffic signs with a recall rate of 0.91 and an accuracy of 0.88. In [26], Li et al. innovatively propose a Perceptual GAN (PGAN) model to deal with the low-resolution object detection problem. In, Meng et al. decompose the original large image into patches and successfully applied Small-Object-Sensitive-CNN (SOS-CNN) to the image pyramid. Both of them efficiently boost traffic sign detection performance, especially for small objects. Tian et.al [18] utilize recurrent attention for multi-scale analysis and local context information, which effectively improved recall and precision by about 1.0%. Yang et al. [17] combine attention network and Fast R-CNN to classify traffic signs robustly according to color features. Lu et al. [27] propose a visual Attention Proposal Model (APM) to locate attention regions and further predict the classification and bounding boxes of objects in each region of interest. Song et al. [20] compress the CNN model to improved efficiency for the small object detection without reducing recognition accuracy. All these works made some improvements in the speed of object detection. Jain et al. [28] use a genetic algorithm (GA) to optimize the number of epochs and hype-parameters during the training period to refine the classification accuracy of traffic targets. Kim et al. [29] report a novel feature embedding scheme with the representative class templates, which perform well on unseen traffic sign recognition.

However, few researchers demonstrate that their work has achieved satisfactory results in proper balance between accuracy and speed, taking into account of real-life traffic sign especially small object detection tasks. Unlike Pascal VOC, COCO, or other common object detection tasks, real-life traffic sign detection needs to deal with challenging objects that make up a much smaller proportion of the image than before. For example, in the Tsinghua-Tencent 100 K dataset, most traffic signs may be only 50×50 pixels or less, scattered across a 2048×2048-pixel image, each one just filling less than 0.1% of the image. Note that even bigger signs with size of 400×400 pixels occupy only 3.8% of the total image area. In other widespread datasets, it may account for close to 20% of each image. As a result, many reported models with superior performance are not considered to be directly applicable for traffic sign detection applications. In order to improve the detection performance, it is important to consider the characteristics of traffic signs when designing the network architecture. Due to the high efficiency and accuracy of the YOLOv3 detector, especially for small targets. We propose an improved end-to-end method based on YOLOv3 model. It has excellent performance in the traffic sign detection tasks.

3 Improvement of Network Structure

3.1 Method Overview

Although the CNN has achieved outstanding achievements in the field of object detection, no model that is well qualified for the task of traffic sign detection, both in terms of detection accuracy and speed. In this paper, we construct a deep neural network based on YOLOv3 for traffic sign detection.

The YOLOv3 model is one of the state-of-the-art object detection systems. It benefits from the advantages of the residual network, which allows the construction of deeper network to improve the nonlinearity in the network, and significantly improves the classification and detection effects. YOLOv3 brings up three feature maps extracted from different scales to predict objects. Small feature maps provide semantic information, and large ones has finer-grained information. The different features are concatenated by stacking adjacent features into different channels using a routing layer. When given an input image, the input image is partitioned

into S × S grids. Each grid cell predicts three bounding boxes. Each predicted bounding box contains 4 coordinates (t_x, t_y, t_w, t_h) to determine the location relative to the center of the grid. It then predicts a confidence score (t_o), which is the probability of the grid will detect an object for each bounding box using logistic regression [24]. Assuming that the offset of the cell from the upper left corner of the image is (c_x, c_y), and that the width and height of the bounding box prior are p_w, p_h, then the predictions correspond to Equ. (1):

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \\ \sigma(t_o) &= P_{r(object)} \times IOU(b, object) \end{aligned} \tag{1}$$

The probability should be 1 when the prediction bounding box coincides with the ground truth location to the cutoff threshold, and 0 otherwise. We can get the confidence value of each bounding box through prediction, and select the bounding box with the highest confidence in each grid cell to predict the object in the image. Each box uses a multi-label classification to predict which classes the bounding box may contain. The binary cross-entropy loss is used for the class predictions during training, instead of using softmax. The YOLOv3 model combines a large number of previous works and performs well in object detection, especially for small targets. The detection speed of this method is one of the fastest algorithms available. To address the characteristics of the traffic sign detection task, we put forward the following improvements on the YOLOv3 network.

Figure 1 provides the architecture of our traffic sign detection method. First, the input image is cropped into 19 subgraphs according to the certain grids, and passed into the detection network of the framework as one batch. Secondly, the optimal bounding boxes and classes of the targets are generated from the prediction results of four different scale prediction branch by non-maximum suppression (NMS). Finally, the results of each subgraph are integrated into the original image as the output of the entire detection network. Next, we will discuss the detailed design of our method.

3.2 Network Pruning

YOLOv3 is a universal object detection model with excellent performance. For the practical scenario task of traffic sign detection mentioned in this article, appropriate deletion of the backbone network is an effective improvement method. The network structure of YOLOv3 feature extractor has 53 stacked convolutions, called Darknet-53, which can be divided into five sections, including 1, 2, 8, 8, and 4 residual blocks respectively (represented as 1–2–8-8-4 in Table 1). In order to obtain the optimal network structure, we scaled down the different parts of the network structure and analyzed the detection performance and model size of the corresponding modified network. The experimental results are shown in Table 1. It is noted that the detection performance of the 1–2–4-4-4 structure is relatively close to the original network, and the amount of model parameters is significantly reduced.

3.3 Four Scale Prediction Branches

It is known that the shallow feature map in CNN contains richer location information, and is more suitable for detecting small targets with low resolution and inconspicuous features. Considering the uneven distribution of the target sizes in the Tsinghua Tencent's 100 K dataset, most of them are small-sized targets. The three scale prediction branches in the original YOLOv3 structure are in the lower levels of the network, and the 9 anchor priors obtained using k-means clustering may be more inclined to large-sized targets because the deep network can provide more semantic information. In order to improve the detection performance of large targets, we added a fourth scale prediction branch to expand the detection range and enrich the feature maps for multi-scale prediction.

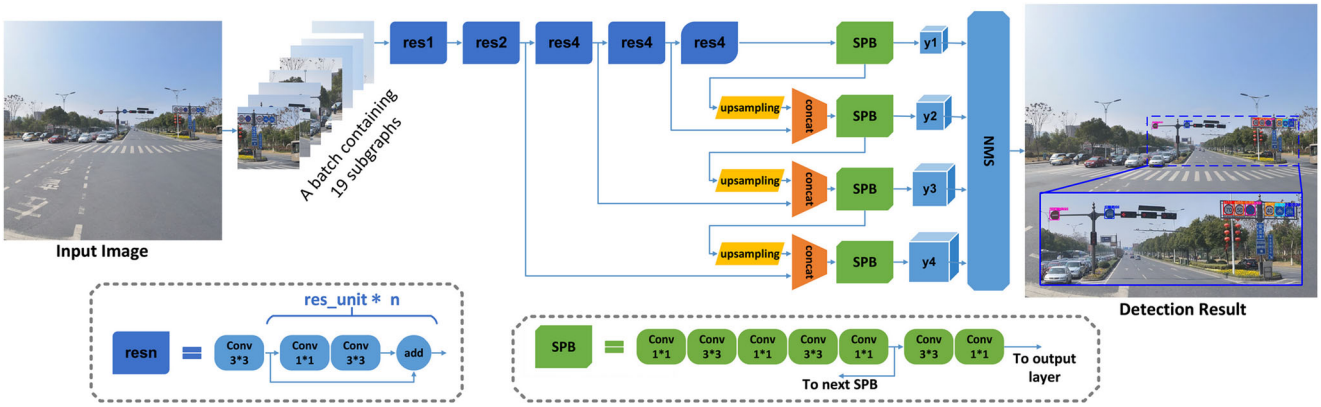


Figure 1 The overview of our traffic sign detection system. The input image is divided into 19 subgraphs as one batch and processed by the convolutional layers, whose architecture is improved from YOLOv3. The final output module combines the predictions of all subgraphs.

Table 1 Comparison of detection performance of the network structures with different degrees of pruning. (In %).

Network structure		All	Small	Medium	Large	Model size
1–2–8-8-4	Recall	90.82	89.25	93.11	85.59	236 M
	Accuracy	90.47	88.29	93.27	85.48	
1–2–4-4-4	Recall	90.80	89.70	92.66	85.84	211 M
	Accuracy	90.45	87.83	93.40	86.84	
1–2–2-2-4	Recall	87.72	85.75	90.67	80.87	198 M
	Accuracy	90.71	89.07	93.24	84.42	
1–4–2-2-4	Recall	88.40	87.81	91.18	83.04	204 M
	Accuracy	90.34	88.51	93.14	83.89	

Traditional K-means clustering method uses the Euclidean Distance function, with larger boxes generate more erroneous clusters than smaller boxes. To this end, what we really want is priors that lead to good IOU scores, independent of the size of the box. The distance function can be computed as Equ. (2):

$$d(box, centroid) = 1 - IOU(box, centroid) \tag{2}$$

Then 12 clusters we get on the Tsinghua-Tencent 100 K dataset were: (16 × 18), (20 × 20), (22 × 24), (26 × 28), (32 × 34), (41 × 44), (53 × 58), (71 × 77), (98 × 107), (127 × 132), (161 × 158), (230 × 211). The fourth scale prediction branch is attached to the end of the second residual block section, while the features maps obtained previously in the network is merged in 2x up-sampling, similar to the connection of intrinsic YOLOv3. In this way, the smallest three anchor boxes are performed by the new fourth scale prediction branch, while the first scale prediction branch can easily predict larger objects using anchor boxes with additional scales. Large-size objects are easier to predict due to the closer proximity of the anchor boxes.

3.4 Loss Function Improvement

The loss function is one important criterion for evaluating the performance of a model. The loss function in YOLOv3 is a simple addition of differences, including coordinate errors, confidence errors, and the classification errors. The loss function can be expressed by the following Equ. (3):

$$Loss = Err_{coor} + Err_{conf} + Err_{cls} \tag{3}$$

where the loss function is simply summed, and the weights of coordinate errors, confidence errors, and classification errors are equal to 1. The confidence is defined as Equ. (4):

$$Confidence = p_r(Object) \times IoU_{pred}^{truth}, p_r(Object) \in \{0, 1\} \tag{4}$$

However, in traffic sign detection task, the targets are substantially divided into three categories: “warning”, “prohibited” and “indication”, with yellow, red, and blue outer circles, respectively. The internal patterns for each category are somewhat different. In general, the models have various misclassifications when predicting the target. In particular, this misclassification is more pronounced when the target is small. Considering the great similarity of the categories in the Tsinghua-Tencent 100 K dataset, it is more challenging to get the network to accurately recognize the categories of traffic signs than to predict the location of the targets. In the loss function of original YOLOv3, the location error weight is equal to classification error, which is unreasonable to directly apply to the traffic sign detection task. In order to improve classification mistakes, we attempt to increase the penalty for class prediction, and the confidence loss and classification loss among each scale prediction are multiplied by the corresponding weights. Therefore, the overall loss function can be defined as Equ. (5):

$$\begin{aligned} Loss \quad &= \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{obj} [(x_i - \hat{x}_i) + (y_i - \hat{y}_i)] \\ &+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\ &+ \lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{obj} \left[-\left(C_i \log \hat{C}_i \right) - (1 - C_i) \log \left(1 - \hat{C}_i \right) \right] \\ &+ \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{noobj} \left[-\left(C_i \log \hat{C}_i \right) - (1 - C_i) \log \left(1 - \hat{C}_i \right) \right] \\ &+ \lambda_{cls} \sum_{i=0}^{S^2} \mathbf{1}_{i}^{obj} \sum_{c \in classes} [-p_i(c) \log(\hat{p}_i(c)) - (1 - p_i(c)) \log(1 - \hat{p}_i(c))] \end{aligned} \tag{5}$$

where: S^2 is defined as the number of grids in the input image; B is defined as the number of bounding boxes generated in each cell; (x, y, w, h) are defined as the coordinates of the center, width, and height of the prediction box; C is defined as the confidence of the prediction box; $p(c)$ is defined as the probability that the object belongs to class c; And the parameter λ_{obj} is the weight of the confidence loss that the predicted target center is within the grid cell, with the value of 5. We set $\lambda_{noobj} = 1$ instead of being equal to λ_{obj} is used to fix the error mentioned in YOLO [22]. As the confidence value of the grid cells that do not contain objects in each image is approximately 0, which distorts the influence of the confidence error of the grid containing objects on the calculation of the gradient of the network parameter. λ_{cls} is the weight of the categorical loss and $\lambda_{cls} = 5$ is chosen in this paper. It is worth mentioning that we adopted the grid search method to determine the weights of each part of the loss function, including λ_{obj}, λ_{noobj}, and λ_{cls}. Specifically, we limit the search range of the three weights to a reasonable range of [0.1, 0.5, 1, 3, 5, 10], then tried all possible parameter combinations. In the experiment, we chose the F1 score

as a moderate function to evaluate the detection performance of the model, which takes both the accuracy and recall into account. Finally, we selected the optimal set of hyperparameters as weights in the loss function and retrained the model.

3.5 Model Construction

In this paper, we have explored several models according to the characteristics of traffic signs, and made some improvements based on the YOLOv3 model. Finally, three networks are constructed to address the challenging traffic sign detection problem. The network structures of the proposed models are illustrated in Fig. 2.

YOLOv3 model is used as a reference comparison group to verify whether the proposed model has improved the performance of traffic sign detection. The network contains 23 residual blocks, divided into 5 residual sections, using as feature extractor and three scale prediction branches to process the obtained features.

The YOLOv3-Pruning network is pruned based on the YOLOv3 network, by removing the third and fourth residual blocks 4 convolutional layers, respectively. And the scale predicted branches remains unchanged. This ensures the feature extraction capability of the network while reducing the number of network parameters to some extent.

The YOLOv3-4SPB network has four scale prediction branches by adding a new one to YOLOv3. The fourth scale prediction branch is connected after the second residual section, which also incorporates the up-sampled features of the third scale prediction branch as in the previous operation. Finally, the YOLOv3-4SPB network in this paper predicts bounding boxes at four different scales: 104 × 104, 52 × 52, 26× 26, and 13 × 13.

The YOLOv3-Final network has not changed the network structure in the feature extraction part but increased penalties for confidence loss and class loss in the calculation of the loss function. We multiplied the confidence loss for the predicted target center within the grid cell and the classification loss by a weight of 5, respectively.

4 Experiments and Discussion

In this paper, we conducted experiments on the Tsinghua-Tencent 100 K dataset which was released in 2016 as a benchmark for large-scale traffic signs of China. [19] The dataset provides 100,000 real-world images with a resolution of 2048 × 2048, and containing 30,000 traffic-sign instances. These images cover large variations in illuminance and weather conditions. As shown in Figure 3(a) and Fig. 3(b), this dataset has an obvious uneven distribution in the number of instances per category and target size. In which, most instances appear in relatively few classes and small traffic-signs are most common. To make a better comparison with other methods, we also ignored classes with fewer than 100 instances, leaving only 45 categories following [19]. Data enhancement is also adopted to balance the severely uneven numbers of instances in the different categories. We used a variety of data enhancement methods, including appropriate color dithering, image blurring, fancy PCA, rotating, and scaling randomly, so that the instances of each class in one epoch during training are above 1000 samples. Following the division method in the original datasets, our training set contains 6105 images, including 15,698 traffic sign instances, and the test set contains 3071 images, including 7812 instances.

All the experiments were completed with the environment of two NVIDIA Tesla K80 GPUs with 12GB memory, Ubuntu 16.04 operating system, cuda9.0, python 3.6. In the

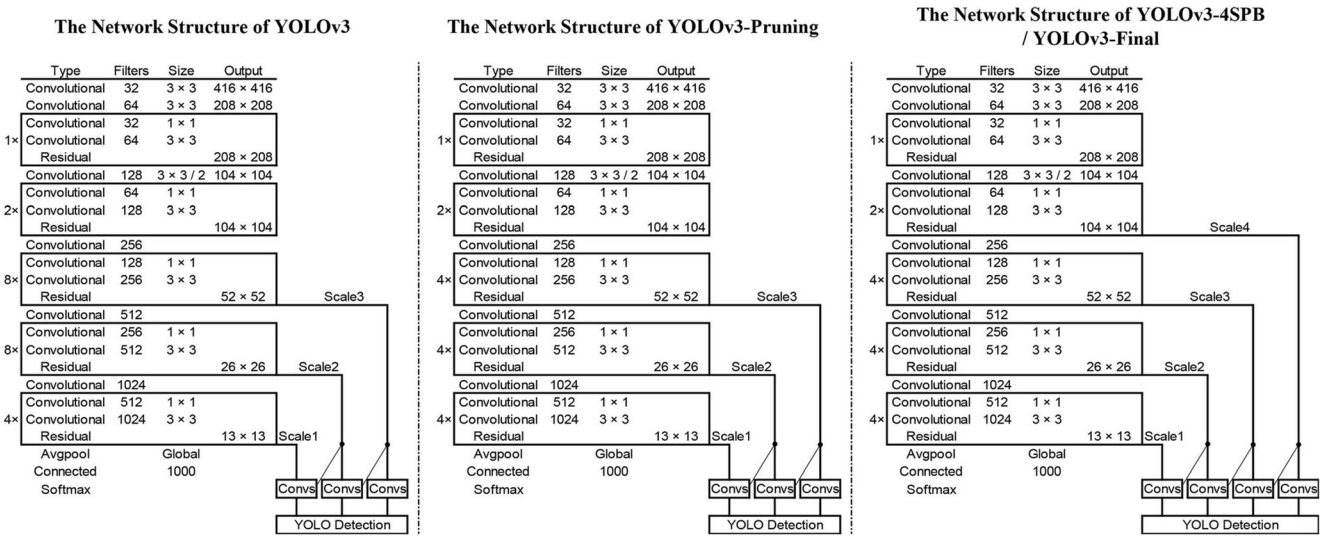


Figure 2 The network structures of YOLOv3, YOLOv3-Pruning, YOLOv3-4SPB, and YOLOv3-Final.

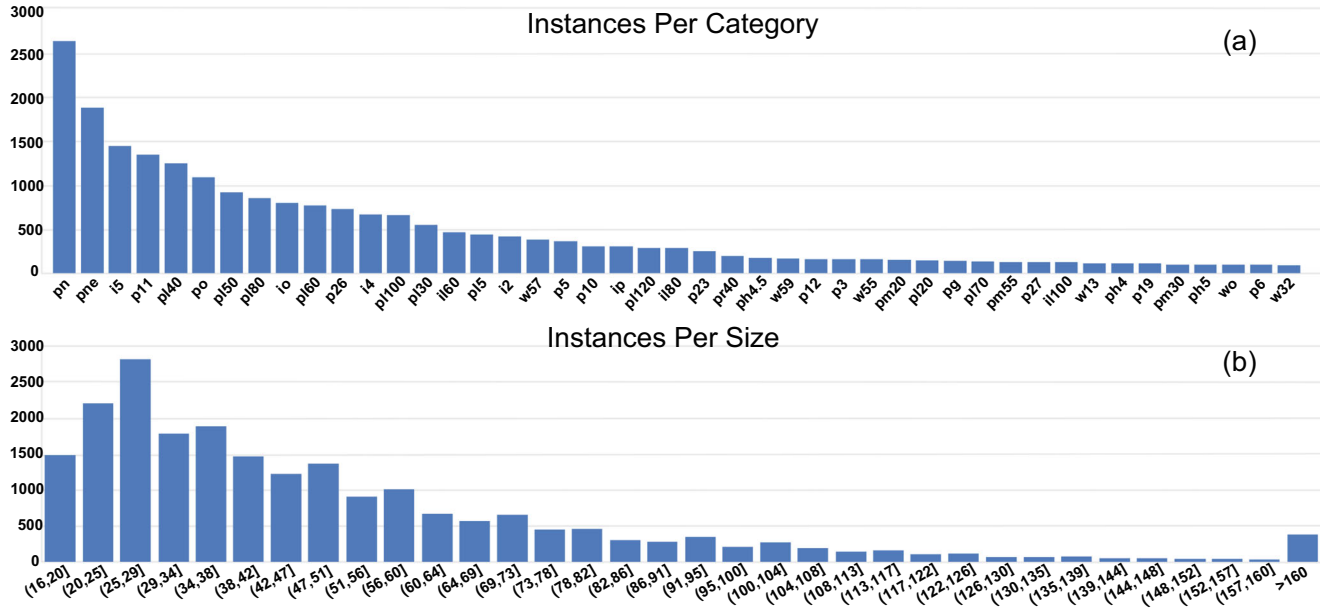


Figure 3 (**a**). The number of instances in each class, for classes with more than 100 instances. [19] (**b**). The number of instances of each size. [19]

process of model training, the original 2048 × 2048 high-resolution images were cropped to subgraphs with a resolution of 416 × 416, which contains all the instances of the original image. The models in our experiments were initialized using some parameters of the published model of Darknet-53 and retrained on the Tsinghua-Tencent 100 K dataset. The training process was divided into three stages. First, all scale prediction branches were trained on the feature extraction network Darknet-53 for 40 epochs, then replaced with our feature extraction networks, and continued training for 60 epochs. The first two stages used Adam optimizer. Finally, all parameters in network layers are released, and additional 30 epochs are trained using Stochastic Gradient Descent (SGD) with momentum of 0.9 and a weight decay of 0.005 for finer tuning.

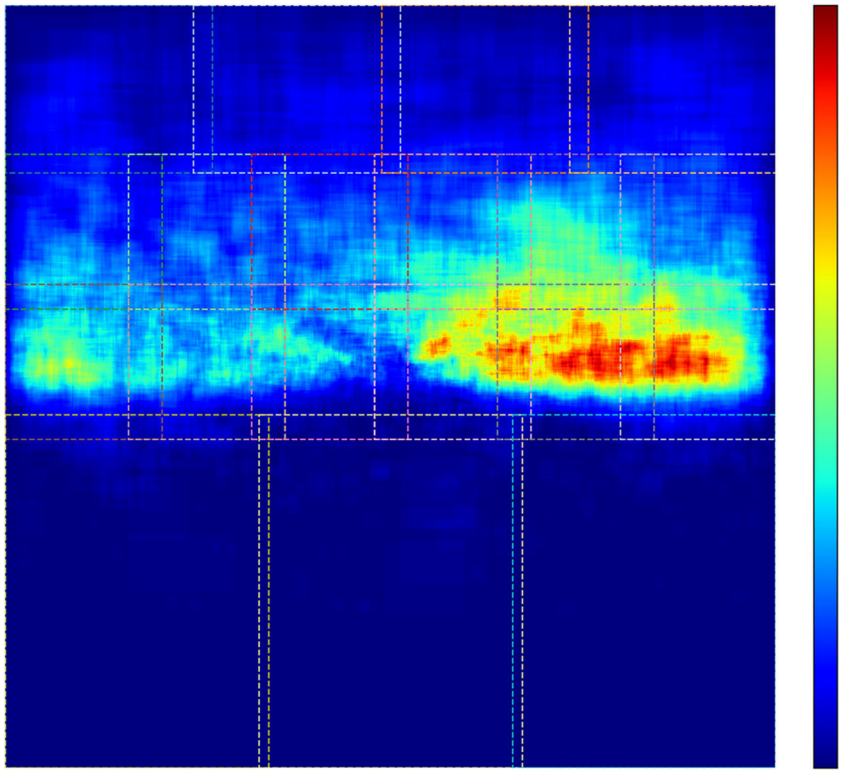


Figure 4 The heat map of traffic sign density in the Tsinghua-Tencent 100 K dataset. The 19 dashed boxes in the figure are the grids used to divide the original 2048 × 2048 high-resolution image into subgraphs for the detection task.

Table 2 Detailed experimental arrangement and related strategies in ablation studies of our proposed method.

	Distribution heat map	Network pruning	Four-scale prediction branch	Loss function balance
YOLOv3	√			
YOLOv3-Pruning	√	√		
YOLOv3-4SPB	√	√	√	
YOLOv3-Final	√	√	√	√

As mentioned earlier, the Tsinghua-Tencent 100 K dataset is a special dataset for traffic sign. In this benchmark, the size of the objects of interest is much smaller than in previous benchmarks. How to detect such small size targets in such a large resolution image is an important issue. If the original 2048 × 2048 high-resolution image is directly resized to a low-resolution image of 416 × 416, this will make the size of the pretty small objects of interest in the image shrink by a factor of nearly 24. It will be a huge challenge for CNN networks to detect the targets with so few pixels information. Therefore, the large, high-resolution image is broken down into small fixed-sized patches before it is fed into the detection network. In this way, the pixel information of the target in the image is retained in several low-resolution subgraphs, but at the cost of reduced detection speed.

In addition, we have analyzed the distribution of traffic sign locations in the images from the Tsinghua-Tencent 100 K benchmark. These images were captured from both vehicles and shoulder-mounted equipment. [19] And the locations of the captured traffic signs were not randomly distributed throughout the image due to the relatively fixed perspective. The heat map of the traffic sign density in the Tsinghua-Tencent 100 K dataset is given in Fig. 4. Instead of even distribution among the image, the traffic signs are concentrated in the upper part of the middle of the image, and few targets can be seen in the lower part of the image. According to the characteristics of the traffic sign distribution, we use non-uniform partitioning when dividing the image into subgraphs. As illustrated in Fig. 4, the grid in areas with densely distribution of signs is tighter than in places with sparse target distribution. In order to ensure the accuracy of edge target recognition, we added some margin when dividing the grid. This non-uniform division method not only ensures the accuracy of traffic sign recognition but also reduces the complexity of network computation. Finally, we obtain one batch containing 19 subgraphs which are resized to 416 × 416 to replace the original high-resolution image for the detection task. The object detection results in all subgraphs are then combined according to the corresponding coordinates and the complete detection results are output. With this non-uniform division method, the YOLOv3-Final model needs about 1.1 s to process these 19 subgraphs. Instead of the conventional division method, such that a high-resolution picture with 2048 × 2048 needs to be divided into 36 sub-images of 416 × 416, and it takes an average of 1.7 s to process these images.

4.1 Detection Performance

We inherited the previous metrics employed for the Microsoft COCO benchmark, and then separated all traffic-signs into three parts of small, medium, and large size. The purpose of dividing the traffic-signs into three categories is to better express the detector's generalization ability and robustness to different size targets. And we made a more detailed comparison between the detection performances of the traffic sign detection model which we constructed for different sizes. In the Tsinghua-Tencent 100 K dataset, even large scale objects occupy only 96–400 pixels, and the average proportion is less than 1% of the entire image with a resolution of 2048 × 2048. Compared to the proportion of each image could approach 20% in other widespread datasets, so this kind of traffic sign detection task can still be considered as the detection of small objects. For binary classification problems, samples can be divided into four types: true positive (TP), false positive

Table 3 The performance comparison of different traffic signs detectors on the Tsinghua-Tencent 100 K dataset. (In %).

Object size		All	Small	Medium	Large	Model size
YOLOv3	Recall	90.82	89.25	93.11	85.59	236 M
	Accuracy	90.47	88.29	93.27	85.48	
YOLOv3-Pruning	Recall	90.80	89.70	92.66	85.84	211 M
	Accuracy	90.45	87.83	93.40	86.84	
YOLOv3-4SPB	Recall	91.44	90.30	92.99	88.27	212 M
	Accuracy	91.91	88.81	95.06	89.52	
YOLOv3-Final	Recall	92.25	90.61	94.29	88.78	212 M
	Accuracy	93.80	91.21	96.31	92.19	

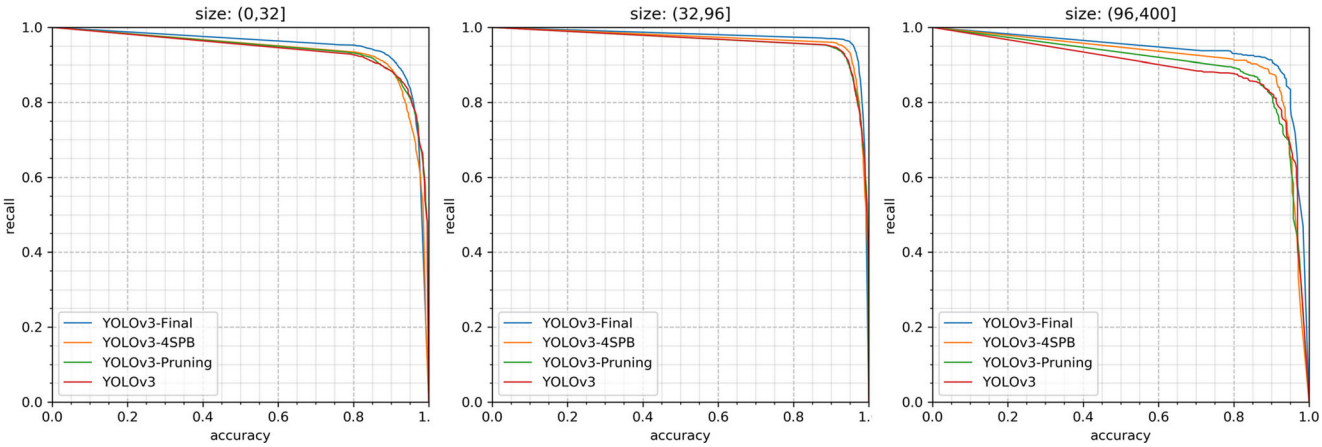


Figure 5 Accuracy-recall curves of simultaneous traffic sign detection performance on Tsinghua-Tencent 100 K, for small, medium, and large signs.

(FP), true negative (TN), and false-negative (FN), according to the combinations of the true class and predicted class of the learned objects.

Accuracy (A) and recall(R) are defined as Equ. (6) and Equ. (7):

$$\mathrm{A} = \frac{TP}{TP + FP} \tag{6}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{7}$$

Since our network is made up of multiple modules, and the improvement of the model is also composed of multiple operations, we performed some ablation studies to verify its effectiveness and achieve the final performance. The detailed experimental arrangement and related strategies are listed in Table 2.

The comparison of experimental results evaluated on test sets for Jaccard similarity coefficient 0.5 are shown in Table 3. From which, it is noted that pruning has a significant impact on medium object detection, with its accuracy rate decreasing from 93.11% to 92.66%. This is because the deletion of the middle 8 residual layers, which reduces the amount of feature information of the medium object extracted by the network. Considering that the detection performance of large targets after pruning is better than that of the initial network, this may be due to overfitting in the intermediate feature extraction of the initial network, and the deeper network learned more correct features after pruning. However, for all objects, we can see that the recall and accuracy achieved by YOLOv3-Pruning are almost the same as YOLOv3: 90.82% vs 90.80%, and 90.47% vs 90.45%, respectively. In other words, YOLOv3-Pruning has comparable detection performance in this task, even though a total of 8 residual layers are removed from networks. This also implies that, for this traffic sign detection task, the YOLOv3 network is over-parameterized. Moreover, the model size of YOLOv3-Pruning is reduced by approximately 11.3%. A smaller model size means less computational resource consumption and shorter response times. It is especially important for mobile platforms such as automotive applications.

Furthermore, we constructed the YOLO-4SPB model by adding a scale prediction branch to YOLOv3-Pruning. Obviously, the YOLOv3-4SPB model outperforms the former in terms of both recall and he accuracy: 91.44% vs 90.80% and 91.91% vs 90.45%. Especially in the aspect of large target detection, the performance of the model has greatly improved:

Table 4 Comparison results of detection performance for different sizes of traffic signs on Tsinghua-Tencent 100 K dataset. (In %).

Object size		All	Small	Medium	Large	Model size	detect time
Fast R-CNN [11]	Recall	56	24	73	86	342 M	–
	Accuracy	50	45	50	55		
Faster R-CNN [12]	Recall	–	50	84	91	357 M	–
	Accuracy	–	24	66	81		
Zhu et al. [19]	Recall	91	87	94	88	418 M	3.0 s
	Accuracy	88	82	91	91		
Song et al. [20]	Recall	–	88	93	89	375 M	2.1 s
	Accuracy	–	85	91	92		
YOLOv3-Final (Ours)	Recall	92	91	94	89	212 M	1.1 s
	Accuracy	94	91	96	92		

Table 5 Comparison results of detection performance for 10 typical classes of traffic signs on Tsinghua-Tencent 100 K dataset. (In %).

Class		i2	i5	io	p10	p19	p26	p6	pm55	w55	w57
Fast R-CNN [11]	Recall	0.32	0.69	0.65	0.51	0.79	0.6	0.54	0.79	0.5	0.56
	Accuracy	0.68	0.71	0.51	0.54	0.67	0.67	0.66	0.57	0.7	0.38
Zhu et al. [19]	Recall	0.82	0.95	0.89	0.95	0.94	0.93	0.87	0.95	0.72	0.79
	Accuracy	0.72	0.92	0.76	0.78	0.53	0.82	0.87	0.6	0.86	0.95
Ours	Recall	0.91	0.95	0.85	0.92	0.97	0.9	0.9	0.95	0.94	0.94
	Accuracy	0.86	0.97	0.89	0.94	1	0.94	0.95	0.95	0.85	0.88

recall and accuracy increased from 85.84% and 86.84% to 88.27% and 89.52%. The poorer detection performance for large target using the YOLOv3 network directly may be due to the fact that small objects are most common in the Tsinghua-Tencent 100 K dataset, whereas the prior anchor boxes obtained using k-means clustering may be closer to the small size. Nevertheless, the fourth scale prediction branch in the YOLO-4SPB model provides more elaborate anchor boxes and achieves good results for large objects, which are in line with our hypothesis.

After fine-tuning the weights of various errors in the loss function, we achieved a significant improvement in the detection performance of the model. As presented in Table 3, the YOLOv3-Final model achieved 92.25% recall and 93.80% accuracy at a Jaccard similarity coefficient of 0.5 when all sizes of traffic signs are considered together,. It is important to note that this improvement is not biased towards the specified scale object. The accuracy and recall are all better than the previous model in terms of performance for each size target. An illustration of accuracy-recall curves for these methods is provided in Fig. 5, which can further demonstrate the effectiveness of the proposed the YOLOv3-Final model in performing well on targets of different sizes. Therefore, the YOLOv3-Final model was selected as the final model for traffic sign detection.

4.2 Comparisons and Discussions

In order to better demonstrate the effectiveness of our proposed method on traffic sign detection, we also evaluated our approach with other state-of-the-art methods of Zhu et al. [19] and Song et al. [20], whose main work was also completed on the Tsinghua-Tencent 100 K dataset. Table 4 shows the detection performance of our results in comparison with previous publications in terms of average recall and accuracy for each target scale. The overall results indicate that our approach achieves remarkable improvements. In particular, our proposed model increases the recall and accuracy by about 4% and 9% when used to detect small traffic signs compared to Zhu et al. [19], and increased by about 3% and 6% compared to Song et al. [20], respectively. In addition, the detection performance for medium and large targets is also some improved. We speculate that this is due to the adoption of up-sampling and fusion methods similar to FPN in the feature extraction layer, which enables local feature fusion between feature maps of different scales and significantly improves the detection performance for small targets. More scale prediction branches facilitate the fusion of shallower features. These shallow features are fused with the up-sampled deep features, which solves the problem that the dimension of deep features is too small. At the same time, more network layers deepened the depth of the network, and improve the effect of

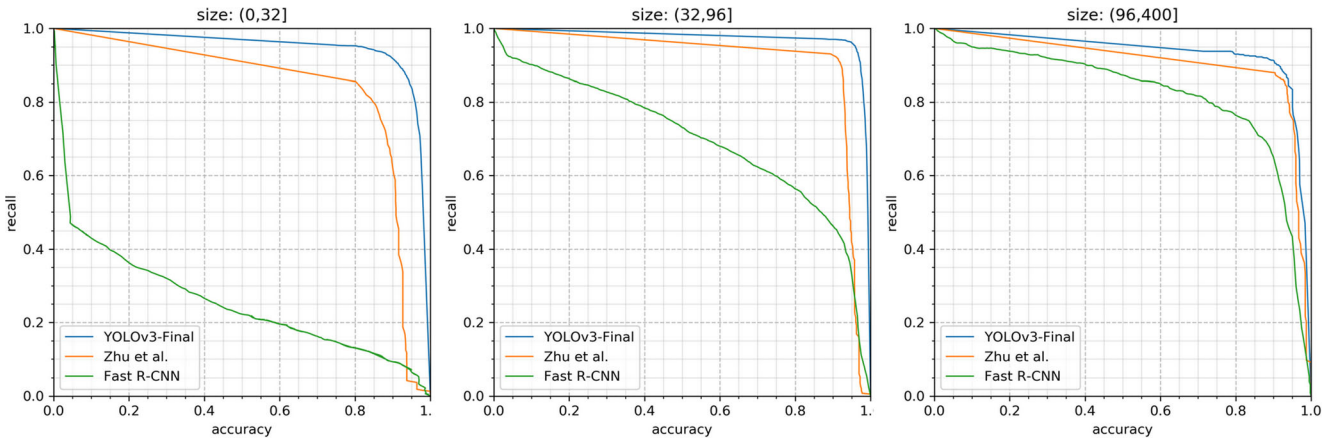


Figure 6 Performance comparison of accuracy-recall curves of Fast R-CNN, zhu et al. [19]. and our approach for different object sizes.

feature expression is. More performance comparisons of accuracy-recall curves for different object sizes are illustrated in Fig. 6. The curves of Fast R-CNN and Zhu et al. are adopted from [19]. We also give accuracy and recall for some typical classes in Table 5. We achieve significant improvements in several categories. This can be further noted by the fact that our proposed model achieves such a large improvement in these three types of targets.

Another advantage of our approach is that both model size and detection speed exceed previous reports. As indicated in Table 4, it can be observed that the model size of YOLOv3-Final is only half of that of Zhu et al. [19], and nearly 43% less than that of Song et al. [20]. The detection speed is also an important factor besides accuracy in practical applications. Especially for this traffic scenario, the shorter response time of the model means that the driver can make corresponding decisions earlier. This reduces the probability of traffic accidents to some extent. In order to locate and classify all traffic signs in an image with 2048 × 2048 pixels, our proposed model requires approximately 1.1 s. In comparison, Song et al. [20] and Zhu et al. [15] take approximately 1.9x and 2.7x times longer to complete the task, respectively. This is mainly benefiting from the excellent feature extractor (Darknet-53) in yolov3. Which has better performance and achieves higher measured floating-point operations per second [24]. In addition, it is related to our trick mentioned above to narrow down the scope of detection using the heat map.

Finally, taking into account factors such as detection accuracy, model size, detection speed, our proposed model not only performs well in various types of traffic signs, especially for small target detection, but also possesses the smallest model size, saves the computational cost and significantly improves detection speed. From the comparison results in these aspects, our approach is more effective in boosting small traffic sign detection than the state-of-the-art methods of Zhu et al. and Song et al..

5 Conclusion

In this paper, we proposed an efficient to address the challenging problem of small traffic sign detection in real-life, which consists of an improved model YOLOv3-Final and its corresponding efficient algorithm. The YOLOv3-Final model is applied to the public Tsinghua-Tencent 100 K dataset and exhibits good robustness for small traffic sign detection. In order to improve the network's ability of extracting the traffic sign features accurately, various optimization strategies are elaborately presented, including network pruning, the fourth scale prediction branch, and loss function modulation. The experimental results demonstrate that the YOLOv3-Final model has a better performance compared to the original YOLOv3 model, and outperforms over state-of-the-art methods in terms of accuracy and speed. Therefore, our proposed method is hopefully used for intelligent traffic sign detection system.

For future work, we will compress the model and optimize the detection algorithm, and add more visual objects to meet the requirements of light-weight and real-time in practical autonomous driving applications.

Acknowledgments This work was supported by National Key Research and Development Project under grant 2019YFC0117302, National Natural Science Foundation of China under grant 62004201, Key Projects of Bureau of International Cooperation Chinese Academy of Sciences under grant 184131KYSB20160018, National Natural Science Foundation of China under grant No. U1831118, and NSFC Youth Fund under grant No. 61704179.

References

1. Liu, H., Liu, Y., & Sun, F. (2014). Traffic sign recognition using group sparse coding. *Information Sciences, 266*, 75–89.
2. Abdi, L., & Meddeb, A. (2018). Spatially enhanced bags of visual words representation to improve traffic signs recognition. *Journal of Signal Processing Systems, 90*(12), 1729–1741.
3. Chen, Y., Zhao, D., Lv, L., & Zhang, Q. (2018). Multi-task learning for dangerous object detection in autonomous driving. *Information Sciences, 432*, 559–571.
4. Maldonado-Bascón, S., Lafuente-Arroyo, S., Gil-Jimenez, P., Gómez-Moreno, H., & López-Ferreras, F. (2007). Road-sign detection and recognition based on support vector machines. *IEEE Transactions on Intelligent Transportation Systems, 8*(2), 264–278.
5. Jang, C., Kim, C., Kim, D., Lee, M., & Sunwoo, M. (2014). Multiple exposure images based traffic light recognition. In *2014 IEEE Intelligent Vehicles Symposium Proceedings,* (pp. 1313-1318).
6. De Charette, R., & Nashashibi, F. (2009). Real time visual traffic lights recognition based on spot light detection and adaptive traffic lights templates. In *2009 IEEE Intelligent Vehicles Symposium,* (pp. 358-363).
7. Cai, Z., Gu, M., & Li, Y. (2012). Real-time arrow traffic light recognition system for intelligent vehicle. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV),* (pp. 1).
8. Bangquan, X., & Xiong, W. X. (2019). Real-time embedded traffic sign recognition using efficient convolutional neural network. *IEEE Access, 7*, 53330–53346.
9. Liu, Z., Du, J., Tian, F., & Wen, J. (2019). MR-CNN: A multi-scale region-based convolutional neural network for small traffic sign recognition. *IEEE Access, 7*, 57120–57128.
10. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* (pp. 580-587).
11. Girshick, R. (2015). Fast r-cnn. *In Proceedings of the IEEE international conference on computer vision*, (pp. 1440-1448).
12. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems,* (Vol. 39, pp. 1137-1149, Vol. 6).
13. Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems,* (pp. 379-387).

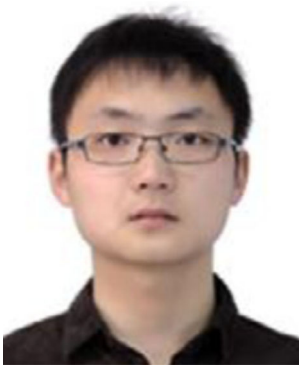
14. Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. J. I. J. O. C. V. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*(2), 303–338.
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.
16. Meng, Z., Fan, X., Chen, X., Chen, M., & Tong, Y. (2017). Detecting small signs from large images. In *2017 IEEE International Conference on Information Reuse and Integration (IRI),* (pp. 217-224).
17. Yang, T. T., Long, X., Sangaiah, A. K., Zheng, Z. G., & Tong, C. (2018). Deep detection network for real-life traffic sign in vehicular networks. *Computer Networks, 136*, 95–104.
18. Tian, Y., Gelernter, J., Wang, X., Li, J., & Yu, Y. (2019). Traffic sign detection using a multi-scale recurrent attention network. *IEEE Transactions on Intelligent Transportation Systems., 20*, 4466–4475.
19. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., & Hu, S. (2016). Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 2110-2118).
20. Song, S., Que, Z., Hou, J., Du, S., & Song, Y. (2019). An efficient convolutional neural network for small traffic sign detection. *Journal of Systems Architecture., 97*, 269–277.
21. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). SSD: Single shot MultiBox detector. *European conference on computer vision*, 21–37.
22. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognitionn,* (pp. 779-788).
23. Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* (pp. 6517-6525).
24. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, doi:1804.02767.
25. Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., et al. (2015). An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*.
26. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., & Yan, S. (2017). Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 1222-1230).
27. Lu, Y., Lu, J., Zhang, S., & Hall, P. (2018). Traffic signal detection and classification in street views using an attention model. *Computational Visual Media, 4*(3), 253–266.
28. Jain, A., Mishra, A., Shukla, A., & Tiwari, R. (2019). A novel genetically optimized convolutional neural network for traffic sign recognition: A new benchmark on Belgium and Chinese traffic sign datasets. Neural Processing Letters, 50(3), 3019-3043.
29. Kim, J., Lee, S., Oh, T.-H., & Kweon, I. S. (2018). Co-domain embedding using deep quadruplet networks for unseen traffic sign recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence,*

Jixiang Wan received the B.S. degree in Hubei University of Technology, Wuhan, China, in 2014, and the M.S. degree in microelectronics from Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China, in 2017. He is currently pursuing a Ph.D. degree in microelectronics science and engineering at University of Chinese Academy of Sciences, Beijing 100,049, China. His main research is on the deep learning algorithm and hardware acceleration.

Wei Ding received the B.S. degree in microelectronics science and engineering from School of Physics and Technology, Wuhan University, Wuhan, China, in 2016. He is currently pursuing a M.S. degree in electronic and communication engineering at University of Chinese Academy of Sciences, Beijing 100,049, China. His main research is on the deep learning hardware acceleration and digital integrated circuit design.

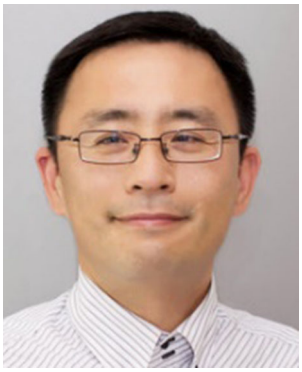
Hanlin Zhu received the B.S. degree in computer science and technology from Tongji University, Shanghai, China, in 2017. He is currently pursuing a M.S. degree in electronics and communication engineering at University of Chinese Academy of Sciences, Beijing 100,049, China. His research interests include machine learning, data mining, time series analysis and artificial intelligence.

Ming Xia received the B.S. degree in from the College of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China, in 2015. He is currently pursuing a Ph.D. degree in microelectronics at University of Chinese Academy of Sciences, Beijing 100,049, China. His current research interests include circuit and system design for deep learning hardware acceleration and image processing.



Zunkai Huang received the B.S. degree in electronics engineering from Tianjin University, Tianjin, China, in 2013, and the Ph.D. degree in microelectronics from Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China, in 2018. He was with the Hiroshima University, Japan, as a special research student, from 2016 to 2017. Since July 2018, he has been with the Shanghai Advanced Research Institute, Chinese Academy of Sciences, where he is currently an assistant professor. His research focuses on CMOS image sensor chip and system, digital image signal processing circuits, and driving circuits for display panels.

Li Tian received the Ph.D. degree in electronics science and technology from Shanghai Institute of Technical Physics of the Chinese Academy of Sciences in 2013. He is currently working as an associate professor at Shanghai Advanced Research Institute, Chinese Academy of Sciences. His research focuses on CMOS image sensor chip, smart vehicle vision systems and digital image signal processing circuits and algorithms.



Yongxin Zhu received the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2001. He was with the National University of Singapore as a Research Fellow from 2002 to 2005, and has been an Associate Professor with the School of Microelectronics, Shanghai Jiao Tong University, Shanghai, China, since 2006. In 2017, he joined the Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, as a Full Professor. His research interests include computer architectures, system-level IC design, and big data processing. He is a professional member of the Association for Computing Machinery. He has served more than 30 conferences and journals as an Editor, Program Chair, Publicity Chair, TPC Member, and Reviewer.

Hui Wang received the Ph.D. degree in Physics from the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China, in 2001. He had a postdoctoral position at IMEC, Belgium, and then worked as an Associate Professor at Shanghai Jiao Tong University, Shanghai, China. In spring 2010, he joined the Shanghai Advanced Research Institute, Chinese Academy of Sciences, as a full professor in Microelectronics. His research interests include high-performance imaging and display panel driving.