An Improved YOLOv4 Lightweight Traffic Sign Detection Algorithm

Jie Cao, Penghui Li, Hong Zhang, Guang Su

***Abstract*—A new lightweight traffic sign detection algorithm based on YOLOv4 has been proposed to address the time-consuming issue of existing algorithms when training more parameters. The algorithm involves several key improvements to YOLOv4's backbone network and feature pyramid. Specifically, MobileNetv3 is used as the backbone feature extraction network, replacing standard convolution with depthwise separable convolution to reduce the number of parameters and computation in both the backbone network and feature pyramid. Furthermore, two SPP modules are added to the feature pyramid part to improve the detection accuracy of the model. The MobileNetv3 network is also improved by dropping the modules after layer 17 and performing PW (PointWise) convolution operations on feature layers to convert their dimensions and connect them with the detection layer. Finally, the a priori frame is initialized by applying the K-means++ algorithm to further enhance the detection accuracy of the algorithm. Experimental results on the CCTSDB dataset indicate that the proposed algorithm achieves a 1.7% increase in mAP (mean average precision), reduces the parameter amount by 197M, and increases detection speed by 25% compared to YOLOv4. These results demonstrate that the improved lightweight algorithm performs exceptionally well in traffic sign detection.**

***Index Terms*—Lightweight algorithms, machine vision, intelligent transportation, traffic sign detection, YOLOv4**

I. Introduction

As driverless technology matures, open road testing under limited conditions has become prevalent both domestically and internationally. However, the detection and localization of traffic signs remains a challenging research topic in the perception of driverless environments. Real-world environments are highly variable due to uncontrollable factors such as lighting conditions, damaged and faded signs, and occlusions caused by adverse weather conditions. These factors present significant obstacles to real-time traffic sign detection. Therefore, there are still many problems that need to be overcome in the field of traffic sign detection. Traffic sign detection techniques have been extensively researched, but they still face significant challenges. Traditional algorithms rely on color or specific shape features, and their adaptability to different environments is limited. Deep learning-based approaches have achieved high accuracy, but they often lack real-time performance and are challenging to apply in practical scenarios. Recent advances in lightweight detection frameworks for mobile and embedded devices have significantly improved detection speed, but they often sacrifice accuracy. In order to develop an effective traffic sign detection method, it is crucial to balance accuracy, real-time performance, and robustness in complex scenarios.

The YOLO series model represents a prominent end-to-end detection model that can predict object categories and generate bounding boxes simultaneously, achieving high detection efficiency and meeting real-time requirements. Ongoing improvements by researchers have led to significant increases in model effectiveness. For example, YOLOv2 [1] incorporates batch normalization to boost detection accuracy, while YOLOv3 [2] improved the backbone network and introduced a multi-scale fusion method to maintain high detection speed. YOLOv4 [3] further enhances accuracy and speed by integrating a CSPNet [4] into the backbone and adding SPPNet [5] and PANet [6] to the neck section. These improvements demonstrate the potential of the YOLO series in achieving high accuracy, speed, and adaptability in complex scenarios.

The YOLO model parameters increase with the number of layers, leading to latency issues for real-time detection. Traffic signs constantly change position and size, and detection algorithms are sensitive to object resolution, making detection performance unpredictable. Low accuracy of lightweight models can limit detection tasks. Finding the right balance among efficiency and effectiveness is challenging. The paper proposes a traffic sign detection model that aims to ensure high accuracy while balancing real-time performance. The model also adopts a lightweight design to better address the aforementioned issues. It optimizes information exchange between feature layers, reduces information loss, and minimizes computational overhead to improve latency. The goal is to enhance detection performance while maintaining accuracy.

This research focuses on improving the YOLOv4 model, which is used to solve the problem of long training time due to the large number of model parameters. The paper's main contributions include: (1) By adopting MobileNetv3-large[7] as the backbone network, the model reduces the number of parameters and computational complexity while significantly improving the inference speed. (2) Implementing depthwise separable convolution in the feature pyramid to further reduce the model's parameters, while adding the SPP module to enhance semantic information expression and algorithmic recognition accuracy. (3) This method improves the MobileNetv3 network by removing some layers after the

Manuscript received May 3, 2022; revised April 22, 2023. This work was supported by Natural Science Foundation of Gansu Province , China (Grant No. 20JR5RA450)

Jie Cao is the President of Lanzhou City College, Lanzhou, China, 730050. (e-mail: caoj@lut.edu.cn).

Penghui Li is a postgraduate student at the Lanzhou University of Technology, Lanzhou, China, 730050. (e-mail: 562267730@qq.com).

Hong Zhang is an Associate Professor at the Lanzhou University of Technology, Lanzhou, China, 730050. (Corresponding author to provide phone: 13669352331; e-mail: zhanghong@lut.edu.cn).

Guang Su is a postgraduate student at the Lanzhou University of Technology, Lanzhou, China, 730050. (e-mail: 2848372862@qq.com).

17th layer and applying PW convolution on three feature layers of different scales. These feature layers are connected to the detection layer to generate the final prediction box. The paper also employs the K-means++ technique to cluster experimental data and determine appropriate prior box sizes for the CCTSDB dataset [8] , resulting in further performance improvements for the YOLOv4 network.

This paper is structured as follows: Section 2 provides an introduction to the YOLOv4 method and the MobileNetv3 algorithm, which is the primary focus of this study. Section 3 presents the proposed improved model, which mainly emphasizes the enhancements made to the MobileNetv3 algorithm, the embedding of the PANet through the SPP module, and the use of K-means++ to determine the optimal anchor box size. In Section 4, we conducted experiments on the CCTSDB dataset to evaluate the performance of our improved algorithm and compared it with previous methods. The experimental results showed that the new model not only reduces the model size but also improves detection speed and accuracy. Finally, we summarized the experimental results and discussed future research directions in the conclusion section.

II. Relate works

Although general object detection methods perform well on many public datasets, traffic sign detection tasks are limited to specific scenarios. In practical applications, they face many challenges mainly due to the small size of traffic signs, the complexity of road scenes, and the high requirements for detection accuracy and speed.

The distinctive colors and shapes of traffic signs make them distinguishable from other objects. However, sign detection presents a challenge due to the small size of signs and complex road scenarios. To address this challenge, researchers have developed specialized algorithms using a combination of color probability models, HOG features, and deep learning techniques. Yang *et al*. [9] used a rapid detection module to establish a color probabilistic model and HOG features for real-time categorization of traffic signs. Similarly, Xu *et al*. [10] developed a traffic sign detection model that extracts color and shape information from the Region of Interest (ROI) containing the extracted traffic sign features. In recent years, deep learning has become one of the popular techniques for traffic sign detection. Shen *et al*. [11] proposed a model based on an improved network structure that significantly improved the accuracy for small-sized traffic signs. Lee *et al*. [12] addressed the challenge of detecting traffic signs in complex scenarios by proposing a detection model based on image segmentation. Yuan *et al*. [13] introduced an attention-based detection approach to refine features of complicated backgrounds and reduce false detections. Moreover, Wang *et al*. [14] presented an improved fast R-CNN algorithm that uses a new sampling method to optimize the network for detecting small targets in traffic images. These techniques have demonstrated improved performance in traffic sign detection under different scenarios, highlighting the importance of specialized algorithms in addressing the unique challenges of sign detection.

Gavrilescu *et al*. [15] used the Faster R-CNN algorithm to train and test 3000 traffic signs and found that it to be superior in accuracy and speed compared to the previous algorithm. Despite achieving high accuracy, the Faster R-CNN algorithm has a natural drawback of slow detection speed, making it unsuitable for real-time traffic sign detection. To address this issue, researchers have turned to one-stage detection algorithms, which have faster detection times. Zhang *et al*. [16] proposed an improved model based on YOLOv3 for traffic sign detection, which incorporates multi-scale spatial pyramid pool blocks in Darknet 53 to learn features more thoroughly, achieving real-time detection of signs. Wang *et al*. [17] upgraded the YOLOv4 algorithm by combining four different feature layers for detection, resulting in increased detection accuracy. As a result, many improved one-stage traffic sign detection algorithms have emerged, reducing the accuracy gap between one-stage and two-stage algorithms and making one-stage detection the mainstream of research. However, detecting traffic signs in real-time remains a significant challenge due to the long training times required when using large numbers of parameters.

The widespread adoption of deep learning networks has brought about a revolution in the field of traffic sign detection. These networks have significantly improved the accuracy and speed of detection beyond traditional algorithms, making them an indispensable tool for modern traffic management. Moreover, deep learning-based detection methods have proven to be highly effective in identifying traffic signs in complex road environments, a feat that was previously considered challenging.

This approach has enormous implications for the development of autonomous vehicles as it provides them with reliable decision-making tools. By accurately detecting traffic signs in real-time, self-driving cars can respond appropriately to changing road conditions, improving their safety and reducing the likelihood of accidents.

To further enhance the robustness of the model, this study utilizes a one-stage deep learning approach to develop a lightweight, highly accurate, and robust detection model. This model is designed to accurately detect traffic signs in various lighting and weather conditions, as well as in scenarios where traffic signs may be partially obscured or have low contrast. The resulting detection model is a crucial step towards achieving safer and more efficient traffic management, and it has the potential to be implemented on a large scale to benefit society as a whole.

III. Improved YOLOv4 Algorithm

In this paper, the YOLOv4-MobileNetv3 model is enhanced in three key areas, resulting in improved performance. Firstly, MobileNetv3 is utilized as the backbone network, which significantly reduces the number of parameters in the network without sacrificing detection accuracy. Secondly, the feature pyramid structure is modified with the integration of SPP, expanding the receptive field of the feature layer and capturing important contextual features, ultimately leading to improved detection performance. In the end, the model uses depthwise separable convolutions to replace some of the convolution operations in the feature pyramid and backbone network, further reducing the quantity of parameters in the model. The improved network structure is shown in Figure 1.

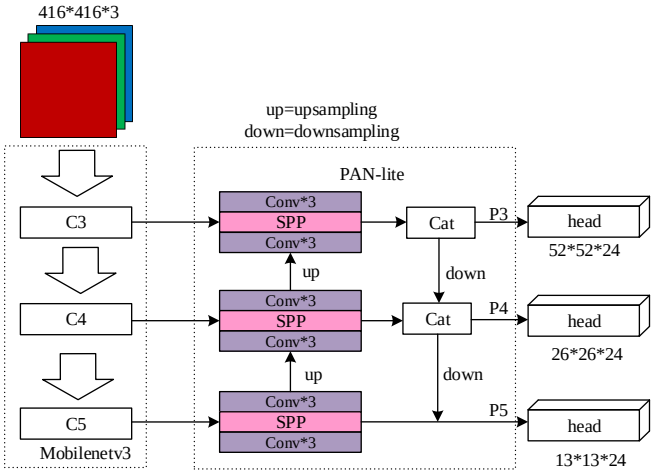


Fig. 1. Improved-YOLOv4 structure. CBH stands for convolutional layer, batch normalization and h-swish activation function. MBN represents a feature extraction module based on MobileBottleNeck.

A. Improved MobileNetv3 Feature Extraction Network

MobileNetv3-large (Neural Architecture Search, NAS) combines the automatic neural network search and the NetAdapt algorithm. The depthwise separable convolution of MobileNetv1 [18] , the inverse residual structure with a linear bottleneck of MobileNetv2 [19] , and the lightweight structure based on SE [20] of MnasNet are the three models that makeup MobileNetv3. The convolution operation in depthwise separable convolution is typically broken down into many phases. Let's consider a convolutional layer with a kernel size of 3×3, which takes 16 input channels and produces 32 output channels. It requires 4086((3×3×16)×32) parameters ((3×3×16)×32) for general convolution. The depthwise separable convolution only requires 656 parameters (3×3×16+(1×1×16)×32). Compared to CSPDarknet53, MobileNetv3 requires fewer parameters and lower operating costs to achieve the same effect. Therefore, MobileNetv3 can replace CSPDarknet53 as the network of YOLOv4 to complete feature extraction.

The backbone feature extraction network in MobileNetv3 adopts the bneck structure, which is the first network to be passed through the input image. In Figure 2, the bneck structure is displayed. The low-dimensional feature map is easily losing information when it passes the ReLU activation function. The information loss is minimal when the ReLU operation is performed in high-dimensional. To balance the channel weights of each feature map, a bneck structure is introduced before the depthwise separable convolution and attention mechanisms. This structure increases the dimensionality of the feature maps and then performs upscaling and depthwise separable convolution on them. Using global average pooling to create the 1×1×N feature map. Two fully connected layers are then applied to alter the weights of each channel. The dimensionality of the resulting feature is reduced and superimposed on the feature of the input bneck structure before outputting it. Multiplying the input feature map generates a feature map with the attention mechanism added.

The detection layer outputs three branches at different scales, and feature extraction takes place in the bottleneck. The three layers are then connected by performing convolutional operations on their dimensions. The layers before average pooling are removed, and convolution is used to compute the feature maps to reduce latency while maintaining high-dimensional features. After feature generation, the layer is removed. The bottleneck mapping layer is no longer needed. This results in a 10ms reduction in overhead.

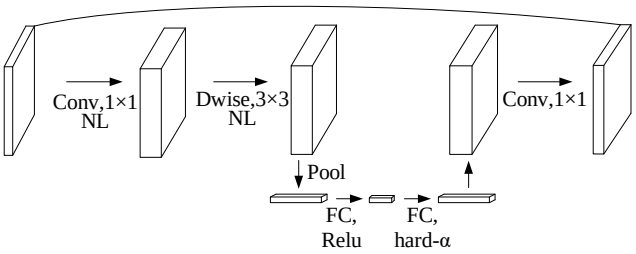


Fig. 2. It mainly implements channel separable convolution, SE channel attention mechanism and residual connection.

B. Improvement of SPP Structure and PANet Module

Different feature layers have different semantic content and contribute differently to the combined output features. The performance of the network suffers when features are extracted only at the output layer, which affects the detection of small objects. To accomplish the task of multi-scale detection, it is necessary to fully exploit the many layers of semantic information. The feature pyramid of YOLOv4 uses PANet and SPP structures, and Figure 3 shows the SPP module. The SPP obtains the output layer of the network. It then applies max-pooling operations with large kernels. The resulting feature maps are concatenated along the channel dimension, resulting in a threefold increase in both the size of the output feature maps and channels. The SPP structure is designed to enlarge the receptive field of the feature layer, capturing more comprehensive contextual features, and enhancing the model's recognition capability.

The SPP structure receives the final layer output from the YOLOv4 network for upsampling. Based on the above, this paper uses SPP structure for three-layer output. It is integrated into the PANet model. This will improve the the representation of output feature messages. This expands the model's ability to perceive objects at different scales, improving its overall robustness.

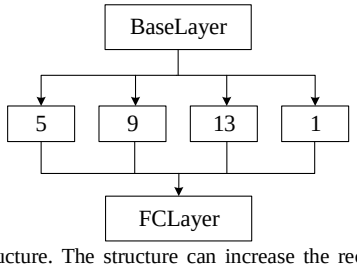
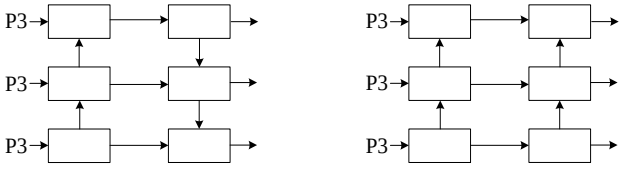


Fig. 3. SPP structure. The structure can increase the receptive field and dynamically change the size of the feature map, so that the model can adapt to images of different resolutions, taking into account both large and small targets

Figure 4 shows the structure of the PANet. The original FPN is complemented by a bottom-up path aggregation network, which further improves the detection performance. This is because the three feature layers are drawn iteratively through the PANet. Thus, more abstract top-level features

and underlying features are fully combined. During forward propagation, horizontal links merge feature maps of the same size, fully leveraging the information from feature layers of various scales and significantly improving the model's detection capability.



(a) PANet (b) FPN

Fig. 4. Magnetization a Characteristic pyramid structure

To improve the image feature extraction process. Both the PANet and the SPPNet combine five consecutive convolutions and three consecutive convolutions. Compared to regular convolution operations, separable convolution significantly reduces the number of parameters. To reduce the model's parameter and memory requirements, the regular convolutions in the SPPNet and PANet have been replaced, as shown in Figure 5.

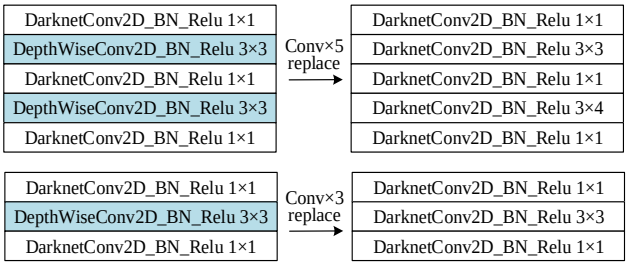


Fig. 5. Depthwise separable convolution instead of ordinary convolution

C. Optimal Anchor Box Size.

The difficulty of target detection in prediction can be reduced by using the anchor boxes generated by clustering. The PASCAL VOC dataset is applied to obtain the predetermined a priori frame of the YOLOv4 network. However, using the original pre-defined anchor frames from the CCTSDB dataset leads to the inability of Yolo Head to select the appropriate target bounding box. This has a significant impact on the ability to detect the target. As a result, the ground truth annotated boxes in the dataset are clustered first in this work.

The YOLOv4 uses K-means to gather the target frames on the dataset. Performance is evaluated using the average Intersection over Union (IoU) between anchor boxes and ground-truth boxes, which measures the degree of overlap between the predicted and actual bounding boxes. The IoU calculation equation is utilized for this purpose.

$$AvgIoU = \sum_{i=1}^{k}\sum_{j=1}^{n_k} iou\left(B_i, C_j\right) \tag{1}$$

$$iou\left(B,C\right) = \frac{B \cap C}{B \cup C} \tag{2}$$

In the equation, n_k denotes the number of centers, denotes the number of target boxes assigned to the kth cluster center, B represents the anchor box associated with the cluster center, and C represents the total number of target boxes assigned to the kth cluster. When K-means clusters ground-truth boxes, the number of cluster centers is predefined. However, in applications, the optimal k value cannot be known in advance, which greatly affects the efficiency of the algorithm. Additionally, k-means clustering requires manual determination of the original clustering centers, and varying initial centers can lead to entirely different clustering outputs.

The K-means++ clustering algorithm was used in this study to cluster object bounding boxes in the CCTSDB dataset. This algorithm is an improvement over the traditional K-means algorithm as it optimizes the selection of initial points by choosing cluster centers more effectively. This reduces clustering bias and leads to better results. By utilizing prior bounding boxes that are suitable for the target dataset, we can enhance the accuracy of sign detection.

The K-means++ algorithm starts by randomly selecting a data point from the dataset as the initial cluster center. Then, the distances between all the data points and the selected center are computed, and the data points are assigned to the closest center. The algorithm then determines the probability of each data point being selected as the next cluster center based on the minimum distance between the data point and its assigned center. The next cluster center is selected as the data point with the highest probability.

$$P\left(x\right) = \frac{D\left(x\right)^2}{\sum_{x \in X} D\left(x\right)^2} \tag{3}$$

Iterate through the above steps until k cluster centers are chosen. Then, employ the K-means++ to compute the final results for the k cluster centers, continuing until there is no further change in the anchor box sizes.

IV. Experiment and Result Analysis

A. Experimental Environment and Parameters

The PyTorch framework is utilized to construct the network models, which are run on a Windows 10 (64-bit) operating system equipped with an Intel i7-10750H CPU, 16GB of memory, an NVIDIA GeForce GTX 2080 GPU, and CUDA version 10.0, as well as CUDNN 7.4.

This experiment utilizes YOLOv4 as the detection algorithm framework and employs the transfer learning method, utilizing pre-trained weights as the fundamental feature extraction model. To ensure training convergence, the initial learning rate was 0.001, label smoothing was 0.005, and the batch size was set to 16. SGDM optimization method was utilized with a CIOU loss function.

B. Dataset Processing

The CCTSDB dataset consists of 15,723 photographs, which are annotated with three categories: directional signs, prohibition signs, and warning signs. During the experiment, the CCTSDB dataset's original photos are converted to the JPG format, while the annotations are transformed into XML files in VOC format, which the YOLOv4 network can read. A 9:1 ratio is used to divide the dataset into training and testing sets.

C. Experimental Process and Result Analysis

(1) Validating the effectiveness of K-means++. To verify whether the detection accuracy of the CCTSDB was optimized by the K-means++ clustering technique. As shown in Figure 6, we choosed k = 2~12 as the cluster centers to

examine the association between k and the average IoU on centers. The AvgIoU number increases as k steadily increases, indicating that the algorithm performs better as k increases. The fluctuation trend of AvgIoU steadily decreases when k is bigger or equal to 9. To achieve better clustering results and reduce the computational complexity of YOLOv4, the optimal value of k should be chosen at the inflection point of the curve, where k is as small as possible. In our study, we selected 9 anchor boxes for predicting the targets. The K-means++ clustering algorithm obtained the following sizes: (6,13), (9,19), (12,26), (15,38), (21,30), (29,44), (41,59), (59,79), (104,142). Figure 7 shows the clustering effect. The centers of each cluster are generally dispersed, and each cluster can be easily identified.

With the same parameter settings, the K-means++ resulted in an average accuracy improvement of 0.9%, indicating a closer alignment between the anchor boxes and the actual targets. Consequently, utilizing the K-means++ clustering algorithm for anchor box determination enhances the accuracy and performance of object detection models. This improvement is presented in Table I.

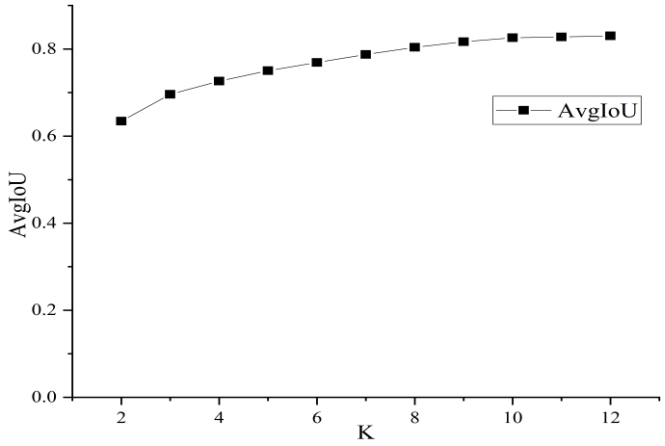


Fig. 6. Curve of AvgIoU increasing with k.

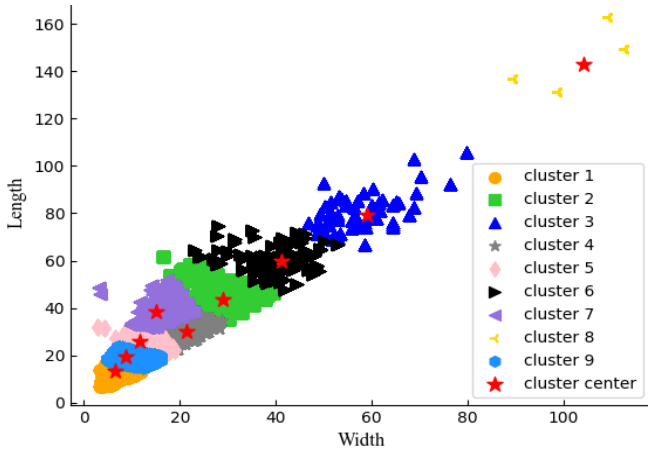


Fig. 7. K-means++ clustering result graph. Clustering effect when k=9.

TABLE I
OPTIMIZATION RESULTS COMPARISON

Model	Backbone	mAP(%)	FPS
YOLOv4	CSPdarknet53	93.9	19
Improved	CSPdarknet53	**94.8**	**20**

the CCTSDB to determine the optimal number of cluster

(2) To test the effectiveness of the improved algorithm, a neural network model that integrates all the aforementioned enhancements is trained and tested on the CCTSDB dataset. The results of the experiments are shown in Table II, which showcases the performance of the proposed method.

The proposed enhanced YOLOv4 algorithm achieved over 10% improvement in accuracy compared to the Faster R-CNN and SSD algorithms, while requiring significantly fewer parameters. This makes it a promising option for real-world applications with limited computing resources. The enhanced YOLOv4 algorithm achieved a modest average accuracy improvement of 1.7%, but offered significant advantages, including a 197M reduction in model size and a 25% increase in detection speed. These outcomes confirm the practicality and effectiveness of the improved algorithm, meeting the requirements for traffic sign detection. As a result, the proposed approach is superior in terms of accuracy, model size reduction, and faster detection speed, making it an advanced option for traffic sign detection.

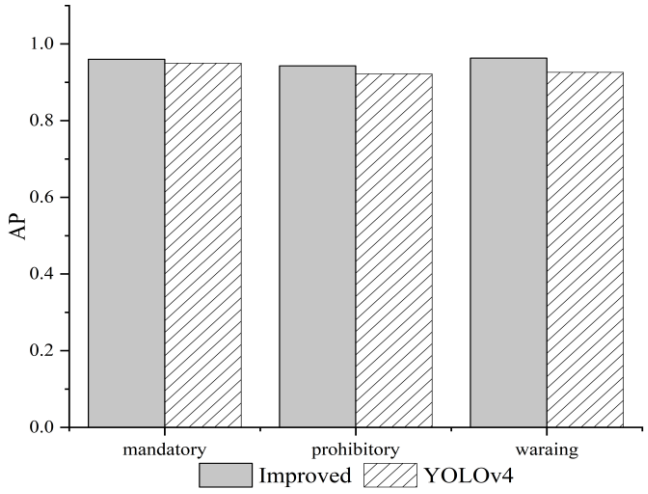
The performance comparison between YOLOv4 and Improved-YOLOv4 based on CCTSDB dataset is presented in Figure 8, showcasing various target detection and P-R performance curves. In Figure 8(a) and 8(b), the detection results of four traffic signs by both models are compared. The results indicate that Improved-YOLOv4 exhibits higher confidence in detection compared to YOLOv4, with more accurate and fewer missed or false detections, leading to higher detection accuracy. These results demonstrate the validity and effectiveness of the proposed approach, meeting the requirements of traffic sign detection. The performance of the improved-YOLOv4 is presented in Figure 8(c) and 8(d), showing higher detection accuracies and P-R curves than the original YOLOv4 model. The proposed approach achieves a mAP of 95.6%, while YOLOv4 achieves 93.9%. Moreover, the improved model outperforms YOLOv4 in terms of precision and recall rates. In the P-R curve comparison of the warning class, the improved model exhibits a larger area, indicating better performance than YOLOv4.

(a). CCTSDB detection results

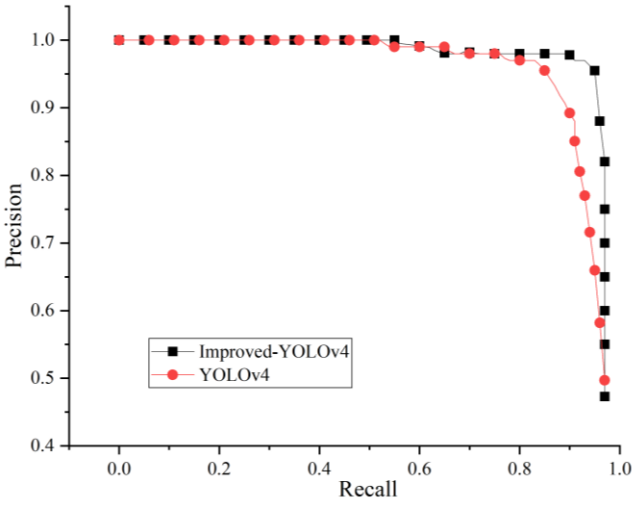
(b). YOLOv4 traffic sign detection results

TABLE II
IMPROVED ALGORITHM PERFORMANCE COMPARISON WITH OTHER NETWORKS

Model	Backbone	Size	mAP(%)	Total Params	FLOPs	FPS	AP-m	AP-p	AP-w
Faster R-CNN	Resnet50	522.9M	84.2	137,078,239	298.21G	10	77.6%	87.4%	87.7%
YOLOv4	CSPdarknet53	244.3M	93.9	64,040,001	60.09G	19	95.9%	93.2%	92.6%
SSD	VGG16	99.8M	86.4	26,151,824	62.64G	23	81.6%	86.1%	91.7%
Improved	Mobilenetv3	**47.3M**	**95.6**	**12,422,749**	**9.149G**	**25**	**96.1%**	**94.3%**	**96.3%**



(c). Improved-YOLOv4 traffic sign detection results



(d). Comparative P-R curves of waring classes

Fig. 8. Example of improved-YOLOv4 and YOOLOv4 model target detection results

(3) The CCTSDB dataset is used to train and evaluate two modified versions of YOLOv4: YOLOv4-M, which uses MobileNetv3 as the backbone network, and YOLOv4-S, which has an improved feature pyramid structure. The performance of these models is compared with the original target detection network, and the results are summarized in Table III.

Based on the data presented in Table III, the YOLOv4-S traffic sign detection network achieves a higher mAP value compared to the original network. Conversely, the YOLOv4-M traffic sign detection network exhibits a lower mAP value than the original network. Notably, the YOLOv4-M detection algorithm employs MobileNetv3 as its backbone network. Its detection accuracy is reduced by 2.1 percentage points, but the model size is reduced by 199.8 M. This shows that MobileNetv3 has a simple structure and strong performance. It can effectively accomplish the feature extraction task, resulting in a significant reduction in parameter count and inference time.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT IMPROVED ALGORITHMS

Model	Backbone	Size	mAP(%)	FPS
YOLOv4	CSPdarknet53	244.3M	94.8%	20
YOLOv4-S	CSPdarknet53	247.5M	**97.1%**	18
YOLOv4-M	Mobilenetv3	**44.5M**	92.7%	**27**
YOLOv4-SM	Mobilenetv3	47.3M	95.6%	25

The YOLOv4-S detection approach, which utilizes an improved feature pyramid structure, demonstrates a 2.3% average accuracy improvement compared to the original network. The method employs the SPP network to expand the perceptual field of the feature layer, enabling the capture of useful contextual features and the enhancement of feature information through multi-scale feature fusion. These advancements lead to a significant improvement in the detection performance of the model.

V. CONCLUSION

To tackle the practical issues of current traffic sign methods with long training time and limited real-time performance when there are many parameters. We propose an effective lightweight model. The model uses the Mobilenetv3+YOLOv4 algorithm to detect traffic signs, and combines multiple SPP modules and K-means++ to improve the accuracy of the model. In comparison to the original model, the improved model offers several key benefits, including rapid detection, enhanced accuracy, and reduced parameter count. The model achieves faster feature extraction by enhancing the performance of the backbone network. By employing depthwise separable convolution, the model significantly reduces the parameter count needed for convolution operations. The proposed method reduces computational complexity, achieving a mAP of 95.6% with fewer parameters and computational effort compared to other models. These results demonstrate that the algorithm surpasses existing models and meets the demands of traffic scene detection.

REFERENCES

[1] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:7263-7271.

[2] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 .2018.

[3] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[4] Wang C Y, Mark Liao H Y, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of cnn[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020:390-391.

[5] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-1916.

[6] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8759-8768.

[7] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:1314-1324.

[8] Zhang J, Huang M, Jin X, et al. A Real-Time Chinese Traffic Sign Detection Algorithm Based on Modified YOLOv2[J]. Algorithms, 2017, 10(4):127-139.

[9] Yang Y, Luo H, Xu H, et al. Towards real-time traffic sign detection and classification[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 17(7):2022-2031.

[10] Xu X, Jin J, Zhang S, et al. Smart data driven traffic sign detection method based on adaptive color threshold and shape symmetry[J]. Future Generation Computer Systems, 2019, 94(5):381-391.

[11] Shen L, You L, Peng B, et al. Group multi-scale attention pyramid network for traffic sign detection[J]. Neurocomputing, 2021, 452(6): 1-14.

[12] Lee H S, Kim K, Simultaneous Traffic Sign Detection and Boundary Estimation using Convolutional Neural Network[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(5): 1652-1663.

[13] Yuan Y, Xiong Z, Wang Q. VSSA-NET: Vertical Spatial Sequence Attention Network for Traffic Sign Detection[J]. IEEE Transactions on Image Processing, 2019,28(7):3423-3434.

[14] Wang F, Li Y, Wei Y, et al. Improved Faster RCNN for Traffic Sign Detection[C]. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020:1-6.

[15] Gavrilescu R, Zet C, Foşalău C, et al. Faster R-CNN: An Approach to Real-Time Object Detection[C]. 2018 International Conference and Exposition on Electrical and Power Engineering (EPE), Iasi, Romania, 2018, 0165-0168.

[16] Zhang H, Qin L, Li J, et al. Real-time detection method for small traffic signs based on Yolov3[J]. IEEE Access, 2020, 8:64145-64156.

[17] Wang H, Yu H. Traffic Sign Detection Algorithm based on improved YOLOv4[C]. 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2020, 9:1946-1950.

[18] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[19] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:4510-4520.

[20] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7132-7141.

Cao Jie, born in Jingyuan, Gansu Province, China in 1966, graduated from the Department of Automation of Gansu University of Technology in 1987 with a bachelor's degree in engineering. In 1994, he obtained a master's degree in engineering from Xi'an Jiaotong University, and his main research direction is intelligent information processing, intelligent transportation system theory and application.In 2004, he was selected as the "555 Science and Technology Talent Project", a cross-century academic and technological leader of Gansu Province. Main research achievements: won the first prize of Gansu Province Teaching Achievement Award, the third prize of Gansu Province Science and Technology Progress Award, and the third prize of Lanzhou City Science and Technology Progress Award.

Penghui Li is currently a master's student in Electronic Information Engineering in the School of Computer and Communication at Lanzhou University of Technology. His research direction is intelligent transportation. He won the third prize in the 17th China Postgraduate Electronic Design Competition in Northwest China. He has won the third artificial intelligence innovation competition (second prize at school level).

Hong Zhang is an associate professor at Lanzhou University of Technology. She received her B.S. degree in Computer and Applications from Lanzhou University of Technology in 2001; her M.S. degree in Communication and Information Systems from Lanzhou University of Technology in 2004; and her Ph.D. degree in Systems Engineering from Lanzhou University of Technology in 2018. Her research interests are Machine Learning, Intelligent Transportation and Big Data Analytics

Guang Su is currently a master's student in Electronic Information Engineering in the School of Computer and Communication at Lanzhou University of Technology. His research direction is intelligent transportation. He published a paper on the prediction of traffic flow in the Journal of Jilin University, China.