IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 December 2023, accepted 5 January 2024, date of publication 9 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3351771

RESEARCH ARTICLE

Research on Road Object Detection Model Based on YOLOv4 of Autonomous Vehicle

PENGHUI WANG[1], XUFEI WANG[1], YIFAN LIU[2], AND JEONGYOUNG SONG[3], (Member, IEEE)

[1]Department of Mechanical Engineering, Shaanxi University of Science and Technology, Hanzhong 723000, China
[2]Department of New Energy, Sanmenxia College of Social Administration, Sanmenxia 472000, China
[3]Department of Computer Engineering, Pai Chai University, Daejeon 35345, South Korea

Corresponding authors: Xufei Wang (wxf@snut.edu.cn) and Jeongyoung Song (jysong@pcu.ac.kr)

This work was supported in part by the Research Foundation of Shaanxi University of Technology under Grant SLGRCQD2321.

ABSTRACT The YOLOv4 network is widely used in object detection tasks as a representative network, but there is also the problem that the complexity of the network model affects the detection speed. In this paper, we propose an improved MV2_S_YE object detection algorithm based on the YOLOv4 network to improve the detection accuracy while increasing the road object detection speed. Firstly, the backbone network CSPDarknet53 of the YOLOv4 network is replaced by the Mobilenetv2 network to reduce the number of parameters of the network; secondly, the channel attention mechanism is introduced, and the SENet module is embedded in the structure of the PANet to optimize the object detection accuracy; finally, the EIOU loss function is used to replace the CIOU loss function to improve the object detection accuracy further. The MV2_S_YE network is obtained and tested on Pascal VOC, Udacity, and KAIST datasets. To evaluate our approach, we compared MV2-S-YE with YOLOv4, YOLOv4-tiny, YOLOv7-tiny and YOLOv8s. The results show that MV2-S-YE 的 mAP@0.5 achieves 80.9%, 66.7%, and 94.8% on the VOC2007, Udacity, and KAIST test sets, respectively, and is higher than YOLOv8s on both the Udacity and KAIST test sets. On the VOC2007 test set MV2-S-YE achieves a detection speed of 45FPS which is higher than YOLOv8s.

INDEX TERMS Object detection, YOLOv4, Mobilenetv2, SENet, EIOU.

I. INTRODUCTION

The application of artificial intelligence in autonomous driving vehicles is getting more and more attention. With the rapid development of computer technology, computer vision technology has become one of the critical technologies for automobile external object detection technology. The object detection technology for autonomous driving systems needs to accurately and in real-time perceive the vehicle's external environment, which is an essential prerequisite for making correct decisions to ensure the vehicle's safe driving [1], [2].

As computer vision technology continues to develop, convolutional neural networks based on deep learning play an increasingly important role in the object detection task, and many excellent object detection algorithms have been successively proposed and applied and have been greatly improved compared with traditional algorithms. Currently, object detection algorithms are usually categorized into one-stage and two-stage detection algorithms. One-stage detection algorithms mainly include regression-based YOLO (You Only Look Once) algorithm [3], SSD (Single Shot Multibox Detector) algorithm [4], etc. These algorithms classify and regress while generating the bounding box and have faster detection speed. Two-stage algorithms mainly include R-CNN (Region-Convolutional Neural Networks) [5], Fast R-CNN [6], Faster R-CNN [7], etc.; these algorithms firstly generate the proposed region and then utilize Convolutional Neural Networks to categorize the proposed area and output the location information, which relative to the one-stage algorithms with more accurate detection results. However, the one-stage detection algorithm extracts features only once for detection, which is relatively faster, but the accuracy will be degraded. The two-stage detection algorithm has higher accuracy but is relatively slow and is not suitable for scenarios and tasks such as vehicle object detection where real-time requirements are high. In object detection in dynamic scenes

The associate editor coordinating the review of this manuscript and approving it for publication was Tomasz Trzcinski.

such as autonomous driving, the algorithm is required to have high detection accuracy and high real-time performance.

To solve the above problems, Redmon et al. [3] proposed the first regression-based one-stage object detection network YOLO in 2016, which eliminates the process of generating candidate regions, the input image can get the position of the object and the confidence probability of the category that the object belongs to after one time of the network, and merges the object border and category regression process into one network, which realizes the end-to-end detection. Redmon et al. later proposed YOLOv2 [8] and YOLOv3 [9] networks.YOLOv2 introduces the a priori frame (Anchor). It uses the K-means clustering method to compute better a priori frame parameters, which improves the detection performance of the network model and strengthens the ability to detect small objects.YOLOv3 refers to the residual idea of changing the backbone network into the Darknet-53, which utilizes a Feature pyramid network (FPN) structure [10] to achieve multi-scale detection and realizes multi-label classification.YOLOv3 has a faster detection speed along with higher detection accuracy. In 2020, Bochkovskiy et al. [11] proposed the YOLOv4 network model with many improvements based on YOLOv3. YOLOv4 combines the Darknet-53 network with the Cross Stage Partial Network (CSPNet) [12] idea to construct the CSPDarknet53 backbone feature extraction network and used SPP and Path Aggregation Network (PANet) [13] in the neck of the network, inherited the detection head of YOLOv3 for multi-scale detection. However, the detection accuracy and speed are improved, but the detection speed is still slow. Haris and Glowacz [14] conducted a comparative study of R-FCN, Mask R-CNN, SSD, RetinaNet, and YOLOv4 networks on the BDD100K dataset to analyze the strengths and limitations of the five networks based on parameters such as detection accuracy and detection speed. The results show that YOLOv4 performs more accurately detecting challenging road objects in complex road scenarios and weather conditions in the same test environment. However, the network model is more significant and not conducive to conducting embedded research on mobile devices. Chen et al. [15] used MobilenetV2 to improve the SSD network with an optimized feature fusion module for vehicle target detection study. Although it improves the detection accuracy and reduces the single inference time of the SSD network, its detection accuracy and real-time performance need to be improved. Cai et al. [16] obtained the YOLOv4-5D network by improving the backbone, improving the feature fusion module, and network pruning for the YOLOv4 network. Although many modifications were made in different modules of the YOLOv4 network structure. However, its network model is large, reaching 91.8 MB, and its mAP@0.5 is only 70.45%, which is not suitable for autonomous vehicle research that relies on high accuracy and fast inference. Wang et al. [17] improved the YOLOv4-tiny algorithm by improving the K-means clustering algorithm and improving the NMS algorithm to enhance the extraction of small target features and optimize the prediction results. However, the mAP@0.5 of the improved YOLOv4-tiny algorithm is only 52.7%, which cannot satisfy the road target detection accuracy requirement.Although the above improvements have improved the detection accuracy, they can still not be applied in real-world scenarios. Wang et al. [18] proposed a new detection network CenterNet-Auto, the backbone network uses RepVGG model, and the average boundary model is proposed, the accuracy and speed of this model still have a gap with the unmanned demand.

Based on the above problems, this paper is based on the YOLOv4 network [19], which currently has better comprehensive performance. Firstly, YOLOv4 is lightly improved using the MobileNet [20] series of networks to obtain three new road object detection networks. The three improved networks are compared and analyzed for the best MobilenetV2_YOLOv4 (MV2-Y) network. Secondly, the object detection accuracy of the YOLOv4 network is improved using SENet (Squeeze-and-excitation Networks) channel attention mechanism [21], and Mobilenetv2_SE_YOLOv4 (MV2_S_Y) network is proposed. Finally, the road object detection network is optimized by replacing the loss function CIOU of the MV2-S-Y network with the EIOU loss function to get the MV2-S-YE network and validated on the PACAL VOC, Udacity, and KAIST datasets.

The main contributions of this paper are as follows:

(1) use a lightweight network MobileNetV2 instead of the backbone network of YOLOv4;
(2) introduce the SENet attention mechanism in the feature fusion network;
(3) improve the YOLOv4 loss function and optimize the function training model using EIOU.

This paper is organized as follows, Section II introduces YOLOv4, MobileNet and SENet networks, and the dataset; in Section III we propose MV2_S_YE and present its general structure. This paper mainly focuses on three improvements to the YOLOv4 algorithm: lightweighting improvement, feature fusion network improvement, and loss function improvement. Section IV gives the experimental results and discussion, which give the performance parameters such as the number of parameters, computation, model size, FPS, Loss curves and mAP. Also, in this section, we compare MV2_S_YE with other state-of-the-art models. Finally, conclusions are drawn in Section V.

II. RELATIONAL WORK

A. YOLOV4 NETWORK

YOLOv4 [19] network is a model obtained by Alexey Bochkovskiy et al. based on YOLOv3 with several improvements, achieving a better balance between detection accuracy and detection speed. The backbone network of YOLOv4, CSPDarknet53, is improved from the Darknet53 of YOLOv3, which draws on the idea of CSPNet (Cross Stage Partial Network) to improve the residual blocks in Darknet53

and introduces the Mish activation function. CSPNet (Cross Stage Partial Network) idea to enhance the residual block in Darknet53, obtained the CSPDarknet53 structure and introduced the Mish activation function.CSPDarknet53 backbone network, compared to the Darknet53 backbone network, reduces the amount of computation simultaneously to ensure accuracy.

B. MOBILENET

MobileNet series network is a lightweight convolutional neural network proposed by Google. The MobileNetV1 [22] (MV1) network was first proposed in 2017, and improvements have been made to this foundation with the successive introduction of the MobileNetV2 [23] (MV2) and MobileNetV3 [24] (MV3) networks.The MobileNet family of networks uses Depthwise Separable Convolution (DSC) [25], which significantly reduces the number of parameters in the network model and improves the speed of network operation.

C. SENET

After the lightweighting of the YOLOv4 network using the MobileNet series of networks, there is a negative impact on the detection accuracy of the network model. Therefore, the SENet module was introduced to improve the detection performance of the convolutional neural network by filtering useless information and enhancing useful features by simulating how humans observe things.

SENet is a one-way attention mechanism, known as Squeeze-and-Excitation Networks, or Compression and Excitation Networks, which is an attention mechanism that focuses on the relationship with the channels. This module enables the network to learn the importance of different channel features in the feature map and weigh the channels according to their importance.

D. DATASETS

In the process of building deep learning models, the quality of datasets impacts the performance of network models. In this paper, three common datasets are selected for the study, and the three datasets are introduced as follows:

(a) The Pascal VOC dataset [26] is a dataset used for the VOC challenge; it is a standard dataset widely used in the field of object detection, including 21504 images with 20 categories. Pascal VOC dataset contains a rich variety and more images of the same object; it is mainly used for image classification and detection and image segmentation tasks.

(b) The Udacity Self-Driving dataset [27] is a dataset for self-driving car algorithm competitions and includes 24423 images with four categories. This dataset has more car and pedestrian objects and rich road scenarios, and it is applied to study the object detection field of autonomous driving.

(c) The KAIST (Korea Advanced Institute of Science and Technology) dataset [28] is a multispectral pedestrian detection dataset, which selects long-wave infrared image data, including three categories of objects and 7600 images. This dataset contains more pedestrian objects and more collected scenes, which can be well used for pedestrian object detection research.

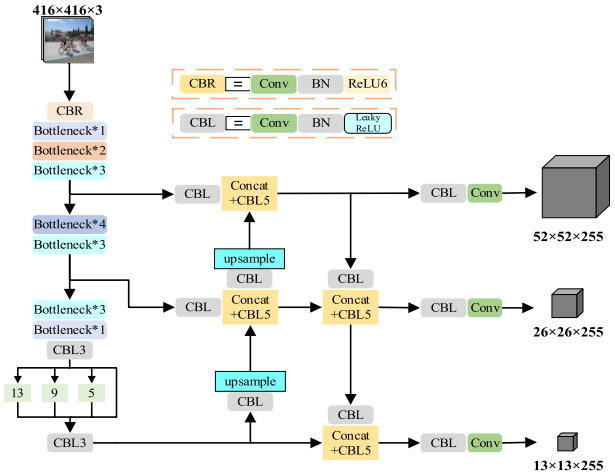


FIGURE 1. MV2-Y network structure.

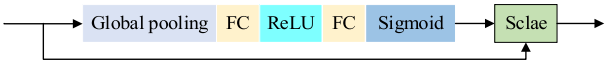


FIGURE 2. SENet structure.

III. IMPROVEMENTS TO THE YOLOV4 NETWORK

A. NETWORK LIGHTWEIGHTING IMPROVEMENTS

By improving the backbone network CSPDarknet53 of YOLOv4 can effectively reduce the training cost and improve the detection speed, the MV1, MV2 and MV3 network models are used to replace the backbone network CSPDarknet53 of YOLOv4 to obtain Mobilenetv1_YOLOv4 (MV1_Y), Mobilenetv2 _YOLOv4 (MV2_Y), and Mobilenetv3_YOLOv4 (MV3_Y) networks, and connects with the subsequent networks according to the input sizes of each layer of the networks, so that the MobileNet family of networks and the following detection networks of the YOLOv4 network can match. Taking MV2 as an example, the structure diagram of the replaced network is shown in Fig. 1.

B. NETWORK ACCURACY OPTIMIZATION

To reduce the computational parameters and improve the detection speed, the MV2 network model is used to replace the backbone network CSPDarknet53 of YOLOv4, sacrificing part of the performance of the original YOLOv4 network model. Some scholars study that the channel attention mechanism SENet can effectively improve the model's performance [20]; in this paper, SENet is embedded into the network model MV2_Y. SENet mainly focuses on the relationship between the channel features and consists of two parts, Squeeze and Excitation, as shown in Fig. 2.

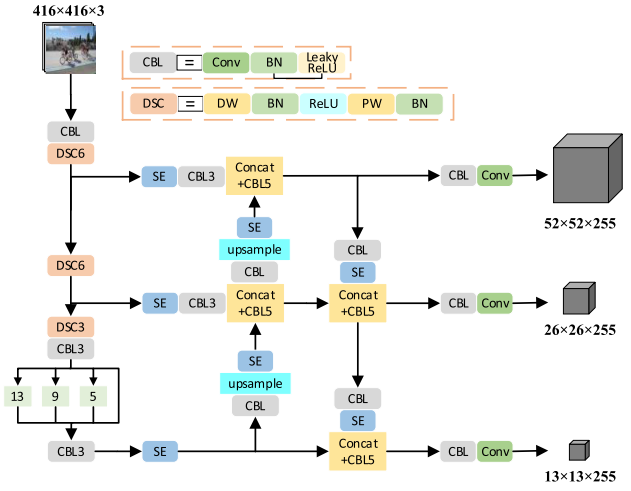


FIGURE 3. MV2-S-Y network structure.

SENet first performs a global average pooling operation on the input feature layer to obtain a feature map of size $1 \times 1 \times C$ (C is the number of channels), after which it predicts the importance of the channel features after two fully connected layers, obtains a weight of size $1 \times 1 \times C$, and uses this weight to multiply it with the corresponding channels of the original feature map, and finally outputs the result.

Embedding the SENet module into the PANet of the MV2_Y network can increase the receptive field and enhance the channel characteristics with a minor parameter cost. As shown in Fig. 3, the SENet module is embedded in the up-sampling and down-sampling processes of the PANet, and the SENet is denoted by SE in Fig. 3, resulting in richer semantic information. Meanwhile, to further lighten the improved network, the MV2_SENet_Y (MV2_S_Y) network model is obtained by using depth-separable convolution to replace the rest of the standard convolution in the model.

C. NETWORK LOSS FUNCTION

In the computer vision-based road object detection task, insufficient light, small objects, and object occlusion in the road environment can cause difficulty detecting image samples. To further improve the detection ability of the object detection network for complex samples, optimization is performed in terms of the loss function of the network.

The EIOU loss function [29] separates the aspect ratio based on the CIOU loss function. It calculates the difference between the width and height of the prediction box and the minimum outer rectangle, respectively, which can reflect the actual difference between the width and height, thus accelerating the convergence speed of the network. In addition, the EIOU loss function also adds the Focal idea; the Focal idea can deal with the uneven problem of sample classification, realize the detection of complex samples, and help improve the network's training effect. The EIOU loss function will judge the current training results of the strengths and weaknesses of the recent training results during the training process of the road object detection network and provide feedback to the network for the adjustment of the parameters.

To further improve the detection accuracy of the MV2-S-Y network, the EIOU loss function is introduced into the MV2-S-Y network instead of the original CIOU loss function in the network to obtain the MV2-S-YE network and train the network.

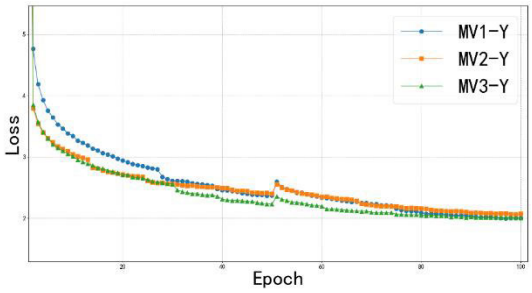


FIGURE 4. Three kinds of network Loss changes during training.

IV. EXPERIMENTATION AND ANALYSYS

A. PLATFORM AND PARAMETER SETTING

1) The hardware platform used for the experiment: Intel Core i9-10900X for CPU, two NVIDIA GeForce RTX 3080 10G graphics cards for GPU, Windows 10 for operating system, TensorFlow2.5 for deep learning framework, and CUDA11.0 and CUDNN8.4 to accelerate the model training process. CUDA11.0 and CUDNN8.4 were used to accelerate the model training process.
2) Parameter settings: The epoch is set to 100, the Batch size is set to 8, the initial learning rate is set to 0.001, the minimum learning rate is set to 0.000001, and the learning rate decay strategy is used with a decay rate of 0.5.
3) The dataset is divided by setting the training set and the test set share to 90% and 10%, respectively [30]. We selected 19,352 samples as the training set and 2,152 samples as the test set in the Pascal VOC dataset, 21,081 samples as the training set and 3,342 samples as the test set in the Udacity dataset, and 6,840 samples as the training set and 760 samples as the test set in the KAIST dataset.

B. EXPERIMENTAL ANALYSYS OF NETWORK IMPROVEMENTS

1) NETWORK TRAINING

To verify the improvement of MV1-Y, MV2-Y, and MV3-Y networks, training was performed on Pascal VOC and Udacity datasets using the experimental platform and parameter settings in IV-A, and the training results showed similar changes in the loss values of MV1-Y, MV2-Y and MV3-Y networks on the two datasets. The variation of Loss values during training on the Pascal VOC dataset was chosen to be plotted, as shown in Fig. 4.

TABLE 1. Comparison of parameters of each network.

Network	Parameters(M)	Calculations(G)	Model size(MB)	FPS ($f \cdot s^{-1}$)
YOLOv4-Tiny	6.06	6.92	22.6	138
YOLOv4	64.43	60.32	245.3	30
MV1-Y	12.43	10.12	47.6	53
MV2-Y	10.55	7.67	40.6	49
MV3-Y	11.47	7.15	44.3	43

TABLE 2. Comparison of detection performance of networks in PASCAL VOC data sets.

Network	mAP@0.5 (%)（P VOC）	mAP@0.5 (%)（Udacity）
YOLOv4-Tiny	77.5	58.6
YOLOv4	83.9	70.1
MV1-Y	80.1	66.1
MV2-Y	79.7	65.1
MV3-Y	78.5	63.8

As seen in Fig. 4, the Loss values of MV2-Y and MV3-Y networks decrease faster than those of MV1-Y at the beginning of training, and then the convergence rate gradually slows down. The fluctuation of the Loss value at the 50th Epoch is because the network in the unfrozen part has not yet learned, so that it will lead to a small increase in the Loss value. The Loss value continues to decrease as training proceeds, and eventually, the Loss value stays near two and stops falling.

2) ANALYSYS OF RESULTS

To verify the improvement effect of MV1-Y, MV2-Y, and MV3-Y networks, YOLOv4-Tiny [31] and YOLOv4 networks were added. After training using the experimental platform and parameter settings in IV-A, the parametric quantities, computational quantities, model sizes, and computing speeds of the models of each network are compared, as shown in Table 1.

As can be seen from Table 1, the complexity of the model after improving the feature extraction network of YOLOv4 decreases dramatically. The number of parameters and computational amount of the model of each improved network are reduced substantially. The number of parameters of MV2-Y is the smallest among the three improved networks, which is only 10.55 M, and it reduces the parameter amount by 83.6 % compared with that of the YOLOv4 network. The MV3-Y has the smallest computation amount of 7.15 G. The model size of MV2-Y is 40.6 MB, which is 204.3 MB less than that of the YOLOv4 network, and the detection speed of MV1-Y is the fastest with 53 frames per second, and MV2-Y is the second fastest with up to 49 FPS.

A comparison of the detection performance of each network on the Pascal VOC dataset and the Udacity dataset is shown in Table 2.

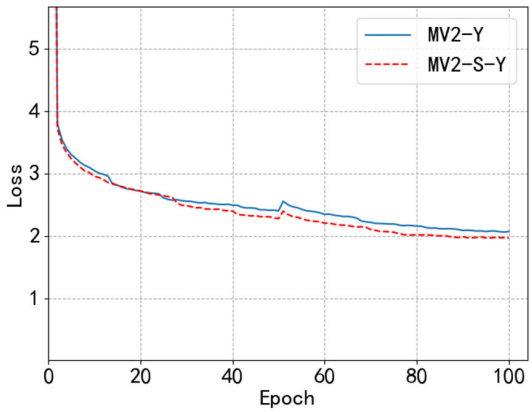


FIGURE 5. Loss change of network during training.

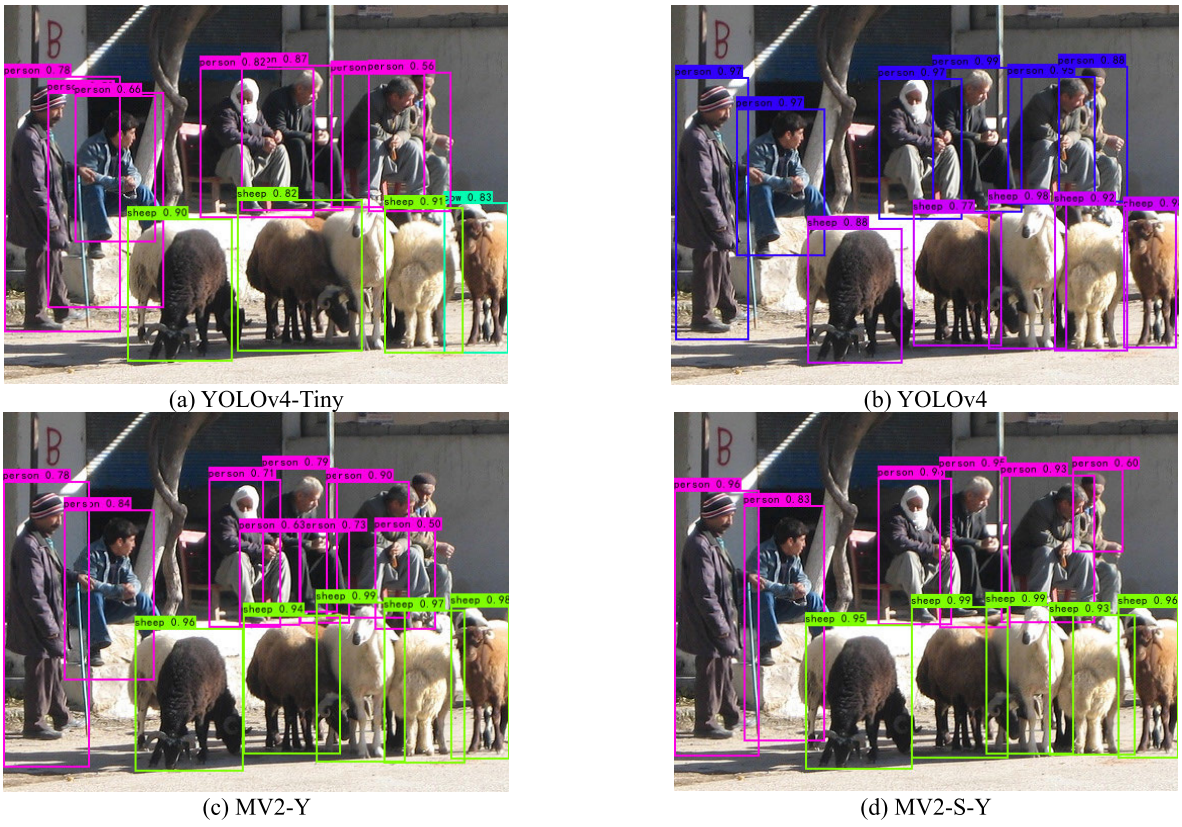
As can be seen from Table 2, in the test results of the Pascal VOC dataset, the mAP@0.5 of the MV1-Y network and MV2-Y network differs by 0.5% and is higher than the detection accuracy of YOLOv4-Tiny and MV3-Y. The mAP@0.5 of the MV2-Y network reaches 79.7 %, which is in the middle of the three improved networks; in the Udacity dataset test results, MV1 performs better as a feature extraction network, with a mAP@0.5 up to 66.1 %. the MV2-Y network has the next best performance in terms of mAP@0.5, which can reach 65.1 %.

The experiments show that although the improved network model has increased detection speed, the enhanced feature extraction network leads to different degrees of degradation in network detection accuracy. From Table 3-6, it can be seen that the MV2-Y network has a better performance in terms of the number of parameters, the amount of computation, and the size of the model, and the detection speed can be up to 49 frames per second, which is 63 % higher than that of YOLOv4. Regarding detection accuracy, the gap between the MV2-Y network and the MV1-Y network is smaller, and the difference in detection accuracy with the YOLOv4 network is 4.2 %. Still, the model of the MV2-Y network is smaller, which is more advantageous in terms of the number of parameters and the amount of computation. In summary, the MV2-Y network was selected for further research.

C. EXPERIMENTAL ANALYSYS OF NETWORK ACCURACY OPTIMIZATION

1) NETWORK TRAINING

To verify whether the performance of the MV2-S-Y network is improved, training is performed on the Pascal VOC,



(a) YOLOv4-Tiny (b) YOLOv4

(c) MV2-Y (d) MV2-S-Y

FIGURE 6. Detection of Pascal VOC samples by four networks.

Udacity, and KAIST datasets using the experimental platform and parameter settings in IV-A, and the training results show that the MV2-Y network and the MV2-S-Y network have similar changes in the loss values on the three datasets. The variation of Loss values during training on the Pascal VOC dataset was selected to be plotted, as shown in Fig. 5.

In the first 30 Epochs, the change in the Loss values of the two networks does not differ much. After 30 Epochs, the Loss values of the MV2-S-Y network converge faster. After the 50th Epoch, when all network layers are unfrozen, both networks' loss values slightly increase and then gradually decrease. The Loss values converge at a comparable rate, but the Loss values of the MV2-S-Y network can connect to smaller values, making the network more accessible to train.

2) ANALYSIS OF RESULTS

The mAP@0.5 is used as a metric to evaluate the model's performance. The performance test is conducted on the Pascal VOC, Udacity and KAIST datasets, and the specific test results are shown in Table 3.

The MV2-S-Y network improved the mAP@0.5 by 1.3 %, 2.2 %, and 0.5 % on the Pascal VOC, Udacity, and KAIST datasets compared to the MV2-Y network. Compared to the YOLOv4-Tiny network model, the mAP values were improved by 4.0 %, 13.5 %, and 3.5 % on the Pascal VOC, Udacity, and KAIST datasets, respectively. This shows that introducing the SENet module to the MV2-S-Y network can effectively improve the detection of the network.

TABLE 3. Comparison of detection results of four networks.

Network	Backbone	mAP@0.5 (%)		
		Pascal VOC	Udacity	KAIST
YOLOv4-Tiny	CSPDarknet53-Tiny	77.5	58.6	91.2
YOLOv4	Darknet53	83.9	70.1	97.8
MV2-Y	MobileNetV2	79.6	65.1	93.9
MV2-S-Y	MobileNetV2	80.6	66.5	94.4

To compare the detection effect of the four algorithms more intuitively, one complex sample image is selected from Pascal VOC for detection, and the detection results of the four algorithms on the sample image at a threshold of 0.5 are shown in Fig. 6, respectively.

The image in Fig. 6 shows two types of objects, human and sheep, resting on the side of the road, where there are six objects for the human and six objects for the sheep, and the specific results of the four networks for detecting samples of the Pascal VOC dataset are shown in Table 4.

As shown in Table 4, the MV2-S-Y network is more accurate. It has a low leakage rate for both types of object detection in the figure, detecting six people and five sheep respectively, with a confidence level of 87.2 and 96.4 for

TABLE 4. Detection results of PASCAL VOC samples by four networks.

Network	Number of humans and sheep	AP (%)	mAP (%)
YOLOv4-Tiny	humans、7	77.0	82.4
	sheep、3	87.7	
YOLOv4	humans、6	95.5	93.1
	sheep、5	90.6	
MV2-Y	humans、8	73.5	85.2
	sheep、5	96.8	
MV2-S-Y	humans、6	87.2	91.8
	sheep、5	96.4	

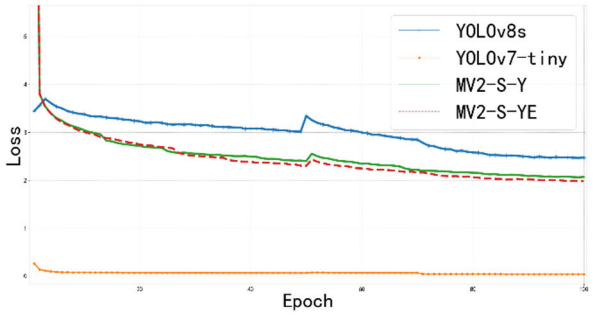


FIGURE 7. Loss changes during training of four networks.

human and sheep objects, respectively, and the mAP reaches 91.8%; the YOLOv4-Tiny algorithm has misdetections for both human and sheep objects and has a high leakage rate, and the confidence level of the detected objects is also lower than that of MV2-S-Y network. Confidence is also lower than that of the MV2-S-Y network.

D. EXPERIMENTAL ANALYSYS OF LOSS FUNCTION OPTIMIZATION

1) NETWORK TRAINING

To verify whether the performance of the MV2-S-YE network is improved, training is performed on the Pascal VOC, Udacity, and KAIST datasets using the experimental platform and parameter settings in IV-A, and the training results show similar changes in the loss values of the YOLOv8s,YOLOv7-tiny, MV2-S-Y network and the MV2-S-YE network on the three datasets. The plotting of the change in loss values during training on the Pascal VOC dataset was selected and is shown in Fig. 7.

From Fig. 7, it can be seen that the four networks stabilize at the 80th-100th epochs of training, MV2-S-YE and MV2-S-Y networks converge at a similar rate at the early part of training, and the MV2-S-YE network converges faster after 30 epochs to around 2.0. YOLOv8s network only converges around 2.5, and there is a gap with MV2-S-YE network.

2) ANALYSIS OF RESULTS

After using the EIOU loss function instead of the CIOU loss function, the number of parameters, computation, model size, and computational speed of the MV2-S-YE, MV2-S-Y, YOLOv8s, and YOLOv7-tiny network models are shown in Table 5.

TABLE 5. Comparison of two kind of network parameters and running speed.

Network	Parameters(M)	Calculations(G)	Model size(MB)	FPS ($f \cdot s^{-1}$)
MV2-S-Y	11.91	7.68	41.2	45
MV2-S-YE	11.91	7.68	41.2	45
YOLOv8s	15.56	8.52	42.7	42
YOLOv7-tiny	9.5	6.55	23.3	143

TABLE 6. Detection results of MV2-S-YE network.

Network	Loss function	mAP@0.5(%)		
		Pascal VOC	Udacity	KAIST
MV2-S-Y	CIOU	80.6	66.5	94.4
MV2-S-YE	EIOU	80.9	66.7	94.8
YOLOv8s	CIOU	81.0	66.7	91.6
YOLOv7-tiny	CIOU	76.6	63.5	88.1

As can be seen from Table 5, MV2-S-YE and MV2-S-Y with the EIOU loss function are the same in terms of performance parameters, and YOLOv8s has a larger number of parameters, computation and model size than MV2-S-YE, and the FPS is smaller than that of MV2-S-YE. YOLOv7-tiny has a higher performance than MV2-S-YE in all the four categories.

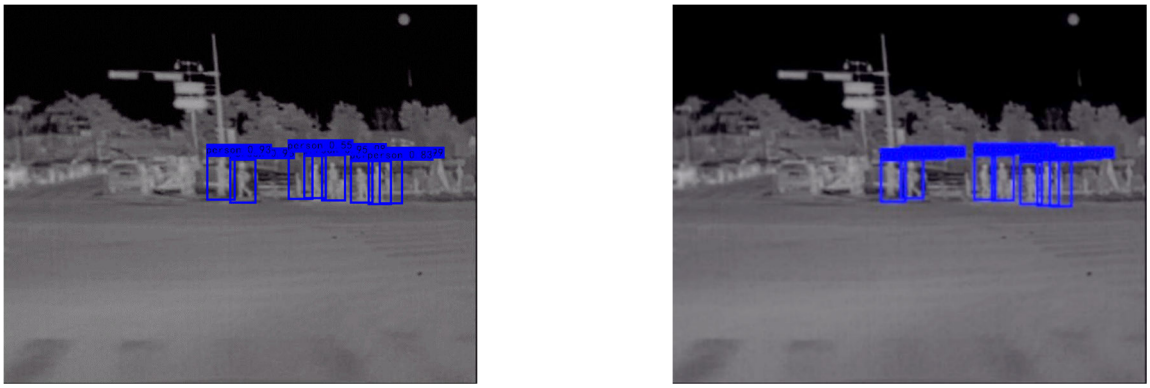
To verify the performance of MV2-S-YE, it was tested on Pascal VOC, Udacity and KAIST datasets using the weight files obtained from training, and the results are shown in Table 6.

From TABLE 6, it can be seen that the MV2-S-YE network using the EIOU loss function improves the mAP of the network model by 0.3%, 0.2% and 0.4% after training on the Pascal VOC, Udacity and KAIST datasets, respectively. Combined with Table 5 it can be concluded that the EIOU loss function can improve the detection accuracy.The mAP of YOLOv8s on Pascal VOC is just 0.2% higher than that of MV2-S-YE, and lower on the other two datasets. mAP of YOLOv7-tiny is lower than that of MV2-S-YE on all three public datasets.

For a more precise comparison of the performance improvement of the object detection network using the EIOU loss function, 1 sample image was selected from the KAIST dataset for detection, and the detection results of the MV2-S-Y and MV2-S-YE networks for the sample image at a threshold of 0.5 are shown in Fig. 8, respectively.

The image in Fig. 8 is a nighttime intersection scene with only 1 class of actual objects and seven pedestrians, and the detection results are shown in Table 7.

As shown in Table 7, the two networks obtained better detection results for distant roadside pedestrians, in which the MV2-S-Y network showed one false detection. The confidence level of the detection results of the MV2-S-YE



(a) MV2-S-Y (b) MV2-S-YE

FIGURE 8. Detection of KAIST samples by two networks.

TABLE 7. Detection results of KAIST samples by two networks.

Network	Number of pedestrians	AP(%)
MV2-S-Y	8	89.8
MV2-S-YE	7	95.7

network was higher than that of the MV2-S-Y network. The localization of the pedestrians was more accurate, which is advantageous for detecting the difficult samples in the complex environment.

V. CONCLUSION

To meet the real-time requirement of object detection speed for the autopilot in-vehicle network model, the Mobilenetv2 network model is used to replace the backbone network CSPDarknet53 of YOLOv4, which makes the network model significantly reduced and improves the detection speed while sacrificing part of the detection accuracy. To compensate for the degradation of the detection performance and to take advantage of the channel attention mechanism SENet module to improve the model performance, the channel attention mechanism SENet is introduced in PANet to optimize the feature extraction capability by adding the channel attention mechanism to assign different weights to each channel. To further improve the detection ability of MV2-S-Y for complex samples, the MV2-S-YE network is obtained by using the EIOU loss function instead of the original CIOU loss function in the network when the introduction of the EIOU loss function does not bring adverse effects such as an increase in the size of the model and computation of the network. To increase the rigor of the work, YOLOv8s and YOLOv7-tiny networks are introduced as comparisons, and MV2_S_YE has advantages in various detection performances.MV2_S_YE network not only has the advantages of high detection accuracy and fast detection speed in complex environments, but also has a powerful real-time monitoring capability.In the future, we intend to continue to optimize the MV2_S_YE network in terms of parameter count, accuracy, and real-time performance and apply it to small embedded devices. Eventually, compared with other network models in the paper, the MV2_S_YE network has a real-time detection speed of up to 45 FPS on the KAIST test set, mAP@0.5 is 94.8%, and improves by 3.2% compared with the YOLOv8s networ. Thus, MV2_S_YE is innovative and can fulfill the requirements of the vehicle network model for object detection.

Our work is still in its early stages, and in the future, we will continue to try to compress the model, utilize channel pruning methods to reduce the number of parameters and increase the detection speed, so as to piggyback the model on low-cost hardware devices and reduce the cost of target detection applications; at the same time, we are considering to continue to look for ways to improve the accuracy of the model. Ultimately, we are committed to applying low-cost, high-precision and high-efficiency models to real-world object detection projects.

REFERENCES

[1] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4316–4336, Jul. 2021.

[2] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, and L. Chen, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 6, pp. 3692–3711, Jun. 2023.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[4] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE Access*, vol. 8, pp. 24344–24357, 2020.

[5] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3500–3509.

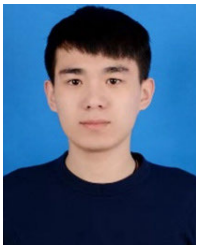
[6] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated RCNN and multi-class deep SORT for multi-object detection and tracking," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 1, pp. 171–181, Feb. 2022.

[7] L. Yang, J. Zhong, Y. Zhang, S. Bai, G. Li, Y. Yang, and J. Zhang, "An improving faster-RCNN with multi-attention ResNet for small target detection in intelligent autonomous transport with 6G," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–9, 2022.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 7263–7271.
[9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[10] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, and Z. Han, "Effective fusion factor in FPN for tiny object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1159–1167.
[11] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[12] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.
[13] L. Tang, Y. Wang, and L.-P. Chau, "Weakly-supervised part-attention and mentored networks for vehicle re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8887–8898, Dec. 2022.
[14] M. Haris and A. Glowacz, "Road object detection: A comparative study of deep learning-based algorithms," *Electronics*, vol. 10, no. 16, p. 1932, Aug. 2021.
[15] Z. Chen, H. Guo, J. Yang, H. Jiao, Z. Feng, L. Chen, and T. Gao, "Fast vehicle detection algorithm in traffic scene based on improved SSD," *Measurement*, vol. 201, Sep. 2022, Art. no. 111655.
[16] Y. Cai, T. Luan, H. Gao, H. Wang, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "YOLOv4–5D: An effective and efficient object detector for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[17] L. Wang, K. Zhou, A. Chu, G. Wang, and L. Wang, "An improved light-weight traffic sign recognition algorithm based on YOLOv4-tiny," *IEEE Access*, vol. 9, pp. 124963–124971, 2021.
[18] H. Wang, Y. Xu, Z. Wang, Y. Cai, L. Chen, and Y. Li, "CenterNet-auto: A multi-object visual detection algorithm for autonomous driving scenes based on improved CenterNet," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 3, pp. 742–752, Jun. 2023.
[19] S. Du, P. Zhang, B. Zhang, and H. Xu, "Weak and occluded vehicle detection in complex infrared environment based on improved YOLOv4," *IEEE Access*, vol. 9, pp. 25671–25680, 2021.
[20] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, "A new image recognition and classification method combining transfer learning algorithm and MobileNet model for welding defects," *IEEE Access*, vol. 8, pp. 119951–119960, 2020.
[21] X. Jin, Y. Xie, X.-S. Wei, B.-R. Zhao, Z.-M. Chen, and X. Tan, "Delving deep into spatial pooling for squeeze-and-excitation networks," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108159.
[22] K. Kadam, S. Ahirrao, K. Kotecha, and S. Sahu, "Detection and localization of multiple image splicing using MobileNet v1," *IEEE Access*, vol. 9, pp. 162499–162519, 2021.
[23] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Embedded control gate fusion and attention residual learning for RGB-thermal urban scene parsing," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 4794–4803, May 2023.
[24] G. Li, H. Fan, G. Jiang, D. Jiang, Y. Liu, B. Tao, and J. Yun, "RGBD-SLAM based on object detection with two-stream YOLOv4-MobileNetv3 in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–11, Jun. 2023.
[25] G. Li, Y. Qiu, Y. Yang, Z. Li, S. Li, W. Chu, P. Green, and S. E. Li, "Lane change strategies for autonomous vehicles: A deep reinforcement learning approach based on transformer," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 3, pp. 2197–2211, Mar. 2023.
[26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[27] X. Wang and J. Song, "ICIoU: Improved loss based on complete intersection over union for bounding box regression," *IEEE Access*, vol. 9, pp. 105686–105695, 2021.
[28] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018.
[29] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.
[30] X. Wang, P. Wang, J. Song, T. Hao, and X. Duan, "A TEDE algorithm studies the effect of dataset grouping on supervised learning accuracy," *Electronics*, vol. 12, no. 11, p. 2546, Jun. 2023.
[31] Y. Yao, L. Han, C. Du, X. Xu, and X. Jiang, "Traffic sign detection algorithm based on improved YOLOv4-tiny," *Signal Process., Image Commun.*, vol. 107, Sep. 2022, Art. no. 116783.

PENGHUI WANG received the B.S. degree in vehicle engineering from the Shaanxi University of Technology, China, in 2021, where he is currently pursuing the M.S. degree in mechanics. His research interests include machine learning, computer vision, and autonomous driving.

XUFEI WANG received the B.S. degree in mechanical engineering from the Shaanxi University of Technology, China, in 1999, the M.S. degree in mechanical engineering from Xinjiang University, China, in 2007, and the Ph.D. degree in computer engineering from Pai Chai University, South Korea, in 2022. Since 2022, he has been a Professor with the Shaanxi University of Technology. His research interests include machine learning and system modeling and simulating.



YIFAN LIU received the master's degree in vehicle engineering from the Shaanxi University of Technology, Hanzhong, China. He is currently teaching with the Sanmenxia College of Social Administration, Sanmenxia, China. His research interests include object detection and autonomous driving.

JEONGYOUNG SONG (Member, IEEE) received the B.S. degree in computer engineering from Hannam University, South Korea, in 1984, and the M.S. and Ph.D. degrees in electrical information and systems from Waseda University, Japan, in 1992 and 1995, respectively. From 1995 to 1997, he was a Researcher of computer science with Cheongun University, South Korea. Since 1997, he has been a Professor with the Computer Engineering Department, Pai Chai University, South Korea. From 2011 to 2012, he was an invited Scholarship Professor with the Department of Electrical Engineering, Idaho State University, USA. His research interests include pattern processing (image, speech, and character) and machine learning.