Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 8
AICT 2024

YOLOv9-SE: A Real-time Vehicle Detection Model Based on Improved YOLOv9

Yangzi Gao*

School of Advanced Manufacturing, Guangdong University of Technology, Jieyang, China

* Corresponding Author

Abstract. With an increasing number of urban vehicles and complex road environments, real-time vehicle detection has become a key technology in autonomous driving, but it faces many challenges. Although traditional two-step target detection algorithms (such as the R-CNN series) have high detection accuracy, their real-time performance is poor, which makes it difficult to meet the needs of vehicle detection. In contrast, one-step detection algorithms such as YOLO stand out for their high speed and higher accuracy. However, the real-time detection of the latest YOLOv9 model in urban vehicle scenarios still needs to be improved. Therefore, this paper have improved the YOLOv9 model, specifically introducing the SENetV1 attention mechanism into the backbone extraction network. The experimental results show that the mAP value of the improved algorithm in vehicle detection is promoted by 5% under the same training conditions. Such an improvement not only enhances the ability to capture relationships between channels, but also improves feature expression capabilities and expands the application of YOLOv9 in autonomous driving.

Keywords: YOLOv9, Real time vehicle detection model, Autonomous driving.

1. Introduction

The progress of vehicle research and development technology has ushered in a new research direction for computer vision, which is an integral technology in autonomous driving, that is, real-time vehicle detection. It requires the processing and analysis of vehicle images. Vehicle inspection plays an indispensable role in real life. However, with more and more vehicles in cities and the increasingly changeable road conditions, real-time vehicle detection is a challenging task. Therefore, a fast and accurate target detection algorithm is of great necessity.

Vehicle inspection methods have been developing in academia and industry for several years. The main problems of vehicle detection are large variations of light, dense occlusion and large variation of object scale. The previous detection methods are mainly divided into the one-step method and the two-step method. At present, the mainstream deep learning detection algorithms of the two-step method include region convolutional neural network (R-CNN) [1], fast region convolutional neural network (FastR-CNN) [2], and faster region convolutional neural network (Faster-R-CNN) [3]. These algorithms all create some candidate regions, extract the features in the candidate boxes, and then classify the targets. In addition, these algorithms have high detection accuracy with a relatively slow speed, and the delay of real-time detection is relatively high, so they can't show good results when applied to real-time vehicle detection. The one-step detection algorithm is more suitable for real-time vehicle detection, because it does not need to create candidate areas and candidate boxes in advance. Instead, the one-step detection algorithm directly predicts them through feature maps, so its detection speed is greatly improved and the delay of real-time vehicle detection is greatly reduced. Moreover, YOLO (You Only Look Once) algorithm with high speed and high accuracy has leapt out, which has become one of the most widely used models in target detection. In recent years, the YOLO algorithm has been continuously improved, and YOLOv3, YOLOv5, YOLOv8, YOLOv9 [4] [5] [6] [7], etc. have come into being. Besides, the YOLOv9 algorithm with higher detection accuracy and faster speed introduces pioneering technologies such as programmable gradient information (PGI) and generalized high-efficiency layer aggregation network (GELAN), and designs a new lightweight network architecture GELAN based on gradient path planning, which can achieve better parameter

utilization only by using traditional convolution operators. The efficiency, accuracy and adaptability of real-time target detection are improved. However, the existing YOLOv9 is not excellent enough in real-time vehicle detection, especially in highway scenarios with high detection speed requirements. It's urgent to require an efficient and high-performance algorithm.

We improved the network structure of YOLOv9, and added the SENetV1 attention mechanism to the backbone extraction network [8]. After the improvement, the experiment achieved effective results. Under the same 100 rounds of training times, the mAP value of the vehicle detection experiment is 5% higher than that of the basic YOLOv9. We have expanded the application of the YOLOv9 algorithm, which can be applied to automobile detection to improve autonomous driving technology and the accuracy of real-time vehicle detection. At the same time, we proposed a new architecture, that is, adding the SENetV1 attention mechanism to the backbone extraction network (Backbone). After that, the ability of the backbone extraction network to capture relationships between channels is enhanced and the overall feature expression ability is improved.

In the following chapters, we will introduce some methods of the current target detection and vehicle detection in the chapter of related work. The architecture of our model and some details of the model in the YOLOv9 algorithm will be explained in the chapter of the YOLOv9 algorithm and its improvement. In addition, we will introduce some experimental data, evaluation indexes, experimental environment and experimental results in the chapter of experimental results. The contribution and future work will be unfolded in the chapter of discussion, with the summary of our work in the conclusion.

2. Related Work

2.1. Object Detection

Object detection is key in computer vision. Its task is to identify what objects are in the image and give the position of the objects. At present, object detection models based on deep learning can be divided into two categories. One is the two-stage models represented by R-CNN, Fast RCNN, and Faster R-CNN. They are characterized by generating some candidate regions in advance, extracting features in candidate boxes, and then classifying targets. The detection accuracy of these algorithms is very high, but the speed is relatively slow. The other is one-stage object detection models represented by SSD, YOLO, EfficientDet, and RetinaNet [9] [10] [11]. The characteristic of these models is that they do not need to create candidate regions and candidate boxes in advance, but directly predict them through feature maps. Their detection speed is greatly improved, but their detection accuracy is not as good as that of two-stage models. Nowadays, detection technology is developing and optimizing, and the accuracy and efficiency of detection models based on deep learning are also in continuous improvement. Besides, YOLOv9 shows good performance in real-time object detection, which makes it feasible to realize real-time vehicle detection based on YOLOv9.

2.2. Vehicle Detection Technology

Real-time vehicle detection technology is indispensable to realizing autonomous driving and building smart cities. With the development of deep learning, there are two types of vehicle detection algorithms, namely one-stage algorithms and two-stage algorithms. Compared with the two-stage algorithm, the one-stage algorithm is faster, which is very practical in vehicle detection scenarios that require real-time speed detection. YOLO algorithm series is the most popular one-stage detection algorithm in current applications, because they guarantee fast and high accuracy simultaneously. At present, the improvement of the detection model mainly focuses on the loss function, feature extraction network and label allocation. In addition to using YOLO and RCNN series for vehicle detection, there are some improved methods to meet the requirements of real traffic scenarios. YOLOv5 is still one of the commonly used detectors for vehicle detection at present due to its good versatility and considerable balance between speed and accuracy. In order to design a lighter vehicle detection model, Dong et al. [12] not only introduced a ghost module based on YOLOv5-S to reduce

model complexity and number of parameters, but also introduced CBAM to reduce redundant noise. Bie et al. [13] incorporated moving inverted bottleneck convolution (MBConv) to reduce computational parameters while improving performance, and introduced BiFPN to enhance feature fusion. Guo et al. [14] integrate Transformer into RetinaNet to improve global modeling capabilities. Xu [15] combined ShuffleNetv2 and BiFPN for accurate and efficient object detection.

However, real-time vehicle detection requires high speed and accuracy. Although Yolov9 has good accuracy and speed, it does not understand the relationship between channels through the attention mechanism, which leads to its vehicle detection effect not reaching expectations. Therefore, the method of adding the SENetV1 attention mechanism to solve low vehicle detection accuracy is put forward in this paper.

3. YOLOv9 Algorithm and Improvement

3.1. Backbone Network: YOLOv9

YOLOv9 is a version of the YOLO (You Only Look Once) series of object detection algorithms. Here are some main features and information about YOLOv9:

1. YOLOv9 introduces PGI. It consists of a master branch (architecture used for inference), an auxiliary reversible branch (generating a reliable gradient to provide back transmission for the master branch), and multi-level auxiliary information (controlling the master branch to learn programmable multi-level semantic information). By assisting the reversible branch to calculate the loss and gradient, the loss of information in layer-by-layer modeling is reduced, and only participates in the calculation loss in training, which does not affect the inference time.

2. Generalized High Efficiency Layer Aggregation Network (GELAN). It is a brand-new network architecture that optimizes the network structure through gradient path planning, which mimics CSPNet and extends ELAN to GELAN to support any compute block. Using only traditional convolution operators, GELAN can achieve better parameter utilization than advanced methods based on depth convolution. Meanwhile, GELAN can exhibit robust and stable performance under different computational modules and depth settings, which can be widely expanded into models suitable for a variety of inference devices.

3. The parameters and calculations of YOLOv9 are reduced and the accuracy is improved. Compared with some previous models, such as YOLOv7 of lightweight and medium models, the parameters of YOLOv9 are reduced by about 10% and the calculations are reduced by 5-15%. However, there is still an improvement of 0.4-0.6% in terms of average accuracy (AP). Compared to the large model YOLOv8-x, YOLOv9-x has a 15% reduction in parameters, a 25% reduction in calculations, and a significant improvement in AP of 1.7%.

The excellent performance of YOLOv9 makes it have wide application potential in various scenarios related to target detection, such as autonomous driving, intelligent monitoring, augmented reality and other fields.

3.2. Improvement Work: Embedding SENetV1 Attention Mechanism

As a stack of SE structures, SENet is not an independent network model, but a module that can be combined with any existing model. The SE structure in SENet mainly consists of Squeeze, Excitation and Scale. By adaptively recalibrating the feature channels, the network can learn the importance level of each feature channel, thus improving the performance of the model.

When multiple SE structures are stacked at different stages in the network, features can be optimized and adjusted from various levels. This stacking method enables SENet to capture and emphasize important feature information in network layers of different depths, enhancing the model's adaptability to different tasks and data.

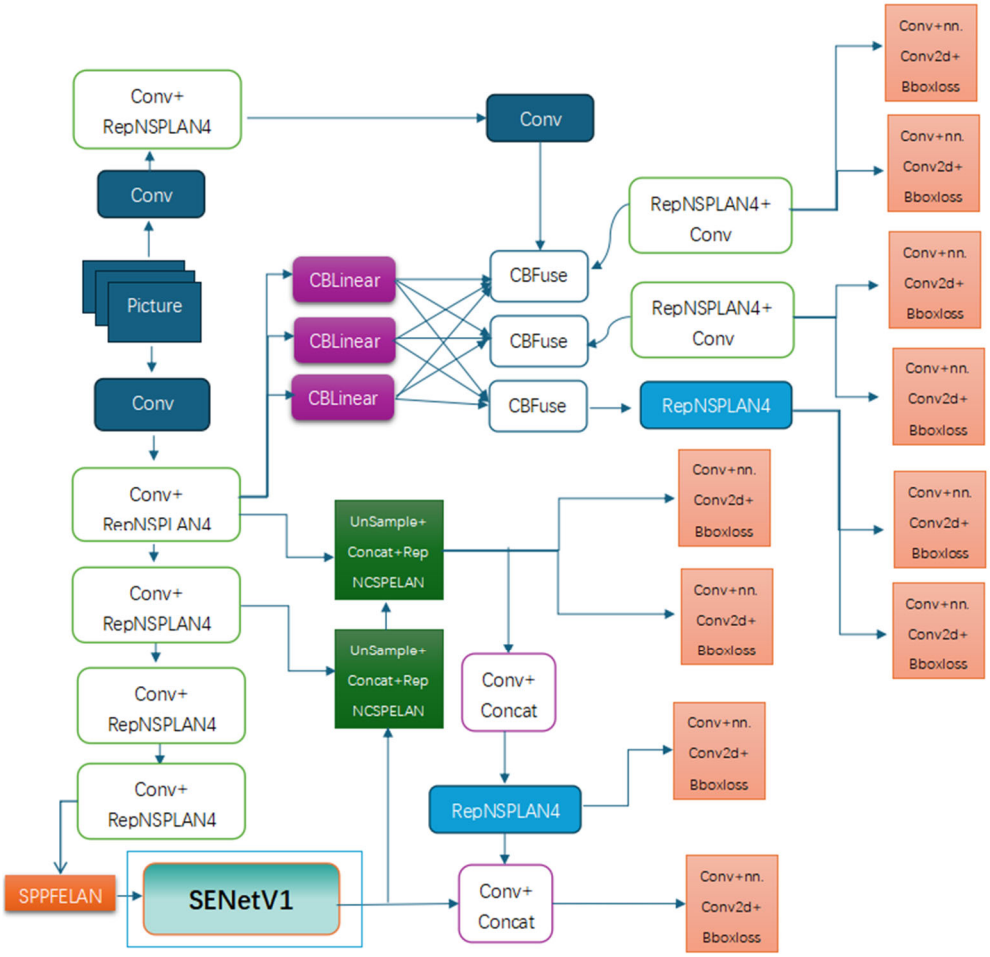

Figure 1. YOLOv9-SE

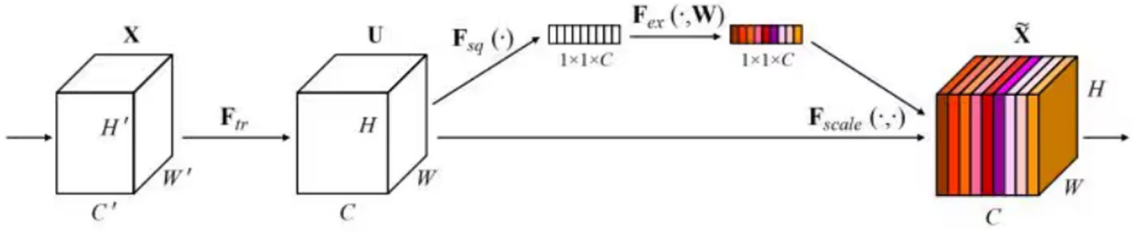

Figure 2. SE Attention Mechanism

Detailed Explanation of SE Structure:

1. Squeeze (Global Information Embedding):

Fsq is the global average pooling using channels. In other words, the global average pooling of channels is used to compress the W×H×C feature map containing global information into a 1×1×C feature vector Z, and the channel features of C feature maps are all compressed into a numerical value. It makes the generated channel-level statistical data Z contain context information, alleviating the channel dependence, which is defined as follows:

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j)$$

2. Excitation (Adaptive Recalibration)

Purpose: To utilize the information aggregated in the compression operation, we comprehensively capture channel dependencies through the Excitation.

The first fully connected layer compresses C channels into C/r channels to reduce the amount of calculation, and then passes a RELU nonlinear activation layer. The second fully connected layer restores the number of channels back to C channels, and then passes Sigmoid activation to obtain the weight s. The final dimension of s is 1×1×C, which is used to describe the weights of C feature maps in the feature map U. r refers to the proportion of compression, with the formula described as follows:

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma\left(\mathbf{W}_2 \delta\left(\mathbf{W}_1 \mathbf{z}\right)\right)$$

3. Scale (Re-weighted)

Purpose: Scale weights the attention weights obtained earlier to the features of each channel. Implementation method: Each feature map in the feature map U is multiplied by the corresponding weight to obtain the final output of the SE module. The formula is described as follows:

$$\tilde{\mathbf{x}}_c = \mathbf{F}_{\text{scale}}\left(\mathbf{u}_c, s_c\right) = s_c \mathbf{u}_c$$

Description of Improvements:

SENetV1 integrates the SE module into the Inception module and the residual error (ResNet) module respectively. Figure 3 shows the structure diagram of two neural network modules, including the Inception module and the residual error (ResNet) module. Each module has its standard form and a modified form. The comparison chart also incorporates the Squeeze-and-Excitation (SE) block to improve performance.

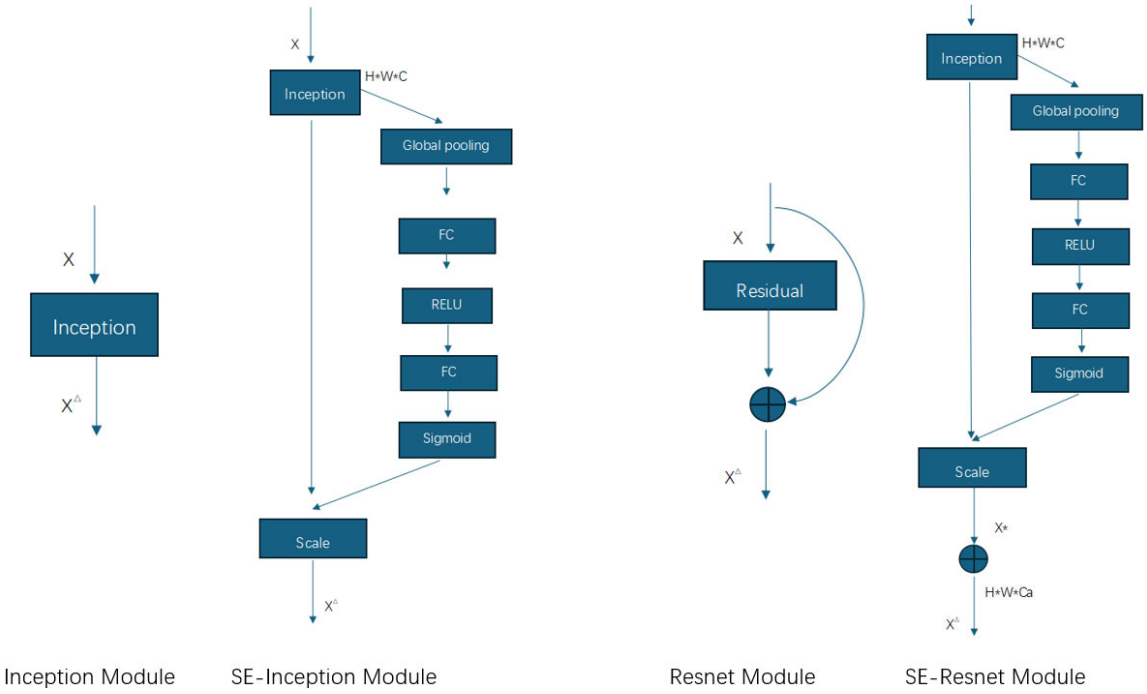

Figure 3. Modules of SE-Inception and SE-Resnet

The parts on the left are the original Inception module (left) and the SE-Inception module (right). The SE-Inception module generates channel-level weights through global average pooling and two fully connected layers (the first using the ReLU activation function and the second using the Sigmoid

function). Then, the SE-Inception module scales the input feature map. The portion on the right shows the original residual module (left) and the SE-ResNet module (right). The SE-ResNet module adds the SE block after the traditional residual connection, likewise using global average pooling and fully connected layers to obtain channel-level weights [16] and scale the output of the residual module. Both modified modules are designed to enhance the network's ability to evaluate the importance of features, thereby improving the performance of the overall model [17].

4. Experimental Results

4.1. Datasets and Evaluation Index

The dataset for this study is from the open source repository "GitHub- MaryamBoneh/Vehicle-Detection", with deep learning and YOLO used for vehicle detection algorithms. There are a total of 1321 images in this dataset, of which there are 1196 images in the train folder, 8 images in the test folder and 117 images in the validation folder. Each image has a resolution of 416×416 pixels. The images have been annotated with categories: cars, motorcycles, trucks, buses and bicycles.

Figure 4. Sample of Dataset

Evaluation Index: mAP

mAP is mean-average accuracy, which is a performance evaluation index commonly used in target detection and image recognition tasks. As a comprehensive evaluation of the performance of the model across multiple categories, mAP is obtained by calculating the average accuracy (AP) of each category and then averaging their average accuracy. Specifically, for each category, the model generates a P-R (precision rate-recall rate) curve based on the prediction results, and then calculates the area under this curve as the AP value of that category [18]. Finally, the AP of all categories is averaged to obtain mAP, with the mAP formula as follows:

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

Where C represents the total number of categories and APi represents the AP value of the i-th class.

4.2. Experimental Environment

Device: NVIDIA GeForce RTX 3090 30GB

CUDA Version: 11.1

Python Environment: 3.8

4.3. Comparison of Vehicle Detection Model Results Based on Improved YOLOv9

The mAP performance value after 100 rounds of training with YOLOv9 without any improvement was 0.7, and the mAP performance value with SENetV1 was 0.75, an increase of 5%. We also did ablation experiments, and compared the roles that different attention modules play in YOLOv9, such as MLCA, GMA, and MSDA. Meanwhile, we tried to add Shape Iou under the condition of adding an attention mechanism. The purpose of adding Shape Iou is to test whether it can cooperate with the attention mechanism to achieve better results. In addition, we tried to add only one SE module (Baseline+SE) for comparison.

The experimental results are as follows:

Table 1. Comparison of Model Results

Method	mAP@0.5
YOLOv3	0.641
YOLOv5	0.650
YOLOv9	0.699
YOLOv9 + MLCA	0.705
YOLOv9 + SE	0.739
YOLOv9 + GMA	0.688
YOLOv9 + MSDA	0.686
YOLOv9 + Shape Iou	0.734
YOLOv9 + GMA + Shape Iou	0.677
YOLOv9 + SE + Shape Iou	0.725
YOLOv9-SE (ours)	**0.750 (↑5%)**

The results show that only adding the SENetV1 has the best effect and the highest lift of 5%. According to the performance of GMA module, this module is not suitable for vehicle detection. The Shape Iou module performs well without other attention mechanisms, but the results are poor when cooperated with other attention mechanisms. After the addition of SE and MLCA modules, the training effect has a slight improvement, but the MSDA has a slight decrease. To sum up, the addition of the SENetV1 module is more suitable for vehicle detection.

Figure 5. Improved Model Detection Effect

5. Discussion

The SENetV1 module is a stack of SE structures, and the SE module is integrated into the Inception module and the ResNet module, respectively, so that the network can be dynamically calibrated according to the characteristics of the channel to improve the learning ability of the network. It is a network structure that improves performance by adjusting channel relationships in convolutional networks. SENet is not an independent network model, but a module that can be combined with any existing model (it can be regarded as a channel-type attention mechanism) [19]. In vehicle detection, the SENet (Squeeze-and-Excitation Networks) structure is of great significance. SENetV1 can adaptively recalibrate features. In the vehicle detection task, it can better pay attention to the important features related to the vehicle and suppress the irrelevant features. For example, it can highlight key feature information such as vehicle outline and license plate, and reduce the interference of complex environmental factors such as trees and buildings in the background, thus improving the efficiency and accuracy of feature extraction. By assigning different weights to the features of different channels, SENetV1 can enhance the ability to express vehicle features. This helps the detection model to more accurately identify various types of vehicles, including vehicles of different colors, sizes, and shapes, as well as vehicles under different lighting conditions and complex scenes, so as to significantly improve the accuracy of vehicle detection. Vehicle detection often faces various complex practical situations, such as bad weather, different shooting angles, partial occlusion, etc. [20]. The SENetV1 structure can make the model pay more attention to the distinguishing features, thus enhancing the resistance of the model to these unfavorable factors and improving the robustness of the model. In short, the SENetV1 structure can improve the efficiency of feature extraction, improve the detection accuracy and enhance the robustness of the model in vehicle detection, which provides str**ong sup**port for achieving more accurate and reliable vehicle detection.

6. Conclusion

After adding the SENetV1 attention mechanism, YOLOv9 has significantly improved the accuracy of vehicle detection with good results. We show that SE modules can be combined together to form a new architecture, namely SENetV1, which enables the network to perform dynamic channel feature

recalibration to improve the network's representation capabilities. A large number of experimental results indicate that this architecture has an effective generalization ability on vehicle-related datasets [21]. After only 100 rounds of training, its mAP value reaches 0.75, an increase of 5%, and only needs to add a small additional computational cost. It is hoped that our model can be applied to more fields and contribute to other fields in addition to vehicle inspection.

References

[1] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[2] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[3] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence 39.6 (2016): 1137-1149.

[4] Redmon, Joseph. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

[5] Wu, Wentong, et al. "Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image." PloS one 16.10 (2021): e0259283.

[6] Glenn, 1: Ultralytics YOLOv8. GitHub. https://github.com/ultralytics/ ultralytics. Accessed 11 Jan 2023.

[7] Wang, Chien-Yao, I-Hau Yeh, and Hong-Yuan Mark Liao. "Yolov9: Learning what you want to learn using programmable gradient information." arXiv preprint arXiv:2402.13616 (2024).

[8] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[9] Liu, Wei, et al. "Ssd: Single shot multibox detector." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016.

[10] Tan, Mingxing, Ruoming Pang, and Quoc V. Le. "Efficientdet: Scalable and efficient object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[11] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[12] Dong, Xudong, Shuai Yan, and Chaoqun Duan. "A lightweight vehicles detection network model based on YOLOv5." Engineering Applications of Artificial Intelligence 113 (2022): 104914.

[13] Bie, Minglin, et al. "Real-time vehicle detection algorithm based on a lightweight You-Only-Look-Once (YOLOv5n-L) approach." Expert Systems with Applications 213 (2023): 119108.

[14] Guo, Feng, et al. "Dense traffic detection at highway-railroad grade crossings." IEEE transactions on intelligent transportation systems 23.9 (2022): 15498-15511.

[15] Xu, Sheng, et al. "An improved lightweight yolov5 model based on attention mechanism for face mask detection." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2022.

[16] Wang, Shufeng, et al. "Non-motor vehicle detection model based on YOLO algorithm." Automotive Engineer. 08 (2024): 8-14. doi: 10.20104/j.cnki.1674-6546.20240223.

[17] Zhang, Yong. Aerial image vehicle detection based on improved YOLOv5. 2023. Henan University, MA thesis.doi: 10.27114/d.cnki.ghnau.2023. 000088.

[18] Bian, Jingchen & Yugui Liu. "Dual channel attention networks." Journal of Physics: Conference Series. Vol. 1642. No. 1. IOP Publishing, 2020.

[19] Burra, Manaswini, et al. "Cross channel interaction based ECA-Net using gated recurrent convolutional network for speech enhancement." Multimedia Tools and Applications (2024): 1-25.

[20] Liu, Yichao, Zongru Shao, and Nico Hoffmann. "Global attention mechanism: Retain information to enhance channel-spatial interactions." arXiv preprint arXiv: 2112.05561 (2021).

[21] Song, Yingkun. Research on complex scene vehicle detection algorithm based on improved YOLO. 2023. Zhejiang Sci-Tech University, MA thesis.doi: 10.27786/d.cnki.gzjlg.2023. 000081.