Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 289-317
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

# A Dangerous Driving Behaviors Detection Method for Car Driver Based on Improved YOLOv7 Model

**Md Tariqul Islam[1*], Akash Joarder[2], Md Niaz Ahmed[2]**

[1]School of Mechatronic Engineering, Mechanical Engineering, China University of Mining and Technology, Xuzhou, China
[2]Department of Manufacturing Engineering and Management, University of Technology Sydney, Sydney, Australia
Email: *tariqulislam@cumt.edu.cn

**How to cite this paper:** Islam, M.T., Joarder, A. and Ahmed, M.N. (2024) A Dangerous Driving Behaviors Detection Method for Car Driver Based on Improved YOLOv7 Model. *Journal of Computer and Communications*, **12**, 289-317.
https://doi.org/10.4236/jcc.2024.1212016

**Received:** December 2, 2024
**Accepted:** December 28, 2024
**Published:** December 31, 2024

## Abstract

The basic theory of YOLO series object detection algorithms is discussed, the dangerous driving behavior dataset is collected and produced, and then the YOLOv7 network is introduced in detail, the deep separable convolution and CA attention mechanism are introduced, the YOLOv7 bounding box loss function and clustering algorithm are optimized, and the DB-YOLOv7 network structure is constructed. In the first stage of the experiment, the PASCAL VOC public dataset was utilized for pre-training. A comparative analysis was conducted to assess the recognition accuracy and inference time before and after the proposed improvements. The experimental results demonstrated an increase of 1.4% in the average recognition accuracy, alongside a reduction in the inference time by 4 ms. Subsequently, a model for the recognition of dangerous driving behaviors was trained using a specialized dangerous driving behavior dataset. A series of experiments were performed to evaluate the efficacy of the DB-YOLOv7 algorithm in this context. The findings indicate a significant enhancement in detection performance, with a 4% improvement in accuracy compared to the baseline network. Furthermore, the model's inference time was reduced by 20%, from 25 ms to 20 ms. These results substantiate the effectiveness of the DB-YOLOv7 recognition algorithm for detecting dangerous driving behaviors, providing comprehensive validation of its practical applicability.

## Keywords

Dangerous Driving Behaviors, Object Detection, YOLOv7, Separable Convolution, CA Attention Mechanism

## 1. Introduction

Due to the swift growth of China's economy, the number of automobiles has significantly increased. In 2023, the total number of cars in the country had reached 336 million [1], with 486 million drivers, 34.8 million newly registered motor vehicles, and 24.29 million newly licensed drivers, a large increase. With the rapid rise in car ownership, the complexity of road traffic has increased correspondingly, resulting in a continuous rise in the rate of traffic accidents [2]. According to the statistics of the Public Security Department, traffic accidents have become one of the main causes of casualties, and there are about 200,000 traffic accidents in China every year, with an average of one person dying every 8 minutes. Traffic accidents are mainly caused by drivers performing unrelated actions while driving [3], which indicates that the driver's subjective interference factors increase the road risk.

China has implemented many road safety measures to reduce traffic accidents caused by dangerous driving behaviors and ensure the safety of people's travel. A lot of manpower and material resources have been invested and a lot of hardware resources have been deployed. For example, surveillance cameras [4] and tachographs [5] are installed in complex traffic sections to encourage drivers to concentrate on driving. However, these monitoring methods have certain limitations, not only is the cost high, but also the monitoring blind spot cannot be avoided, it is difficult to achieve all-around real-time monitoring, and the problem cannot be fundamentally solved. Therefore, how to monitor the driver's real-time driving status and reduce the possibility of accidents has become an urgent problem to be solved. In recent years, with the rapid development of deep learning technology, remarkable results have been achieved in mature applications in the fields of image segmentation [6], speech processing [7], and object detection [8], which provide new ideas for solving the problem of real-time monitoring of drivers' driving behavior.

This paper explores the development of a deep learning-based system for recognizing dangerous driving behaviors. It first outlines the overall framework of the recognition system, addressing the challenges encountered in identifying such behaviors within the vehicle environment. The paper introduces an improved version of the YOLOv7 object detection algorithm, incorporating deep separable convolution and a CA attention mechanism, along with enhancements to the bounding box loss function and clustering methods. These improvements are validated through comparative experiments on a custom dataset. The research offers valuable insights into the integration of deep learning for enhancing road safety by accurately detecting dangerous driving behaviors.

## 2. Literature Review

Deep learning-based detection technology is a non-contact detection method that collects data on the driver's driving behavior by installing a camera inside the vehicle. The data is then fed into a specific algorithm to analyze whether the driver has committed a violation. This method does not directly interfere with the normal driving of the driver and has the advantages of lower cost and easier implementation.

In 2012, Zhao *et al.* [9] used an artificial neural network based on the principle of multi-layer perceptron to quickly identify violations such as eating snacks and using mobile phones while driving. However, because the skin tone areas of the human face and hands tend to be connected, this results in a relatively low accuracy of the recognition effect. In 2015, Zhang *et al.* [10] developed a method to detect the behavior of drivers using mobile phones. Firstly, the AdaBoost algorithm is used to detect the driver's face, and then the FB Error feature point selection algorithm is used to screen out the key feature points on the face. Using the feature points obtained by these filters, it is possible to track the face of the driver when the driver's head is turned. By analyzing the position of the head, it is possible to identify areas where the driver's hand may be located when making a phone call. An adaptive skin tone detection algorithm is applied to the area to assess whether a driver is holding a phone in the area. This method provides an effective technical means to identify and prevent distractions while driving. In 2016, Mo *et al.* [11] developed a deep learning-based computer vision model capable of identifying 12 different types of human behaviors, and this process does not require any prior knowledge. To enhance the generalization ability and robustness of the model, the team also adopted some algorithms to optimize the performance of the neural network. The recognition accuracy of this model reached 81.8%, which indicates that the use of a convolutional neural network model with supervised learning is an effective method for human behavior recognition. In 2018, Xia *et al.* [12] proposed a method to detect distracted driving behavior using convolutional neural networks. In this method, the driver's upper body is divided into 9 key points, and the data is collected using the Alpha Pose system. Then, with these 9 key points as the center, they are fused and passed to the next convolutional layer, which uses the VGG16 network and the ResNet50 network for simulation processing. On the other hand, Xing *et al.* [13] captured 7 different images of driving behavior through a camera, and they first segmented the original image to extract the driving area in the image and then used the trained CNN model for behavior detection. Both methods demonstrate the effectiveness and potential of utilizing deep learning techniques in driving behavior recognition. In 2019, Xiong *et al.* [14] conducted an in-depth study on the problem of drivers using mobile phones to detect and proposed a new detection method based on deep learning. Firstly, the progressive calibration network is used to track the driver's face in real-time, to accurately determine the position of the driver's hand when making a phone call. Next, the neural network algorithm identifies the driver's call behavior in the identified candidate area. The result is a recognition accuracy of up to 96.56% and detection at 25 frames per second, showing excellent performance and practicability. In 2021, Ni Chengrun *et al.* [15] proposed a ResNet-LSTM architecture with ResNet-LSTM as the core to realize the comprehensive processing of spatial feature extraction, temporal information modeling and behavior recognition, introduced a traditional convolutional attention mechanism into the LSTM network, and proposed a global temporal attention module, which uses coefficient weighting to emphasize key features and suppress irrelevant features, thereby optimizing the modeling process of temporal information. Experiments verify the performance of the ResNet-LSTM network on the public dataset UCF101, with an accuracy rate of more than 95%. In 2023, Zhang *et al.* [16] improved the YOLOv5s algorithm to detect six unsafe driving behaviors of drivers, including playing with mobile phones with the left hand, playing with mobile phones with the right hand, drinking water, making phone calls with the left hand, making phone calls with the right hand, and taking both hands off the steering wheel, which is of great significance for improving road safety.

## 3. YOLO Series Object Detection Algorithms

The core idea of YOLO series object detection is to transform the object detection task into a regression problem [17], firstly, the input image is divided into several grids, and then it is assumed that the detected target center is one of these grids, then this grid is responsible for predicting the target, and the image is extracted by convolutional neural network, and then through regression prediction for each grid, the bounding box is output, and each predicted bounding box is represented by five quantities, which are the center position of the object (x, y), high h, wide w, and confidence. **Figure 1** shows the implementation process of the YOLO algorithm, firstly, the input image is divided into a grid of S × S. Each grid element is then responsible for identifying the target where its center point is located; The final output includes the coordinates of the bounding box of the target object, the confidence level, and the category information. In addition, the algorithm helps the network better learn the position information of objects with the help of a series of preset anchor frames, and the preliminary size of these anchor frames is obtained by clustering and analyzing the existing label images so that it can more accurately predict the position of the target object.

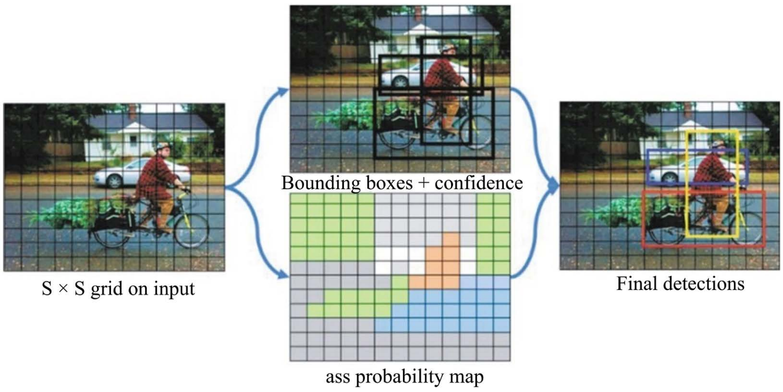


**Figure 1.** YOLO target detection process.

The architecture of the YOLO object detection algorithm consists of three main parts: the feature extraction network, the feature fusion layer, and the detection decoupling head. Firstly, the algorithm uses the feature extraction network to extract key information from the input image, such as the texture and edge of the image. Then, through the feature fusion layer, the algorithm integrates features from different levels to obtain richer feature representations, enhance the model's ability to recognize objects of various sizes and complexity and improve the model's understanding and representation ability. The detection decoupling head converts the feature information into the output of object detection. It mainly outputs the location and category information of the target, including the position and size of the bounding box and the confidence of the target category. Figure 2 shows the network structure.

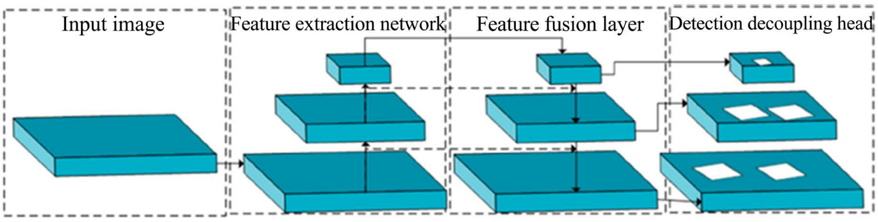


Figure 2. YOLO series network architecture.

In the object detection task, directly predicting the size and center position of the bounding box may lead to too large a space of the solution, which makes it difficult for the model to converge stably, thus wasting a lot of computing resources. To solve this problem, the YOLO series algorithm designed an anchor frame strategy, which generates a series of anchor frames at a fixed position in the image. Then, by learning the position of the objective, adjust the preset box. It not only reduces the computational complexity but also improves the speed of model convergence, effectively improving the efficiency and accuracy of detection. Figure 3 shows how it works.

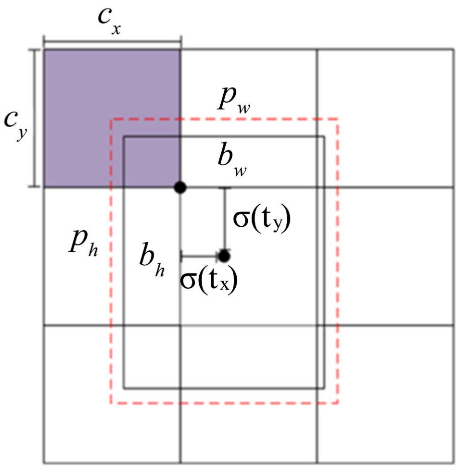


Figure 3. The prediction principle of YOLO's bounding box.

$$b_x = \sigma\left(t_x + c_x\right) \tag{3.1}$$

$$b_y = \sigma\left(t_y + c_y\right) \tag{3.2}$$

$$b_w = p_w e^{t_w} \tag{3.3}$$

$$b_h = p_h e^{t_h} \tag{3.4}$$

where $p_w$ and $p_h$ represent the width and height of the anchor frame, respectively; $b_w$ and $b_h$ represent the width and height of the prediction box, respectively; $c_x$ and $c_y$ represent the coordinate values in the upper-left corner of the bounding box; $t_x$ and $t_y$ represent the center offsets of the prediction and anchor boxes; σ(t) represents the sigmoid function, the value range is 0 to 1, and the YOLO series network learns the offset of $t_x, t_y, t_w, \textbf{and } t_h$ through the training sample.

## 4. Making and Processing Dataset

In the dangerous driving behavior recognition model, a reasonable data set is crucial to the accuracy of the model recognition. To make the dangerous driving behavior recognition model converge to a good range as soon as possible, the model is preliminarily trained by the existing public dataset so that the model tends to a relatively ideal convergence state. Therefore, when the dangerous driving behavior dataset is trained, more image features can be learned with fewer iterations, which can improve the model's ability to detect dangerous driving behaviors. In general, the selection of datasets requires that the test set and the training set are independently distributed, the amount of data is abundant, and the labels are correct. This paper selects PASCAL VOC as the pre-trained dataset, which is rich in data and has a distribution of targets, including 5 categories and 20 sub-categories of people, animals, vehicles, and indoor pictures, with 10,728 training sets and 10,775 test sets, with a total of 52,090 sample frames and an average of 2.41 targets per target, which meets the basic requirements of this paper for the pre-trained dataset, and obtains the pre-trained weight file of the dangerous driving behavior recognition model through pre-training. Figure 4 shows the specific data classes of the PASCAL VOC dataset.

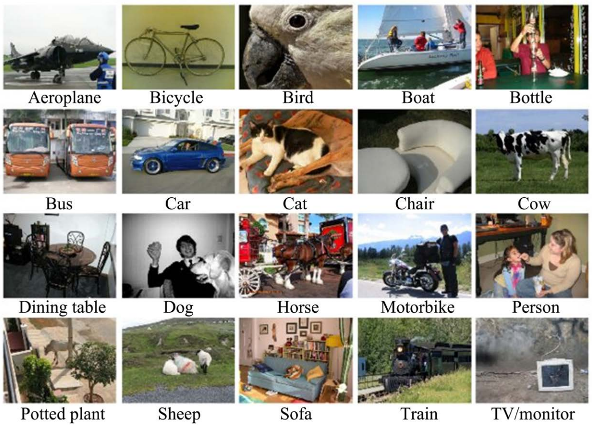


Figure 4. PASCAL VOC dataset category.

### 4.1. Production of Dataset

This paper involves five dangerous driving behaviors: making phone calls, playing with mobile phones, smoking, drinking, and eating, as shown in Figure 5. To get pictures of the below five dangerous driving behaviors, the camera is mounted on the A-pillar on the passenger side, which can completely capture all the movements of the driver's upper body. A total of 15,860 images were used for five dangerous driving behaviors at different illumination levels, including 9510 images in the training set, 4750 in the test set, and 1600 in the verification set.

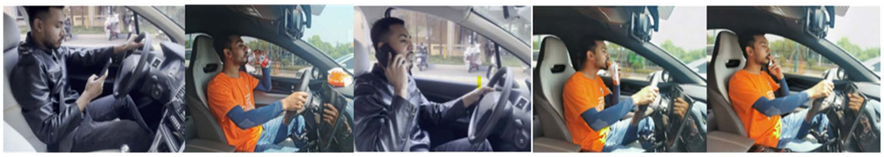


Figure 5. Five dangerous driving actions.

### 4.2. Data Annotation

Make a dataset for the above five kinds of dangerous driving behaviors, after the data is collected, the dangerous driving behaviors in the picture need to be labeled. we need to correctly label each dangerous behavior of each picture. In this paper, the online annotation tool of Roboflow is selected, and the dangerous driving behaviors are labeled by using the annotation tool, and the labeling categories are divided into five categories, namely making phone calls, playing mobile phones, smoking, drinking water, and eating. The labeling process is shown in Figure 6.

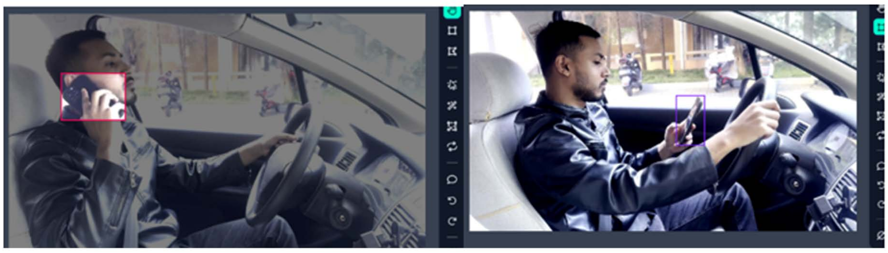


Figure 6. Roboflow annotation tool.

To improve the accuracy of the dangerous driving model, it is necessary to avoid missing and wrong labels when using the annotation tool, and different pictures of dangerous driving behaviors are stored in different folders. When labeling, you need to pay attention to selecting the YOLO labeling box and the folder address where the dangerous driving behavior is saved, and then labeling.

## 5. DB-YOLOV7 Network Structure Design

This section will analyze and improve the YOLOv7 network, so that the improved network is more suitable for the recognition of dangerous driving behaviors and further improve the recognition accuracy of the model. This paper improves the

YOLOv7 network through four aspects: firstly, by introducing deep separable convolution and CA attention mechanism into the YOLOv7 network; Secondly, the bounding box loss function and the clustering method of the initial anchor box of YOLOv7 were optimized. In the following, we will first analyze the structure of the YOLOv7 network, and then make improvements to the YOLOv7 network.

## 5.1. YOLOv7 Network Structure Analysis

YOLOv7 target detection algorithm [18]-[20] inherits the network architecture of YOLOv5 and integrates many advantages of the YOLO series of algorithms. The YOLOv7 algorithm obtained by further optimizing the YOLOv5 algorithm has the following changes compared with YOLOv5, namely the detection method of the auxiliary head, re-parameterization, and efficient aggregation network, etc., the overall network structure is shown in Figure 7. The performance of the YOLOv7 algorithm on the official data set is better than that of all previous YOLO series algorithms. Therefore, this article selects the YOLOv7 algorithm as the research object of the dangerous driving behavior recognition algorithm.

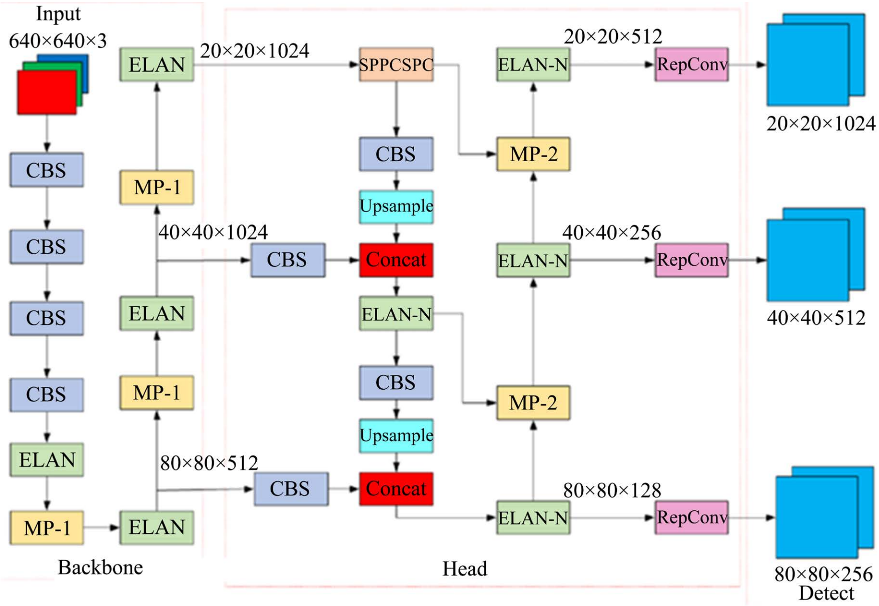


Figure 7. Network structure of YOLOv7.

The YOLOV7 network is mainly composed of three parts: input, Backbone, and Head. After the image is input, convolution, normalization, and other operations are performed in the Backbone module to complete feature extraction, and three feature maps of different sizes are output. The input is then processed in the Head module. Feature maps are fused, and finally, prediction and classification are performed. As shown in Figure 7, there are various structural blocks in the YOLOv7 network structure diagram. Their structure and functions are introduced as follows:

- CBS structure

The basic convolution module in the YOLOv7 network is the CBS module, which mainly consists of three parts: the ordinary convolution module part, batch normalization, and SiLu activation function, as shown in Figure 8.

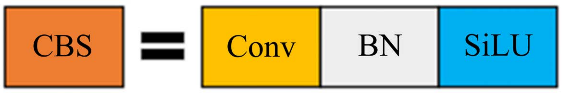


Figure 8. CBS convolution structure diagram.

- ELAN structure

The ELAN structure is constructed by superimposing multiple CBS layers, as Figure 9 shows. This module's main function is to perform continuous convolution processing on the input feature map and then splicing and fusion operations on the output feature map. This method effectively solves the problem of slow model convergence when increasing model depth.

- MP structure

Figure 10 shows that the MP structure has two branches. The left side is first down-sampled through the maximum pooling operation to achieve the effect of dimensionality reduction and parameter reduction. Then, it is changed through a 1 × 1 convolution operation. The number of channels of the feature map. On the right side, the number of channels is first changed through 1 × 1 convolution, and then feature extraction is performed through 3 × 3 convolution. The step size of the convolution operation here is 2, and the down-sampling effect is also achieved. Finally, the results of the two branches are spliced in the channel dimension. This structure can effectively improve the feature extraction and fusion capabilities of the network.

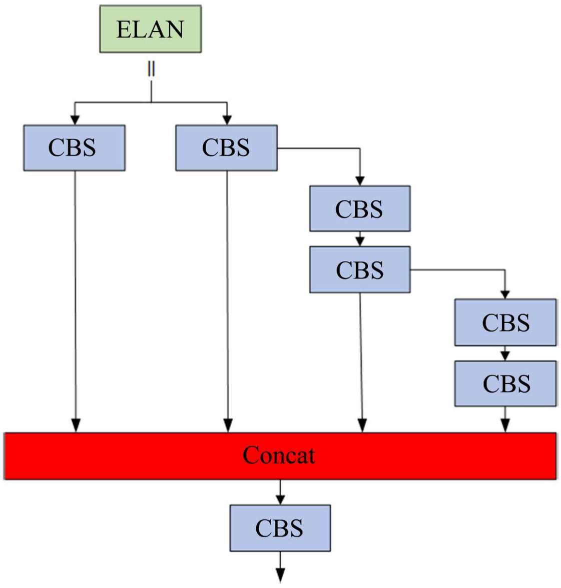


Figure 9. ELAN structure.

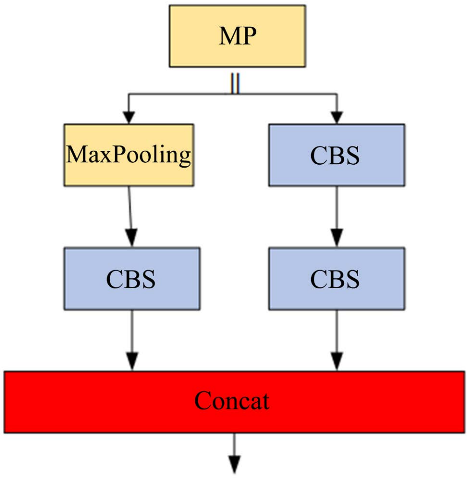


**Figure 10.** MP structure.

- SPPCSPC structure

The SPPCSPC module is improved from the SPP structure. It is obtained by adding a branch to the SPP structure. Its structure is shown in **Figure 11**. When entering SPPCSPC, a convolution operation is first performed, and then the traditional SPP operation is performed. Perform merge processing. This processing method can make the network have richer gradient combinations, improve the learning ability of the network, and reduce the amount of calculation.

- ELAN-N structure

The ELAN-N structure is similar to the ELAN structure. The main difference is the number of outputs on the right branch. The ELAN module is spliced with three branches, while the ELAN-N module chooses a five-branch structure, as shown in **Figure 12**. ELAN-N modules have a denser branching structure that helps improve the network's learning capabilities.

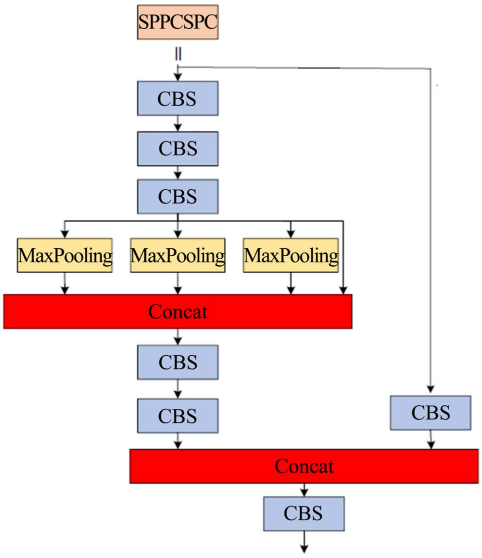


**Figure 11.** SPPCSPC structure.

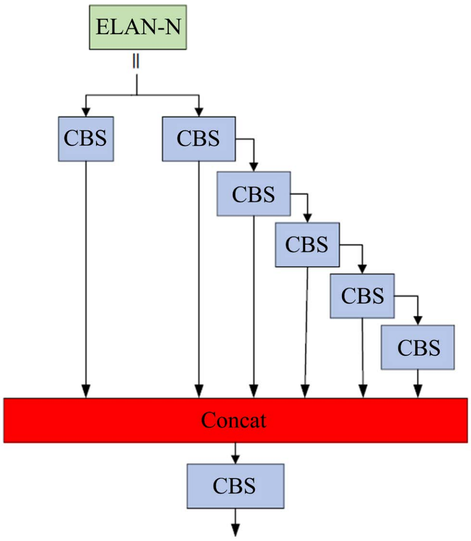
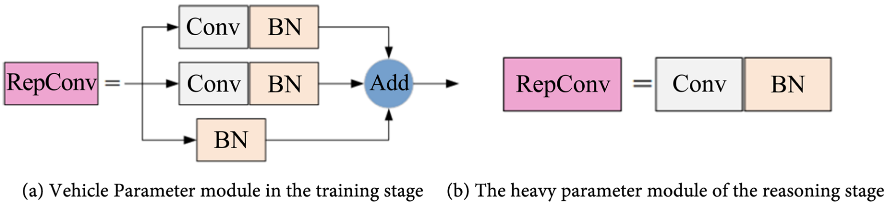


Figure 12. ELAN-N structure.

- RepConv structure

The RepConv heavy-parameterization structure is adopted in the YOLOv7 network structure, as shown in Figure 13. This structure is different in the training and inference processes. During the inference process, the RepConv structure uses a 3 × 3 convolution to replace the 1 × 1 and 3 × 3 convolutions in the three branches during training. This replacement operation can improve the inference speed of the model without affecting the accuracy. and reduce memory.



(a) Vehicle Parameter module in the training stage (b) The heavy parameter module of the reasoning stage

Figure 13. RepConv structure.

## 5.2. Improved Structural Block Design Integrating Depth-Wise Separable Convolution

Depth-wise separable convolution [21]-[23] technology is implemented by decomposing ordinary convolution into two stages: depth convolution and point-wise convolution, as shown in Figure 14. The characteristic of depth convolution is that the convolution kernel and the number of channels of the input data are matched one-to-one. The number of input and output channels is the same. The point-by-point convolution operation is similar to the standard convolution operation, and the channel number of each channel can be processed. The convolution results are combined. Using depth-wise separable convolution can greatly reduce the number of parameters and reduce the model size without losing model accuracy, which fully meets the needs of dangerous driving behavior detection.

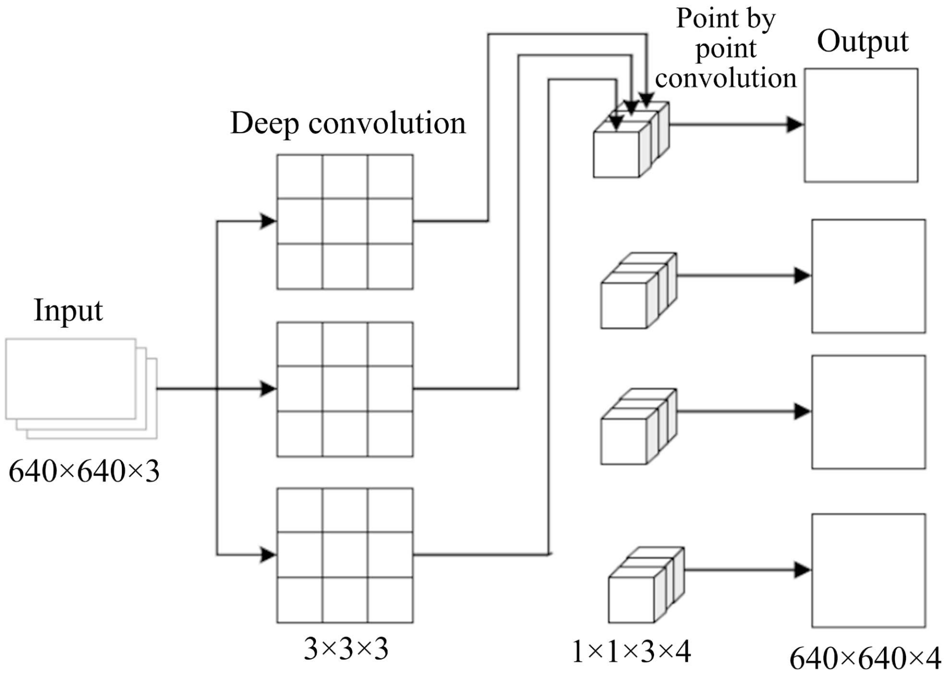


Figure 14. Deep separable convolution.

As can be seen from the above figure, if the input image is 640 × 640 × 3 and the output result is 640 × 640 × 4, the parameter stc of the ordinary convolution is stc = 3 × 3 × 3 × 4 = 108; The parameter quantity dsc of depth-wise separable convolution is the sum of the parameters of depth-wise convolution and point-wise convolution: dsc = 3 × 3 × 3 + 1 × 1 × 3 × 4 = 39. When the input and output are the same, the number of parameters using depthwise separable convolution is much smaller than that of traditional convolution. Therefore, the introduction of depth-wise separable convolution can significantly reduce the number of parameters of the model while maintaining effective feature extraction capabilities. In addition, the SiLu activation function is used in the CBS basic module of YOLOv7, as follows:

$$SiLu(x) = \frac{x}{1+e^{-x}} \tag{5.1}$$

Due to the existence of exponential form in the SiLu function, the derivation calculation requires a large amount of calculation and the calculation cost is too high, as shown in Figure 15. To further reduce the amount of calculation, this paper uses the Hard-Swish activation function [24] to replace the SiLu activation function in the CBS structure.

The hard-Swish activation function is a commonly used piecewise function, its expression is as follows:

$$Hard-Swish(x) = \begin{cases} 0 & if\ x \le -3 \\ x & if\ x \ge +3 \\ x \times (x+3)/6 & otherwise \end{cases}.$$

Compared with the SiLu activation function, the Hard-Swish activation function has no upper bound on the function value and is a relatively smooth function. Its function curve and derivative curve are shown in **Figure 16**. It can satisfy the nonlinear regression of the model and effectively reduces the calculation amount of the model. However, the Hard-Swish activation function is not differentiable when x = ±3. To ensure the invertibility of the function in the continuous space, it is necessary to program the function and write case conditions. Based on the above comparison, this article selects the Hd-Swish activation function.

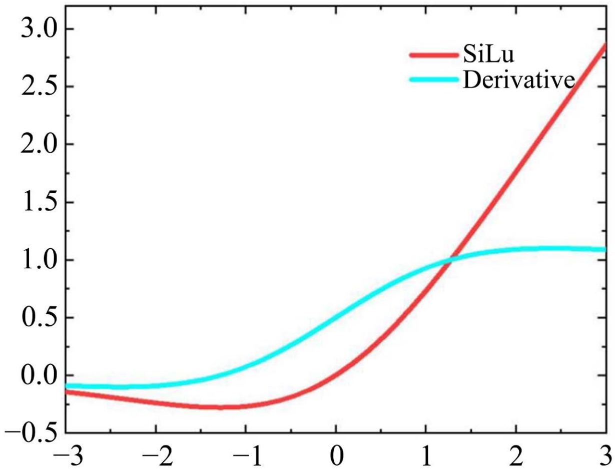


**Figure 15.** SiLu activation function.

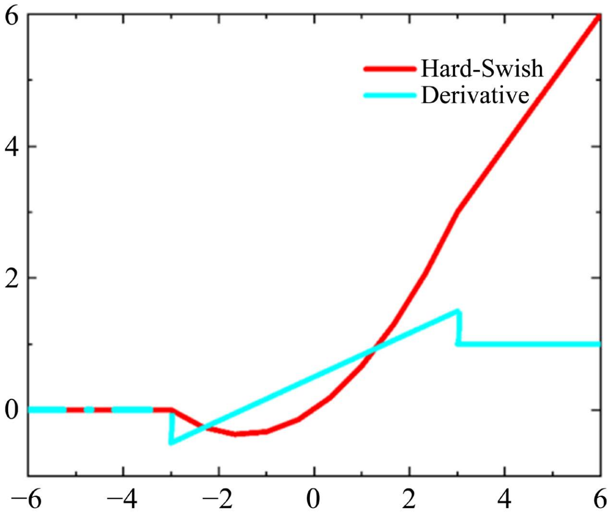


**Figure 16.** Hard-Swish activation function.

The ELAN structure is used many times in the backbone network of YOLOv7. The ELAN structure is stacked by multiple CBS layers. Depth-separable convolution and Hard-Swish activation functions can be used to replace the ordinary convolution and SiLu activation functions in the original structure. The specific improvement details are shown in **Figure 17**.

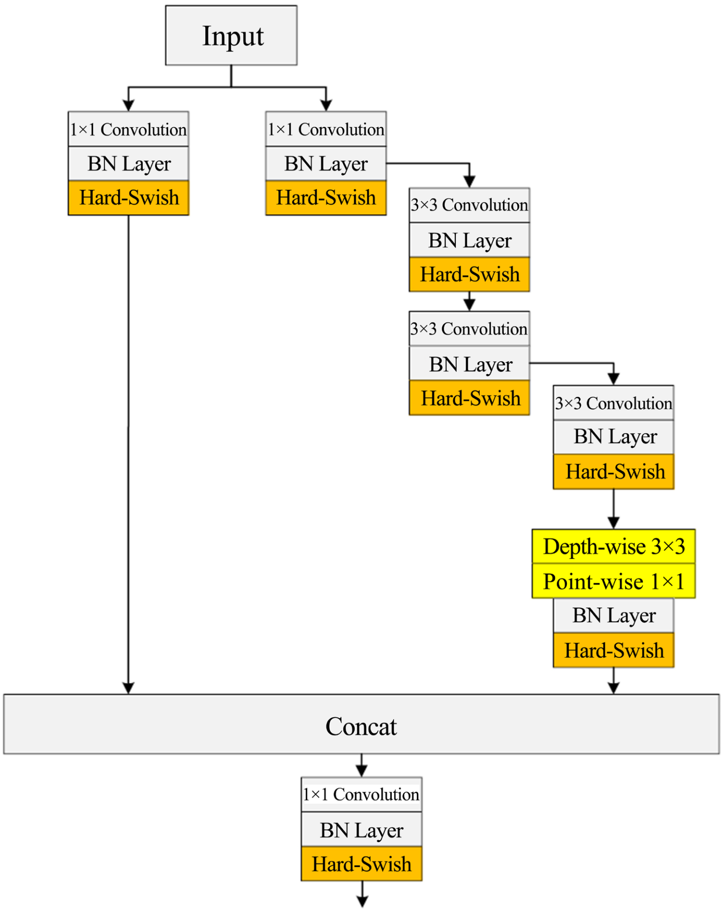


**Figure 17.** Improved ELAN Structural block.

Replace the SiLu activation function in the ELAN structure with the Hard-Swish activation function. Use depth-separable convolution at the end of the merge channel to replace the original 3 × 3 ordinary convolution, as shown in the yellow area in the figure above. Through this replacement, reduce the calculation amount of the model and speed up the reasoning speed of the model.

### 5.3. Improvement of Anchor Frame Calculation Method

The anchor box is used as a candidate frame in the field of target detection. It does not have a fixed size and aspect ratio, but the setting of the anchor box parameters has a crucial impact on the accuracy of target detection. The anchor boxes of the YOLO series of algorithms are all derived from the public data set COCO. The COCO data set is divided into eighty categories. However, the data studied in this article only has five categories. If the initial anchor box is too different from the size of the detected object, if it is large, it will lead to large adjustments during model training, which is not conducive to network convergence and affects model accuracy. Therefore, the initial anchor box of YOLOv7 is not completely suitable for the data set of dangerous driving behaviors. The dimensions of the default anchor box of YOLOv7 are shown in Table 1.

Table 1. The size of the default anchor frame for YOLOv7.

| Characteristic diagram | Anchor frame 1 | Anchor frame 2 | Anchor frame 3 |
|---|---|---|---|
| 80 × 80 | (12, 16) | (19, 36) | (40, 28) |
| 40 × 40 | (36, 75) | (76, 55) | (72, 146) |
| 20 × 20 | (142, 110) | (192, 243) | (459, 401) |

In the YOLOv7 network, the K-means clustering algorithm is used by default as the clustering method of the initial anchor frame. The specific implementation process is as follows: First, k clustering centers are randomly selected from the dangerous driving behavior data set samples produced in this article $\{C_1, C_2, \cdots, C_k\}$; Then, calculate the Euclidean distance from the remaining sample points to the cluster center and classify them according to the proximity principle. Update the position of the cluster center and repeat the above iterative process multiple times until the end condition is reached and the sample clustering is completed.

The calculation process of the objective of minimizing the square error E of the K-means clustering algorithm is as follows:

$$E = \sum_{i=1}^{k} \sum_{x \epsilon c_i} \left\| x - \mu_i \right\|_2^2 \tag{5.2}$$

where, $\mu_i$ is the mean vector of $c_i$, and the expression is as follows:

$$\mu_i = \frac{1}{|c_i|} \sum_{x \epsilon c_i} x \tag{5.3}$$

However, the allocation of the initial clustering center of the K-means algorithm is random. If the selected initial clustering center is far from the optimal clustering center, it will affect the convergence speed of the network and the accuracy of model identification. To further improve the model performance, this paper chooses the K-means++ clustering algorithm [25] to re-cluster the collected dangerous driving behavior data sets. The K-means++ clustering algorithm is improved based on the K-means clustering algorithm. It solves the problem caused by the K-means clustering algorithm's random allocation of initial frames. The specific implementation process is as follows: First, in the input data set Select a cluster center from X; then, calculate the shortest distance $D(x)$ from the remaining sample points to the cluster center; then, use formula 5.4 calculate the probability $P(x)$ corresponding to each sample point, and select the point with the highest probability as the next clustering center.

$$P(x) = \frac{D(x)^2}{\sum_{x \epsilon X} D(x)^2} \tag{5.4}$$

Although the amount of calculation is increased and more time is consumed when selecting the initial clustering center, by accurately selecting the clustering center, the clustering stage can converge quickly. This reduces the overall calculation time and obtains a more suitable Anchor frame for dangerous driving behavior recognition and detection, significantly improving the accuracy of the model. During the clustering process, the convergence relationship curve between different k values and IoU is shown in **Figure 18**.

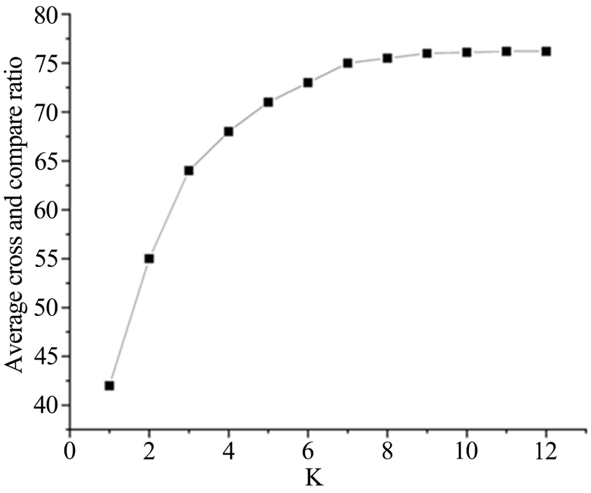


**Figure 18.** Convergence curve of clustering center *k* and IoU.

As seen from the above figure, when the value of *k* ranges from 0 to 9, the upward trend of the curve is faster. When the value of *k* is greater than 9, the curve tends to be stable. Therefore, this article selects 9 clustering centers to ensure accuracy and improve computational efficiency as much as possible. The specific dimensions of the anchor box are shown in **Table 2**.

**Table 2.** Size of anchor frame after K-Means++ clustering.

| Characteristic diagram | Anchor frame 1 | Anchor frame 2 | Anchor frame 3 |
|---|---|---|---|
| 80 × 80 | (17, 23) | (23, 35) | (30, 45) |
| 40 × 40 | (38, 62) | (53, 77) | (60, 96) |
| 20 × 20 | (80, 130) | (102, 180) | (140, 208) |

As can be seen from **Table 2**, the gap between the 9 prior frames obtained by this article through the K-means++ clustering algorithm is large, which has a good clustering effect on dangerous driving behavior and effectively solves the YOLOv7 initial anchor problems caused by box training of models.

## 5.4. CA Attention Mechanism

The attention mechanism can increase the weight of learning important information during the learning process of the network. For example, in the data set of dangerous driving behavior recognition, the attention mechanism can improve the neural network's attention to dangerous driving behavior information and reduce the impact of background information on model training. influence and improve model recognition performance [26] [27]. The CA attention mechanism has the characteristics of high efficiency, flexibility, compatibility, etc. Its structure is shown in Figure 19.

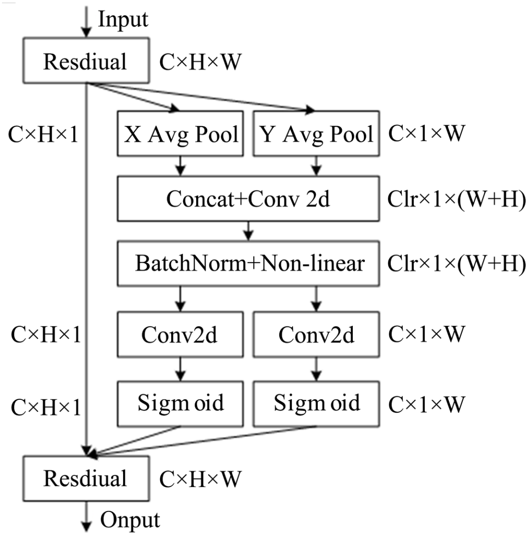


Figure 19. Construction of coordinate attention.

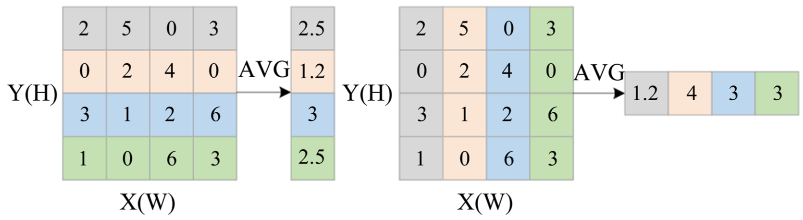
The specific implementation process of the CA attention module is as follows:

- To obtain accurate position information, the CA attention mechanism performs global average pooling on the input feature map, that is, average pooling in the horizontal and vertical directions. The calculation process is as follows:

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \le j < W} x_c(h,i) \tag{5.5}$$

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j,w) \tag{5.6}$$

Among them, $Z_c^h$ and $Z_c^w$ represent the pooling results of the height and width of the $c$ channel respectively and two feature maps are generated. The specific coordinate information embedding operation process is shown in Figure 20.



(a) Horizontal direction (b) Vertical direction

Figure 20. A coordinate information embedding.

- Splice the feature maps in the spatial dimension, and then perform convolution operation and sigmoid activation function, etc., and finally obtain two attention vectors $g^h$ and $g^h$, the calculation formula is as follows:

$$f=\delta\left(F_1\left(\left[z^h,z^w\right]\right)\right),\ f\epsilon R^{\frac{C}{r}\times(H+W)\times 1} \tag{5.7}$$

$$g^h=\delta\left(F_h\left(f^h\right)\right),\ f^h\epsilon R^{\frac{C}{r}\times H\times 1} \tag{5.8}$$

$$g^w=\delta\left(F_w\left(f^w\right)\right),\ f^w\epsilon R^{\frac{C}{r}\times 1\times W} \tag{5.9}$$

Among them, $\delta$ is a nonlinear activation function, $F_1$ represents the matrix connection of $z^h$ and $z^w$, $r$ is the scaling factor, and $f$ is the output result after the above operation, which contains the intermediate features of horizontal and vertical information.

- Convolve the feature maps of $g^h$ and $g^w$ obtained through a broadcast transformation with the input feature map $x_c$ after residual processing to obtain the final attention feature $y_c$. The calculation process is as follows:

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \tag{5.10}$$

This article adds a CA attention module at the connection between the Backbone and the head part of the YOLOv7 network. Since the Backbone layer of YOLOv7 outputs three feature maps of different sizes, it is necessary to add three CA attention modules at the connection between the Backbone and the head part. The specific adding location is shown in **Figure 21**. Due to the different sizes of the input feature maps, to reasonably allocate computing resources, this paper chooses CA structures with different convolution kernels for calculation. For the feature maps of 20 × 20 × 512 and 40 × 40 × 1024, a convolution of 3 × 3 is used. Use a convolution of 7 × 7 for the feature map of 80 × 80 × 512.

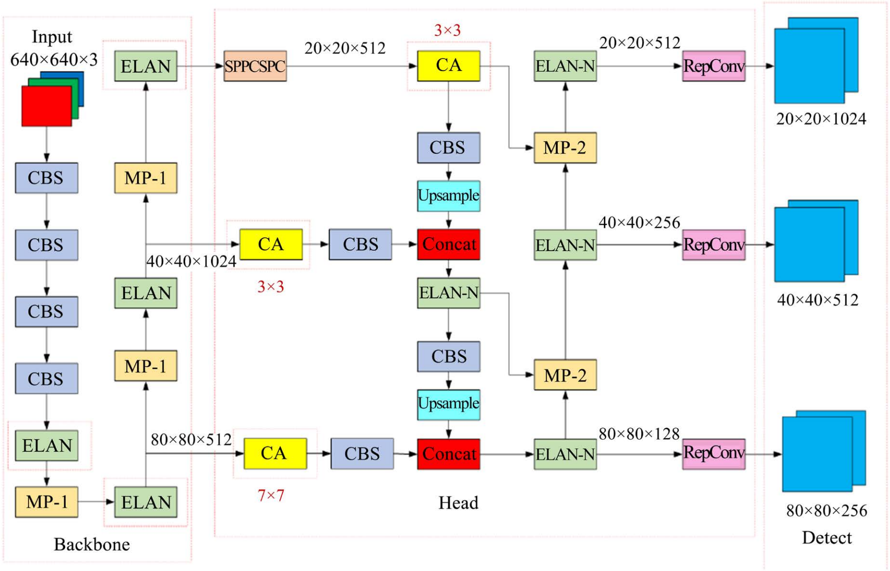


**Figure 21.** YOLOv7 network structure with CA attention mechanism.

## 5.5. WIoU Bounding Box Loss Function

The bounding box loss function is crucial in the target detection task. It directly affects the learning ability of the model. An appropriate bounding box loss function can effectively guide the model to reduce the difference between the predicted box and the real box, thereby improving the target detection model. overall performance. The original YOLOv7 algorithm uses the CIoU bounding box loss function, but this loss function has some shortcomings in the actual operation process. This article will use the WIoU bounding box loss function [28] instead of CIoU.

$$L_{CIoU} = 1 - IoU + \frac{d^2}{c^2} + \alpha v \tag{5.11}$$

As shown in **Figure 22**, $c$ represents the diagonal length of the minimum circumscribed rectangle between the predicted box and the actual box; $d$ is the Euclidean distance between the center points of the two boxes; $v$ represents the similarity between the predicted box and the actual box; $\alpha$ is expressed as an adjustment factor used to balance the impact of $v$ in the total loss. The expressions for $v$ and $\alpha$ are as follows:

$$v = \frac{4}{\pi^2}\left(arctan\frac{x_g}{y_g} - arctan\frac{x}{y}\right)^2 \tag{5.12}$$

$$\alpha = \frac{v}{1 - IoU + v} \tag{5.13}$$

Among them, $x_g, y_g, x, y$ are the width and height of the prediction box and target box respectively, as shown in **Figure 23**.

As can be seen from the above, the calculation of the CIoU loss function is more complicated and requires more information to be considered. Therefore, the training process of the model requires a large amount of data, which is not friendly to the training of small-scale data sets. Secondly, the penalty term of the aspect ratio is introduced by the CIoU loss function. When the aspect ratio of the predicted frame and the real frame are the same or similar, the penalty term will be invalid, causing the CIoU loss function to be unstable.

The WIoU v3 bounding box loss function selected in this article is a bounding box loss based on a dynamic non-monotonic focusing mechanism. It can adaptively adjust the weight according to the difficulty of sample identification to improve the overall performance of the detector.

The WIoUv3 loss function is improved based on WIoUv1 and WIoUv2. The specific implementation process is as follows:

$$L_{WIoU_{V1}} = R_{WLoU} L_{IoU} \tag{5.14}$$

$$R_{WIoU} = \exp\left(\frac{d^2}{\left(W_g^2 + H_g^2\right)^*}\right) \tag{5.15}$$

Among them, $R_{WIoU} \epsilon \left[1, e\right]$ is used to amplify the score of ordinary quality anchor the boxes; $L_{IoU} \epsilon \left[0, 1\right]$ the purpose is to reduce the $R_{WIoU}$ of high-quality anchor boxes; the benefits of this processing method When the overlap between the anchor frame and the target frame is high, the distance between the center points of the two frames is mainly considered, effectively solving the problem of CIoU. The problem of the failure of the penalty term; $W_g$ and $H_g$ represent the width and height of the minimum bounding box respectively; "*" means stripping the parameters from the calculation graph, effectively eliminating factors that hinder the convergence gradient in $R_{WIoU}$; $W_i$ and $H_i$ respectively Indicates the width and height of the intersection, as shown in Figure 22.

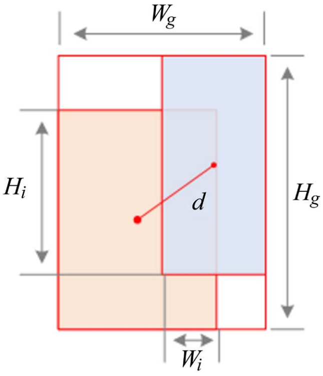


Figure 22. WloU loss function diagram.

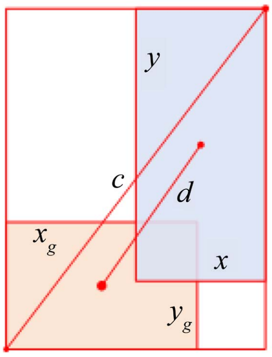


Figure 23. CIoU loss function diagram.

The WIoUv₂ loss function adds a monotonic static focusing mechanism based on WIoU$_{v1}$. It is derived from a monotonic focusing mechanism (FM) for cross-entropy, which allows the model to focus on difficult samples and reduce the number of simple samples. The influence of the function loss value further improves the classification performance of the WIoU loss function. The expression is as follows:

$$L_{WLoU_{V2}} = L_{IoU}^{\gamma*} L_{WIoU_{V1}} \tag{5.16}$$

among them, $L_{IoU}^{\gamma*} \epsilon \left[0,1\right]$, represents the monotonic focusing coefficient introduced to adjust the gradient; since $L_{IoU}$ will gradually become smaller during model training, and the gradient gain will decrease accordingly, resulting in slower model convergence. To solve this problem, the mean value of $L_{IoU}$ is used as the normalization factor to balance the gradient gain in this way. The specific implementation process is as follows:

$$L_w = \left( \frac{L_{IoU}^*}{\overline{L_{IoU}}} \right)^{\gamma} L_{WIoU_{v1}} \tag{5.17}$$

among them, $\overline{L_{IoU}}$ represents the dynamic average, which can maintain the gradient gain at a high level, effectively solving the problem of slow convergence of the model. Based on the research on WIoU$_{v1}$ and WIoU$_{v2}$, WIoU$_{v3}$ uses the dynamic non-monotonic focusing mechanism to further improve the allocation strategy of allocating gradient gains. The specific implementation is as follows:

$$L_{WIoU_{V3}} = rL_{WIoU_{V1}} \tag{5.18}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{5.19}$$

$$\beta = \left( \frac{L_{IoU}^*}{\overline{L_{IoU}}} \right) \tag{5.20}$$

among them, $\beta$ represents the abnormality of the anchor box, which is also called the outlier degree. When the value of $\beta$ is larger, it means that the quality of the current anchor box is lower, and a smaller gain $r$ will be assigned to it, which also directly affects the classification performance. Since the value of $\overline{L_{IoU}}$ changes dynamically, the quality evaluation criteria of the anchor frame are also dynamically adjusted. WIoU$_{v3}$ can select the most appropriate gradient enhancement strategy according to the conditions of training to improve the classification performance of the network. $\alpha$ and $\delta$ are two hyperparameters. The mapping between different hyperparameters and gain gradients is shown in Figure 24.

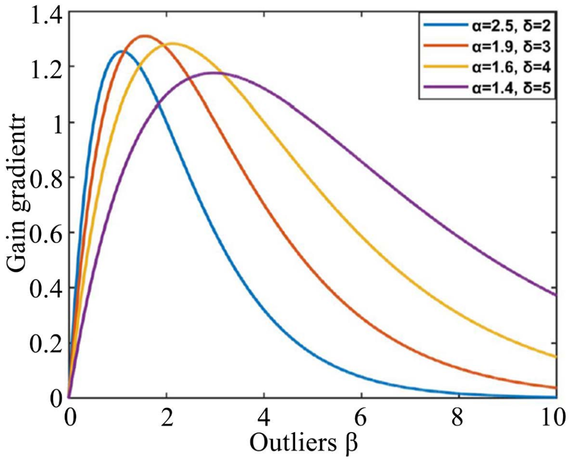


Figure 24. Dispersion and gain gradient relationship.

## 6. DB-YOLOv7 Network Model Verification and Result Analysis

In this paper, the DB-YOLOv7 network model is trained on the Window10 platform using the self-made dangerous driving behavior dataset, the Python 3.8 environment is configured, and CUDA10.2 + CUDNN7.6.5 is used under the Pytorch1.8.1 framework to RTXA4000 the graphics card for acceleration. DB-YOLOv7 was trained on dangerous driving behavior recognition, using the

Adams optimizer, the batch size was set to 8, the initial learning rate was 0.001, and the natural exponential decay was carried out. The dataset is divided according to the training set:validation set:test set = 6:3:1. The training process is shown in **Figure 25**. As can be seen from the figure, the model tends to converge after 300 training iterations, so the model at this time is taken as the optimal model in this paper, and subsequent performance test experiments are carried out.

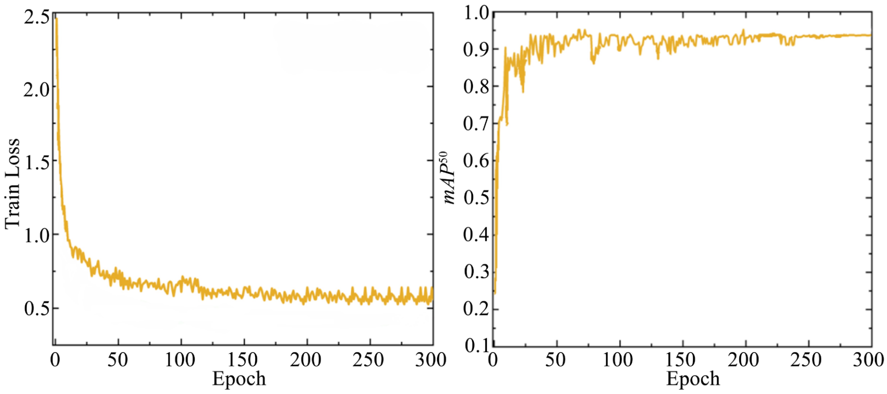


**Figure 25.** The DB-YOLOv7 model training curve.

## Evaluation Indicators for Object Detection

The evaluation indexes of object detection mainly include Precision, Recall, F-score, and Average Class Accuracy mAP. Precision represents the proportion of all positive samples predicted that are labeled as positives, and the expression is as follows:

$$Precision = \frac{TP}{TP+FP} \tag{6.1}$$

Where $TP$ refers to the number of samples correctly predicted to be a positive class, and $FP$ refers to the number of negative samples incorrectly labeled as a positive class. Recall refers to the proportion of data that the model correctly identifies as positive samples out of all data that are positives, and is expressed as follows:

$$Recall = \frac{TP}{TP+FN} \tag{6.2}$$

$FN$ represents the number of positive samples incorrectly predicted to be negative.

Both precision and recall are affected by the number of positive and negative samples in the sample, and considering only the precision and recall are not enough to evaluate the performance of the model correctly, therefore, the $F-score$ is introduced, which combines the two evaluation parameters of precision and recall, and its definition formula is as follows:

$$F_\beta = \left(1+\beta^2\right) \times \frac{precision \times recall}{\left(\beta^2 \times precision\right) + recall} \tag{6.3}$$

In different $\beta$ values, $F_\beta$ the focus on precision and recall is different, when $\beta =$ 1, the $F-score$ evaluation measures pay the same attention to precision and recall, referred $F1-score$. When $\beta = 2$ indicates that the evaluation index pays more attention to the recall rate and is called the $F2-score$. Both precision and recall are crucial for the prediction evaluation of this paper. Therefore, the $F1-score$ was selected as the evaluation index.

During the evaluation of the object detection model, the sensitivity of the model can be adjusted by setting different confidence thresholds. Only if the confidence level of the prediction exceeds this threshold, the prediction is considered successful. Different confidence thresholds directly affect the accuracy and recall of the model. The P-R curve is plotted with recall as the horizontal axis and precision as the vertical axis, as shown in Figure 26, and the average accuracy AP can be calculated by integrating the P-R curve.

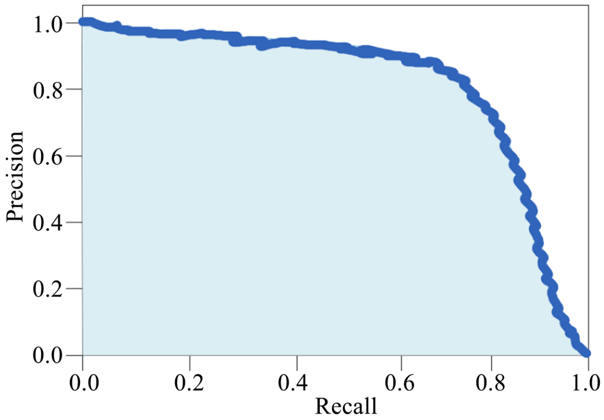


Figure 26. P-R curve.

Average accuracy AP is defined as follows:

$$AP = \int_0^1 P(R)\,dr \tag{6.4}$$

$AP^{50}$ refers to the effectiveness of the prediction target when the $IoU \geq 0.5$ of the prediction box and the dimension frame is $AP^{50}$ is a case-specific AP metric that reflects the model's ability to predict positive samples when that condition occurs. mAP, on the other hand, refers to the evaluation of multiple categories, which is the average of the AP values of each category, and provides a comprehensive way to evaluate the model's performance across all categories. mAP integrates the detection accuracy of the model in all categories, which is a key indicator to evaluate the performance of the object detection model. The formula is as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{6.5}$$

## 7. Experimental Design and Analysis of Results

To verify the superiority of YOLOv7 as the basic network and the effectiveness of the YOLOv7 network improvement method, two sets of experiments were designed for analysis.

- Experiment 1:

To verify the comprehensive performance of the DB-YOLOv7 model proposed in this paper, before training the model, the DB-YOLOv7 network model was pre-trained on the PASCAL VOC dataset for 100 rounds to obtain the basic object detection model, and then six types of objects were randomly selected from the dataset. These images contain 1532 detection targets, and comparative experiments were carried out in YOLOv7 and DB-YOLOv7 in this paper, with the evaluation index of $AP^{50}$, the relevant experimental data are shown in Table 3 and Table 4, and the inference output of the model is shown in Figure 27.

Table 3. Comparison of VOC dataset recognition accuracy before and after YOLOv7 improvement.

| Category | DB-YOLOv7 | YOLOv7 |
|---|---|---|
| mAP$^{50}$ | 75.5 | 74.1 |
| Bird | 74.5 | 73.2 |
| Dog | 78.8 | 77.5 |
| Cat | 79.6 | 78.7 |
| Person | 71.8 | 71.7 |
| Couch | 72.1 | 71.2 |
| Car | 76.2 | 72.3 |

Table 4. Reasoning time before and after YOLOv7 improvement.

| Algorithm model | Reasoning time |
|---|---|
| DB-YOLOv7 | 24 ms |
| YOLOv7 | 28 ms |

It can be seen from Table 3 and Table 4 that the DB-YOLOv7 model is superior to the YOLOv7 basic model in terms of detection accuracy and speed. Experiments show that on the PASCAL VOC test data set, the model achieves an average detection accuracy of 75.5%, which is 1.4% higher than the original YOLOv7 model. In particular, the recognition accuracy of cars is improved by about 4%, and the inference time is also improved to a certain extent. Compared with the original model, the 28 ms is reduced to 24 ms. Therefore, the dangerous driving behavior recognition model DB-YOLOv7 proposed in this article can have better recognition accuracy and speed in the target detection task, meeting the design requirements of the dangerous driving recognition model.

- Experiment 2:

To deeply verify the impact of the improvement strategy proposed in this article on the performance of the YOLOv7 network, this article designed and conducted a series of ablation experiments. Select the self-made dangerous driving behavior data set in this article for fine-tuning training. The test set uses 1600 normal lighting images. With the hardware conditions and other parameter settings remaining consistent, 300 rounds of training are performed. The experimental results use precision, recall, $F1-score$, and inference time as indicators to evaluate model performance. The model identification results are shown in Table 5, and the intuitive performance of the ablation experiment results is shown in Figure 28.

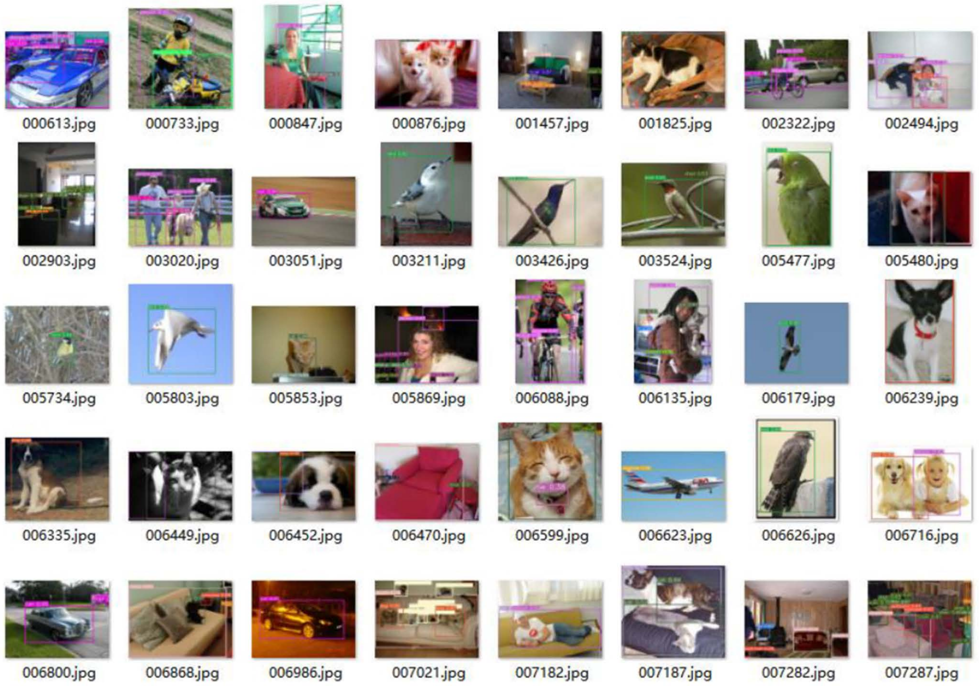


Figure 27. DB-YOLOv7 model recognition results.

Table 5. Comparison of ablation results of DB-YOLOv7.

| Network model | Precision | Recall | F1-score | Inference time (ms) |
|---|---|---|---|---|
| YOLOv7 | 91.32 | 90.41 | 90.86 | 25 |
| YOLOv7 + DSC | 90.03 | 90.48 | 90.25 | 22 |
| YOLOv7 + DSC + CA | 92.56 | 91.02 | 91.78 | 22 |
| YOLOv7 + DSC + CA + WIoU | 93.65 | 92.42 | 92.98 | 21 |
| DB-YOLOv7 | 94.03 | 93.03 | 93.52 | 20 |

As can be seen from the table above, the performance of the DB-YOLOv7 model has been greatly improved compared to the YOLOv7 model. The precision rate has increased from 91.32% to 94.03%; the recall rate has increased from 90.41% to 93.03%; and the $F1-score$ has increased from 90.86 % to 93.52%; the inference time also dropped from 25 ms to 20 ms, verifying the effectiveness of this method in improving the YOLOv7 network. The performance of DB-YOLOv7 in identifying five types of dangerous driving behaviors is shown in Table 6 below.

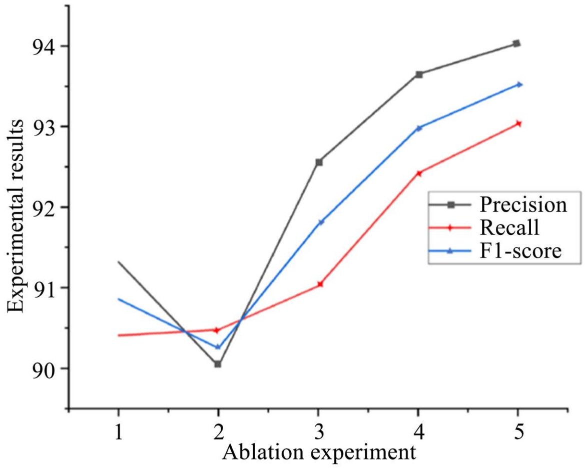


Figure 28. DB-YOLOv7 Performance test result.

Table 6. Performance of DB-YOLOv7 model on a self-made dataset.

| Behavior type | Precision | Recall | F1-score | Inference time (ms) |
|---|---|---|---|---|
| Smoking | 90.06 | 90.37 | 90.21 | 23 |
| Phone call | 97.38 | 97.42 | 97.40 | 19 |
| Play with phone | 96.57 | 96.68 | 96.62 | 20 |
| Drinking | 96.25 | 96.42 | 96.33 | 19 |
| Eating | 94.83 | 95.08 | 94.90 | 20 |

As can be seen from the above table, the DB-YOLOv7 network model has the best recognition performance for call behavior, with an accuracy rate of 97.38%, a recall rate of 97.42%, and an inference time of only 19 ms. The recognition performance of smoking was slightly weaker, with a recognition accuracy of 90.06%, a recall rate of 90.37%, and an inference time of 23 ms. DB-YOLOv7 has a recognition accuracy of more than 90% for the five dangerous driving behaviors designed in this paper, which meets the needs of dangerous driving behavior recognition.

## 8. Conclusion and Future Works

This paper presents a deep learning-based approach for the recognition of dangerous driving behaviors, to enhance road safety by accurately detecting potentially hazardous actions during vehicle operations. We proposed an improved version of the YOLOv7 object detection algorithm, which integrates deep separable convolution and a CA attention mechanism, along with optimizations to the bounding box loss function and clustering methods. The resulting DB-YOLOv7 network demonstrated substantial improvements in both recognition accuracy and inference speed.

Experimental results confirm that the DB-YOLOv7 model significantly outperforms the original YOLOv7 in terms of detection accuracy, with an average increase of 4%. Furthermore, the inference time was reduced from 25 ms to 20 ms, highlighting its efficiency for real-time applications in monitoring dangerous driving behaviors. These results not only validate the effectiveness of the proposed approach but also underline the potential for real-world implementation in enhancing road safety.

Despite these promising results, there are still several challenges and opportunities for further improvement. One key limitation is the dataset size and diversity; while the custom dataset used in this study covers various dangerous driving behaviors, it is still relatively small compared to the diverse range of driving conditions and environmental factors encountered in real-world scenarios. Future research should focus on expanding the dataset by incorporating more diverse driving situations, including different weather conditions, traffic densities, and geographic locations. This would improve the generalizability and robustness of the model across various environments. However, this study has made significant strides in utilizing deep learning for dangerous driving behavior detection, there remains considerable potential for future improvements and the integration of complementary technologies to create a more comprehensive and efficient road safety system.

## Acknowledgements

We gratefully acknowledge Professor Liu Xinhua for his invaluable assistance in creating a comprehensive dataset that significantly enhanced the quality of our work. His guidance and expertise were instrumental in shaping our research outcomes. Thank you for your unwavering support.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] https://www.gov.cn/lianbo/bumen/202401/content_6925362.htm

[2] Wang, F., Wang, J., Zhang, X., Gu, D., Yang, Y. and Zhu, H. (2022) Analysis of the Causes of Traffic Accidents and Identification of Accident-Prone Points in Long Downhill Tunnel of Mountain Expressways Based on Data Mining. *Sustainability*, **14**, Article 8460. https://doi.org/10.3390/su14148460

[3] Wang, K. (2020) Research on Driver Distraction and Fatigue Driving Behavior Monitoring Methods Based on Deep Learning. Master's Thesis, Wuhan University of Technology. (In Chinese)

[4] Wang, H. (2020) Application of Radar Camera Data Fusion in Intelligent Assisted Driving. Master's Thesis, Jilin University. (In Chinese)

[5] Toyoda, M., Yokoyama, D., Komiyama, J. and Itoh, M. (2017) Road Safety Estimation Utilizing Big and Heterogeneous Vehicle Recorder Data. 2017 *IEEE International Conference on Big Data* (*Big Data*), Boston, 11-14 December 2017, 4841-4842. https://doi.org/10.1109/bigdata.2017.8258561

[6] Yuan, Y., Du, Y., Miao, S.Q., *et al.* (2023) Image Segmentation Method and Research

Status Based on Deep Learning. China Computer Users Association Network Application Branch. *China Computer Users Association Network Application Branch* 27 *th* 2023 *Proceedings of the Annual Conference on New Network Technologies and Applications*, Beijing Key Laboratory of Information Service Engineering, Beijing Union University.

[7] Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D. and Rorden, C. (2016) Revealing the Dual Streams of Speech Processing. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 15108-15113. https://doi.org/10.1073/pnas.1614038114

[8] Shivappriya, S.N., Priyadarsini, M.J.P., Stateczny, A., Puttamadappa, C. and Parameshachari, B.D. (2021) Cascade Object Detection and Remote Sensing Object Detection Method Based on Trainable Activation Function. *Remote Sensing*, **13**, Article 200. https://doi.org/10.3390/rs13020200

[9] Zhao, C., Gao, Y., He, J. and Lian, J. (2012) Recognition of Driving Postures by Multiwavelet Transform and Multilayer Perceptron Classifier. *Engineering Applications of Artificial Intelligence*, **25**, 1677-1686. https://doi.org/10.1016/j.engappai.2012.09.018

[10] Zhang, B., Wang, W.J., Wee, M. and Cheng, B. (2015) Detection Handheld Phone Use by Driver Based on Machine Vision. *Journal of Jilin University*, **45**, 1688-1695.

[11] Mo, L., Li, F., Zhu, Y. and Huang, A. (2016) Human Physical Activity Recognition Based on Computer Vision with Deep Learning Model. 2016 *IEEE International Instrumentation and Measurement Technology Conference Proceedings*, Taipei, 23-26 May 2016, 1-6. https://doi.org/10.1109/i2mtc.2016.7520541

[12] Xia, H.S., Shen, H. and Hu, W. (2019) Distracted Driving Behavior Recognition Based on Human Body Key Points. *Computer Technology and Development*, **29**, 5. (In Chinese)

[13] Xing, Y., Tang, J., Liu, H., Lv, C., Cao, D., Velenis, E., *et al.* (2018) End-to-End Driving Activities and Secondary Tasks Recognition Using Deep Convolutional Neural Network and Transfer Learning. 2018 *IEEE Intelligent Vehicles Symposium* (*IV*), Changshu, 26-30 June 2018, 1626-1631. https://doi.org/10.1109/ivs.2018.8500548

[14] Xiong, Q., Lin, J., Yue, W., Liu, S., Liu, Y. and Ding, C. (2019) A Deep Learning Approach to Driver Distraction Detection of Using Mobile Phone. 2019 *IEEE Vehicle Power and Propulsion Conference* (*VPPC*), Hanoi, 14-17 October 2019, 1-5. https://doi.org/10.1109/vppc46532.2019.8952474

[15] Ni, C.R. (2021) Driver Driving Behavior Detection and Identification Based on Deep Learning. Master's Thesis, Suzhou University. (In Chinese)

[16] Zhang, Z.Y. (2023) Research on Detection Methods of Dangerous Driving Behavior of Motor Vehicle Drivers Based on Deep Learning. Master's Thesis, Hangzhou University of Electronic Science and Technology. (In Chinese)

[17] Zhang, X. (2023) Research on Key Technologies of Belt Conveyor Foreign Matter Detection System. Master's Thesis, China University of Mining and Technology. (In Chinese)

[18] Wang, W., Chen, J., Huang, Z., Yuan, H., Li, P., Jiang, X., *et al.* (2023) Improved Yolov7-Based Algorithm for Detecting Foreign Objects on the Roof of a Subway Vehicle. *Sensors*, **23**, Article 9440. https://doi.org/10.3390/s23239440

[19] Cao, L., Zheng, X. and Fang, L. (2023) The Semantic Segmentation of Standing Tree Images Based on the YOLOv7 Deep Learning Algorithm. *Electronics*, **12**, Article 929. https://doi.org/10.3390/electronics12040929

[20] Li, S., Wang, S. and Wang, P. (2023) A Small Object Detection Algorithm for Traffic Signs Based on Improved YoLOv7. *Sensors*, **23**, Article 7145. https://doi.org/10.3390/s23167145

[21] Chollet, F. (2017) Xception: Deep Learning with Depthwise Separable Convolutions. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1800-1807. https://doi.org/10.1109/cvpr.2017.195

[22] Wang, Z.Y. and Wang, G.Z. (2023) Application of Improved Lightweight YOLOv5 Algorithm in Pedestrian Detection. *Frontiers of Data &Computing*, **5**, 161-172. https://cstr.cn/32002.14.jfdc.CN10-1649/TP.2023.06.015

[23] Yin, A., Ren, C., Yan, Z., Xue, X., Zhou, Y., Liu, Y., *et al.* (2023) C2S-RoadNet: Road Extraction Model with Depth-Wise Separable Convolution and Self-attention. *Remote Sensing*, **15**, Article 4531. https://doi.org/10.3390/rs15184531

[24] Li, S.Z. and Liu, W. (2023) Small Target Detection Model in Aerial Images Based on YOLOv7X+. https://doi.org/10.21203/rs.3.rs-3052166/v1

[25] Liu, K., Sun, Q., Sun, D., Peng, L., Yang, M. and Wang, N. (2023) Underwater Target Detection Based on Improved Yolov7. *Journal of Marine Science and Engineering*, **11**, Article 677. https://doi.org/10.3390/jmse11030677

[26] Yang, T.H. (2023) Video Behavior Recognition Model Based on Self-Attention Mechanism. Master's Thesis, China University of Mining and Technology. (In Chinese)

[27] Gao, Y.P., Yan, W.H. and Pan, X. (2024) Hyperspectral Image Classification Based on Convolutional Neural Network and Attention Mechanism. *Journal of Optoelectronics Laser*, **35**, 483. (In Chinese)

[28] Du, X.Q., Cheng, H.C., Ma, Z.H., Lu, W.W., Wang, M.X., Meng, Z.C., Jiang, C.J. and Hong, F.W. (2023) DSW-YOLO: A Detection Method for Ground-Planted Strawberry Fruits Under Different Occlusion Levels. *Computers and Electronics in Agriculture*, **214**. http://dx.doi.org/10.2139/ssrn.4511691