Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 16, No. 2, 2025

Driver Distraction Detection Algorithm Based on High-Order Global Interaction Features

Chuanghui Zhang*

Department of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China

* **Corresponding author**

Abstract: Distracted driving behavior is one of the main causes of road traffic safety problems. In view of the problems of high model complexity, unstable detection performance and high hardware cost of existing distraction detection algorithms, this paper proposes a driver distraction detection algorithm based on high-order global interaction features. First, the backbone network is reconstructed using the C3-HB module, in which the HorNet recursive gated convolution is used to learn the long-range dependencies of the image and obtain high-order features, and the bottleneck in the C3 module is replaced by Hornet Block, which improves the detection accuracy of small targets and complex scenes; secondly, the global parallel attention mechanism PGAM is designed to enhance the perception ability of global interaction features and reduce local information loss; finally, the loss function of YOLOv5s is replaced by α-CIOU, which effectively balances the difficult and easy samples and further optimizes the detection effect of driver distraction. Experiments show that the model proposed in this paper achieves an mAP50 of 97.32% on the State Farm dataset and the self-built dataset. While improving the detection accuracy, the number of parameters is only 8.68M and the computational complexity is 12.80GFLOPs, showing excellent comprehensive performance and is suitable for tasks that require real-time and high precision.

Keywords: Global Interaction; YOLO; Distracted Driving; Attention Mechanism.

1. Introduction

Traffic safety [1] has always been a key area of global concern. Driver distraction [2] is one of the major causes of traffic accidents. Distracted driving refers to the driver's attention being diverted to activities other than driving tasks, such as using a mobile phone, eating, and talking to passengers. These behaviors reduce the driver's ability to respond to road conditions, thereby increasing the risk of traffic accidents. With the rapid development of artificial intelligence and computer vision technology, more and more vehicles are equipped with driver distraction behavior detection equipment and automatic driving capabilities, so that when the driver has distracted driving behavior, the vehicle can automatically identify and control the vehicle, ensuring the driver's safety and the risk of injury to pedestrians on the road. Therefore, an accurate and real-time driver distraction behavior detection algorithm is crucial for traffic safety.

At present, driver distraction detection research mainly focuses on the following three aspects: first, driver behavior [3] detection based on image and video analysis, focusing on features such as facial expressions, eye movement trajectories, and head posture. Second, using sensor data to analyze the driver's physiological and behavioral characteristics, such as heart rate, hand movements, etc. Third, combining multimodal data fusion technology to improve the accuracy and robustness of detection. Fourth, using classification tasks to identify driver actions and determine driver distraction.

Researchers around the world have devoted a lot of energy to the study of distracted behavior and published a large number of related papers in international conferences and journals. Tran et al. [4] proposed a driver behavior detection system based on binocular cameras. After using binocular cameras and performing data fusion, the recognition accuracy is significantly higher than that of monocular cameras. However, the detection accuracy requires too high equipment performance and the hardware cost is relatively high. Peng et al [5] improved the pose estimation algorithm OpenPose and input the skeleton confidence map, skeleton affinity field and original image output by the algorithm into the VGG19 deep learning network to classify distracted and abnormal driving behaviors. Although distracted detection was achieved, the real-time performance of the algorithm was slightly insufficient. Ren et al. [6] used a graph convolutional network to extract driver posture features and combined it with a target detection algorithm to classify distracted driving behaviors. The accuracy rate reached 93% on the StateFarm dataset. Lou et al. [7] proposed a lightweight network based on YOLOv5, which integrated the attention mechanism to enhance the algorithm's attention to the target of interest, improved the accuracy of distracted driving behavior recognition, and ensured the real-time detection speed. Du et al. [8] proposed a YOLO-LBS model, which enhanced the path aggregation network to amplify multi-level feature fusion and context information propagation, and expanded the 9 different types of distracted driving in the public StateFarm dataset to 14 categories, introduced night scenes, improved the robustness of detection, and ensured accuracy and real-time performance.

To solve the above problems, this paper studies a driver distraction detection algorithm based on high-order global interaction features from the perspective of the accuracy of the driver distraction detection algorithm and the lightweight model. The main work is as follows:

(1) A global parallel attention mechanism PGAM is designed to enhance the perception of global dimension interaction features, reduce the loss of local information, and improve the accuracy of target detection;

(2) In the backbone network, an improved C3-HB module is designed. The HorNet residual structure enables the high-order interaction module to extract more discriminative features, avoid the problem of missing small targets such as

cigarettes and phones, and improve the detection performance in scenarios with passenger interference;

2. YOLOv5

YOLOv5 [9] is a highly efficient object detection algorithm developed by Ultralytics and is one of the latest versions of the YOLO (You Only Look Once) series. It is implemented on the PyTorch framework and is lightweight, high-precision, and fast for inference. YOLOv5 further improves detection performance by improving network structure, data augmentation, and training strategies while keeping computing resource requirements low. It supports a variety of model sizes (such as YOLOv5s, YOLOv5m, YOLOv5l, etc.), and users can choose the suitable model according to their actual needs. YOLOv5 is widely used in industrial inspection, autonomous driving, security monitoring and other fields.

The main advantage of YOLOv5 is its ease of use and flexibility. It provides a complete training and inference toolchain that allows users to easily train their own datasets and deploy them on a variety of hardware platforms. In addition, YOLOv5 supports a variety of export formats (such as ONNX, CoreML, TensorRT, etc.) for easy integration with other deep learning frameworks. Although YOLOv5 is not the official version of the YOLO series, its open-source community is active, with continuous optimizations and updates, making it one of the popular choices in the field of object detection.

The corresponding specific network structure diagram is shown in Fig. 1. The structure of YOLOv5 is based on a one-stage detector design, which is mainly composed of three parts: Backbone, Neck, and Head. Backbone uses CSPDarknet53 as the feature extraction network to reduce the computational effort and improve the feature fusion capability through Cross Stage Partial Connections (CSP) [10]. The Neck part uses a multi-scale feature fusion structure combining PANet (Path Aggregation Network) [11] and FPN (Feature Pyramid Network) [12], which can effectively capture target information at different scales. The Head part is responsible for generating the final detection result, including the target category, bounding box coordinates, and confidence level.

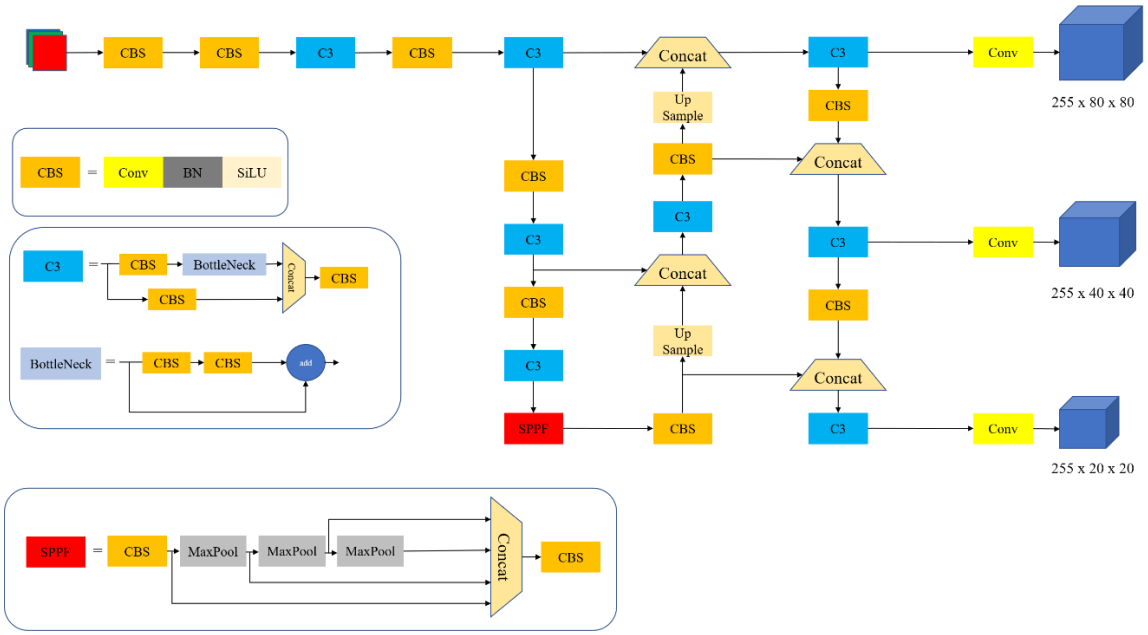

Fig. 1 YOLOv5 network structure diagram

3. Design of distraction detection algorithm

3.1. Network structure

This paper uses YOLOv5s as the baseline and optimizes and improves it to make it better for distraction detection. The network structure diagram is shown in Fig. 2. It consists of Backbone, Neck and Head parts. The C3-HB (C3-HornetBlock) module is designed to reconstruct the backbone network, improve the model's expression ability in complex tasks, and use high-order spatial interactions to enhance feature representation capabilities. The PGAM (Parallel Global Attention Mechanism) attention mechanism is designed to enhance the expression of global features. In the Head part, multi-scale target detection is performed on the feature map extracted by the backbone network to achieve driver distraction detection.

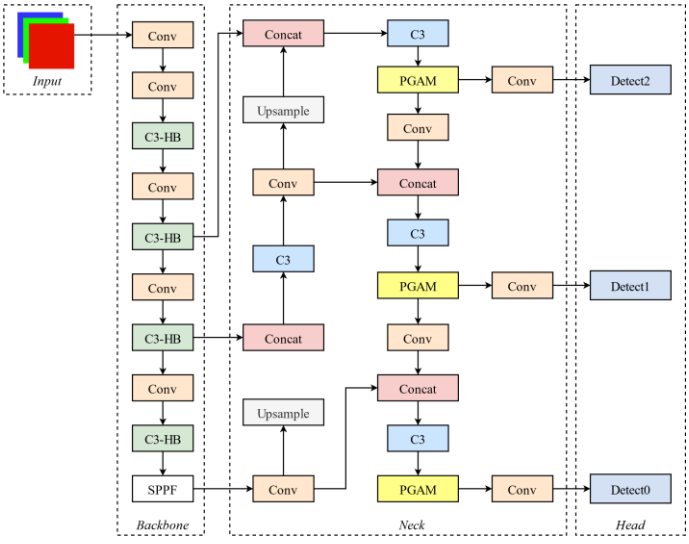

Fig. 2 Proposed network architecture

3.2. PGAM

In the traditional driver distraction behavior detection method, due to the lack of effective features of small objects such as cigarettes and mobile phones, there is a problem of missed detection, and the perception of the global dimension is insufficient. To improve the detection results, this paper adds PGAM in the feature fusion stage (PANet), which helps to optimize the information flow of feature maps of different resolutions by enhancing the global interaction of features of different scales. The traditional global attention mechanism GAM (Global Attention Mechanism) [13] includes a spatial attention module [14] and a channel attention module [15]. The implementation details of PGAM are shown in Fig. 3.

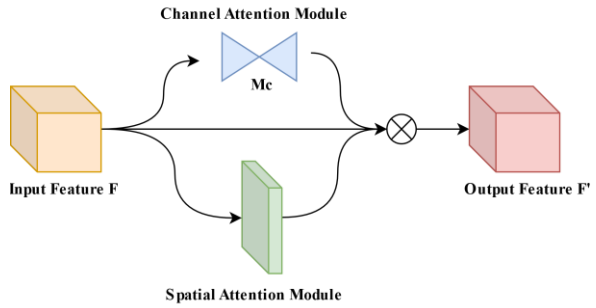

Fig. 3 Global Parallel Attention Mechanism

The channel attention submodule uses a three-dimensional arrangement to retain information in three dimensions. A two-layer MLP is used to amplify the cross-dimensional channel-space dependencies. The internal details of the channel attention submodule are shown in Fig. 4. The usual practice is to perform nonlinear transformations on the input data through multiple neurons to capture the relationship between different dimensions. First, the input data is passed to the first layer of MLP. The neurons in this layer extract features and learn the dependencies between channels. The output of the first layer will serve as the input of the second layer. Here, the activation function Sigmoid is added to introduce nonlinearity, so that the model can better express complex relationships. Finally, the output layer maps the processed features back to the shape of the original data, thereby enhancing the association between spatial features.

The design of the spatial attention submodule focuses on spatial information. The internal details of the attention submodule are shown in Fig. 5. First, the feature map is sent to the channel submodule to assign corresponding weight information to the channel; then, after being combined with the feature map, the feature map is sent to the spatial attention submodule to strengthen the spatial features related to the feature information; finally, the feature map of the fused channel attention submodule is output, and the predicted box generated on the feature map is classified and regressed.

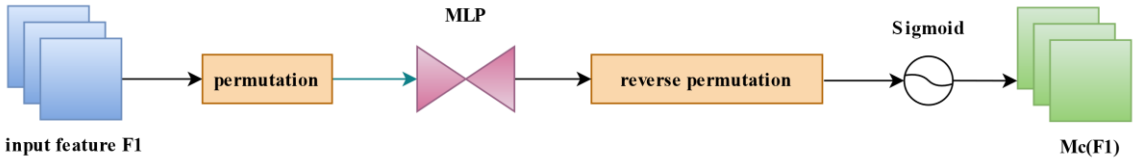

Fig. 4 Channel Attention Submodule

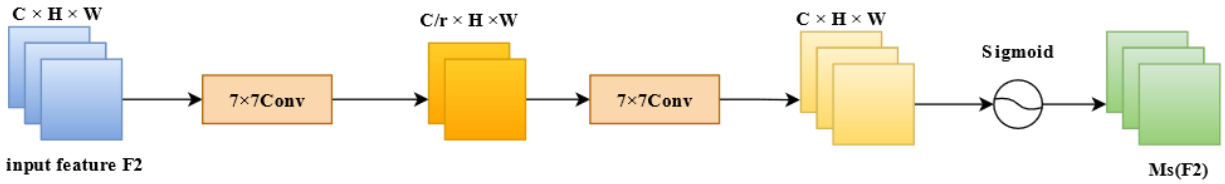

Fig. 5 Spatial attention submodule

3.3. C3-HB

In order to avoid the problem of missing small objects such as cigarettes and mobile phones, the high-order interaction module of HorNet can be used to extract more discriminative features and improve the detection effect of small targets and complex scenes [16]. Under the premise of lightweight, the detection accuracy is effectively improved by adjusting the depth or complexity of the residual block. The enhanced channel interaction mechanism of HorNet helps to optimize the multi-scale target detection performance.

This paper designs a C3-HB (C3-HornetBlock) module, and the specific implementation is shown in Fig. 6. The Backbone of YOLOv5 uses the C3 module in the CSP (Cross Stage Partial Network) structure. This paper replaces the Bottleneck with the HorNet residual block. HorNet uses recursive gated convolution (gnconv) to learn the long-range dependencies of images and obtain high-order features [16]. The architecture of recursive gated convolution is shown in Fig. 6(C).

Assuming that the two-dimensional input features of the recursive gated convolution are $X \in R^{H\times W\times C}$, a set of projected features p_0 and $\{q_k\}_{k-0}^{n-1}$ can be obtained, which can be calculated as follows:

$$\left[p_0^{HW\times C_0}, q_0^{HW\times C_0}, \cdots, q_{n-1}^{HW\times C_{n-1}}\right] = \varphi_{in}(X) \tag{1}$$

where $\varphi_{in}(\bullet)$ denotes the projective operation and then the gated convolution is recursively executed by:

$$p_{k+1} = f_k(q_k) \odot g_k(p_k)/\alpha \tag{2}$$

where α represents the scaling parameter, $f_k(\bullet)$ represents the deep convolution operation, and $g_k(\bullet)$ represents the linear projection layer for the number of channels to adjust.

$$g_k = \begin{cases} Identity, k = 0 \\ Linear(C_{k-1}, C_k), 1 \le k \le n-1 \end{cases} \tag{3}$$

Finally, the output is obtained by projecting features on the last recursive layer. In the residual structure, the residual connection is used to retain the original feature information. The batch normalization (BN) layer is used to increase the speed of training and convergence of the block network and

prevent the gradient from disappearing. The SiLU activation function enhances the expressive power of the model through nonlinear transformation. Fig. 6 shows the architecture of the HorNet-based residual structure, and the process is as follows:

$$\rightarrow Conv+BN+SiLU \rightarrow HorNet \rightarrow Conv+BN+SiLU \quad (4)$$

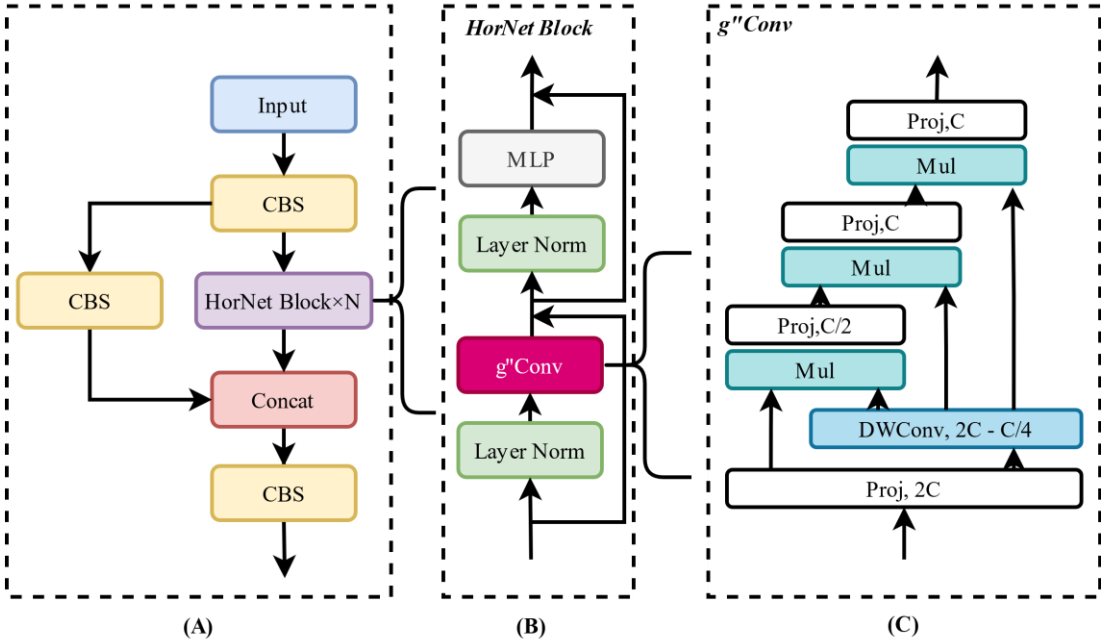

Fig. 6 C3-HB module

4. Experimental results and analysis

4.1. Experimental results and analysis

The computer used in the experiment is configured with Windows 10 operating system, CPU Intel Core i5-13400F 2.50GHz, GPU Nvidia Geforce RTX 4060Ti, and 16G memory. The software environment used is Pytorch 2.1.0, Cuda 11.8, and Python 3.9.18. The model uses the default hyperparameters of YOLOv5, the initial image is 640×640 pixels, the batch size is 32, the learning rate is set to 0.01, the weight decay regularization term coefficient is 0.0005, the momentum factor is 0.937, and a total of 300 rounds are iterated, and the optimal result is selected for analysis.

The experimental data of the algorithm proposed in this experiment include the StateFarm dataset and some self-built datasets. Because this experiment is mainly based on the driver's behavior categories such as drinking water, smoking, making phone calls, etc., the StateFarm dataset was processed and screened.

4.2. Experimental results and analysis

Select accuracy (precision, P), recall (recall, R), and mean average precision (mAP) as indicators to measure the accuracy of target detection. The calculation formulas are as follows. The calculation formula is shown in formula (7); the size of the parameter reflects the scenarios in which the model can be applied.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$AP = \frac{1}{r}\sum_{i=1}^{r} P_i \quad (6)$$

$$mAP = \frac{1}{SUM}\sum_{j=1}^{SUM} AP_j \quad (7)$$

Among them, TP represents samples with positive detection results, FP represents negative samples with positive detection results, r represents all possible values of recall rate, and SUM represents the total number of categories.

4.3. Ablation experiment

Ablation experiments are performed to explain the causality of the working principle of the model and verify that each step of improvement is feasible and effective. The ablation experiments in this paper are shown in Table 1. The YOLOv5s network is used as the baseline to test the performance indicators of each improved module in the network.

Table 1. Ablation experiment

Model	Precision/%	Recall/%	mAP50/%	Params (M)	GFLOPs
YOLOv5s(baseline)	96.90	95.40	96.40	1.77	4.10
YOLOv5s+PGAM	96.91	95.60	96.56	6.23	6.80
YOLOv5s+C3-HB	96.59	95.45	97.23	5.68	10.50
YOLOv5s+PGAM+C3-HB	97.98	95.70	97.32	8.68	12.80

In this study, the PGAM attention mechanism was first added, and the model accuracy and recall rate were improved. Compared with the YOLOv5n algorithm, the mAP value increased by 0.16 percentage points, which shows that the global attention mechanism takes into account the feature information of three dimensions: channel, spatial width and height, which is conducive to enhancing feature extraction and avoiding information loss, thereby improving detection accuracy. However, the model has a large amount of calculation, which reduces the speed to a small extent; secondly, the C3-HB module was added to replace the original C3 module, and the mAP value of the model increased by 0.83 percentage points. Although the recursive operation of the gated convolution increases the complexity of the model, it better captures the long-range dependencies

of the image and obtains high-order features, enhancing the expression ability of the network. The PGAM attention mechanism and the C3-HB module were added to the model at the same time, and the mAP value increased by 0.92 percentage points. This shows that the improved model improves the detection accuracy while ensuring low computational complexity, showing excellent comprehensive performance, and is suitable for tasks that require real-time performance and high precision.

4.4. Comparative experiment

Comparative Experiment refers to a systematic comparison of different algorithms, models, technologies or methods to evaluate their performance differences and advantages and disadvantages on specific tasks or problems.

In order to further analyze the robustness of the model, this study conducted comparative experimental evaluations on the proposed model with YOLOv5 (YOLOv5n, YOLOv5s, YOLOv5m) models of different sizes, DenseNet, ResNet-101, and related models recently proposed by domestic and foreign scholars. By comparing multiple models, the performance of the model in this paper under different scales and different architectures can be comprehensively evaluated, and its robustness and practicality can be further verified. The comparison results are shown in Table 2.

Table 2. Comparative experiment

Model	Precision/%	Recall/%	mAP50/%	Params(M)	GFLOPs
YOLOv5n	88.90	83.70	88.3	1.76	4.10
YOLOv5s	96.90	95.40	96.40	7.02	15.80
YOLOv5m	97.00	95.90	96.10	20.87	47.90
YOLO	91.40	83.80	89.80	5.40	8.20
YOLO-LBS	93.80	92.30	96.30	1.80	6.80
Dri-CGN+obj	93.20	92.13	93.00	20.01	14.22
DenseNet	96.61	95.40	95.56	14.30	14.80
ResNet-101	97.51	95.80	97.50	44.23	50.12
Ours	97.98	95.70	97.32	8.68	12.80

As shown in Table 2, in the comparison with ResNet-101, although ResNet-101 performs well, its model complexity is relatively high, resulting in a decrease in real-time performance. In the comparison with YOLOv5n, YOLOv5s and YOLOv5m, the improved model not only surpasses the confidence of YOLOv5m, but also reduces the number of model parameters by half, and the computational complexity is significantly reduced. Compared with the YOLO model proposed by Lou, although the model complexity of this paper is slightly higher, the accuracy, recall rate and mAP50 are improved by 6.58, 11.9 and 7.52 respectively. Compared with the YOLO-LBS model proposed by Du, the accuracy, recall rate and mAP50 of the model proposed in this paper are improved by 4.18, 3.4 and 1.02 respectively.

In summary, the model proposed in this paper performs outstandingly in the comparison algorithms, with mAP50 reaching 97.32%. While improving the detection accuracy, the number of parameters is only 8.68M, and the computational complexity is 12.80GFLOPs. Compared with the excellent YOLOv5m and ResNet-101, the model complexity is significantly reduced, and the efficient real-time performance is maintained. The method proposed in this paper finds a better balance between model complexity and detection accuracy, ensuring a lower computational cost while improving detection performance.

4.5. Visual comparison of experimental results

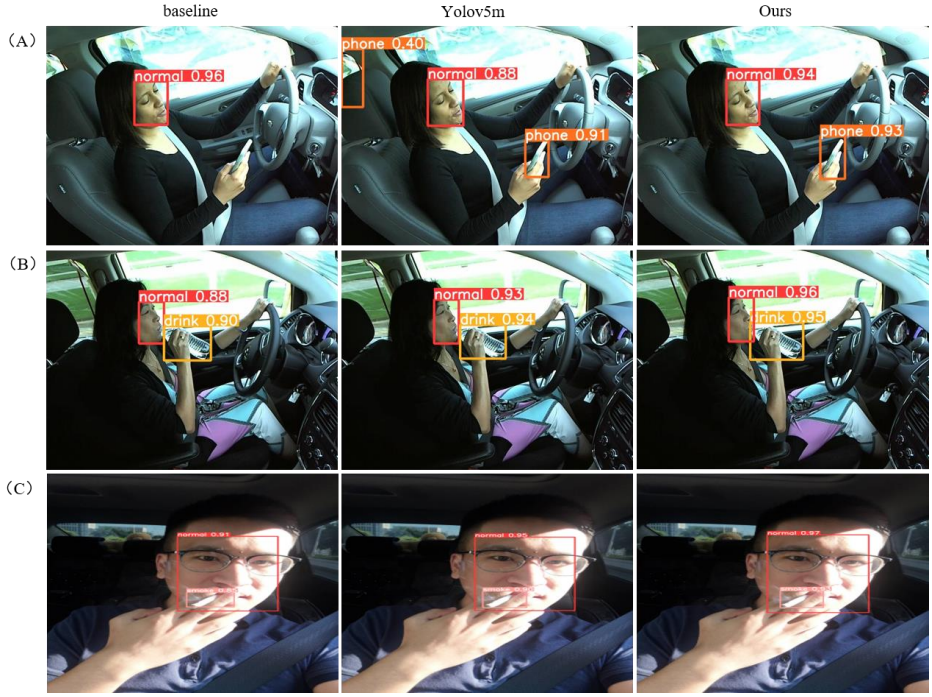

Fig. 7 Visual comparison of distraction detection

In order to verify the model detection effect, this section gives the visualization results. The figure shows the

visualization detection results of YOLOv5s (Baseline), the algorithm proposed in this paper, and the YOLOv5m algorithm on the StateFarm dataset and the self-built dataset. The first column of the detection results is the result of the Baseline detection, the second column is the result of the YOLOv5m algorithm detection, and the third column is the result of the model detection proposed in this paper. The visualization results are compared for distracting behaviors such as making phone calls, drinking water, and smoking.

As can be seen from Figure 7, the method proposed in this paper shows relatively satisfactory detection results. In Figure 7(A), the baseline method missed the driver's distracted behavior of answering and making phone calls, and the YOLOv5m method had the problem of false detection. The method proposed in this paper avoided the above problems and improved the detection confidence of faces and mobile phones. In Figure 7(B), this group of pictures shows the detection of the driver's distracted behavior of drinking water by three methods. All three methods have achieved basic functions, and the detection confidence of the method proposed in this paper is the highest. Compared with the baseline method, the detection confidence of the driver's face and drinking water behavior has increased by 0.08 and 0.05 respectively. In Figure 7(C), this group of pictures shows the detection of the driver's distracted behavior of smoking. Compared with the other two methods, the method proposed in this paper shows good detection performance and the highest confidence.

5. Conclusion

This paper proposes a driver distraction behavior detection algorithm based on high-order global interaction features, and proposes an innovative solution to the shortcomings of existing deep learning-based detection methods in terms of computational complexity and generalization ability. By designing the C3-HB module, this paper uses the recursive gated convolution mechanism of HorNet to effectively learn the long-range dependencies in the image, significantly improving the detection accuracy of small targets and complex scenes. At the same time, the designed global parallel attention mechanism PGAM enhances the model's perception of global interaction features, reduces the loss of local information, and thus improves the overall detection effect. The weights of difficult and easy samples are effectively balanced, further optimizing the detection performance.

Experimental results show that the model proposed in this paper achieves an accuracy of 97.32% on both the StateFarm dataset and the self-built dataset, and can provide high-precision driver distraction behavior detection while ensuring low computational complexity. This model not only improves the detection accuracy, but also shows excellent comprehensive performance, and is suitable for driving safety monitoring tasks that require real-time performance and high precision. Future research can further explore how to enhance the generalization ability of the model in more complex scenarios, or optimize the operating efficiency of the algorithm on different hardware platforms so that it can be more widely used in actual safe driving systems.

Acknowledgment

1). Key Special Project for Science and Technology Strategy Research in Shanxi Province (No.202304031401011)

2). Graduate Innovation Project of Shanxi Province (No.2024SJ364)

References

[1] Khan, M.N. and S. Das, Advancing traffic safety through the safe system approach: A systematic review. Accident Analysis & Prevention, 2024. 199: p. 107518.

[2] Kashevnik, A., et al., Driver distraction detection methods: A literature review and framework. IEEE Access, 2021. 9: p. 60063-60076.

[3] Bouhsissin, S., N. Sael, and F. Benabbou, Driver behavior classification: A systematic literature review. IEEE Access, 2023. 11: p. 14128-14153.

[4] Tran, D., et al. Real-time detection of distracted driving using dual cameras. in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2020. IEEE.

[5] Peng, Q., C. Zheng, and C. Chen. A dual-augmentor framework for domain generalization in 3d human pose estimation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[6] Ren, W., N. Jin, and L. OuYang, Phase space graph convolutional network for chaotic time series learning. IEEE Transactions on Industrial Informatics, 2024.

[7] Lou, C. and X. Nie, Research on lightweight-based algorithm for detecting distracted driving behaviour. Electronics, 2023. 12(22): p. 4640.

[8] Du, Y., et al., Optimizing road safety: Advancements in lightweight YOLOv8 models and GhostC2f design for real-time distracted driving detection. Sensors, 2023. 23(21): p. 8844.

[9] Jocher, G., et al., ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. Zenodo, 2022.

[10] Wang, C.-Y., et al. CSPNet: A new backbone that can enhance learning capability of CNN. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.

[11] [Wang, K., et al. Panet: Few-shot image semantic segmentation with prototype alignment. in proceedings of the IEEE/CVF international conference on computer vision. 2019.

[12] Gong, Y., et al. Effective fusion factor in FPN for tiny object detection. in Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021.

[13] Liu, Y., Z. Shao, and N. Hoffmann, Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561, 2021.

[14] Zhu, X., et al. An empirical study of spatial attention mechanisms in deep networks. in Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[15] Qin, Z., et al. Fcanet: Frequency channel attention networks. in Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Rao, Y., et al., Hornet: Efficient high-order spatial interactions with recursive gated convolutions. Advances in Neural Information Processing Systems, 2022. 35: p. 10353-10366.