Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 8, No. 1, 2023

Real-time Monitoring System for Driver Phone Usage Based on Improved YOLOv5s

Jialing Liu, Baofeng Wang

School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China

Abstract: In response to the impact of driver's violations, such as using a mobile phone, on vehicle safety during the driving process, we propose an improved real-time monitoring algorithm based on YOLOv5s with lightweight optimization. Firstly, we replace the C3 module (CSP Bottleneck with 3 convolutions) in the backbone network of YOLOv5s with a lightweight Ghost Module to reduce the model's parameter count, enhance detection speed, and maintain inference accuracy unaffected, thus meeting the requirements of real-time monitoring. Secondly, we introduce the RepConv (Receptive Field Block) module into the Feature Extraction Network (PANet) structure to increase the neural network's receptive field for input images and further reduce the model's computational load. Experimental results show that the improved network achieves an mAP@0.5 of 95.7%, a detection speed of 140 FPS, and a model size reduction to 10.6MB, meeting the demand for real-time and reliable detection on embedded devices.

Keywords: Real-time monitoring; Embedded; YOLOv5s; Ghost-Module; RepConv.

1. Introduction

The control of the driver is crucial for the safe operation of a vehicle, and normal driving behavior greatly ensures the safety of driving. However, in actual driving, we often witness drivers engaging in violations, such as using mobile phones, which poses a significant threat to road safety [1]. According to relevant studies, approximately 25% to 50% of traffic accidents are caused by improper behaviors of drivers, such as using mobile phones while driving [2]. In order to prevent drivers from being unable to respond promptly to unexpected situations due to phone usage, real-time monitoring of driver behavior becomes particularly important.

Currently, detection algorithms for driver phone usage can be broadly categorized into two types: those based on mobile signal detection and those based on computer vision detection. Scholars such as Ascariz [3] and Jie Yang [4] have proposed methods to identify whether drivers are using phones through mobile signal detection. However, this approach is susceptible to interference from passenger mobile signals, leading to high detection errors. Wei Minguo [5] and others proposed a method to detect phones by extracting F-B Error information to obtain facial features. However, this method has lower robustness and is prone to be affected by factors such as lighting, resulting in the failure of the detection algorithm. Wang Dan [6] and other researchers decomposed the action of a driver making a phone call into a series of sub-actions with certain temporal relationships and detected whether the driver is using a phone in a video through statistical analysis. However, this method is easily interfered with by external factors, leading to detection failures. Wu Chenmou [7] and colleagues proposed a method based on human body pose estimation to estimate the three-dimensional coordinates of eight skeletal nodes in the upper body, using spatial analysis of coordinates to determine if the driver is using a phone. However, this method may mistakenly classify other postures similar to phone usage as phone call actions.

Given the current shortcomings in mainstream detection methods, such as poor robustness, complex algorithm structures, and slow inference speeds, this paper proposes an approach to achieve real-time detection of driver phone usage. The method focuses on achieving high detection accuracy while simultaneously improving the algorithm's detection speed to some extent. This is accomplished through appropriate enhancements to the YOLOv5s algorithm, which is a novel and advanced research direction [8-10]. The proposed improvements involve replacing the original C3 module and Conv module in the backbone network with a more lightweight Ghost-Module and introducing RepConv into the Neck structure to enhance the model's field of view sensitivity. These modifications reduce the redundancy in model information processing and enhance the model's field of view sensitivity. In summary, these improvements enable the detection method to adapt to complex driving scenarios and achieve better real-time detection performance on embedded devices.

2. Related Work

2.1. The overall structure of YOLOv5

YOLOv5 is a one-stage object detection algorithm that integrates numerous advantages from the previous YOLO series, demonstrating outstanding accuracy and real-time performance in the field of object detection. It encompasses multiple versions, such as YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, allowing for flexible adjustments of the network's depth and width through configuration file tuning. Moreover, YOLOv5 boasts remarkable portability and has been widely applied and deployed in practical production environments. The network structure of YOLOv5 is illustrated in Figure 1.

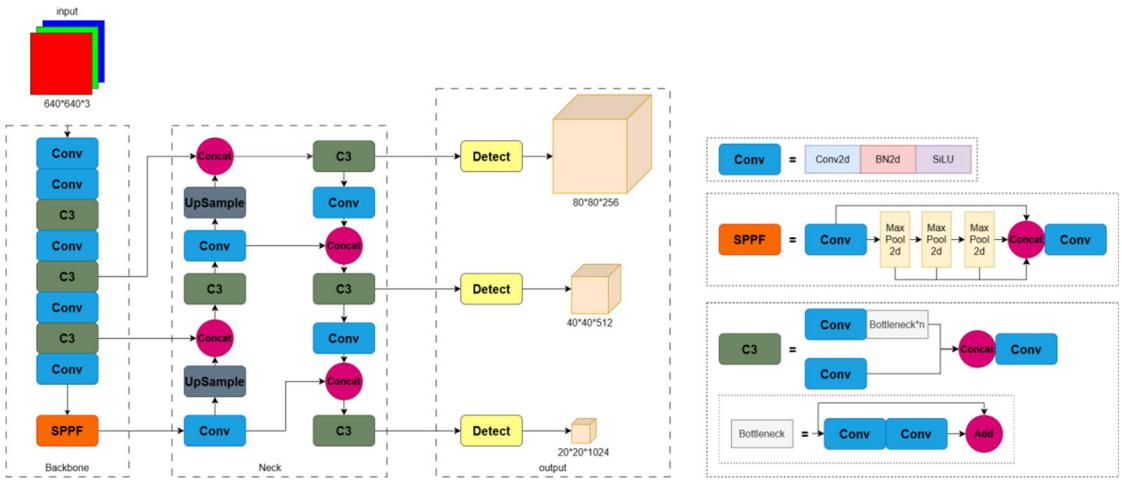

Figure 1. Network structure of YOLOv5

2.2. Loss Function

DYOLOv5 employs three types of loss functions, namely classification loss, localization loss, and confidence loss, to assess the algorithm's detection performance. Among them, Intersection over Union (IoU) is a simple function used to calculate the localization loss. It evaluates the matching degree of two bounding boxes by measuring their overlap and then utilizes Non-Maximum Suppression (NMS) to filter out the optimal detection results.

However, in practical applications, it has been observed that the IoU function has some deficiencies, making it challenging to meet the requirements for localization loss in complex environments. Consequently, various improved versions of localization loss calculation methods have emerged, with common ones including Generalized IoU (GIoU), Distance IoU (DIoU), and Complete IoU (CIoU). Among these, CIoU is a loss function that considers additional optimization strategies and exhibits higher accuracy in evaluation.

YOLOv5 precisely adopts the CIoU function as the localization loss function for the network model. The expressions for IoU and CIoU are shown in equations (1) and (2):

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

$$Loss_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{2}$$

3. Improved YOLOv5s Network Model

As a currently highly regarded object detection algorithm, YOLOv5s has made significant progress by integrating numerous optimization strategies. However, due to its large number of parameters, relatively low model inference speed, and substantial model file size, real-time detection on embedded devices has become a challenge. Recognizing this issue, this paper addresses it by designing an improved version of the YOLOv5s detection algorithm, aiming to more effectively meet the real-time requirements of driver behavior detection tasks.

References are cited in the text just by square brackets [1]. (If square brackets are not available, slashes may be used instead, e.g. /2/.) Two or more references at a time may be put in one set of brackets [3, 4]. The references are to be numbered in the order in which they are cited in the text and are to be listed at the end of the contribution under a heading References, see our example below.

3.1. Backbone Network Improvement

In order to achieve efficient deployment on low-computational-power devices, lightweight and low-latency characteristics are particularly important for convolutional neural networks. Recently, Han and his colleagues introduced a new neural network called GhostNet[15]. This network is based on the Ghost-Module, and it is constructed by stacking Ghost bottlenecks, resulting in an efficient and lightweight network structure. Verified experimentally, GhostNet maintains high accuracy while exhibiting superior computational efficiency and inference speed. Compared to other models, GhostNet achieves the highest throughput on GPUs and lower latency on CPUs and ARM.

Building upon this foundation, this paper introduces the Ghost bottleneck structure and Ghost-Module into the backbone network. While maintaining detection accuracy with minimal decline, this significantly enhances the lightweight and low-latency characteristics of the network model.

3.1.1. Ghost-Module and Ghost bottleneck

The Ghost bottleneck, as the core structure of the lightweight GhostNet, demonstrates outstanding performance. Its architecture is implemented based on the Ghost-Module, a phantom convolution module.

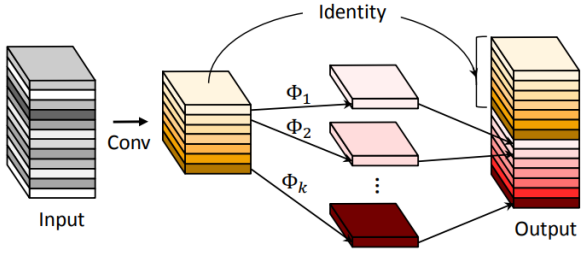

Figure 2. Ghost-Module network architecture diagram.

From Figure 3, it is evident that the Ghost-Module first generates some feature maps through a regular convolution, then performs a cheap operation on the generated feature maps to produce redundant feature maps. The convolution used in this step is a depth-wise separable convolution. Finally, the feature maps generated by the regular convolution are concatenated with the feature maps generated by the cheap operation. By stacking Ghost-Modules, the Ghost bottleneck structure is obtained.

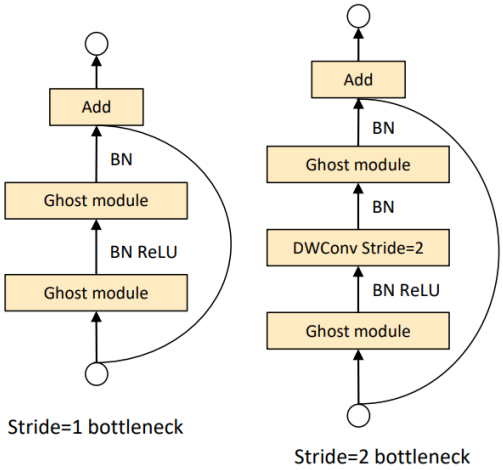

Figure 3. Ghost bottleneck network architecture diagram.

As shown in the figure, the Ghost bottleneck appears to be similar to the Basic Residual Block in ResNet, incorporating multiple convolutional layers and a shortcut. The Ghost bottleneck primarily consists of two stacked Ghost modules. The first Ghost-Module serves as an expansion layer, increasing the number of channels, while the second Ghost-Module reduces the number of channels to match the shortcut path. Then, the shortcut connects the inputs and outputs of these two Ghost modules.

In this paper, Ghost-Modules and Ghost bottleneck structures are introduced into the backbone network of YOLOv5, optimizing the design to create a more lightweight feature extraction network structure.

3.2. Feature Pyramid Network Improvement

In object detection tasks, multi-scale features play a crucial role in encoding objects with scale variations. Common strategies for multi-scale feature extraction include the use of classical top-down and bottom-up feature pyramid networks. The YOLOv5 network model, however, adopts a bidirectional fusion backbone network known as the Path Aggregation Network (PANet), which operates in a top-down and bottom-up manner to enhance the structure of the feature pyramid. Additionally, a "short-cut" path is added between the bottom and top layers to shorten the fusion path of high and low-level features.

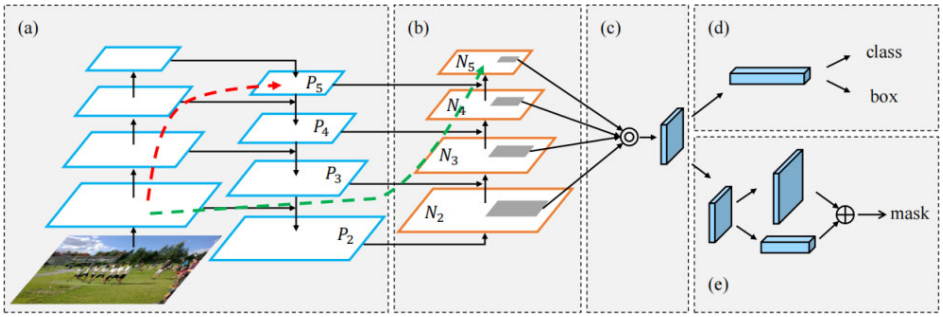

Figure 4. PANet Network Structure.

In this paper, on the basis of the PANet feature extraction network, the Reparameterizable Convolution (RepConv) module [17] is introduced. This module dynamically adjusts the shape and parameters of the convolution kernel according to the network's requirements. In comparison to regular convolution operations, it can adapt to a broader range of tasks and network architecture needs. Specifically, during training, this module functions as a multi-branch module, and during inference, the multi-branch module is equivalently transformed into a single-path module.

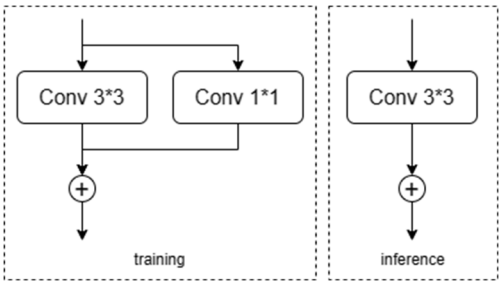

Figure 5. RepConv Network Structure.

The module has been introduced into the Neck structure of YOLOv5, providing lightweight processing for the feature extraction network. The improved PANet network structure is illustrated in Figure 6.

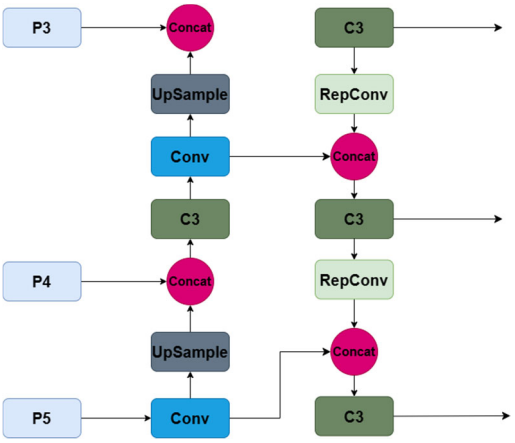

Figure 6. The improved Neck section structure.

4. Experiment

4.1. Dataset and Experimental Environment

This experiment utilized a self-built dataset, consisting of 1400 photos captured from real driving scenes and 221 photos collected from the internet. The dataset comprises a total of 1621 images captured by smartphones, including targets of various scales. These photos cover diverse scenarios such as real driving environments, streets, and subways, enhancing the complexity of the photo environment and the diversity of the data. Following the principle of an 8:1:1 ratio, the photos were divided into training, validation, and test sets. The dataset was further categorized into three classes: head, body, and phone.

The hardware setup required an Nvidia Geforce RTX 3060 graphics card. The experiment was conducted on a system running the Windows 10 operating system with 16GB of RAM and equipped with a 12th Gen Intel(R) Core(TM) i5-12400F processor. The deep learning framework PyTorch 1.12.0 was employed in the experimental environment.

4.2. Evaluation Metrics

In order to quantitatively evaluate and analyze the detection performance of YOLOv5s for detecting drivers using mobile phones, this paper selected detection accuracy, detection speed, and model size as performance evaluation metrics. Specifically, these include precision, recall, mean average precision (mAP), and model file memory occupancy (MB). In this paper, we use mAP at an IoU threshold of 0.5 as the evaluation metric. These metrics can be calculated using the following formulas:

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$mAP = \frac{\sum_{i=1}^{c} AP_i}{c} \tag{7}$$

$$FPS = \frac{N}{t} \tag{8}$$

4.3. Model Training

In this paper, the models before and after improvement were analyzed under the same hardware facilities and parameter conditions. The analysis was based on the Loss curve representative of the experimental results. During the model training process, the network parameters were set as shown in Table 1.

Table 1. Network parameters

Parameters	Value
Iteration Rounds	150
Batch Size	8
Learning Rate	0.01
Weight Decay	0.0005
Image Size	640*640
Momentum	0.937

4.4. Comparative Experiments

To effectively demonstrate the superior performance of the proposed algorithm for real-time detection of drivers using mobile phones, it is compared with current mainstream lightweight detection algorithms YOLOv5s, FasterNet, and RepViT on the self-built dataset, using the same hardware and training parameters. The results are shown in Table 2.

Table 2. Compares with current mainstream detection algorithms

Algorithm Model	P/%	R/%	mAP/%	Memory Usage Ratio(MB)
YOLOv5s	98.1	93.1	96.3	14.1
FasterNet_t0	98.2	93.5	96.2	6.42
RepViT_m1	97.7	93	95.8	12.0
Ours	97.1	92.5	95.7	10.6

After analyzing the comparative results, we found that the improved algorithm achieves an average precision of 95.6%. Compared to the current mainstream detection algorithms, there is a slight decrease in accuracy, but there is a significant improvement in terms of FPS and model memory usage.

Through the comparison, it is observed that although the algorithm in this paper has a slight decrease in detection accuracy, this is understandable in the process of lightweighting the model, and the degree of decrease is within an acceptable range. In terms of lightweighting, the algorithm proposed in this paper is significantly better than the current mainstream detection algorithms.

4.5. Dropping Experiment

Conducting Ablation Experiment to confirm the effectiveness of the proposed improvements in this paper. This aims to comprehensively evaluate the performance impact of each improvement on the algorithm. The introduced enhancements include incorporating Ghost-Module and RepConv. The experimental results are presented in Table 3.

Table 3. Compares with current mainstream detection algorithms

Ghost-Module	RepConv	P/%	R/%	mAP	Memory Usage Ratio(MB)
×	×	98.1	93.1	96.3	14.1
√	×	96.7	94	94.3	10.4
×	√	98.3	94.4	96.9	14.3
√	√	97.1	92.5	95.7	10.6

4.6. Model Deployment

For a more direct understanding of the detection performance of the algorithm before and after improvement, the improved algorithm model was deployed on the RK3399Pro development board for inference detection. The deployment experimental results are shown in Table.

Table 4. Comparison between Baseline and Improved Models on RK3399Pro

Algorithm Model	FPS/(frame/s)
YOLOv5s	16
Ours	20

The detection results are shown in the following images:

(a) Daytime driver using a mobile phone.

(b) Nighttime driver using a mobile phone.

Figure 7. Detection Results of the Improved Model

As shown in Figure 7(a) and (b), the improved network model can accurately and efficiently complete the detection of drivers using mobile phones in complex environments, including daytime and nighttime scenarios.

5. Summary

This paper addresses issues with existing real-time detection algorithms for detecting drivers' use of mobile phones, such as high memory usage and poor real-time performance, by proposing a series of improvement strategies. By introducing Ghost-Module into the backbone network, the feature extraction is enhanced, simultaneously reducing the model's parameter count, suppressing invalid information input, and improving the model's focus on target information. In the Neck section, the RepConv module is introduced to lightweight the feature extraction network. The experimental results demonstrate that the improved YOLOv5s algorithm for real-time detection of drivers' use of mobile phones, as proposed in this paper, has a significant advantage over other mainstream detection algorithms. Compared to the original algorithm, this algorithm has undergone substantial lightweight processing while meeting the requirements for real-time detection. However, due to the relatively lower accuracy of the modified model, future work will focus on improving both accuracy and detection speed to achieve better detection performance on embedded devices.

References

[1] World Health Organization.Global status report on road safety[R].Geneva:WHO,2018[2019-09-15].

[2] H K M B, SKOV M B, THOMASSEN N G.You can touch, but you can’t look: interacting with in-vehicle systems [C]//Conference on Human Factors in Computing Systems. Florence, Italy: CHI, 2008:1139-1148.

[3] BEISSEL S, BELYTSCHKO T. Nodal integration of the element-free Galerkin method[J].Computer Methods in Applied Mechanics and Engineering, 1996,139(1-4): 49-74.

[4] CHENG Y M, ZHANG Y H, CHEN W S. Wilson non-conforming element in numerical manifold method[J]. Commun. Numer. Meth, 2002, 18(12): 877-884.

[5] Hou Yuqingyang, Quan Jicheng, Wang Hongwei. Overview of Deep Learning Development[J]. Journal of Shipborne Electronic Engineering, 2017, 37(4): 5-9.

[6] Wang Dan. Driver Phone-Calling Behavior Detection Based on Computer Vision [D]. Beijing: Beijing Institute of Technology, 2015.

[7] Wu Chenmou, Fang Zhijun, Huang Zhengneng. Active Driving Behavior Analysis Algorithm Based on Monocular Camera [J]. Journal of Shandong University (Engineering Science), 2018, 48(5): 69-76.

[8] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time ObjectDetection[J]. IEEE, 2016: 444-453.

[9] Redmon J, Farhadi A. YOLO9000:Better,Faster,Stronger[C]// IEEE.IEEE, 2017:6517-6525.

[10] YANG Y Z. Drone-view obiect detection based on the improved YOLO5(CJ/Proceedings of the lEEE ntemational Conference on Electrical Engineering, Big Data and Algorithms.Changchun: IEEE.2022: 612-617.

[11] HE K M, ZHANG X Y. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9): 1904-1916.

[12] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8759-8768.

[13] REZATOFIGHI H,TSOI N,GWAK J,et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.IEEE, 2019:658-666.

[14] ZHENG Z,WANG P,LIU W,et al. Distance-IoU loss: Faster and better learning for bounding box regression [C]// Proceedings of the AAAI conference on artificial intelligence, 2020.AAAI,2020:12993-13000.

[15] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2019). GhostNet: More Features From Cheap Operations. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1577-1586.

[16] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate Attention for Efficient Mobile Network Design. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13708-13717.

[17] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). RepVGG: Making VGG-style ConvNets Great Again. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13728-13737.

[18] Lee, J., Park, S., Mo, S., Ahn, S., & Shin, J. (2020). Layer-adaptive Sparsity for the Magnitude-based Pruning. International Conference on Learning Representations.

[19] Liu, S., Huang, D., & Wang, Y. (2019). Learning Spatial Fusion for Single-Shot Object Detection. ArXiv, abs/1911.09516.

[20] [20].Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO Series in 2021. ArXiv, abs/2107.08430.

[21] Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H.P. (2016). Pruning Filters for Efficient ConvNets. ArXiv, abs/1608.08710.

[22] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C. (2017). Learning Efficient Convolutional Networks through Network Slimming. 2017 IEEE International Conference on Computer Vision (ICCV), 2755-2763.

[23] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16091-16101.

[24] Zhao Min, Yang Guoliang, Wang Jixiang, Gong Zhipeng. Improving the real-time detection algorithm for safety helmets in YOLOv7 tiny [J]. Radio Engineering, 2023, v.53; No.411 (08): 1741-1749.

[25] Xiong Qunfang, Lin Jun, Yue Wei, et al. Driver Phone Use Detection Method Based on Deep Learning [J]. Control and Information Technology, 2019(6): 5. DOI: 10.13889/j.issn.2096-5427.2019.06.400.

[26] ZHANG S, ZHU X, LEI Z, et al. S^3FD: Single Shot Scale-Invariant Face Detector[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE Computer Society, 2017：192-201.

[27] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(6):1137-1149.

[28] Zhang J, Chen Z, Liu W, et al. A field study of work type influence on air traffic controllers' fatigue based on data-driven PERCLOS detection[J]. International journal of environmental research and public health, 2021, 18(22):11937.