Journal of Real-Time Image Processing (2024) 21:75
https://doi.org/10.1007/s11554-024-01455-3

RESEARCH

Driver fatigue detection based on improved YOLOv7

Xianguo Li[1,2] · Xueyan Li[1] · Zhenqian Shen[1] · Guangmin Qian[3]

Received: 30 January 2024 / Accepted: 25 March 2024 / Published online: 13 April 2024

Abstract
Fatigue driving is one of the main reasons threatening road traffic safety. Aiming at the problems of complex detection process, low accuracy, and susceptibility to light interference in the current driver fatigue detection algorithm, this paper proposes a driver Eye State detection algorithm based on YOLO, abbreviated as ES-YOLO. The algorithm optimizes the structure of YOLOv7, integrates the multi-scale features using the convolutional block attention mechanism (CBAM), and improves the attention to important spatial locations in the image. Furthermore, using the Focal-EIOU Loss instead of CIOU Loss to increase the attention on difficult samples and reduce the influence of sample class imbalance. Then, based on ES-YOLO, a driver fatigue detection method is proposed, and the driver fatigue judgment logic is designed to monitor the fatigue state in real-time and alarm in time to improve the accuracy of detection. The experiments on the public dataset CEW and the self-made dataset show that the proposed ES-YOLO obtained 99.0% and 98.8% mAP values, respectively, which are better than the compared algorithms. And this method achieves real-time and accurate detection of driver fatigue status. Source code is released in https://www.github/driver-fatigue-detection.git.

Keywords Fatigue detection · YOLOv7 · CBAM · Focal-EIOU loss · Fatigue judgment logic

1 Introduction

In recent years, with the rapid growth of transportation vehicles such as cars and subways, various traffic accidents have occurred frequently, resulting in a large number of property losses and casualties, as well as posing a huge safety hazard to society. Studies have shown that in addition to overloading and speeding, fatigue driving is one of the three main causes of traffic accidents. According to statistics, fatigue driving accounts for 20% of the causes of traffic accidents [1]. When people are in the state of fatigue, their attention and reaction speed will decrease significantly, thereby increasing the risk of traffic accidents. Therefore, the research and design of accurate and real-time detection algorithm and method for driver fatigue detection, real-time and accurate judgment of driver fatigue state, and alarm prompt when fatigue is detected, improve driving safety, timely prevent traffic accidents caused by fatigue, reduce the accident rate, and improve the level of road safety. The research holds high social benefits and practical value.

At present, the detection methods of fatigue driving are mainly divided into three categories: detection methods based on driver physiologic characteristics, detection methods based on vehicle motion characteristics and methods based on driver's facial features.

The fatigue detection methods based on physiologic characteristics mainly use sensors to collect the physiologic signals of drivers, and judge the fatigue state by observing changes in relevant parameters. Common physiologic signals include electro-oculography (EoG) [2], electroencephalogram (EEG) [3], electrocardiogram (ECG) [4], electromyogram (EMG) [5], etc. Lin et al. [6] designed a wearable EoG data collection device to estimate blink frequency and subsequently estimate fatigue state. Dogan et al. [7] designed a new manual modeling learning framework using EEG for fatigue state detection. Zhang et al. [8] used heartbeat signals and blink signals captured by antennae for fatigue detection. Chen et al. [9] proposed a new minimum spanning tree for feature extraction and then feature fusion to detect driver

✉ Zhenqian Shen
shenzhenqian@tiangong.edu.cn

[1] Tiangong University, Tianjin 300387, China

[2] Tianjin Key Laboratory of Optoelectronic Detection Technology and System, Tianjin 300387, China

[3] Tianjin Railway Traffic Operation Group Co, Tianjin 300392, China

fatigue state. This method has high accuracy, but it is easy to interfere with the driver, so it is not widely used.

The detection methods based on vehicle motion characteristics judge driver fatigue based on vehicle state changes. Li et al. [10] proposed a driver fatigue detection method based on steering wheel angle. The method uses decision tree classifier to extract approximate entropy features from signals recording steering wheel angle, with an accuracy of 82.7%. Forsman et al. [11] found that the variability of lateral lane position can be obtained from the measured change of steering wheel angle through the transfer function, so the variability of steering wheel can be used as an indicator to detect fatigue driving. The detection results of this method are susceptible to factors such as driving conditions and driving habits.

The detection methods based on drivers' facial features usually use image processing techniques to analyze the captured driver images. By detecting the driver's eye and mouth states, calculating the frequency of blinking and yawning, the driver's fatigue state can be determined. The calculation of blinking frequency depends on the reliable recognition of the driver's eye state, which is the key to fatigue detection. Yi et al. [12] proposed an eye-based fatigue detection algorithm that fuses multiple eye features to detect the fatigue state. This method has fast-detection speed and high accuracy, but it needs to use the Dlib tool to mark the facial feature points, which is complex to operate. Jia et al. [13] designed three networks to detect the driver's face, head and eye-mouth state, and then detect the fatigue state. Du et al. [14] used convolutional neural network to establish two models based on heart rate and percentage of eyelid closure over the pupil over time (PERCLOS), and realized multi-modal fusion fatigue detection. Sun et al. [15] proposed a multi-stream facial feature fusion convolutional neural network that improves the fatigue detection performance at low-quality input, but it needs to collect signals such as EOG, which will interfere with the driver and increase the detection cost. Mateusz Knapik and others [16] proposed a driver fatigue detection method based on thermal imaging yaw detection. First, the corners of the eyes were detected to achieve face alignment, and then the yawning thermal model was proposed to detect yawning, so as to realize fatigue state recognition. But the low resolution of the thermal image will affect the accuracy of the detection. Yang et al. [17] used 3D convolution and bidirectional long short-term memory networks for spatiotemporal feature extraction. This method can effectively distinguish yawning and similar facial movements, but cannot detect images with low resolution.

In summary, the detection method based on driver's facial features has become a research hotspot for fatigue detection, but there are still shortcomings such as complex detection process, low accuracy rate, and susceptibility to light influence. Therefore, this paper researches and proposes a driver's Eye State detection algorithm based on YOLO named ES-YOLO (hereinafter referred to as the ES-YOLO eye state detection algorithm), along with a driver fatigue detection method. The main contributions are mainly in three aspects:

1. The convolutional block attention mechanism (CBAM) is used to replace the 3×3 ordinary convolution and 1×1 ordinary convolution at specific locations of the feature extraction, feature fusion, and detection head module, using YOLOv7-tiny as the basic framework. This can better extract the deep and shallow features, improve the ES-YOLO model's attention to important spatial locations in the image, and improve the accuracy of eye state detection.
2. The Focal-EIOU Loss is used to replace the CIOU Loss, so that the ES-YOLO model can more effectively solve the category imbalance problem and increase the attention to the difficult samples, so as to improve the training efficiency and the accuracy of eye state detection.
3. The driver fatigue detection method based on ES-YOLO is proposed, and the driver fatigue judgment logic is designed. This method can monitor the fatigue status in real time and provide timely alarms, improving the accuracy and practical value of fatigue detection.

2 Network structure of ES-YOLO eye state detection algorithm

YOLOv7 is one of the widely recognized target detection algorithms, which is a typical representative of one-stage detection algorithms, and it performs well in terms of high accuracy and high real-time performance. Compared with two-stage detection algorithms such as Faster R-CNN, YOLOv7 has a higher detection speed and better real-time performance. YOLOv7 contains three different sizes of network structures, namely YOLOv7-tiny, YOLOv7, and YOLOv7-X. Compared with YOLOv5, YOLOv6 [18], and YOLOv8, the YOLOv7 [19] algorithm is more mature and faster to detect, which is more suitable for engineering applications. The algorithm in this topic needs to take into account real-time and lightweight, so YOLOv7-tiny is chosen for improvement.

We analyze the structural characteristics and optimization methods of the YOLOv7-tiny network, and propose the ES-YOLO algorithm for detecting the eye state rapidly and accurately to further determine the fatigue state. The overall network structure of ES-YOLO is shown in Fig. 1, which mainly contains three parts: the feature extraction module, the multi-scale feature fusion module, and the multi-scale detection head module. First, the feature extraction module extracts low-level

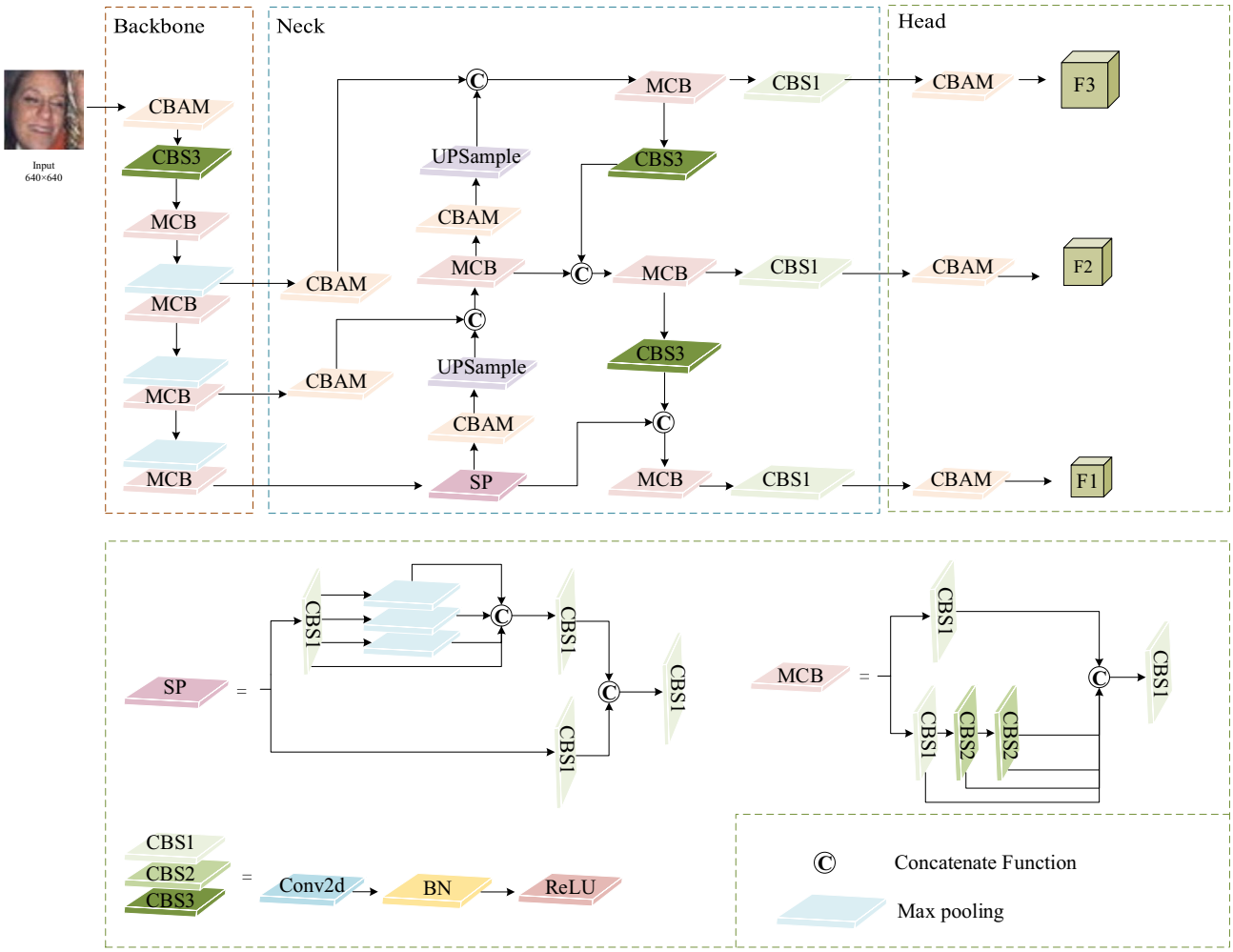

Fig. 1 Overall network structure of ES-YOLO

spatial information and high-level semantic information from the input images and delivers them to the feature fusion module. Then, the feature fusion module fuses the extracted multilevel features to ensure that the network can better perceive targets at different scales. Finally, the multiscale-detection head is responsible for generating dense-bounding boxes and predicting the category scores, and the final inference is obtained by employing Non-Maximum Suppression (NMS).

The components of ES-YOLO are mainly composed of three building blocks: standard convolution (CBS), pyramid pooling (SP), and multi-standard convolution (MCB). CBS has three forms, CBS1, CBS2 and CBS3, which respectively represent convolution modules with convolution kernel 1 and step 1, convolution kernel 3 and step 1 and convolution kernel 3 and step 2.

2.1 Convolutional block attention module (CBAM)

To make the model better focus on key regions and features, capture important information in the image more comprehensively, and improve the performance of the model, we introduce the CBAM attention module. Convolutional block attention module (CBAM) [20] is a simple and effective attention module for feedforward neural networks, as shown in Fig. 2. It contains two independent sub-modules, namely the Channel Attention Module (CAM) and Spatial Attention Module (SAM), which can focus on channel information and position information of objects. For the input feature maps, attention operations can be performed sequentially in the channel and spatial dimensions.

Given an intermediate feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ map as input, CBAM can sequentially infer a one-dimensional channel

attention map $\mathbf{M_c} \in \mathbb{R}^{C\times1\times1}$ and a two-dimensional spatial attention map $\mathbf{M_s} \in \mathbb{R}^{1\times H\times W}$. The overall CBAM attentional process can be expressed as

$$\begin{aligned}\mathbf{F'} &= \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F} \\ \mathbf{F''} &= \mathbf{M_s}(\mathbf{F'}) \otimes \mathbf{F'}\end{aligned} \tag{1}$$

where $\otimes$ denotes element-by-element multiplication, during multiplication, the channel attention value is broadcast along the spatial dimension, $\mathbf{F''}$ is the final output. Figures 3 and 4 depict the computation process of each attention map. The following describes the details of each attention module.

A channel attention map is generated using inter-channel relationships. Since each channel of the feature map is considered as a feature detector, channel attention focuses on 'what' is meaningful given an input image. To compute the channel attention efficiently, the spatial dimensions of the input feature maps are compressed. The channel attention mechanism utilizes both average pooling and max pooling features to aggregate spatial information, significantly enhancing the representational capacity of network. The channel attention mechanism can be expressed as

$$\begin{aligned}\mathbf{M_c}(\mathbf{F}) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathbf{F}))) \\ &= \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{avg}}))+\mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{max}})))\end{aligned} \tag{2}$$

where σ is the sigmoid function, $\mathbf{W_0} \in \mathbb{R}^{C/r\times C}$, $\mathbf{W_1} \in \mathbb{R}^{C\times C/r}$. CAM compresses the spatial information of feature maps by using global max pooling and global average pooling to obtain two different spatial context descriptors: $\mathbf{F^c_{avg}}$ and $\mathbf{F^c_{max}}$, then forward them to the shared network to generate channel attention maps $\mathbf{M_c} \in \mathbb{R}^{C\times1\times1}$. The shared network consists of a multilayer perceptron (MLP) [21] with one hidden layer. To reduce parameter overhead, the hidden-activation size is set to $\mathbb{R}^{C/r\times1\times1}$, where r is the reduction ratio. After applying the shared network to each descriptor, use element-by-element summation to merge the output feature vectors. The channel attention process is shown in Fig. 3.

Spatial attention maps are generated using spatial relationships of features. Unlike channel attention, spatial attention focuses on the 'where' as the information component and is complementary to channel attention. Spatial attention can be expressed as

$$\begin{aligned}\mathbf{M_s}(\mathbf{F}) &= \sigma(f^{7\times7}([\mathrm{AvgPool}(\mathbf{F});\mathrm{MaxPool}(\mathbf{F})])) \\ &= \sigma(f^{7\times7}([\mathbf{F^s_{avg}};\mathbf{F^s_{max}}]))\end{aligned} \tag{3}$$

where $f^{7\times7}$ represents a convolution operation with the filter size of 7×7. First, aggregate channel information of a feature map by using two pooling operations, generating two 2D maps: $\mathbf{F^s_{avg}} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{F^s_{max}} \in \mathbb{R}^{1\times H\times W}$, each denotes

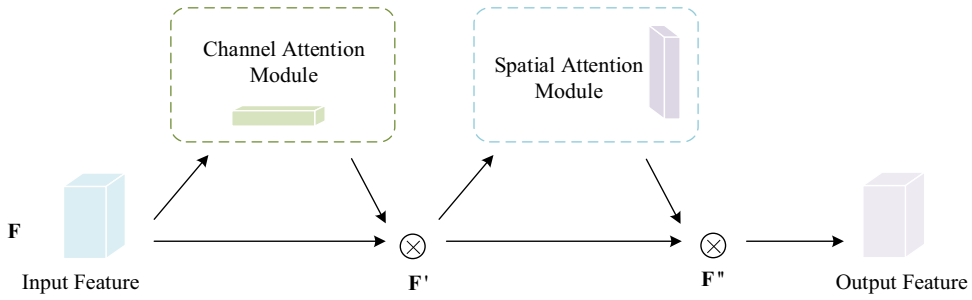

Fig. 2 Structure of CBAM attention mechanism

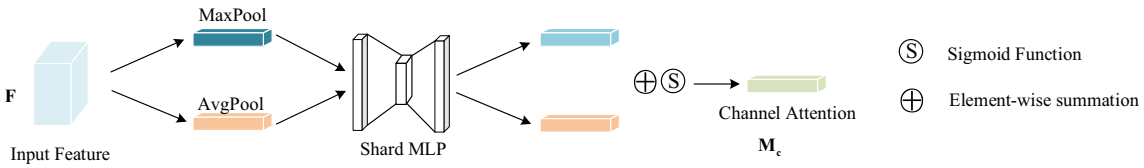

Fig. 3 Channel attention

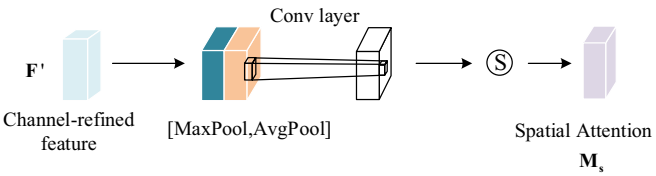

Fig. 4 Spatial attention

average-pooled features and max-pooled features across the channel. Following that, the feature maps are connected and convolved through a 7×7 convolutional layer, and then normalized using the sigmoid function to obtain the spatial attention feature map. The spatial attention process is illustrated in Fig. 4.

We replace the 1×1 ordinary convolution at specific locations of Backbone in the YOLOv7-tiny network with the CBAM attention mechanism, which enhances the feature representation between different channels and enables the network to better capture relevant information in different channels. Second, the spatial attention in CBAM can help the network focus on important regions in the image, helping the network better understand the spatial structure in the image. Replacing the 1×1 ordinary convolution at specific locations in the Neck part with the CBAM attention mechanism, the channel attention mechanism in the CBAM helps the model to integrate the multi-scale features more efficiently, so that the model pays more attention to the important channels in each feature map. The spatial attention mechanism helps the model to focus more on important spatial positions in the image, particularly for small or occluded targets, so that the network pays more attention to the key areas containing critical information. Lastly, replacing the 3×3 normal convolution in the Head part with the CBAM attention mechanism improves the performance of the network in target detection.

2.2 Focal-EIOU loss

To solve the imbalance problem between high-quality and low-quality samples (the intersection size of different regions on the union of predicted-truth boxes and ground-truth boxes), we introduce the Focal-EIOU Loss. Focal-EIOU Loss makes the regression process more focused on high-quality anchor boxes, which can effectively solve the imbalance problem between high-quality and low-quality samples, accelerate the convergence of the model, and improve the detection accuracy. The loss function of YOLOv7-tiny consists of three parts: target confidence loss, classification loss, and regression loss. The confidence and classification losses are calculated using binary cross-entropy, while the regression loss is calculated using the CIOU Loss [22]. However, the CIOU Loss only considers the aspect ratio difference between the predicted bounding box and the ground truth, neglecting the true relationships between the width and height of the bounding box and the ground truth.

To solve this problem, reference [23] optimized the CIOU Loss and proposed the EIOU Loss. The EIOU Loss regresses the actual width and height of the predicted box instead of using the aspect ratio, thereby eliminating the negative impact of the uncertainty in the aspect ratio. The formula for calculation is as follows

$$L_{\mathrm{EIOU}} = 1 - \mathrm{IOU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{\mathbf{gt}})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{\mathrm{gt}})}{(w^c)^2} + \frac{\rho^2(h, h^{\mathrm{gt}})}{(h^c)^2} \tag{4}$$

where h^c and w^c are the width and height of the smallest enclosing box covering the two boxes, $\mathbf{b}$ and $\mathbf{b}^{\mathbf{gt}}$ denote the central points of predicted box and the ground truth bounding box, w and h represents the width and height of the prediction bounding box, w^{gt} and h^{gt} represents the width and height of the ground truth bounding box.

Based on the EIOU Loss, this paper introduces the Focal-EIOU Loss [23] in the proposed ES-YOLO network to make the regression process focus on high-quality anchor boxes and to reduce the drastic fluctuations of the loss value caused by the training of the model on low-quality samples. The calculation formula for Focal-EIOU Loss is

$$L_{\mathrm{Focal-EIOU}} = \mathrm{IOU}^{\gamma} L_{\mathrm{EIOU}} \tag{5}$$

where γ is a parameter to control the degree of inhibition of outliers. In this article, γ is taken as 0.5.

3 Driver fatigue detection method based on ES-YOLO

We use the PERCLOS method to determine driver fatigue status, PERCLOS originated from an experiment conducted by Wierwille and colleagues, which proved that the eye closure time reflects fatigue to some extent. Based on this, the Carnegie Mellon Institute, after repeated experiments and demonstrations, proposed the PERCLOS, which is used to describe the fatigue condition. Numerous experiments have proved that the PERCLOS method is the most accurate detection method, and it is also the only driver fatigue detection method recognized by the National Highway Traffic Safety Administration (NHTSA) in the United States. There are three measurements of PERCLOS, which are P70, P80, and EM. Among them, P80 is considered to be the most reflective of human fatigue [24, 25], and this paper adopts the P80 measurement to make the driver fatigue state to judge.

We evaluate the fatigue state of drivers by studying their eye closure status and analyzing their videos. A driver fatigue detection method based on ES-YOLO is designed using PERCLOS, and the flowchart is shown in Fig. 5. This method mainly consists of two parts: eye closure state detection and statistical discrimination. When driver fatigue is detected, a fatigue driving alert is issued.

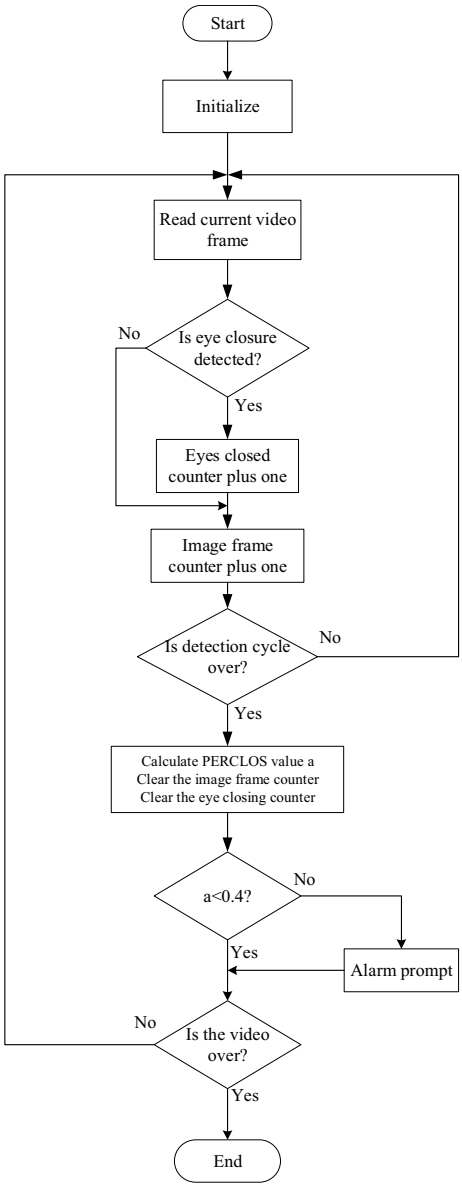

Fig. 5 Flowchart of driver fatigue detection method based on ES-YOLO

4 Experiment analysis

We use driver driving videos provided by subway companies and public datasets: the University of Texas at Arlington Real-Life Drowsiness Dataset (UTA-RLDD) [26], and the CEW dataset [27], to create a driver fatigue detection dataset, which includes a total of 7954 images. The performance of the ES-YOLO eye state detection algorithm and driver fatigue detection method proposed in this paper are tested using publicly available and self-made dataset.

4.1 Experimental environment and settings

The hardware test environment is: Intel(R) Core (TM)i7-7700 CPU @ 3.60 GHz, 32G memory, two NVIDIA GeForce GTX 1080Ti GPU. Software environment is Windows 10 operating system, CUDA11.1, Python3.8.18 and PyTorch1.9.1.

The dataset is divided into training, validation and test sets in the ratio of 7:2:1 and trained using SGD optimizer. The input image size is 640×640 on both CEW dataset and self-made dataset, epoch is set to 80 and 100 respectively, bath size is set to 64, and data enhancement is carried out using methods such as Mixup [28]. The initial-learning rate was set to 0.01, momentum to 0.937, and weight decay to 0.0005. To improve the accuracy of the model for eye localization, the IOU threshold in this paper was set to 0.55. All tests were performed on the same hardware and software platforms.

4.2 Evaluation metrics

The accuracy of the model was assessed using precision (P), recall (R), mean average precision (mAP), and the speed of the model was assessed in terms of Number of Parameters (Params), Detection Time, and number of floating point operations (FLOPs), and the reconciled average of precision and recall (F1 score), and false rate (FPR).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{7}$$

$$\text{AP} = \int_0^1 P(R)\text{d}r \tag{8}$$

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} \text{AP}(\text{classes}_i) \tag{9}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{11}$$

where TP represents the number of positive samples, when detecting eye closure, the true positive sample is when the predicted bounding box of eye closure matches the ground truth box and the IOU is greater than the specified threshold; FP represents the number of false positive samples, FN represents the number of negative samples, TN is the number of negative samples correctly predicted, n is the total number of target categories. In this paper, n is 2.

4.3 ES-YOLO experiments and result analysis

To validate the effectiveness and superiority of the proposed algorithm in this paper, a series of experiments were conducted on the CEW dataset and self-made dataset, followed by ablation experiments and analysis.

4.3.1 Comparison between different algorithms

Table 1 demonstrates the comparison between ES-YOLO and other target detection algorithms on the CEW dataset. It can be seen that compared to the original YOLOv7-tiny, ES-YOLO exhibits minimal increases in parameter count, time, and FLOPs, while achieving a 0.7% and 1.0% improvement in precision for open and close eye detection, respectively. Compared to YOLOv5-n, ES-YOLO achieves 0.7% and 1.2% higher precision on open and close kinds with a small increase in detection time. Compared to YOLOv5-m, the proposed algorithm achieves the same accuracy as it does on the kind open, and 1.2% higher precision on the kind close, with a parameter count of only 28% and FLOPs of 27% in comparison. In addition, in terms of mAP, although ES-YOLO is 0.1% behind YOLOv7, it only has 16% of the parameter count and 12% of the FLOPs of YOLOv7. Compared to YOLOv8s, ES-YOLO achieves a higher precision of 0.3% and 1.6% in the open and close categories, respectively, while also exhibiting lower parameter count and processing time. Compared to Faster R-CNN, the algorithm proposed in this paper is numerically 1.0% higher in terms of precision on close kind and in terms of recall, the algorithm in this paper is 3.8% higher on open kind. The mAP of our algorithm surpasses Faster R-CNN by 1.4%. Our algorithm outperforms Faster R-CNN in terms of parameters, time, and FLOPs. Moreover, our algorithm demonstrates comparable performance to other target detection algorithms in terms of F1 score. For FPR, ES-YOLO achieves the lowest false detection rate.

Figure 6 shows the detection results of ES-YOLO and other algorithms on the CEW dataset. It can be observed from the figure that the proposed ES-YOLO algorithm exhibits less false positives in detecting closed eyes compared to the other three algorithms. Compared with YOLOv7, our algorithm has higher confidence and more accurate judgment. As shown in the left figure of Fig. 6c, YOLOv7 has misjudged and incorrectly identified open as close.

To demonstrate the accuracy of the algorithm more intuitively, Fig. 7 shows the PR curves on the CEW dataset.

In addition, experiments were conducted on a self-made dataset, and the results are presented in Figs. 6, 8 and Table 2 Compared to the original YOLOv7-tiny, ES-YOLO showed an increase of 0.2% in precision for the open category, and although the precision for the close category decreased by 0.9%, the recall rates for both open and close categories, respectively improved by 0.2% and 1.5%. In comparison to YOLOv5-n, although the precision for the close category decreased by 0.7%, the precision and recall rates for the open category were improved by 0.1%, and the recall rate for the close category also increased by 0.5%. In terms of mAP, the proposed algorithm in this paper demonstrated a 0.3% improvement over YOLOv5-m. Compared to YOLOv7, the precision rates for the open and close categories for the proposed algorithm were numerically higher by 1.6% and 2.6% respectively, and the recall rates were also higher by 2.3% and 3.3% for open and close categories. Compared to YOLOv8s, the algorithm presented in this paper achieves a higher precision of 0.2% and 0.8% in the open and close categories, respectively, and also exhibits a higher recall of 0.3% and 0.7%, respectively. Compared with Faster R-CNN, the precision of the algorithm proposed in this paper is numerically higher by 0.9% and 0.5% in open and close categories. In terms of recall, the algorithm in this paper is higher by 0.4% and 1.3% in open and close categories, respectively. From the perspective of mAP, our algorithm is 1.8% higher than Faster R-CNN. From the right image of Fig. 6b, it can be observed that YOLOv7-tiny has missed detections, while our algorithm demonstrates higher accuracy.

4.3.2 Ablation experiment

To visually observe the impact of different improvement methods on model performance, ablation experiments

Table 1 Comparison of ES-YOLO and other algorithms on CEW dataset

Algorithms	*P*		*R*		mAP	Params (M)	Time (ms)	FLOPs (G)	F1	FPR
	Open	Close	Open	Close						
Faster R-CNN	97.8	97.3	94.5	95.8	97.6	41.3	220	230.1	0.96	2.5
YOLOv5-n	96.4	97.1	99.3	97.9	99.1	1.7	5.5	4.1	0.98	3.3
YOLOv5-m	97.1	97.1	98.5	97.1	99.0	21.2	13.4	47.9	0.97	2.9
YOLOv7-tiny	96.4	97.3	98.8	96.7	99.1	5.7	4.3	13.0	0.97	3.2
YOLOv7	97.5	98.3	97.8	96.7	99.1	36.5	23.1	103.2	0.98	2.1
YOLOv8s	96.8	96.7	97.1	97.5	99.0	11.2	13.0	28.4	0.97	3.3
Ours	97.1	98.3	98.3	94.7	99.0	6.0	7.0	13.2	0.97	2.3

(a)YOLOv5-n detection results

(b) YOLOv7-tiny detection results

(c)YOLOv7 detection results

(d) ES-YOLO detection results

Fig. 6 Comparison of the detection results of ES-YOLO with other algorithms

were conducted. To ensure the rigor of the experiments, 100 epochs were set on the same training platform, and after training was completed, testing was conducted on the test set of a self-made dataset. The experimental results are shown in Table 3. In the table, ✓ represents the location where changes were made. ✓ next to CBAM indicates the replacement of specific 3×3 ordinary convolutions in the Backbone, specific 1×1 ordinary convolutions in the Neck, and three groups of 1×1 ordinary convolutions in the Head with the CBAM attention mechanism. ✓ next to Focal-EIOU indicates the usage of Focal-EIOU Loss to replace the CIOU Loss in YOLOv7-tiny. P and R represent the precision and recall for the close category. It can be observed that the model achieves the best detection performance when both improvement methods are used simultaneously.

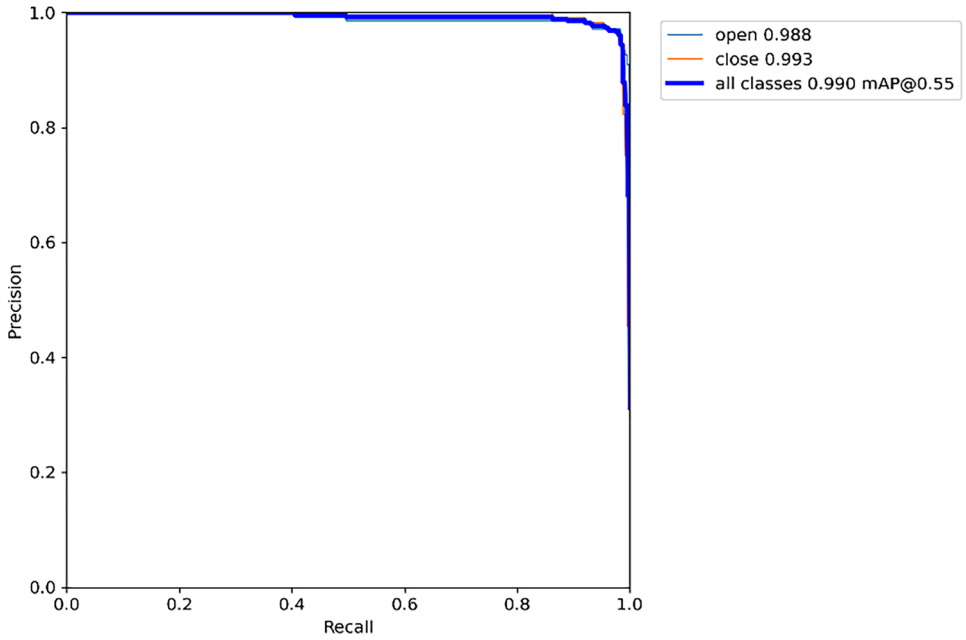

Fig. 7 PR curve of CEW dataset

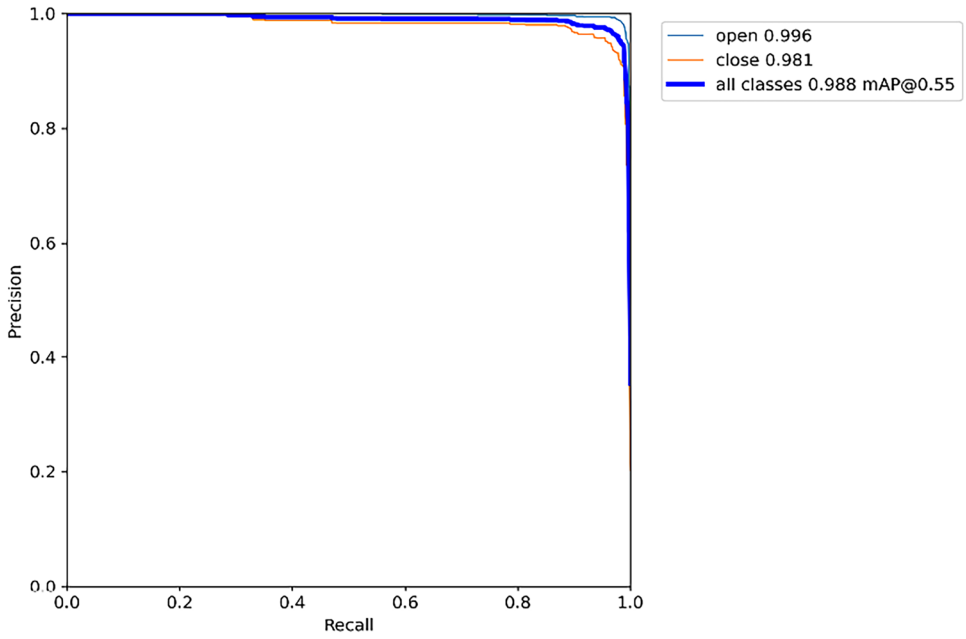

Fig. 8 PR curve of self-made dataset

Table 2 Comparison of ES-YOLO and other algorithms on self-made dataset

Algorithms	*P*		*R*		mAP	F1	FPR
	Open	Close	Open	Close			
Faster R-CNN	98.7	95.2	98.5	94.2	97.0	0.96	3.1
YOLOv5-n	97.5	96.4	98.8	95.0	98.7	0.97	3.1
YOLOv5-m	97.3	96.7	99.0	95.7	98.5	0.97	3.0
YOLOv7-tiny	97.4	96.6	98.7	94.0	98.5	0.97	3.0
YOLOv7	96.0	93.1	98.6	91.8	97.8	0.95	5.5
YOLOv8s	97.4	94.9	98.6	94.8	98.5	0.96	3.9
Ours	97.6	95.7	98.9	95.5	98.8	0.97	3.3

To visually assess the impact of the CBAM attention mechanism at different positions in the network on model performance, ablation experiments on the CBAM attention position were conducted on a self-made dataset, and the experimental results are shown in Table 4. Here, P and R represent the precision and recall for the close category.

The first line is the experimental results of the baseline. The second set of experimental data is only replacing the 3×3 ordinary convolutions at certain locations in the Backbone with the CBAM attention mechanism. The third set of experimental data is only replacing the 1×1 ordinary convolutions at certain locations in the Neck with the CBAM attention mechanism. The fourth set of experimental data is replacing all three sets of 3×3 ordinary convolutions in the Head part with CBAM attentional mechanisms. The fifth set of experimental data is replacing 1×1 ordinary convolutions and 3×3 ordinary convolutions at certain locations in the Backbone and Neck parts with CBAM attentional mechanisms, and at the same time to replacing all 1×1 ordinary convolutions in the Head part with CBAM attentional mechanisms. It can be seen that the fourth set of experiments, i.e., choosing to replace the convolutions at specific locations in the Backbone, Neck, and Head sections, is the best in all other metrics, even though the running time is slightly longer.

4.4 Experiment and result analysis based on ES-YOLO driver fatigue detection method

To verify the real-time and effectiveness of the driver fatigue detection method based on ES-YOLO proposed in this paper, the driver driving videos provided by Metro are detected. Table 5 shows the experimental data of the proposed method for eye closure state detection based on ES-YOLO and eye closure state detection based on the original YOLOv7 algorithm, respectively. Due to the differences in drivers' individual facial features, in Video 1, YOLOv7 produces a large number of misjudgments for the eye closure category, which in turn leads to fatigue misjudgments and multiple false alarms, while ES-YOLO can accurately determine the eye closure state. In Videos 2 and 3, YOLOv7 also has a small number of misjudgments. In addition, in terms of detection time, for the same video, the algorithm in this paper requires less detection time. Figure 9 illustrates the practical application effect of the driver fatigue detection method based on ES-YOLO.

Table 3 Ablation experiments conducted on self-made dataset

CBAM	Focal-EIOU	P	R	mAP	Params (M)	Time (ms)
		94.8	96.5	98.9	5.7	4.3
✓		95.5	94.5	98.7	6.0	5.2
	✓	95.4	95.9	98.9	6.0	4.5
✓	✓	95.7	95.5	98.8	6.0	7.0

Table 4 CBAM attention position ablation experiments

Backbone	Neck	Head	P	R	mAP	Time (ms)
			96.6	94.0	98.5	4.7
✓			95.6	93	98.5	4.8
	✓		92.5	92.8	97.9	4.4
		✓	95.2	96.7	98.7	4.5
✓	✓	✓	95.7	95.5	98.8	7.0

Table 5 Tests of real-time driver fatigue detection and alarm system on different models

Video	Model	Frames with eyes closed	Number of closed eye frames detected	Number of alarms	The number of alarms should be reported	Detection time (ms/frame)
Video1	YOLOv7	268	317	5	3	34.9
	Ours		270	3		23.2
Video2	YOLOv7	118	120	1	1	35.3
	Ours		117	1		24.4
Video3	YOLOv7	154	168	2	2	34.8
	Ours		150	2		23.3

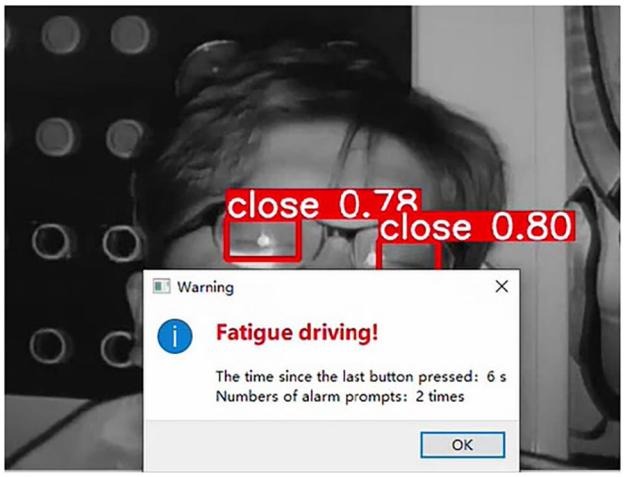

Fig. 9 Actual application effect of driver fatigue detection method based on ES-YOLO

5 Conclusion

First of all, this article proposes an improved driver eye state detection algorithm ES-YOLO based on YOLOv7. CBAM attention mechanism is used to replace 3×3 ordinary convolutions at specific positions in the Backbone, so that the network can better capture the relevant information in different channels. In addition, the 1×1 ordinary convolution at some specific positions of the Neck is replaced by the CBAM attention mechanism, which improves the feature fusion ability of the network. Furthermore, the three groups of 1×1 ordinary convolutions in the head are all replaced by the CBAM attention mechanism, which improves the performance of the network in the target classification task. To address the problem of the imbalance in the number of categories in the self-made dataset, Focal-EIOU Loss is used to replace the original CIOU Loss, improving the detection accuracy of network. On this basis, using PERCLOS, this paper proposes a driver fatigue detection method based on ES-YOLO, and designs the driver fatigue judgment logic to improve the accuracy of driver fatigue state detection. Tests on CEW data sets and self-made data sets show that ES-YOLO eye state detection algorithm meets the requirements of high accuracy and high real-time detection for fatigue detection in actual scenes. Ablation experiments prove the rationality and effectiveness of the algorithm. The driver fatigue detection method based on ES-YOLO can timely and accurately detect driver fatigue and issue alerts, with a small parameter volume and fast-detection speed, making it easy to port to hardware devices to build a driver fatigue monitoring system, and has high practical value. Future research will focus on further improving the accuracy of the algorithm to better meet practical application requirements. However, when the driver wears glasses or the face rotates at a large angle, the accuracy of the algorithm will decrease. In future, further research can be conducted on the human eye state to adapt to the situation of drivers wearing glasses and improve the robustness of algorithms.

Acknowledgements This work was supported by the Tianjin "Project+Team" Key Training Special Project under Grant XB 202007 and Tianjin Transportation Technology Development Project Plan under Grant 2021-36.

Author contributions All authors contributed to the study conception and design. The data analysis was done by Li, X.G. The systems and experiments were designed by Li, X.Y. The paper was revised by Shen, Z.Q., and the research was conducted by Qian, G.M.

Funding This study was supported by The Tianjin "Project + Team" Key Training Special Project (XB 202007); Tianjin Transportation Technology Development Project Plan (2021-36).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interests The authors declare no competing interests.

References

1. Sikander, G., Anwar, S.: Driver fatigue detection systems. A review. IEEE Trans. Intell. Transp. Syst. **20**, 2339–2352 (2019)
2. KołOdziej, M., Tarnowski, P., Sawicki, D.J., Majkowski, A., Rak, R.J., Bala, A., Pluta, A.: Fatigue detection caused by office work with the use of EOG signal. IEEE Sens. J. **20**, 15213–15223 (2020)
3. Jap, B.T., Lal, S., Fischer, P., Bekiaris, E.: Using EEG spectral components to assess algorithms for detecting fatigue. Expert Syst. Appl. **36**, 2352–2359 (2009)
4. Zhao, L., Li, M., He, Z., Ye, S., Qin, H., Zhu, X., Dai, Z.: Data-driven learning fatigue detection system: a multimodal fusion approach of ECG (electrocardiogram) and video signals. Measurement **201**, 111648 (2022)
5. Mashayekhi, M., Moghaddam, M.: EMG-driven fatigue-based self-adapting admittance control of a hand rehabilitation robot. J. Biomech. **138**, 111104 (2022)
6. Lin, B., Wu, P., Chen, C.: 2D/3D-display auto-adjustment switch system. IEEE J. Biomed. Health Inform. **22**, 799–805 (2018)
7. Dogan, S., Tuncer, I., Baygin, M., Tuncer, T.: A new hand-modeled learning framework for driving fatigue detection using EEG signals. Neural Comput. Appl. **35**, 14837–14854 (2023)
8. Zhang, J., Wu, Y., Chen, Y., Wang, J., Huang, J., Zhang, Q.: Ubi-fatigue: toward ubiquitous fatigue detection via contactless sensing. IEEE Internet Things J. **9**, 14103–14115 (2022)
9. Chen, J., Wang, H., Hua, C.: Electroencephalography based fatigue detection using a novel feature fusion and extreme learning machine. Cognitive Syst. Res. **52**, 715–728 (2018)
10. Li, Z., Li, S., Li, R., Cheng, B., Shi, J.: Driver fatigue detection using approximate entropic of steering wheel angle from real driving data. Int. J. Robot. Autom. **17**, 495 (2017)
11. Forsman, P., Vila, B., Short, R., Mott, C., Dongen, H.: Efficient driver drowsiness detection at moderate levels of drowsiness. Accid. Anal. Prev. **50**, 341–350 (2013)

12. Yi, Y., Zhou, Z., Zhang, W., Zhou, M., Yuan, Y., Li, C.: Fatigue detection algorithm based on eye multifeature fusion. IEEE Sens. J. **23**, 7949–7955 (2023)
13. Jia, H., Xiao, Z., Ji, P.: Real-time fatigue driving detection system based on multi-module fusion. Comput. Graph. **108**, 22–33 (2022)
14. Du, G., Zhang, L., Su, K., Wang, X., Teng, S., Liu, P.: A multimodal fusion fatigue driving detection method based on heart rate and PERCLOS. IEEE trans. Intell. Transp. Syst. **23**, 21810–21820 (2022)
15. Sun, Z., Miao, Y., Jeon, J., Kong, Y., Park, G.: Facial feature fusion convolutional neural network for driver fatigue detection. Eng. App. Artif. Intell. **126**, 106981 (2023)
16. Knapik, M., Cyganek, B.: Driver's fatigue recognition based on yawn detection in thermal images. Neurocomputing **338**, 274–292 (2019)
17. Yang, H., Liu, L., Min, W., Yang, X., Xiong, X.: Driver yawning detection based on subtle facial action recognition. IEEE Trans. Multimedia **23**, 572–583 (2021)
18. Li, C., Li, L., Jiang, H., et al: YOLOv6: A single-stage object detection framework for industrial applications. arXiv (2022) https://doi.org/10.48550/arXiv.2209.02976
19. Wang, C., Bochkovskiy, A., Liao, H.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: IEEE computer society conference on computer vision and pattern recognition, 2023. CVPR2023. IEEE, pp 7464–7475 (2023)
20. Woo, S., Park, J., Lee, J., Kweon, I.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer vision—ECCV 2018. ECCV 2018. Lecture notes in computer science, vol. 11211. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
21. Tolstikhin, I., Houlsby, N., Kolesnikov, A., et al: MLP-mixer: an all-MLP ARCHITECTURE FOR VISION. arXiv (2021) https://doi.org/10.48550/arXiv.2105.01601
22. Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., Zuo, W.: Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans. Cybern. **52**, 8574–8586 (2022)
23. Zhang, Y., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T.: Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing **506**, 146–157 (2022)
24. Mu, Z., Jin, L., Yin, J., Wang, Q.: Research on a driver fatigue detection model based on image processing. Comput. Intell. Neurosci. (2022). https://doi.org/10.22967/HCIS.2022.12.017
25. Zhou, M., Zhang, H., Zhang, H., Yi, Y.: An improved random forest algorithm-based fatigue recognition with multiphysical feature. IEEE Sens. J. **23**, 26195–26201 (2023)
26. Ghoddoosian, R., Galib, M., Athitsos, V.: A realistic dataset and baseline temporal model for early drowsiness detection. In: IEEE Computer society conference on computer vision and pattern recognition workshops, 2019. CVPRW2019. IEEE, pp 178–187 (2019)
27. Liu, Z., Jiang, C., Li, S., Wu, M., Cao, W., Hao, M.: Eye state detection based on weight binarization convolution neural network and transfer learning. Appl. Soft Comput. **109**, 107565 (2021)
28. Zhang, H., Cisse, M., Dauphin, Y., Lopez-Paz D.: mixup: beyond empirical risk minimization. arXiv (2017) https://doi.org/10.48550/arXiv.1710.09412