IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 January 2024, accepted 3 March 2024, date of publication 7 March 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3374726

RESEARCH ARTICLE

Distracted Driving Behavior and Driver's Emotion Detection Based on Improved YOLOv8 With Attention Mechanism

BAO MA[1], ZHIJUN FU[1], (Member, IEEE), SUBHASH RAKHEJA[2], DENGFENG ZHAO[1], WENBIN HE[1], WUYI MING[1], AND ZHIGANG ZHANG[1]

[1]College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[2]Department of Mechanical and Industrial Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

Corresponding author: Zhijun Fu (2019003@zzuli.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62073298; in part by the Special Application for Key Scientific and Technological Project of Henan Province under Grant 232102221036; in part by the Key Research and Development Projects of Henan Province, in 2022, under Grant 221111240200; in part by the Doctoral Fund of Zhengzhou University of Light Industry under Grant 2019BSJJ002; and in part by the Henan Center for Outstanding Overseas Scientists under Grant GZS2023011.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT An improved YOLOv8 detection method is proposed for detecting distracted driving behavior and driver's emotion. Unlike the commonly used YOLOv8 method, an attention mechanism named MHSA and a CNN module are synthesized to ensure improved performance in terms of accuracy and convergence, where MHSA is used to detect distracted driving behavior and CNN is used to detect driver's emotion. The FER2013 dataset and collected dataset are used to train the improved YOLOv8. The training results show that the proposed YOLOv8 demonstrates improved performance compared with the commonly used YOLO based methods. Finally, the validity of the proposed YOLOv8 method is illustrated through implementations in Jetson Nano platform, where the TensorRT and DeepStream methods in the Jetson Nano device are used to optimize the volume and operational speed of the proposed YOLOv8 method, respectively. Test results show that the proposed YOLOv8 method can yield better real-time and accuracy properties.

INDEX TERMS YOLO, multi-head self-attention, CNN, visual object classes, distracted driving behavior, driver's emotion.

I. INTRODUCTION

Traffic accidents account for nearly 1.2 million human fatalities worldwide each year with even greater number of non-fatal injuries [1]. Traffic accidents involving commercial vehicles also cause excessive property damages and environmental risks, especially when transporting hazardous cargos [2]. The developments in various vehicle stability enhancement and driver-assist systems have contributed to a steady decline in road accidents. The economic and social costs of fatal as well as non-fatal traffic accidents, however, continues to be excessive and unacceptable. Advancements in active safety devices have helped reduce the number and severity of accidents attributed to road- and vehicle-related factors, while the driver-related behavioral factors constitute the primary causal factors. These include reduced cognition, inattention/distractions, changes in emotions, fatigue, poor assessment of road or traffic condition, over speeding, and inadequate following distance [3]. It has been reported that nearly 60% of road accidents are directly caused by factors related to driver, while these contribute to about 95% of all accidents [4]. The human driver behavior and emotions are thus key factors to be considered in designing safer road-vehicle systems [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Rabiul Islam.

Some of the advance driver assist systems (ADAS) incorporate monitoring of abnormal driving behaviors and provide some form of warning [6], while the implementations have been highly challenging as these require accurate and real-time identifications. Monitoring of one or more measures of driver's behavior has been attempted via the Vehicle Ad Hoc Network (VANET) [7]. Al-Sultan [8] expanded the VANET by integrating a real-time probabilistic model based on the dynamic Bayesian network (DBN). The dynamic driver behavior was inferred by incorporating contextual information related to the driver, vehicle, and the environment to detect four behavioral measures, namely, drunkenness, fatigue, aggressive and normal. Owing to the large number of parameters and considering that VANET operates as a mobile AD hoc network (MANET), the contextual information transmission to a real-time probabilistic model via dedicated short-range communication (DSRC) imposes considerable computational demand. Reducing the number of parameters is thus vital for realizing a computationally efficient network.

The computational efficiency may be further enhanced by making use of optimal combinations of image as well non-image features recognitions related to desired behavioral factors. For instance, Shahverdy et al. [9] proposed that a distracted behavior may be related to non-image vehicle motion features such as acceleration, throttle, speed, and engine rpm (revolutions per minute) Shahverdy et al. [10] further introduced a lightweight one-dimensional convolutional neural network (CNN) based on vehicle motion signals to classify driver behaviors in a computationally efficient manner. A number of studies have employed multi modal physiological signals [11], driving performance indicators [12], and eye movements [13] to detect negative emotions and behaviors of drivers. For instance, an algorithm to by using Changes in electroencephalogram (EEG) bands have been used to detect level of driver fatigue [14]. In addition, Lin et al. [15] proposed an EEG-based driver fatigue estimator on the basis of a linear regression model to predict alertness level of the drivers. The model was formulated using digital band power spectrum, correlation and principal component analyses of the EEG signals. Lethaus and Rataj [16] utilized the eye movement features to predict driver behavior in view of anticipated actions.

Advances in image processing techniques such as machine vision can facilitate identifications of negative emotions and abnormal driving behavior in an efficient manner. Ucar and Oguchi [17] used the following distance together with path deviation to detect a distracted driving behavior. Zheng et al. [18] proposed an improved CornerNet Sade scheme to detect driver distractions caused by smoking or eating while driving. A machine vision based distracted driving behavior detection method via a fast Region-CNN (R-CNN) model was developed in [19]. The key driving behavior features were analyzed using the class activation mapping method.

Real-time object detection network, 'you only look once' (YOLO) [20] is considered as a major breakthrough in the object detection regime that solved the object detection as a simple regression problem. The network is many times faster than the popular two-stage detectors like Faster-RCNN, while it compromises the accuracy. A number of alternate YOLO networks with different architectures have evolved to improve detection accuracy. For instance, Qin et al. [21] developed a method, called ID-YOLO, to judge the distracted driving behavior by detecting the primary object being observed by the driver. Hnewa and Radha [22] developed an integrated multiscale domain adaptive YOLO (MS-DAYOLO) framework for real-time object detection in a highly efficient manner. The proposed network solved the domain shift problem encountered by many deep learning applications at a substantially faster rate. The deep learning models employing YOLOv3-tiny, YOLOv3-tiny-3l, YOLO-fastest, YOLO-fastest-xl have been reported in [23] for detecting distractions such as head turning, drowsiness, eating, and phone usage. The algorithm performed at a rate of 30 frames/s when implemented on the NVIDIA GPU-based embedded platform. The relative real-time performance of different YOLO object detection models, including YOLOv5, YOLOv6, YOLOv7, and YOLOv8 have been reported in a number of studies [24], [25], [26]. These have shown superior performance of YOLOv8 compared to the other models in terms of accuracy and efficiency, while it required relatively fewer parameters. The studies have also emphasized the importance of considering requirements of specific tasks when selecting a model.

This study proposes a methodology for detecting emotions and distracted driving behavior using an improved YOLOv8 network. A multi-head self-attention (MHSA) module [27], [28], [29] together with convolutional neural network (CNN) module [30], [31], [32] is synthesized to detect distracted driving behavior and driver's emotions. In addition, the improved YOLOv8 network is optimized using TensorRT [33] and DeepStream [34] methods for deployment on the Jetson Nano. The main contributions of this paper include: (i) YOLOv8 network integrating the MHSA module is presented so as to minimize information loss and enhance accuracy of behavior recognition; (ii) a methodology for accurate detections of driver's emotions is developed by integrating the CNN coupled with the region of interest (ROI) in the YOLOv8 network; and (iii) the proposed improved YOLOv8 is optimized for speed and minimal number of parameters for implementation on the Jetson. Nano platform.

The remaining paper is organized as follows. The distracted driving behavior and driver's emotions detection methods based on an improved YOLOv8 are presented in Section II. The experimental implementations, methods and results obtained from the proposed network are presented and discussed in Section III. Verification of the proposed scheme is presented in Section IV considering the experimental, and the major findings are briefly summarized in Section V.

II. DISTRACTED DRIVING AND DRIVER'S EMOTION DETECTION

Figure 1 illustrates the framework developed for real-time identifications of distracted driving behavior and driver's emotions. The framework is based on an improved YOLOv8, which is realized by integrating MHSA (multi-head self-attention) and CNN modules to enhance detection accuracy and computational efficiency. The framework is a lightweight detection algorithm for real-time detections. The framework is based on distractions caused by drinking, smoking and phone usage, in addition to various emotions reflecting anger, fear, disgust, etc. Figure 4 shows the integration of MHSA and CNN modules into the YOLOv8 network, which is presented as the improved YOLOv8 in the detection framework in Fig. 2. The annotation files were built using the dataset constructed according to the distracted driving behavior and driver's emotions information. The labeled datasets were subsequently preprocessed and iteratively trained. The detailed methods used to build the framework are described below.

A. IMPROVED YOLOv8 MODEL

1) ATTENTION MECHANISMS MODULE

Owing to relatively large number of parameters associated with the YOLOv8, the accuracy as well speed of detections has been of concern [35]. In this study, the MHSA module is integrated into the backbone network of YOLOv8 to realize real-time detections. MHSA is a multi-head self-attention mechanism used to model dependencies at different locations in an input sequence. It divides the input sequence into multiple heads and computes the attention weighting of each head to yield the final weighted output. Each head possesses its own query, key, and value, which are derived from the input sequence through linear transformations (usually a full-join layer). The attention weights are subsequently used obtain the two weighted sum leading to output. You can see how MHSA works in Fig.1. Furthermore, Figure 3 describes a summary of the MHSA process.

$$b_r = \sum_{i=1}^{T} \widehat{a}_{T,i} \cdot v_i \tag{1}$$

$$\widehat{\alpha}_{T,i} = softmax(\alpha_{T,i} \tag{2}$$

$$\alpha_{T,j} = \frac{q_T^{Tranpose} \cdot k_i}{\sqrt{d_{q,k}}} \tag{3}$$

$$q_i = W^Q a_i \tag{4}$$

$$k_i = W^K a_i \tag{5}$$

$$v_i = W^V a_i \tag{6}$$

$$a_i = W x_i \tag{7}$$

The input to MHA consists of three vectors: a query vector, a key vector and a value vector. For a given query vector, the MHA weights and sums the key vectors, the weights are calculated from the similarity between the query and key vectors, and the resulting weighted sum is multiplied by the value vector for output.

In the MHSA design, each head focuses on a particular input within the sequence, and it represents complex functions. The similarities among different head are thus generally obtained from the dot products or bilinear relations to yield:

$$\text{MultiHead}(Q, K, V) = Concat(\text{head}_i, \ldots, \text{head}_h)W^o \tag{8}$$

where Q, K and V represent the query, key and value vectors, respectively. The transformation matrix W^o is used to transform the output of each head (head_i; $i = 1,..,h$). The self-attention mechanism is employed to compute head_i by an *Attention* function, such that:

$$\text{head}_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{9}$$

where QW_i^Q, KW_i^K and VW_i^V are the query, key and value transformation matrices of the i^{th} head, respectively. The *Attention* in MHA (multihead attention) is computed from the self-attentive mechanism, as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{10}$$

where d_k is the dimension of the key vector and superscrip 'T' represents the matrix transpose. The *softmax* similarity, normalized by the dimension, yields the weight of each key vector. The computed key weights applied to the value vector are summed to yield *Attention* output.

The MHSA layer is inserted after the convolutional layer of the BackBone of YOLOv8 (Fig. 4), so as to ensure that the input to the backbone network is correctly passed to the MHSA layer, and the output of the MHSA layer is correctly passed to subsequent layer. The forward propagation in YOLOv8's transmits the input to the MHSA layer at the appropriate location through the backbone network.

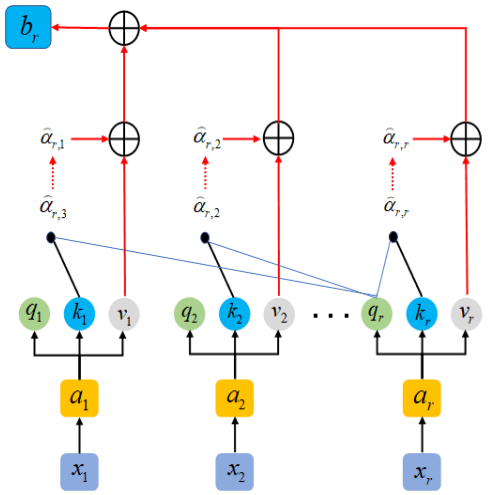


FIGURE 1. Structure of the MHSA attention mechanism.

2) EMOTION DETECTION USING CNN MODULE

The convolutional neural network (CNN) possesses many advantages over YOLO in feature learning, local feature capture, data expansion, pre-training, and adaptability to classification tasks [36]. In facial emotion detection, YOLO

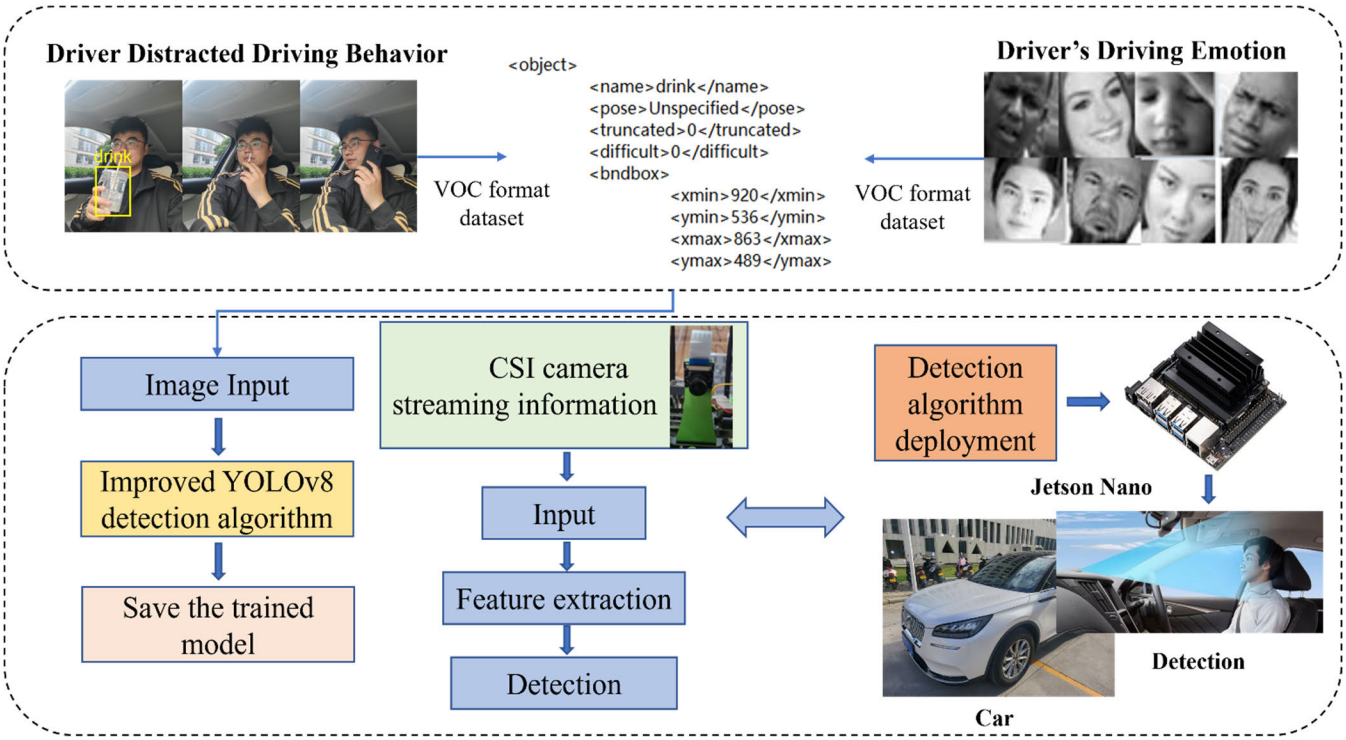


FIGURE 2. Framework designed for identifying distracted driving and driver's emotions.

(You Only Look Once) is a target detection framework belonging to convolutional neural networks (CNN), whereas facial expression detection focuses on recognizing and classifying emotions conveyed through facial expressions. In contrast, YOLO is designed to locate and classify objects within an image by dividing it into a grid and predicting bounding boxes and class probabilities for each grid cell. Although both involve CNNs, their goals and outputs differ. The CNN architecture for facial expression detection is likely optimized for facial feature extraction and emotion classification, while YOLO is geared towards detecting various objects in an image. The advantages of CNN include its ability to automatically learn feature representations, model spatial relationships in images, and perform end-to-end training on large-scale datasets, making it highly effective in tackling complex visual tasks. CNN is considered especially suitable for tasks requiring higher sensitivity and fine-grained feature extractions such as facial emotion detections with greater classification accuracy.

It should be noted that specific feature graphs are used within each layer of YOLOv8 network to perform target detections. Furthermore, each layer is responsible for detecting targets of a specific size and aspect ratio. For instance, one detection layer may be used for detecting small objects, while another for large objects. The driving emotion detection module, developed using CNN, is thus inserted in a YOLOv8 layer in order to discriminate among different emotions. Combining the CNN-based sentiment classification model with YOLOv8 involves two steps. Firstly, sentiment analysis is conducted using textual data, with an independent training of the CNN model for sentiment analysis. Subsequently, YOLOv8 is applied for object detection on images or videos, identifying various objects within the scene. If emotions are associated with specific regions or objects in the images (e.g., emotions expressed in text within the image), the sentiment predictions from the CNN model are then linked to the detected objects in the image, associating sentiment labels with specific regions or objects, such as emotions expressed in detected text. Finally, the results from both models are integrated, providing a comprehensive understanding of the content by combining sentiment information with detected objects. As shown in Fig.4, the CNN convolution module is inserted into the head detection head of YOLOv8 to form an emotion detection layer, and the ROI region in the object detection algorithm can be used efficiently. YOLOv8 effectively shares ROI so that the same feature map can be shared among different detection layers, including the emotion detection layer. These sharing permits realize real-time detection of driver's emotions.

The CNN network considered for emotions detection comprises of several layers, as shown in Fig. 5. These include: (i)an input layer that receives 48 $\times$ 48 pixels grayscale facial images; and (ii) a series of convolution layers consisting of 32 and 64 filter layers of 3 $\times$ 3 convolutions, and 128 spiral filter layers of 3 $\times$ 3 convolutions. Considering the nonlinearity in the network, each convolutional layer is activated by the ReLU function. The feature graph dimension following each convolution layer is further reduced by a maximum pool layer

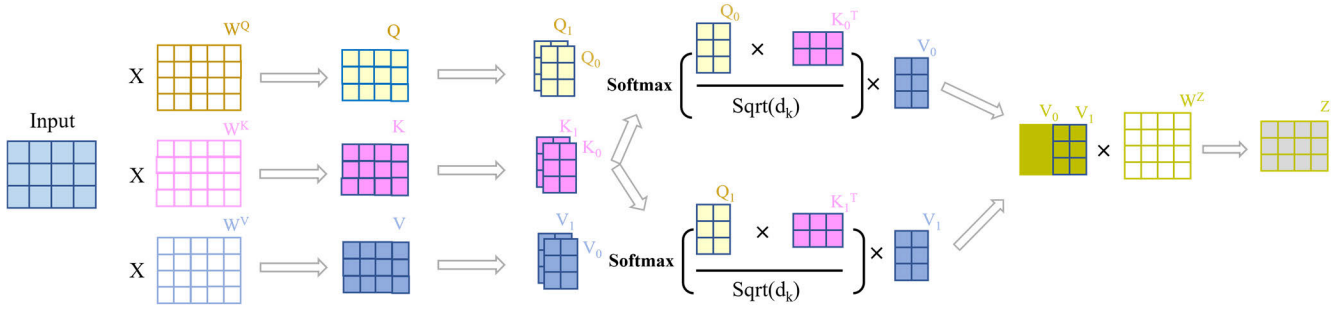


FIGURE 3. Summary of MHSA workflow.

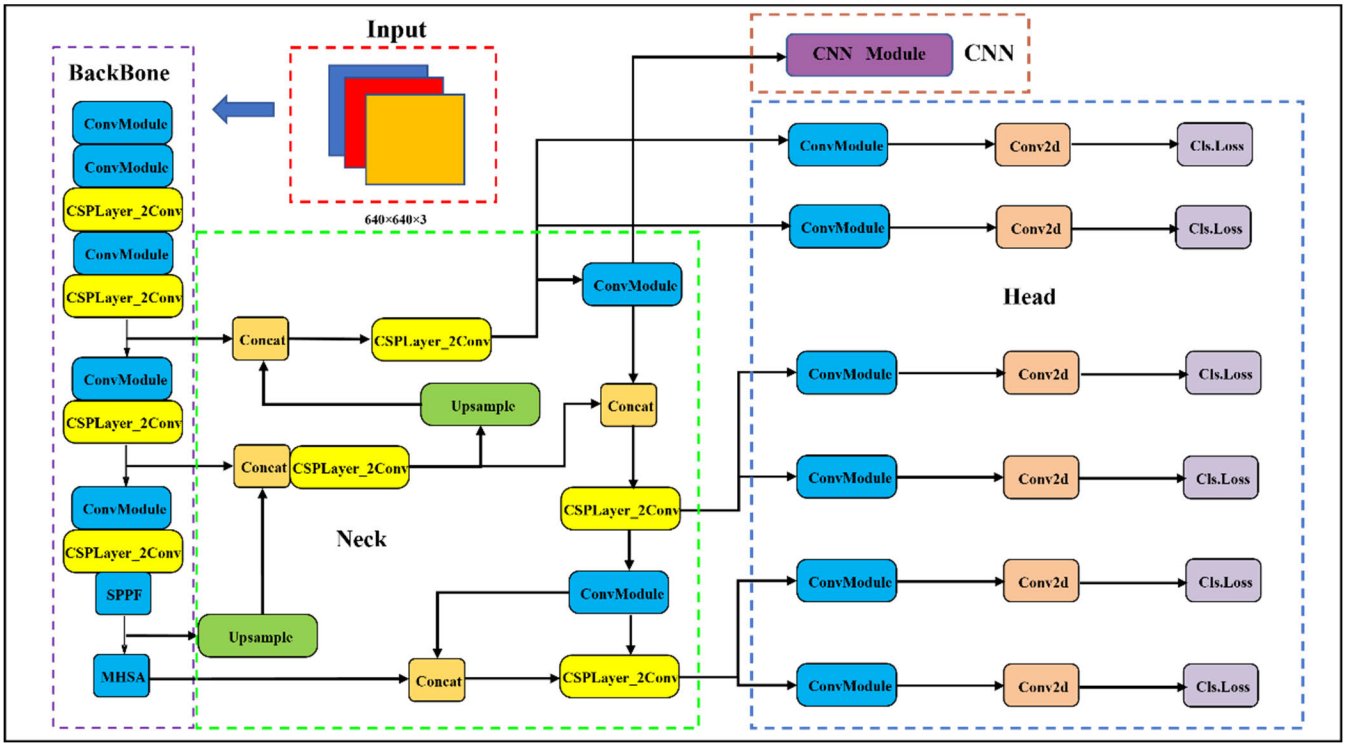


FIGURE 4. Improved YOLOv8 network integrating multi-head self-attention (MHSA) and convolutional neural network (CNN).

with a pool size of 2 × 2. The reduced feature map is subsequently flattened to 1-D vectors, which are fed into the fully connected layer consisting of 64 units with ReLU activations. The layer is subject to regularization at a rate of 0.5 to avoid overfitting. Finally, the softmax activation function generates the probability of a class of emotion through classification of the fully connected layer with the number of units equal to the number of classes (7: anger, disgust, fear, happy, sad, surprised, normal). The model uses the cross-entropy loss function for multi-class classifications and is optimized using the Adam optimizer [37].

III. EXPERIMENT DESIGN AND DATA ANALYSIS

A. EXPERIMENTAL PLATFORM

The experimental platform, designed in the study, consists of two stages. In the first stage, the hardware and software were developed for identifying distracted driving behavior and driver's emotions, as described in section II. The second stage involved deployment of, the algorithm in the Jetson Nano developer hardware, which included camera serial interface (CSI). Table 1 summarizes the configurations of the designed experimental platform in a laboratory environment.

B. DATASET CONSTRUCTION

1) DISTRACTED DRIVING BEHAVIOR DATASET

The data collections were performed in the laboratory using the video stream and photo information from an i-phone. A total of 20 subjects participated during data collection for the distracted driving scenarios (drinking, smoking and phone usage). The image data corresponding to each scenario were annotated using software to generate XML (extensible

markup language) files as described in [38]. As an example, Fig. 6 illustrates the images and XML file generated using LabelImg. Each file contains location coordinates and scenario classification. Subsequently, a dataset containing the distracted driving behavior was established, for the purpose of training.

2) DRIVER'S EMOTION DATASET

The dataset relevant to chosen emotions was taken from [39]. The study reported a comprehensive facial emotion recognition (FER) dataset to train and evaluate emotion identification algorithms. The dataset contains a diverse collection of different human facial expressions images, which are regarded as indicators of various human emotional states. The images in the FER dataset were labeled considering seven basic emotions: anger, disgust, fear, happiness, neutrality, sadness, and surprise. It includes still images and frames extracted from video sequences, and provides a comprehensive representation of emotions in a variety of contexts.

3) DATASET INTEGRATION

The datasets established for distracted driving behavior and driver's emotion were integrated to build a single compressive dataset containing 10 subsets, as seen in Table 2 together with the number of images (sample size) for each classification.

C. MODEL TRAINING AND RESULTS

The integrated dataset was divided into training and a test subset considering 9:1 proportion. A unity value of the pre-training weight was used so as to test the effectiveness of the improved YOLOv8 network. The SGD (stochastic gradient descent) optimizer was used to adjust the learning rate of each training parameter, while the learning rate decline mode was set to *cos*. The hyperparameter method was used to optimize training parameters for simultaneous training of different layers. The initial learning rate was taken as 0.01 with 640 × 320 pixels input image size and the batch size of 16. Furthermore, multi-threaded data reading was used to enhance training speed. The training model was compiled using the *Adam* optimizer together with a cross-entropy loss function for classification. The optimizer permitted updating of the weights during training. The Epoch was set to 300 and 220 for the driving distractions and driver's emotion data subsets, respectively, and the corresponding weight files were saved.

Training results are shown in Fig.7 in terms of the loss values as a function of the epochs. The training loss for the distracted driving detection includes box, cls, dfl losses in addition to the total loss, as seen in Fig. 7(a). The individual as well total losses decline rapidly with increasing epoch, suggesting convergence of the improved YOLOv8. Training loss of the emotion data subset also decreases rapidly with increasing number of iterations, as seen in Fig.7 (b), which shows the convergence of the facial emotion recognition method using the proposed CNN module.

D. QUANTIZATION AND DEPLOYMENT OF MODULES

Quantization is a process of approximating the continuous values of a signal to a finite number of discrete values and is essential for optimizing models for reducing the computing demand. In this experiment, the INT8 precision of the TensorRT is used to linearly quantify the improved YOLOv8 model. This further reduces the size of the model engine file. Fig. 8 shows the process of modeling quantification. The quantization process involves converting valid values and weights from the original FP32 format to the lower precision INT8 format using linear mapping.

In the quantization process, the convolution layer operation is performed to obtain the INT32-bit activation value. Then, this INT32 activation value is re-quantized back to the INT8 format as the input of the next layer. In the last layer of the network, inverse quantization is used to convert the activation value back to the original FP32 format.

$$X_{int} = clip(\left[\frac{X}{S}\right] + Z; -2^{b-1}, 2^{b-1} - 1) \tag{11}$$

The conversion from FP32 to INT8 format can be expressed using (11), where X represents the original FP32 value, Z represents the zero point of the mapping, S is the scale factor, [.] is a mathematical function for approximate rounding, rounding up, rounding down, etc.; X_{int} is an integer value after quantization. The rounding function can be approximate rounding, rounding up, rounding down, etc. Equation (12) shows the functions used for rounding.

$$clip(x; a, c) = \left\{ \begin{array}{ll} a, & \text{if } x < a, \\ x, & \text{if } a \le x \le c, \\ c, & \text{if } x > c. \end{array} \right\} \tag{12}$$

$$X = S(X - Z) \tag{13}$$

$$X = S(clip\left[\frac{X}{S}\right] + Z; -2^{b-1}, 2^{b-1} - 1) - Z) \tag{14}$$

when $Z = 0, X_{\min} = -2^{b-1}S, X_{\max} = (2^{b-1} - 1)S$.

Notably, equations (13) and (14) can be used to implement the inverse quantization process that converts the quantized value back to its original FP32 format. By applying linear quantization with INT8 precision, the size of the model engine file is reduced while still maintaining acceptable accuracy. This allows for higher peak performance on deployed hardware and introduces minimal additional computational overhead.

Several different types of behaviors and emotions were tested using the trained model, and the test results are shown in Fig.9. As can be seen that the driver distracted driving behavior and driver's emotion can be accurately detected.

IV. ANALYSISI OF RESULTS

A. COMPARISON MODEL PARAMETERS

On the same experimental platform, the improved algorithm model was compared with several YOLO series models. The results for the number of parameters and the size of the weight

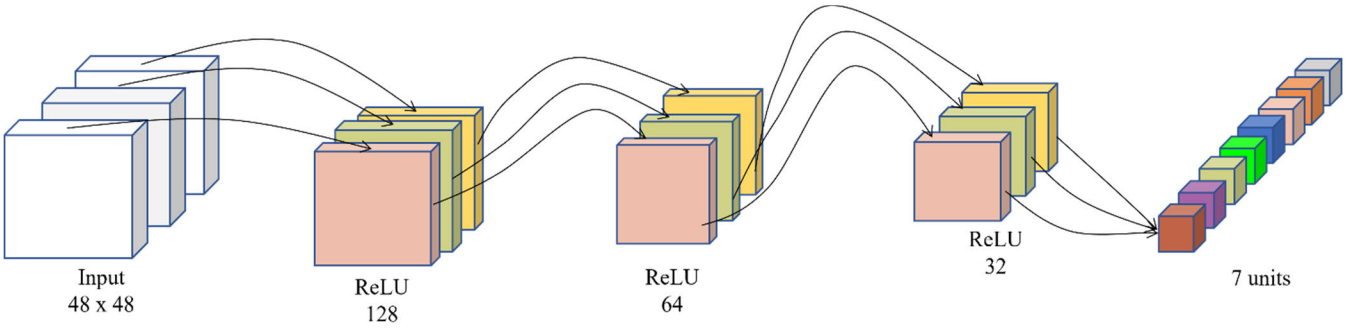


FIGURE 5. Layered structure of CNN module using for emotion recognition.

TABLE 1. Configuration of the designed experimental platform.

Stage	Related configurations	
1. Development of improved YOLOv8 network	Software	Pycharm
	Programing language	Python
	CPU	I7-13700K
	GPU	NVIDIA RTX 4070Ti
	Running memory	16 GB
	Deep learning framework	Pytorch
2. Deployment of improved YOLOv8 network on the Jetson-Nano hardware	Hardware	Jetson Nano, CSI camera
	Programing language	Python' C
	CPU	4-core ARM A57
	GPU	128-core Maxwell
	Running memory	4GB

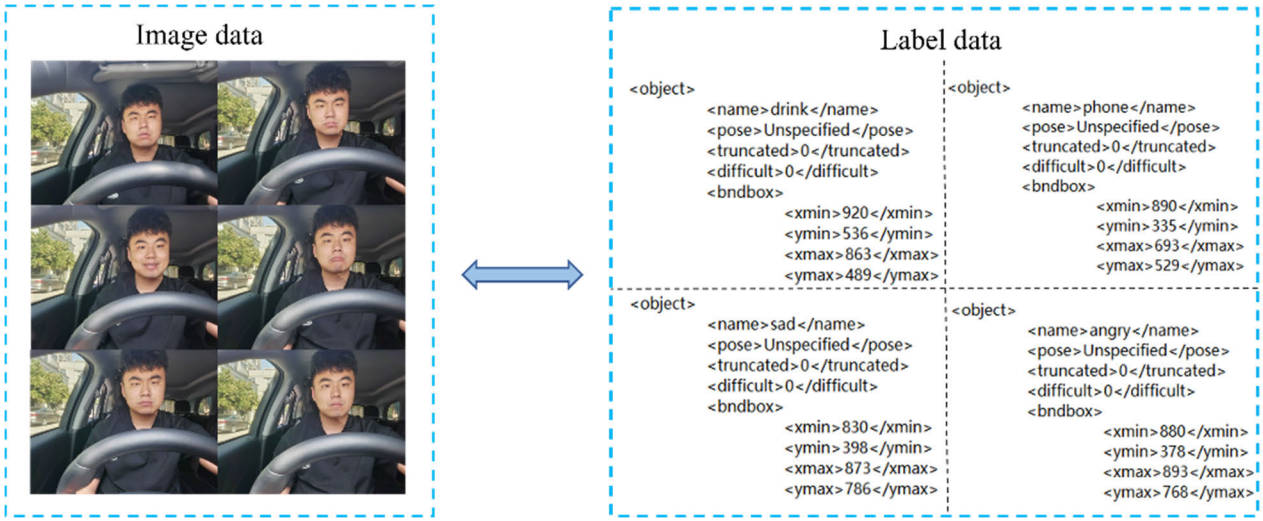
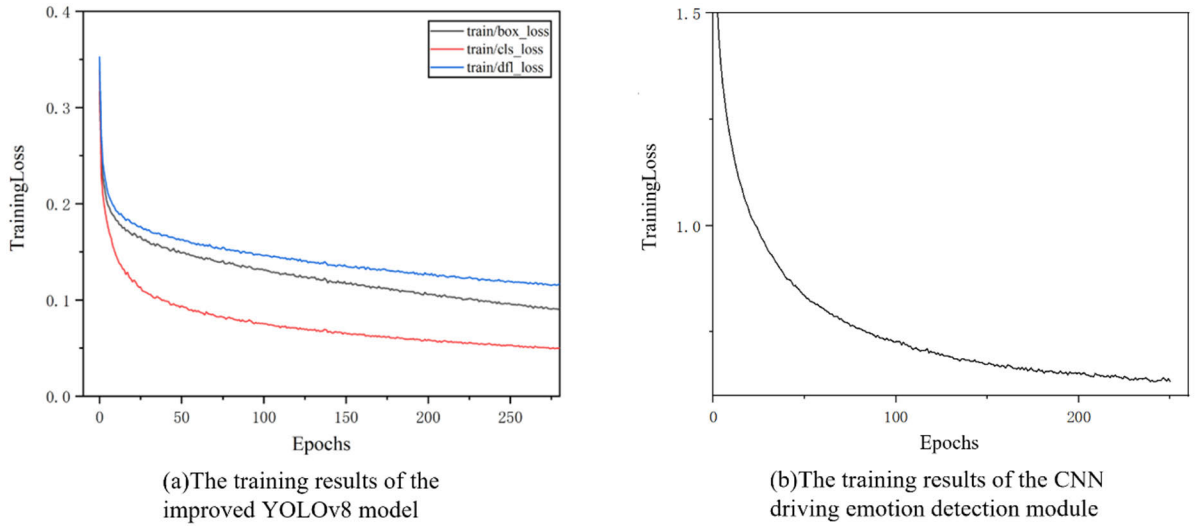


FIGURE 6. An example image and the resulting XML file generated by LabelImag doftware.

files are shown in table 5. As can be seen from the table, model parameters and weight parameters of the improved network model are lower than those of YOLO series models. In terms of model parameters, the maximum reduction is 79.11%, and the lowest is 62.15%. In terms of weight parameters, the maximum reduction is 79.33% and the lowest is 62.47%. In summary, the number of parameters and weight parameters of the improved model are lower than those of YOLO series models and a lightweight network is realized.

B. COMPARISON OF MODEL DETECTION ACCURACY

In order to verify the accuracy and effectiveness of the proposed model, accuracy (Precision), recall rate (Recall) and F1 value are used as measurement indicators. Specific



(a)The training results of the improved YOLOv8 model

(b)The training results of the CNN driving emotion detection module

FIGURE 7. The training loss vs epochs: (a) distracted driving data subset; and (b) emotion recognition data subset.

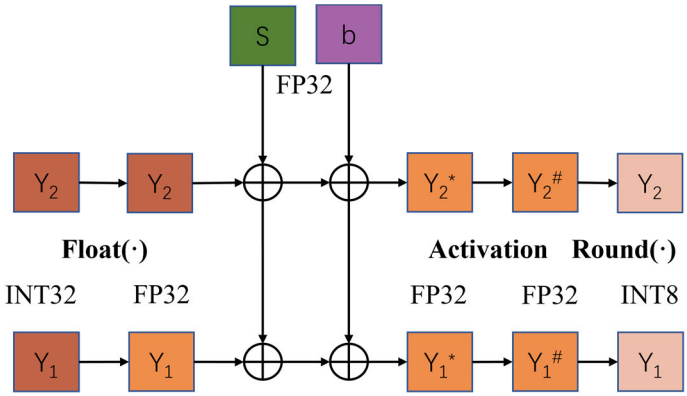


FIGURE 8. INT8 convolutional layer inference flow.

TABLE 2. Subsets of the overall dataset and the sample sizes.

Classification	Category	Sample size
Anger	emotion	710
Disgust	emotion	736
Fear	emotion	798
Happy	emotion	830
Sad	emotion	810
Surprised	emotion	799
Normal	emotion	870
Drink	distracted driving	920
Phone	distracted driving	855
Smoke	distracted driving	875

calculation methods are expressed as

$$\text{precision}(P) = \frac{TP}{TP + FP} \tag{15}$$

$$\text{recall}(R) = \frac{TP}{TP + FN} \tag{16}$$

$$F_1 - \text{score} = \frac{2P * R}{P + R} \tag{17}$$

where P is the accuracy rate, R is the recall rate, F_1 is the harmonic mean of the accuracy rate and the recall rate. TP indicates that positive samples are correctly detected, FP indicates false detection, FN indicates missed detection and TN indicates that negative samples are correctly detected. Specific representation is shown in Table 3.

AP_i is the area enclosed by the $P - R$ curves of single-category data, such that

$$AP_i = \int_0^1 P_i(R_i) dR_i \tag{18}$$

To verify the effect of using the MHSA attention mechanism module to improve accuracy, the following comparative experiments were conducted using the same network to train and validate the YOLOv8 before and after using the MHSA attention mechanism module. The results are shown

TABLE 3. Expression of specific indicators.

Label		Real result	
		Positive sample	Negative sample
Prediction result	Positive sample	TP	FN
	Negative sample	FP	TN

TABLE 4. MHSA attention mechanism module before and after use comparison table.

Serial number	Model	MHSA	mAP
1	Original YOLOv8	×	0.757
2	Improved YOLOv8	√	0.814

FIGURE 9. Detection results.

TABLE 5. Comparison table of the number of parameters and weight file size.

Serial number	Model	Model parameters	Weight parameters
1	YOLOv3	52.2 MB	235.1 MB
2	YOLOv4	28.8 MB	129.5 MB
3	YOLOv5	38.5 MB	179.8 MB
4	YOLOv7	43.5 MB	156.4 MB
5	Original YOLOv8	49.8 MB	198.2 MB
6	ImprovedYOLOv8	10.9 MB	48.6 MB

in Table 4. $\surd$ means that the MHSA attention mechanism module is used, $\times$ means that the MHSA attention mechanism module is not used.

The mean Average Precision (mAP) represents the average of all APs and the average detection accuracy of all categories of objects, which is defined by

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{19}$$

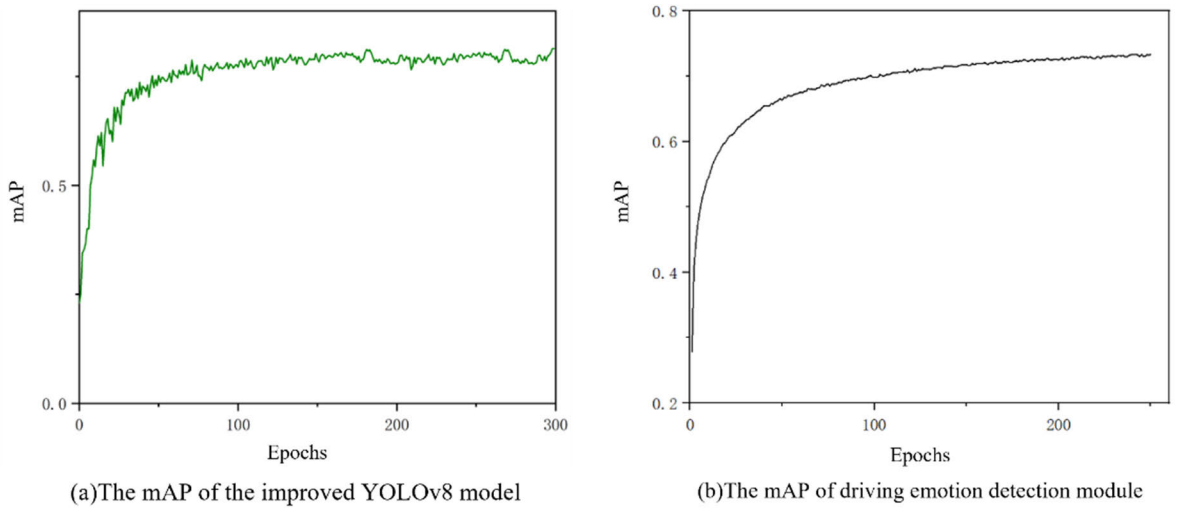


FIGURE 10. The mAP of the driver distracted driving behavior and driving emotion method.

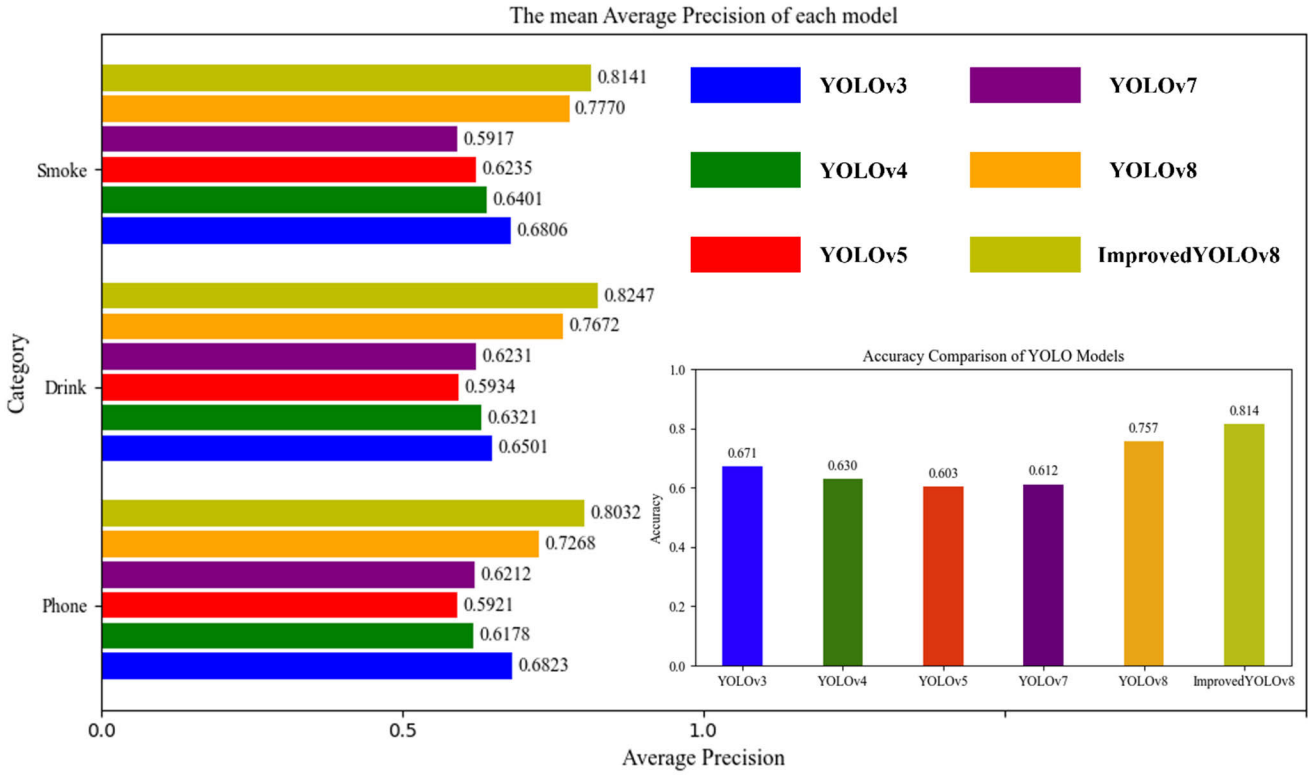


FIGURE 11. A comparison of the results of multiple detection algorithms.

TABLE 6. Comparison of FPS.

Serial number	Model	FPS
1	Original YOLOv8	20
2	Improved YOLOv8	25

In (19), the *mAP* (Mean Average Precision) is the combined weighted average of all values in (11), and the guaranteed accuracy is obtained by combining all values in (18).

During the training process, the mixed data sets of distracted driving behavior and driver's emotions were iterated 300 times and 220 times on the same hardware, respectively. The *m*A*P* of distracted driving behavior and driver's emotion model is shown in Fig.9. It can be seen that the accuracy of distracted driving behavior and driver's emotion finally reached 0.814 and 0.733, respectively, indicating that the model can accurately classify distracted driving behavior and driver's emotion.

In addition, Figure 11 shows the *m*A*P* of the improved YOLOv8, YOLOv3, YOLOv4, YOLO5s, YOLOv7 and the original YOLOv8 under the same conditions. We can see that the accuracy of the improved YOLOv8 is higher than other detection algorithms. At the same time, Table 6 compares the FPS of the original model with that of the improved model, and it can be intuitively seen that the detection speed has increased from 20 (FPS) of the original models to 25FPS.

To sum up, compared with other detection algorithms, our improved detection algorithm not only ensures the detection accuracy, but also makes the model lighter and has higher real-time performance. This meets the industry's need for real-time performance and accuracy. It offers the possibility of safe driver assistance deployment.

V. CONCLUSION

In this paper, we propose an improved algorithm based on YOLOv8 to detect distracted driving behavior and driver's emotion. Three conclusions can be drawn from the training and experimental results of the proposed improved YOLOv8 algorithm: (1) The proposed algorithm adopts the MHSA (Multi-Head Self-Attention) structure, that is, the multi-head attention-self-attention mechanism module is inserted into the fully connected layer of YOLOv8. Compared with the existing attention mechanism module methods, the proposed method has higher accuracy. (2) Compared with the existing method of simultaneously detecting the distracted driving behavior and driver's emotion with a single detector, the proposed CNN Convolutional neural network module insertion method has model-free and adaptive characteristics, and shows improved performance in terms of convergence speed and accuracy. (3) The proposed improved YOLOv8 method was deployed on the Jetson Nano platform, where TensorRT and DeepStream methods performed well in terms of model volume and computing speed. In our future work, we will focus on designing driver detection models that consider more applicable scenarios.

REFERENCES

[1] J. J. Rolison, S. Regev, S. Moutari, and A. Feeney, "What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records," *Accident Anal. Prevention*, vol. 115, pp. 11–24, Jun. 2018.

[2] A. Iranitalab, A. Khattak, and G. Bahouth, "Statistical modeling of cargo tank truck crashes: Rollover and release of hazardous materials," *J. Saf. Res.*, vol. 74, pp. 71–79, Sep. 2020.

[3] Y. Zhang, L. Jing, C. Sun, J. Fang, and Y. Feng, "Human factors related to major road traffic accidents in China," *Traffic Injury Prevention*, vol. 20, no. 8, pp. 796–800, 2019.

[4] E. Petridou and M. Moustaki, "Human factors in the causation of road traffic crashes," *Eur. J. Epidemiol.*, vol. 16, pp. 819–826, Jan. 2000.

[5] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3017–3032, Dec. 2015.

[6] C. Marina Martinez, M. Heucke, F.-Y. Wang, B. Gao, and D. Cao, "Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 666–676, Mar. 2018.

[7] H. Hasrouny, A. E. Samhat, C. Bassil, and A. Laouiti, "VANet security challenges and solutions: A survey," *Veh. Commun.*, vol. 7, pp. 7–20, Jan. 2017.

[8] S. J. Al-Sultan, "Context aware drivers' behaviour detection system for VANET," Ph.D. dissertation, Softw. Technol. Res. Lab., De Montfort Univ., Leicester, U.K., 2013.

[9] M. Shahverdy, M. Fathy, R. Berangi, and M. Sabokrou, "Driver behavior detection and classification using deep convolutional neural networks," *Expert Syst. Appl.*, vol. 149, Jul. 2020, Art. no. 113240.

[10] M. Shahverdy, M. Fathy, R. Berangi, and M. Sabokrou, "Driver behaviour detection using 1D convolutional neural networks," *Electron. Lett.*, vol. 57, no. 3, pp. 119–122, Feb. 2021.

[11] C. Hieida, T. Yamamoto, T. Kubo, J. Yoshimoto, and K. Ikeda, "Negative emotion recognition using multimodal physiological signals for advanced driver assistance systems," *Artif. Life Robot.*, vol. 28, no. 2, pp. 388–393, May 2023.

[12] T.-C. Lin, S. Ji, C. E. Dickerson, and D. Battersby, "Coordinated control architecture for motion management in ADAS systems," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 2, pp. 432–444, Mar. 2018.

[13] J. A. Renshaw, J. E. Finlay, D. Tyfa, and R. D. Ward, "Understanding visual influence in graph design through temporal and spatial eye movement characteristics," *Interacting Comput.*, vol. 16, no. 3, pp. 557–578, Jun. 2004.

[14] S. K. L. Lal, A. Craig, P. Boord, L. Kirkup, and H. Nguyen, "Development of an algorithm for an EEG-based driver fatigue countermeasure," *J. Saf. Res.*, vol. 34, no. 3, pp. 321–328, Aug. 2003.

[15] C.-T. Lin, R.-C. Wu, T.-P. Jung, S.-F. Liang, and T.-Y. Huang, "Estimating driving performance based on EEG spectrum analysis," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 19, pp. 1–10, Dec. 2005.

[16] F. Lethaus and J. Rataj, "Do eye movements reflect driving manoeuvres?" *IET Intell. Transp. Syst.*, vol. 1, no. 3, p. 199, 2007.

[17] S. Ucar and K. Oguchi, "Demo: Distracted driving behavior detection to avoid rear-end collisions," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Nov. 2021, pp. 115–116.

[18] W. Zheng, Q. Q. Zhang, Z. H. Ni, Z. G. Ye, Y. M. Hu, and Z. J. Zhu, "Distracted driving behavior detection and identification based on improved cornernet-saccade," in *Proc. IEEE Int. Conf. Parallel Distrib. Process. Appl., Big Data Cloud Comput., Sustain. Comput. Commun., Social Comput. Netw.*, Dec. 2020, pp. 1150–1155.

[19] J. Wang, Z. Wu, F. Li, and J. Zhang, "A data augmentation approach to distracted driving detection," *Future Internet*, vol. 13, no. 1, p. 1, Dec. 2020.

[20] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022.

[21] L. Qin, Y. Shi, Y. He, J. Zhang, X. Zhang, Y. Li, T. Deng, and H. Yan, "ID-YOLO: Real-time salient object detection based on the driver's fixation region," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15898–15908, Sep. 2022.

[22] M. Hnewa and H. Radha, "Integrated multiscale domain adaptive YOLO," *IEEE Trans. Image Process.*, vol. 32, pp. 1857–1867, 2023.

[23] Y.-S. Poon, C.-Y. Kao, Y.-K. Wang, C.-C. Hsiao, M.-Y. Hung, Y.-C. Wang, and C.-P. Fan, "Driver distracted behavior detection technology with YOLO-based deep learning networks," in *Proc. IEEE Int. Symp. Product Compliance Eng.-Asia*, Nov. 2021, pp. 1–5.

[24] M. Sozzi, S. Cantalamessa, A. Cogato, A. Kayad, and F. Marinello, "Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms," *Agronomy*, vol. 12, no. 2, p. 319, Jan. 2022.

[25] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved YOLO-V3 model," *Comput. Electron. Agricult.*, vol. 157, pp. 417–426, Feb. 2019.

[26] L. Huang, Q. Fu, M. He, D. Jiang, and Z. Hao, "Detection algorithm of safety helmet wearing based on deep learning," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 13, p. e6234, Jul. 2021.

[27] Q. Guo, J. Huang, N. Xiong, and P. Wang, "MS-pointer network: Abstractive text summary based on multi-head self-attention," *IEEE Access*, vol. 7, pp. 138603–138613, 2019.
[28] X.-B. Jin, W.-Z. Zheng, J.-L. Kong, X.-Y. Wang, M. Zuo, Q.-C. Zhang, and S. Lin, "Deep-learning temporal predictor via bidirectional self-attentive encoder–decoder framework for IoT-based environmental sensing in intelligent greenhouse," *Agriculture*, vol. 11, no. 8, p. 802, Aug. 2021.
[29] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, and B. Zheng, "Age-invariant face recognition by multi-feature fusionand decomposition with self-attention," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 1s, pp. 1–18, Feb. 2022.
[30] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.
[31] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.
[32] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," 2023, *arXiv:2304.00501*.
[33] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.
[34] Z. Zhang, "Improved Adam optimizer for deep neural networks," in *Proc. IEEE/ACM 26th Int. Symp. Quality Service (IWQoS)*, Jun. 2018, pp. 1–2.
[35] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, 2023.
[36] G. Akyol, A. Kantarcı, A. E. Çelik, and A. Cihan Ak, "Deep learning based, real-time object detection for autonomous driving," in *Proc. 28th Signal Process. Commun. Appl. Conf. (SIU)*, Oct. 2020, pp. 1–4.
[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
[38] C. Zhang, J. Cheng, L. Li, C. Li, and Q. Tian, "Object categorization using class-specific representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4528–4534, Sep. 2018.
[39] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, Jan. 2018.

BAO MA received the bachelor's degree in software engineering from Nanyang University of Technology, in 2021. He is currently pursuing the master's degree with the Mechanical and Electrical Engineering College, Zhengzhou University of Light Industry. His research interests include deep learning and embedded systems.

ZHIJUN FU (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of Science and Technology Beijing, China, in June 2013. From 2010 to 2012, he was a Joint Training Doctoral Student with the University of Concordia University, Canada. He is currently a Professor with the College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include vehicle intelligent control and neural network control.

SUBHASH RAKHEJA is currently a Professor of mechanical engineering with the Department of Mechanical and Industrial Engineering, Concordia University, Montreal. His research interests include advanced transportation systems and highway safety, human responses to workplace vibration, and driver-vehicle interactions. He is a fellow of the American Society of Mechanical Engineers (ASME) and the Canadian Society of Mechanical Engineers (CSME). He is the Concordia Research Chair of Vehicular Aerodynamics. He is serving as an Editor for the *International Journal of Industrial Ergonomics* and an Associate Editor for *SAE Journal of Commercial Vehicles* and *International Journal of Heavy Vehicle Systems*.

DENGFENG ZHAO received the Ph.D. degree in mechanical engineering from Jilin University, China, in 2003. He has been engaged in bus design, development, and technology research at Yutong Bus Company Ltd., for more than ten years. He is currently an Associate Professor with the College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include active safety control technology and comfort technology of electric vehicles.

WENBIN HE received the Ph.D. degree in mechanical engineering from Tsinghua University, in July 2014. He is currently a Professor with the College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, China. His research interests include vehicle intelligent control, mechanical design and optimization, and driving behavior safety.

WUYI MING was born in Hubei, China, in 1981. He received the bachelor's degree in engineering from Huazhong Agricultural University, in 2002, the master's degree in engineering from Zhengzhou University, in 2010, and the Ph.D. degree in engineering from Huazhong University of Science and Technology, in 2014. He is currently an Associate Professor with the School of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry. In addition, he also serves as the Deputy Director of the Guangdong Province Manufacturing Equipment Intelligent Engineering Technology Research Center. His main research direction is digital technology and equipment for dificult-to-process materials. He has done in theoretical and applied research in EDM and deep learning.

ZHIGANG ZHANG received the Ph.D. degree from the Department of Engineering Mechanics, Dalian University of Technology, Dalian, Liaoning, China, in December 2015. He is currently a Lecturer with the College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include multibody system dynamics and control, vehicle dynamics, and vehicle-induced whole-body vibration.