IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 July 2022, accepted 3 August 2022, date of publication 8 August 2022, date of current version 11 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3197146

RESEARCH ARTICLE

AB-DLM: An Improved Deep Learning Model Based on Attention Mechanism and BiFPN for Driver Distraction Behavior Detection

TAIGUO LI[1], YINGZHI ZHANG[1], QUANQIN LI[2], AND TIANCE ZHANG[1]
[1]School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
[2]Children's Rehabilitation Department, Shaanxi Kangfu Hospital, Xi'an 710065, China

Corresponding author: Taiguo Li (leetg@mail.lzjtu.cn)

This work was supported in part by the Gansu Provincial Science and Technology Program under Grant 21JR7RA303, and in part by the Youth Science Fund of Lanzhou Jiaotong University under Grant 2020002.

ABSTRACT Driver distraction behavior causes a large number of traffic accidents every year, resulting in economic losses and injuries. Currently, the driver still plays an important role in the driving and control of the vehicle due to the low level of vehicle automation and the immature development of autonomous driving. Therefore, it is vital to research distraction detection for drivers. However, in realistic driving scenarios with uncertain information, they are still some challenges in efficient and accurate driver distraction detection. In this paper, an improved deep learning model based on attention mechanisms and bi-directional feature pyramid networks (BiFPN) is proposed to identify driver distractions. Firstly, an improved data augmentation strategy is introduced to increase the data diversity to enhance the generalization capability of the model. Secondly, the squeeze-and-excitation (SE) attention mechanism layer is used after the C3 module of the original backbone network to enhance the important feature information and suppress the minor feature information. Finally, the BiFPN module is introduced into the neck network to better achieve multi-scale feature fusion without increasing the calculation amount too much. The experimental results show that the method proposed in this paper has an average mean accuracy rate (mAP) of 0.956 on the test set. Compared to the original model the mAP has improved by 13.2%. The detection speed of the model is 71 frames per second, and the memory occupation is 15.9 MB. This method has the advantages of high recognition accuracy, fast detection speed, and small memory occupation of the model, which are important for achieving engineering deployment.

INDEX TERMS Driver distraction, attention mechanism module, BiFPN module, deep convolutional neural network, driving behavior.

I. INTRODUCTION

According to a report published by the World Health Organization [1], approximately 1.35 million people worldwide are killed in traffic accidents each year. Approximately 20 to 50 million people will suffer serious disabilities. The report also states that distracted driving by drivers is one of the major causes of traffic accidents. Data from the National Highway Traffic Safety Administration (NHTSA) [2] also shows that distracted driving is blamed for 80% of road crashes and 16% of highway fatalities.

With the rapid development of artificial intelligence and vehicle automation, the automotive industry worldwide is gradually moving towards digitalization and autonomous driving [3]. However, so far, autonomous driving has not matured in terms of vehicle control and the level of automation is still not high. At the same time, fully autonomous driving (L5 level) has not yet been achieved at this stage. Therefore, the driver is still required to remain focused at all times while the vehicle is in motion. And the driver will be ready to operate the vehicle in case of an emergency [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

A study by the National Transportation Safety Board found that 37 of the crashes in Uber's self-driving test vehicles occurred while the vehicle was moving automatically. The accidents were caused by driver distraction behavior failing to take over the operation of the vehicle on time [5]. In the foreseeable future, both manual and automated driving will still require the driver to keep a close eye on the vehicle's driving status and thus ensure that it operates safely. Therefore, real-time detection of driver distraction is very important to improve transportation safety. Alerting the driver when a distraction is detected or activating a vibration at the seat to warn the driver, thus improving the driver's concentration [6]. This will provide ideas for improving the advanced driver assistance system (ADAS) [7]. At the same time, for the further development of the automobile in the direction of safety. The ultimate purpose of scientific research is also to reduce accidents and ensure the driving safety of people and vehicles on the road [8].

The NHTSA defines distracted driving as ''any activity that diverts attention from driving''. The Centers for Disease Control and Prevention (CDC) also defines distracted driving more broadly [9]. A driver is considered distracted when his or her attention is taken away from the driving task and focused on other activities. These distraction behaviors can be divided into three categories: cognitive distractions, manual distractions, and visual distractions.

Today, smartphones are a major source of distraction for drivers, and research into the detection of distractions such as driver use of mobile phones is now coming into focus. A study in The New England Journal of Medicine estimated the risk of crashing for different categories of distracting behavior [10]. The highest category is ''talking on a mobile phone'' and the second highest is ''holding an object other than a mobile phone''. Driver distraction has the most direct impact on the normal operation of a vehicle at the operational level. Some behaviors even include cognitive distractions and visual distractions that have a broader impact on a holistic level. Behavioral distractions are therefore extremely dangerous for safe driving. Therefore, monitoring drivers' behavioral attention is particularly important for transportation safety.

The detection of distracted driving can be classified into three categories: physiological signals [11], eye attention tracking [12], and computer vision [13]. Physiological signals will change when the driver performs different driving operations [14]. However, most detection methods based on physiological signals are not practical and feasible, because intrusive physiological acquisition sensors often compromise the driver's driving experience. The limitation of the eye attention tracking detection method is that the detection algorithm will ignore the complexity of the driving scene [15]. For example, when the driver passes a large intersection slowly or merges into the lane at the blind spot of vision, or when the vehicle turns around. These situations require the driver's eyes to constantly look left and right at the surrounding situation. This is likely to be judged as distracted driving due to not looking straight ahead for a long time.

The advantage of the driver distraction detection method based on computer vision is that it is a non-invasive detection method. It won't affect the driver's normal driving. How to extract valuable features from the images obtained by computer vision is very important for driver distraction detection [16]. Traditional machine learning methods have limitations when it comes to extracting features, and manual feature extraction is less efficient and less effective. Deep learning methods are highly efficient for feature extraction and also enable an effective fusion of feature information. It is also favored by researchers because of its high accuracy and speed of detection. This is why mainstream driver distraction detection is now gradually shifting to deep learning. However, the focus of most researchers is still on classifiers [17]. The driver distraction detection method based on classifiers feeds the whole image into the network, which results in feature redundancy and excessive calculation amount. This method needs more computation, but it makes little contribution to the improvement of model detection accuracy.

After recognizing the issues presented by the whole image input, researchers gradually try to use part of the image as input to train the network model. But they still didn't change the underlying method. Although these solutions solve the excessive calculation amount and feature redundancy, it introduces new problems. That is when local images are extracted, such as heads, hands, and arms, which may not contain discriminatory feature information. At the same time, they still face another problem. Most distractions are complex, involving multiple body parts and various object objects (e.g. water glasses, mobile phones, center console). Due to the complexity of the movements, multiple body parts are involved when distractions occur. A single body part is less relevant to other parts of the body as a local image input.

After reading relevant papers and extensive research, it is found that the YOLO model [18] transforms the problem of object detection into a regression problem. Direct grid-to-area mapping is achieved through meshing, reducing redundancy feature information and calculation amount. So try to use advanced object detection methods to solve the above-mentioned problems. After comprehensive consideration, the current mainstream YOLOv5s model is selected as the base model. Improvements and optimizations are made based on the YOLOv5s model structure. The whole image input will lead to feature redundancy, and the local picture input may lead to non-discriminative feature information. Therefore, the attention mechanism module is introduced and added to the proposed model. This will more accurately select the key part of the discriminative information as the input. Also, using the BiFPN [19] module will better enable the fusion of feature information at multiple scales. Improve the detection accuracy to a greater extent. When making the dataset, all the objects associated with the distraction behavior will be selected as label data. In this way, the person's body part, the shape of the movement, and the objects at the time of the behavior occurring are all highly correlated.

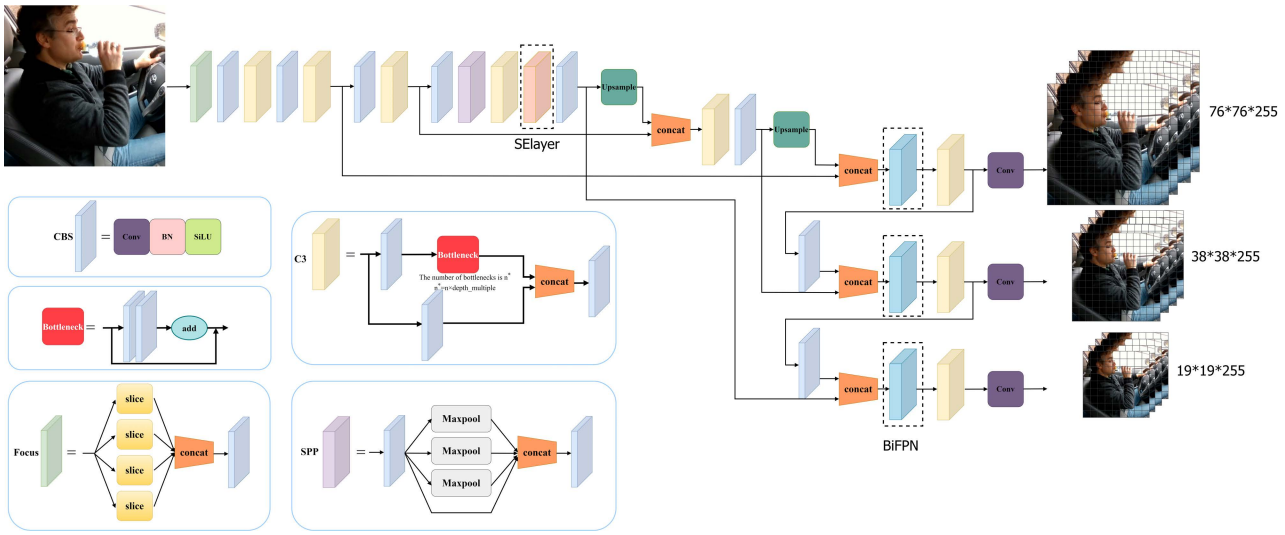


FIGURE 1. The pipeline of the proposed AB-DLM network.

The main contributions of our work are summarized below:

- This paper proposes a novel driver distraction detection method based on object detection. Compared with advanced classifier methods, the method proposed in this paper has certain advantages. By adding the attention mechanism, this method can well address the limitations of feature redundancy in existing work. At the same time, the BiFPN module is used to strengthen the exchange of feature information and improve the performance of the driver distraction detection model.
- This paper develops an improved deep learning model based on attention mechanism and BiFPN (AB-DLM). This model uses cspdarknet53 as the backbone network and adds the SE attention mechanism module to realize multi-channel information fusion. The BiFPN module is added to the neck of the model as a means of fusing multi-scale feature information. An improved data augmentation strategy is introduced to diversify the dataset and improve the detection performance and generalization capability of the model without increasing any calculation amount.
- In experiments, a comprehensive comparison is conducted between the proposed method (AB-DLM) and the SOTA lightweight detection networks including YOLOv4-tiny, PP-YOLO Tiny, YOLOX-S, etc. The results demonstrate that the AB-DLM model has better detection performance and lightweight deployment capability. Comparisons are also made with other SOTA classification methods to verify the advantage and feasibility of our work.

The pipeline of the AB-DLM network proposed in this paper is shown in Fig. 1

II. RELATED WORK

Driver distraction behavior detection (DDBD) based on computer vision has been widely developed due to the progress of feature extraction [22], [27], supervised learning [28], [29], deep learning [32], [34], and related research subfields such as object detection [18] and human behavior recognition [22], [23]. The DDBD can be categorized based on different technology into two categories: traditional machine learning based detection and deep learning based detection.

A. TRADITIONAL MACHINE LEARNING BASED DETECTION

In the early days, driver distraction behavior was detected through manually extracted features and traditional machine learning methods. At the same time, traditional machine learning methods are based on non-deep learning algorithms. Berri *et al.* [20] extracted features manually and used a support vector machine (SVM) model to detect distraction. At the same time, a genetic algorithm was used to optimize the super parameters. Their purpose was to detect the use of mobile phones in the front image of drivers. Similarly, Artan *et al.* [21] also employed SVM to achieve driver behavior detection using mobile phones. They differed from Berri by using a near-infrared camera system for the front windscreen area of the vehicle. Craye and Karray [22] first extracted four features from the driver by using a Kinect camera to obtain an image of the overall range of the driver, including head orientation, facial expression, eye gaze and closure, and arm position. The above features were constructed into a feature representation capable of assessing driver distraction, and then the distracting behavior was classified by an AdaBoost classifier. Zhang *et al.* [23] created a dataset that contained four categories of driving activities. These were: safe driving, manipulating the gearstick, talking on the phone, and eating. In addition, they evaluated the recognition accuracy of four different classifiers, and the accuracy of these models was 85.09%, 87.16%, 37.21%, and 90.63% respectively. Zhao *et al.* [24] also used this dataset [23] to extract features based on a right-angle gradient pyramid and used these features to train a multilayer

perceptron classifier. The final model was tested with an acc of 94.75%. Yan *et al.* [25] extracted a pyramidal histogram of gradients from the driver's motion history images based on the temporal nature of the data. A random forest classifier was applied to the extracted features with a classification accuracy of 96.56%. However, manual extraction of features takes a lot of time, requires specialist knowledge and achieves poor results.

B. DEEP LEARNING BASED DETECTION

In recent years, due to the development of computer configuration, a favorable environment has been established for the advancement of deep learning methods. The researchers considered a deep learning approach to solve driver behavior detection. [26]. Mofid *et al.* [27] attempted to employ a body part segmentation model and a skin color segmentation model to obtain images of key body parts. Baheti *et al.* [28] proposed an improved VGG-16 model by replacing two convolutional layers with fully connected layers. The activation function was modified to LeakyRelu, and the dropout layer was added. The improved VGG-16 model can still achieve 95.54% classification accuracy in the case of fewer parameters. Similar to the study [28], Masood *et al.* [29] used the VGG-19 model to achieve distracted driver behavior detection in 10 categories. The dataset he used was the Statefarm dataset provided by kaggle. Their model was eventually able to achieve a classification accuracy of 99. Eraqi *et al.* [30] created a publicly available American University in Cairo driver distraction dataset (AUCD2). This dataset has the same 10 categories as the SFD3 dataset. They used the Alexnet model and the InceptionV3 model, which were pre-trained through ImageNet. Both models were applied to specific downstream tasks, including RGB image, face image and hand image segmentation. They fine-tuned the model, while using a genetic algorithm. The accuracy of the model was 95.98%. However, due to the large size of the model, this was difficult to achieve for real-time detection.

Rangesh and Trivedi [31] proposed a DCNN model to detect the hand-related driving behavior of drivers. Deo and Trivedi [32] set up multiple cameras in the cab to acquire images of the driver from multiple angles. Multiple DCNNs are used to extract the state features of the driver's body parts in the images. The long-term memory model is used to predict the driver's takeover state in L3 autonomous vehicles. Guo *et al.* [33] proposed to enhance the low-light images through regularized illumination optimization and deep noise suppression. By using this approach, the model has still achieved good recognition results when faced with driving scenes with low-light images. Zhao *et al.* [34] proposed a distraction detection model based on an adaptive spatial attention mechanism. The model extracted images through adaptive discriminative space and cropped them through three sub-networks in turn. Then, a multi-scale feature representation was extracted. Finally, a k-nearest neighbor classifier was used for classification. The existing work on deep learning-based methods still has limitations. Inputting the entire image into the model leads to feature redundancy. Partial image input to the model leads to insufficient feature information, resulting in low recognition accuracy of the model. At the same time, the existing work does not take into account the engineering deployment when designing and proposing the model, so it does not take into account the performance and lightness of the model.

III. PROPOSED METHOD

A. IMPROVED YOLOv5s ALGORITHM

The previous version of the YOLOv3 model was complex, and at the same time did not perform well on medium and large object detection. Although YOLOv4 weighs model performance and detection speed, the overall detection accuracy also needs to be improved. Therefore, considering the deployment cost, detection speed, and accuracy rate, YOLOv5s is selected as the baseline.

The overall improvement idea is as follows:

The first step is to improve the data augmentation strategy. Through the mixup data augmentation strategy [35], the background complexity of the training image is higher, and the diversity of the training set is increased. The detection performance and robustness of the model are improved without changing the network structure and increasing the calculation amount.

Next, the SE module is introduced and added to the last layer of the YOLOv5s backbone network. The SE module processes the given intermediate feature map to obtain the global receptive field. The spatial information squeeze is performed to get the weight information of all channels. Finally, the new feature map is reweighted and output.

Then, the BiFPN feature pyramid network is introduced in the neck section. BiFPN can be understood as a bi-directional feature fusion pyramid network with weights. The output of the feature maps of three sizes is preceded by a concat operation. The BiFPN module is added after these three ''concat''. BiFPN can not only realize bidirectional feature information flow but also add cross-connection. Therefore, the feature information flow at the same level can integrate more features. Furthermore, BiFPN will not produce too much calculation amount compared with traditional FPN and PANet.

B. IMPROVED DATA AUGMENTATION STRATEGY

In a limited amount, the diversity of input images is increased by data augmentation. The trained model will also have better robustness and generalization ability. Among them, geometric transformation and pixel-based color transformation are two common forms of data augmentation. Geometric transformations are usually horizontal and vertical flips, multi-angle rotations, scaling, and panning. Colour transformations include adjustment of image brightness, contrast, and saturation. These two types of data augmentation are not very effective on the object in action behavior detection for car drivers. Mixup will arbitrarily blend two training images with

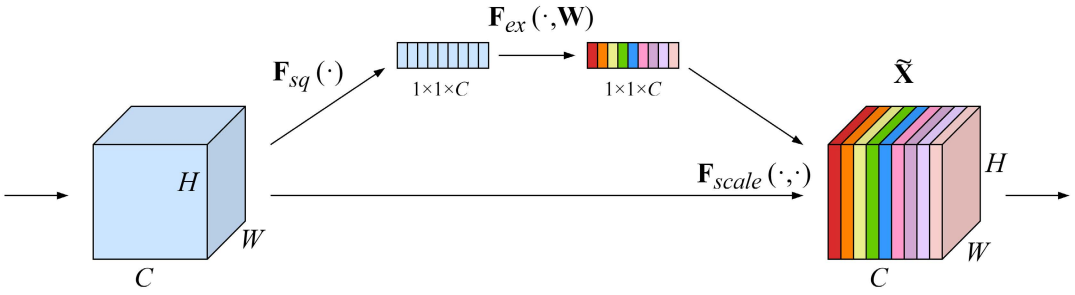


FIGURE 2. SE module structure.

pixels for regularisation, resulting in a single image with the two input labels. Mosaic data augmentation randomly crops four images and puts them together in sequence to form a single image. In this work, combine these two data augmentation methods to maximize the mixing of different contextual information, increase the diversity and contextual complexity of the dataset, and achieve a pure performance gain.

The mixup method is to randomly select two images from the training set and obtain their time series data (x_i, y_i) and $(x_j, y_j)(i \neq j)$, where x_i and x_j are the two images, an y_i and y_j are the corresponding labels of the images, both of which are one-hot encoding. Finally, a new image is obtained by actuarial computation, $\tilde{x}$ and $\tilde{y}$ are the corresponding newly generated images and their labels, respectively, calculated as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \tag{2}$$

$\lambda \in [0, 1]$ follow the $Beta(\alpha, \alpha)$ distribution. Where α is a hyper-parameter designed to control the strength of interpolation between feature-target pairs. stay $\alpha \to 0$, the data enhancement effect is close to failure.

C. SE MODULE

The attention mechanism is derived from the research on human vision. It is widely used in many types of machine learning tasks such as natural language processing and image recognition [36]. SE, an attention mechanism module, does not require the redesign of the network structure and tuning of parameters. Therefore, it can be directly inserted into the existing deep learning network model. The SE module has three main components, squeeze, excitation, and reweight. Its overall structure is shown in Fig 2. After a given input feature map, compression is performed in the spatial dimension to obtain information about the weights in each channel. The obtained weights are again multiplied with the original input to finally obtain the new feature map.

In the original YOLOv5s backbone network, the convolutional layer mainly calculates the feature information at the adjacent position of each feature map. Since each channel in the feature map contains different feature information, the convolutional layer ignores the correlation mapping between the channel information. With the addition of the attention mechanism module SE, multi-scale feature fusion is achieved to enhance the training of feature information between channels. It is well documented in the literature [37] that the attention mechanism module SE optimizes the learning of class-specific feature information in feedforward neural networks to effectively improve detection and classification performance. In this work, the attention mechanism module SE is added to the last layer of the backbone network. This improved backbone network, with the spatial and channel attention mechanism modules, enhances the focus on

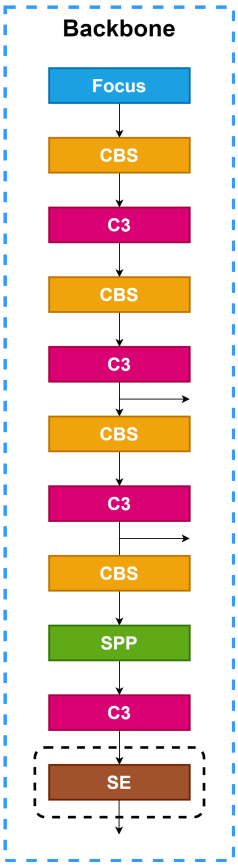


FIGURE 3. Introduction of the SE attention module.

important feature information while suppressing the focus on secondary information. The action behavioural information is extracted more effectively, ultimately enabling the trained model to better classify and detect specific classes of behavior.

The first part is squeeze, where information within the same channel is aggregated along the spatial dimension. This is done by globally averaging the pooling of a feature map given input $H \times W \times C$ to obtain a multidimensional statistic $\mathbf{z} \in \mathbb{R}^C$. The c-th element is:

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{3}$$

where $u_c(i, j)$ is the element in the row i and column j of the input data.

The second part is the excitation, where the features are input to the first fully connected layer (FC1) and then the $1 \times 1 \times C$ features are reduced to $1 \times 1 \times \frac{C}{r}$. This is followed by a ReLU and then input to the second fully connected layer (FC2). The $1 \times 1 \times \frac{C}{r}$ feature is restored to the C dimension, while a sigmoid activation function is picked up. The calculation process is as follows.

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \tag{4}$$

where $\mathbf{s}$ is the weight describing the importance of each channel in the input feature map $\mathbf{U}$. $\mathbf{W}_1 \in \mathbb{R}^{(C/r) \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{(C/r) \times C}$ are the parameters of the FC1 layer and FC2 layer respectively. r is a weakening factor to balance the recognition performance of the model with the calculation amount. In this work, r is used for 16. $\sigma(\cdot)$ and $\delta(\cdot)$ are the sigmoid activation function and the ReLU activation function, respectively.

The third part is the reweight. The weights $\mathbf{s}$ are multiplied on the channel with the original input feature map $\mathbf{U}$ to form the final output new feature map $\tilde{\mathbf{X}}$, calculated as follows:

$$\tilde{\mathbf{x}}_c = \mathbf{F}_{scale}(\mathbf{u}_c, s_c) = \mathbf{u}_c \cdot s_c \tag{5}$$

$$\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \cdots \tilde{\mathbf{x}}_C] \tag{6}$$

Traditional CNN channels are unable to exchange information among themselves and have the limitation of the local receptive field. The SE module, through global average pooling and fully connected layers, obtains the weights of each channel. This enhances the effective information association between channels. Moreover, the model will not add too much calculation amount when inserting the SE module.

D. BIFPN MODULE

After the input images have been extracted by the YOLOv5s backbone network, they need to be processed by the neck network and output to the detection layer. In the original YOLOv5s network structure, PANet is used as the neck network, and its network structure is shown in Fig. 4. Through bottom-to-top and top-to-bottom dual-path aggregation, feature fusion of bottom feature information and top strong semantic information is achieved, while shortening the information path before the bottom and top. Feature layers of the same length and width in upsampling and downsampling are stacked, which also in turn ensures features and information of small objects.

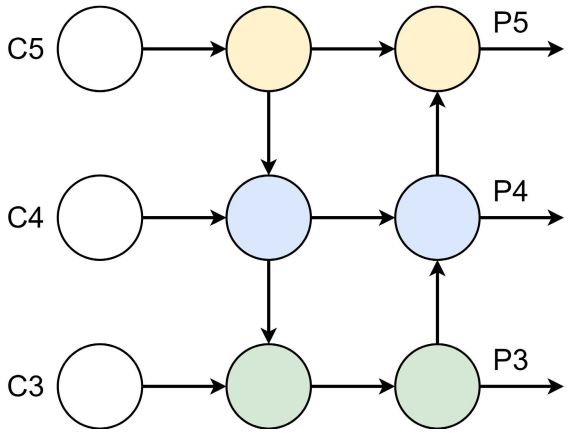


FIGURE 4. PANet network structure.

In Fig. 5, C is the input feature map and P represents the output feature map.

The neck PANet structure in the original YOLOv5s structure is a simple two-way fusion in the feature pyramid. Although the shallow information transfer and the strong semantic information of the high-level feature map can be achieved for fusion, the two parts are fused with a direct summation operation. There is no associated weighting design. To address this issue, this work introduces the BiFPN module, whose structure is shown in Fig. 5. The BiFPN module improves the structure of PANet in the original Neck. Firstly, the intermediate node inside the top and bottom edges, i.e. a node between C5 to P5 and C3 to P3 in Fig. 4, is removed. This is because nodes with a more homogeneous input and output contribute less to a network that fuses different features. Secondly, residual connections are added by skipping the deleted nodes. An additional edge is added using residuals only if there are intermediate nodes and their input and output nodes are in the same hierarchy. Only three scales of feature information fusion in YOLOv5s. Remove the middle node of the two edges. Only the nodes between C4 and P4 are left with the same level of input and output. Therefore a residual connection is added to the line C4 to P4. The aim is to fuse more features without increasing the calculation amount. Finally, the BiFPN is used directly as a base unit that can be stacked repeatedly and added to the network. Whereas the PANet has only one top-down and bottom-up path, the BiFPN treats a pair of paths as a feature layer that can be stacked repeatedly to fuse more high-level features. However, as different features have different resolutions, this contributes differently to feature fusion. For this reason, the BiFPN structure adds additional weight to each input during feature congruence and the network gradually learns the importance of each input feature during the training process.

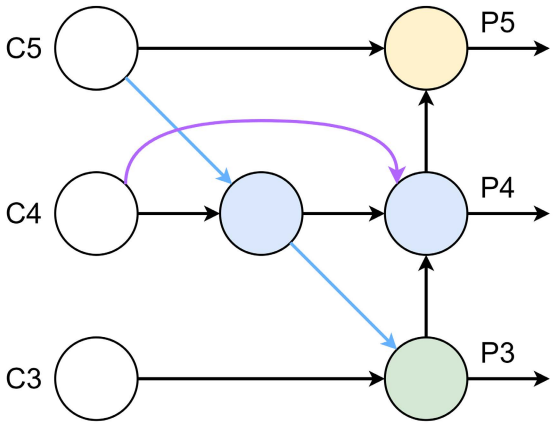


FIGURE 5. BiFPN module structure.

Based on this, three different weighting methods are experimented with as follows.

$$O = \sum_i w_i \cdot I_i \tag{7}$$

$$O = \sum_i \frac{e^{w_i}}{\sum_j e^{w_j}} \cdot I_i \tag{8}$$

$$O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} \cdot I_i \tag{9}$$

where O is the node output value and I_i is the input value from the node i. w_i is the weight of the input node i. Where j is the sum of the number of input nodes. Where $\varepsilon = 0.0001$ is a small quantity that guarantees the stability of the value.

FPN is a traditional top-to-bottom feature fusion network that is limited by one-way information communication. Based on this, the original YOLOv5s structure of PANet in Neck is a simple bi-directional feature fusion network implementing both bottom-to-top and top-to-bottom to achieve bi-directional information communication.

In this work, it is proposed to add a BiFPN module with better performance to the neck. The improved structure is shown in Fig 6. In the original Neck structure on the left side of the figure, concat is directly connected to the C3 module. Feature maps of different scales are output through three C3 modules. One layer of BiFPN modules is stacked before the last three C3 modules in the original Neck. Because the BiFPN module can be regarded as a basic unit. The right part of Fig.6 is the improved Neck structure which contains various modules and operations. The three inputs from the backbone are input to two concat operations and a CBS module. These feature maps go through a series of upsamples, CBS modules and C3 modules. After the concat operation, the feature maps of three different scales will be output to the detection head through the BiFPN module, the C3 module and the convolution operation in turn. Among them, the large-size and medium-size feature maps will also have branch outputs to the CBS module before being input to the convolution operation.

E. LOSS FUNCTION

In terms of the loss function, the model is guided and trained by adopting three parts: classification loss function (cls_{loss}), localization loss function ($giou_{loss}$), and confidence loss function (obj_{loss}). The binary cross-entropy loss function is used to calculate the class probability and object confidence score loss, and GIOU LOSS is used as the regression loss for the bounding box.

$$cls_{loss} = -\sum_p [y_p \log(\hat{y}_p) + (1 - y_p) \log(1 - \hat{y}_p)] \tag{10}$$

$$giou_{loss} = -\frac{1}{\sum_p 1} \sum_p (1 - iou_p) \tag{11}$$

$$obj_{loss} = -\sum_i [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{12}$$

where y is the prediction category, p is the index of predicted cases, and i is the total index of prediction frames. Where iou_p is the intersection ratio of the predicted example box p to the real box

IV. EXPERIMENTS

A. DATA COLLECTION

Driver Monitoring Dataset [38] (DMD) for attention and alertness analysis is an extensive, diverse and comprehensive set of driver behavior monitoring data. It includes both real and simulated driving scenarios, which include distracted driving behaviors and driver fatigue. The footage was captured from two Intel Realsense D415 and one Intel Realsense D435 camera. The DMD fills a gap in the multi-purpose dataset for driver monitoring and extends the idea of driver behavior detection.

The dataset used in this paper is mainly produced by filtering and optimizing based on DMD data. By cropping and framing the videos with RGB information of body parts, 9335 images of distracted driving behavior of drivers are obtained. Of the 9335 collected images, 7335 are used as the training set, 1000 validation set, and 1000 are related to the test set. If only the images of the DMD dataset are used for model training and verification, the generalization ability and robustness of the model will be affected. To address this issue, 500 StateFarm data [39] images and 500 self-captured images are added to the test set used in the experiments. It is important to note that the input to the model is not a sequence of multiple images. Examples and quantities of the final collected experimental datasets are shown in Tables 1 and 2.

B. EXPERIMENTAL ENVIRONMENT

To ensure the efficiency of the improved YOLOv5s model training and testing, the experimental environment configuration is shown in Table 3.

To ensure reasonable parameter settings, we follow the parameter settings provided by the YOLOv5 official project.

The relevant training parameters are set as follows: initial learning rate 0.01, termination learning rate 0.2, Batch_Size of 32, and the number of epochs is 300.

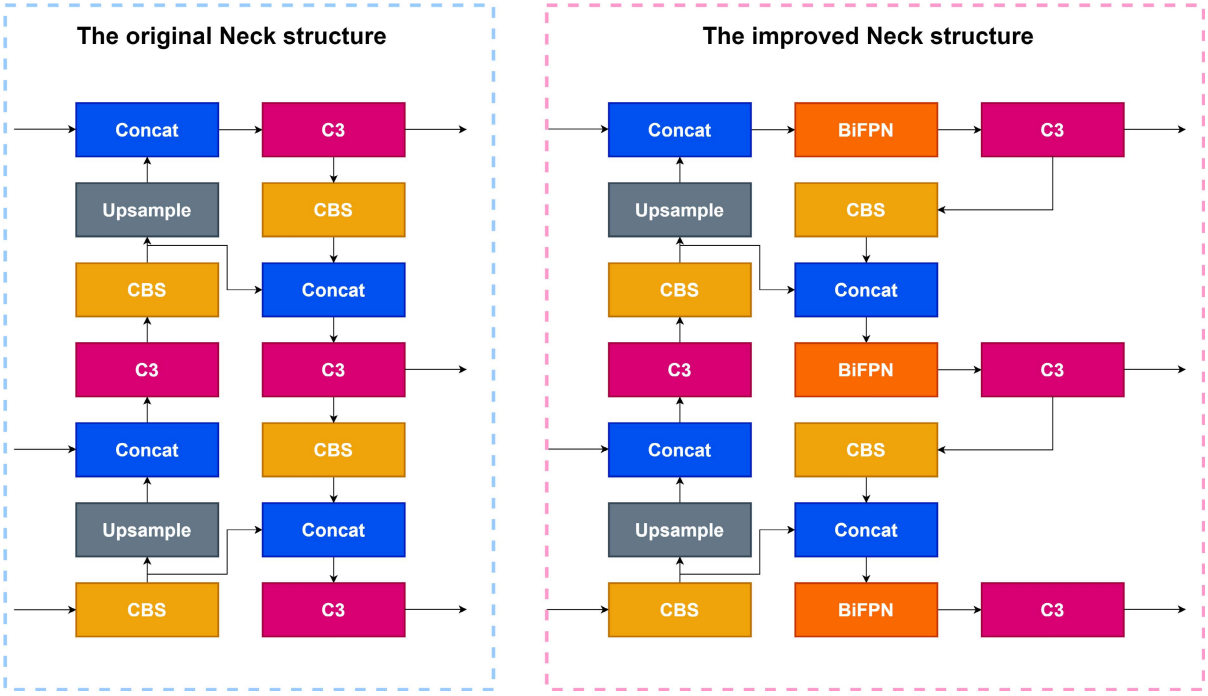


FIGURE 6. Comparison of improvements to the Neck section.

TABLE 1. Sample driver distraction behavior.

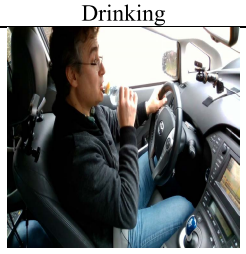
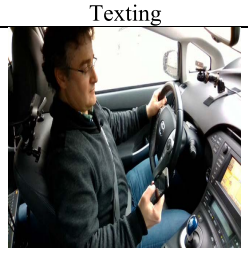
Operating the radio	Drinking	Talking on the phone	Texting	Hair and make up
				
C0	C1	C2	C3	C4

TABLE 2. Number of the dataset.

Dataset name	C0	C1	C2	C3	C4	Total
Training set	1490	1460	1465	1481	1439	7335
Validation set	200	200	200	200	200	1000
Test set	400	400	400	400	400	2000
Total	2090	2060	2065	2081	2039	10335

C. METRICS

In measuring the performance of the model, Precision, Recall, Average Precision (AP), mean Average Precision (mAP), Intersection-Over-Union (IOU), W_S (weight size), the number of floating-point operations GFLOPs and detection speed FPS are used as the relevant metrics for model performance evaluation.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

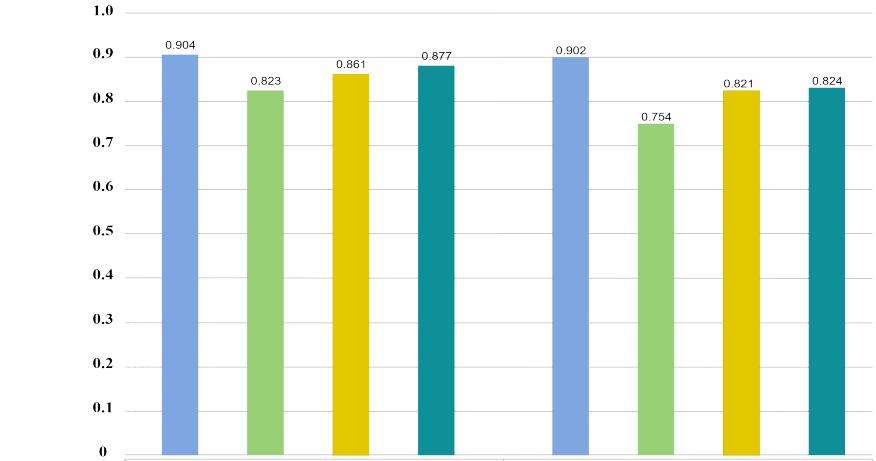
$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where, TP: True Positive, FP: False Positive, FN: False Negative.

The AP is the mean value of the accuracy in predicting a category. The mAP represents the ratio of the summed AP values for all single categories to the number of categories. The value of mAP is generally calculated at IOU = 05.

$$AP = \frac{\sum P}{N} \tag{15}$$

$$mAP = \frac{\sum AP}{NC} \tag{16}$$



	Ours	Baseline
Precision	**0.904**	**0.902**
Recall	**0.823**	**0.754**
F1	**0.861**	**0.821**
mAP	**0.877**	**0.824**

FIGURE 7. The results using the improved data augmentation strategy.

TABLE 3. Hardware and software configuration.

Operating systems	Linux Ubuntu 20.04.2 LTS
CPU	Intel Xeon Gold 6142
GPU	NVIDIA Geforce RTX 3080（10G）
Memory	16G×2
Solid State Drives	480G
Pytorch	V1.10
CUDA	V11.2

where N is the number of images and NC is the number of sample types.

The IOU is the overlap between the resulting prediction frame and the true frame.

$$IOU = \frac{A \cap B}{A \cup B} \tag{17}$$

where A is the prediction box and B represents the ground-truth box. The numerator is the intersection of the two frames and the denominator is the union of the two boxes.

The W_s is the size of the memory occupied by the model. GFLOPS (Giga Floating-point Operations Per Second) is used to measure the complexity of an algorithm or model. FPS is the number of pictures processed per second.

D. TEST RESULTS USING THE IMPROVED DATA AUGMENTATION STRATEGY

To verify that the proposed data augmentation strategy has improved the model detection performance, we evaluated the method on the validation set. It was also compared with the original data augmentation strategy. The results are shown in Fig. 7.

In the original baseline, the lowest Recall value is 0.754, but Precision remains at a high level of 0.902. Therefore, the F1-score does not reach a high value of 0.821 due to the low Recall value. The mAP is often an important reference indicator when measuring the detection performance of a model. The baseline mAP value is only 0.824, which is not sufficient to achieve high accuracy distracted driving detection. By using an improved data augmentation strategy for the original model, an improvement in model detection performance can be achieved. As shown in Fig. 7, these four indicators have been improved due to the introduction of the improved data augmentation strategy. The Precision has increased from 0.902 to 0.904, a small increase but still a high value, which reflects the high accuracy of the model. Recall also reached 0.823 from the previous 0.754. Recall increased by 6.9% The significant increase in Recall also resulted in a 4% increase in F1-score. The data augmentation can significantly improve the recall value. At the same time, the mAP was increased from the original 0.824 to 0.877. The mAP is an important model measure. The 5.3% improvement in mAP proves that the proposed data augmentation is an effective optimization.

E. ABLATION EXPERIMENTS

In ablation experiments, the performance of each model is analyzed by comparing the original model with the improved model on a unified dataset.YOLOv5s is the original model, and YOLOv5s+SE and YOLOv5s+BiFPN are the improved network models proposed in this paper.

TABLE 4. Precision of the networks with different model.

Model	C0	C1	C2	C3	C4	AP
YOLOv5s	0.975	0.916	0.938	0.861	0.822	0.902
YOLOv5s+SE	0.984	0.915	0.933	0.898	0.827	0.911
YOLOv5s+BiFPN	0.989	0.918	0.942	0.903	0.831	0.916

TABLE 5. Recall of the networks with different model.

Model	C0	C1	C2	C3	C4	AR
YOLOv5s	0.929	0.635	0.606	0.764	0.834	0.754
YOLOv5s+SE	0.926	0.748	0.804	0.751	0.831	0.812
YOLOv5s+BiFPN	0.930	0.767	0.811	0.762	0.833	0.820

TABLE 6. F1-score of the networks with different model.

Model	C0	C1	C2	C3	C4	F1-score
YOLOv5s	0.951	0.75	0.736	0.809	0.828	0.821
YOLOv5s+SE	0.954	0.823	0.863	0.818	0.828	0.858
YOLOv5s+BiFPN	0.958	0.835	0.871	0.826	0.831	0.865

TABLE 7. Precision of the networks with different model.

Model	C0	C1	C2	C3	C4	AP
YOLOv5s+SE+BIFPN	0.997	0.968	0.928	0.841	0.962	0.939
YOLOv5s+SE+Mixup	1	0.933	0.928	0.98	0.813	0.931
YOLOv5s+BiFPN+Mixup	0.996	0.933	0.933	0.98	0.816	0.932

TABLE 8. Recall of the networks with different model.

Model	C0	C1	C2	C3	C4	AR
YOLOv5s+SE+BIFPN	0.999	0.748	0.632	0.778	0.968	0.825
YOLOv5s+SE+Mixup	0.956	0.847	0.792	0.831	0.943	0.874
YOLOv5s+BiFPN+Mixup	0.965	0.847	0.854	0.831	0.943	0.888

TABLE 9. F1-score of the networks with different model.

Model	C0	C1	C2	C3	C4	F1-score
YOLOv5s+SE+BIFPN	0.998	0.843	0.752	0.808	0.964	0.878
YOLOv5s+SE+Mixup	0.977	0.888	0.854	0.899	0.873	0.901
YOLOv5s+BiFPN+Mixup	0.980	0.889	0.891	0.899	0.874	0.909

Tables 4, 5, and 6 show the performance of the models in terms of Accuracy, Recall, and F1-score respectively. C0-C4 in the table represent each of the five distracted driving behaviors. AP is the macro-recision (average precision value across all categories). AR is the average Recall across all categories. F1-score is calculated from the P and R values across all categories. By comparing the data in the three tables, both the YOLOv5s+SE and YOLOv5s+BiFPN models have higher AP and F1-score than the original model YOLOv5s. Among them, YOLOv5s+BiFPN is higher than the original model in both precision and F1-score for all types of behaviors.

In experiments, the three methods for improving the model are paired two by two. The mixup data augmentation strategy is used to combine SE and BiFPN in turn, and SE with BiFPN.

The following three models YOLOv5s+SE+BiFPN, YOLOv5s+Mixup+SE, and YOLOv5s+Mixup+BifPN are

TABLE 10. Comprehensive performance of the networks with different model.

Model	Precision	F1-score	mAP	W_S/MB	GFLOPs	V/FPS
Baseline	0.902	0.821	0.824	13.7	16.4	90
AB-DLM	**0.974**	**0.940**	**0.956**	**15.9**	**17.8**	**70**

eventually obtained. The same experiments as in the previous section are still tested using the uniform data set to derive the performance of the model, as shown in Tables 7, 8, and 9. The YOLOv5s+Mixup+BiFPN model is the best performer. By using the data augmentation strategy and the BiFPN module, the model showed significant improvements in detection performance. This resulted in improved accuracy, Recall, and F1-score. At the same time, the remaining two models also performed better than the original model. The effectiveness of the three improved methods proposed in this paper is verified.

The experiments in this section compare the original model YOLOv5s with the AB-DLM model. The results are shown in Table 10, which displays the performance of the two models in terms of Precision, F1-score, mAP, GLOPs, and FPS, respectively.

The AB-DLM model achieves a maximum mAP of 0.956, a 13.2% improvement over the original model YOLOv5s mean accuracy of 0.824. The F1-score also improved by 11.9% from the original 0.821 to 0.940. At this point, the model detection performance reached a maximum level. With the addition of both modules, there is a certain increase in calculation amount and model weights. However, compared to the YOLOv5s model, the model weights increased by only 2.2MB and the GFLOPs by only 2.4. A significant accuracy improvement is achieved at the expense of a slight memory weight gain and an increase in the calculation amount for the model. The improved model has an FPS value of 70, which still maintains a high value for fast detection, while also easily meeting the needs of real-time monitoring.

F. COMPARISON WITH OBJECT DETECTION ALGORITHMS

To further verify the efficiency and classification accuracy of the improved algorithm for driver distraction detection, the algorithm in this paper is compared with the mainstream object detection algorithms at this stage. A unified test set is used, with the same experimental parameters, and the same hardware and software environment set. The algorithm is used to ensure the consistency of the algorithm variables except for the comparison experiments.

The results of the comparison with other object detection algorithms are shown in Fig 8. The algorithms in this paper are trained with YOLOv3-tiny, YOLOv4-tiny, Mobilenet-SSD, PP-YOLO Tiny, YOLOX-S, and YOLOv5m for 300 iterations simultaneously. The earlier Mobilenet-SSD network model has the lowest mAP value. The traditional YOLO series, YOLOv3-tiny, and YOLOv4-tiny, which are both lightweight networks of YOLO, have a higher mAP value than Mobilenet-SSD. The PP-YOLO series developed by Baidu also has the same lightweight network, PP-YOLO Tiny, which has an improved mAP value compared to YOLO's tiny series. On this basis, the S model in the relatively new YOLOX series performs better than the above models. The performance of the improved model is better than yolov5m, and the memory occupation is smaller. The detection performance of these most advanced mainstream object detection models is still lower than the algorithmic model proposed in this paper.

The visualization of the distracted behavior detection results is shown in Fig 9.

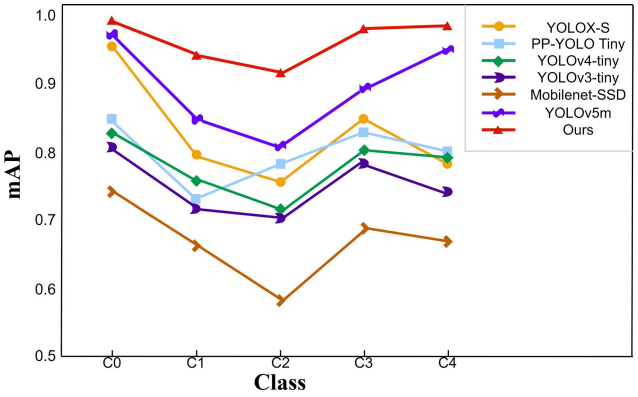


FIGURE 8. Comparison of different object detection models.

G. COMPARISON WITH OTHER METHODS

To further analyze the performance of the algorithms in this paper, and to achieve widely recognized comparative results. The proposed improved model is compared with SOTA methods by employing the data used in our study. Compared with the six methods proposed by Eraqi *et al.* [30], Mase *et al.* [40], Mase *et al.* [41] and Zhao *et al.* [34]. The comparison results are shown in Table 11. The computation and extraction of experimental results for AP and F1-score among these five SOTA methods are done by Eraqi *et al.* [30], Mase *et al.* [40], Mase *et al.* [41] and Zhao *et al.* [34]. When comparing models, AP is still more meaningful for object detection or classification reference. Because the model is always the most important for the accuracy of behavior detection, it is important to compare the AP of all methods when comparing. By observing the results in the table, it is concluded that among all the models, the recognition accuracy of the algorithm model proposed in this paper is the best.

Most of the research on the detection of driving behaviors by researchers is to classify the driving behaviors made by

TABLE 11. Results of method comparison.

Researchers	Method	AP(%)	F1 score(%)
Eraqi[30]	Ensemble Inception V3	91.62	90.33
Mase[40]	Inception V3-BiLSTM	-	93.10
Mase[41]	C-SLSTM	93.70	93.40
Zhao[34]	ResNeXt50-S3F	94.72	94.33
Zhao[34]	ResNeXt101-S3F	97.32	97.16
Ours	AB-DLM	**97.40**	94.03

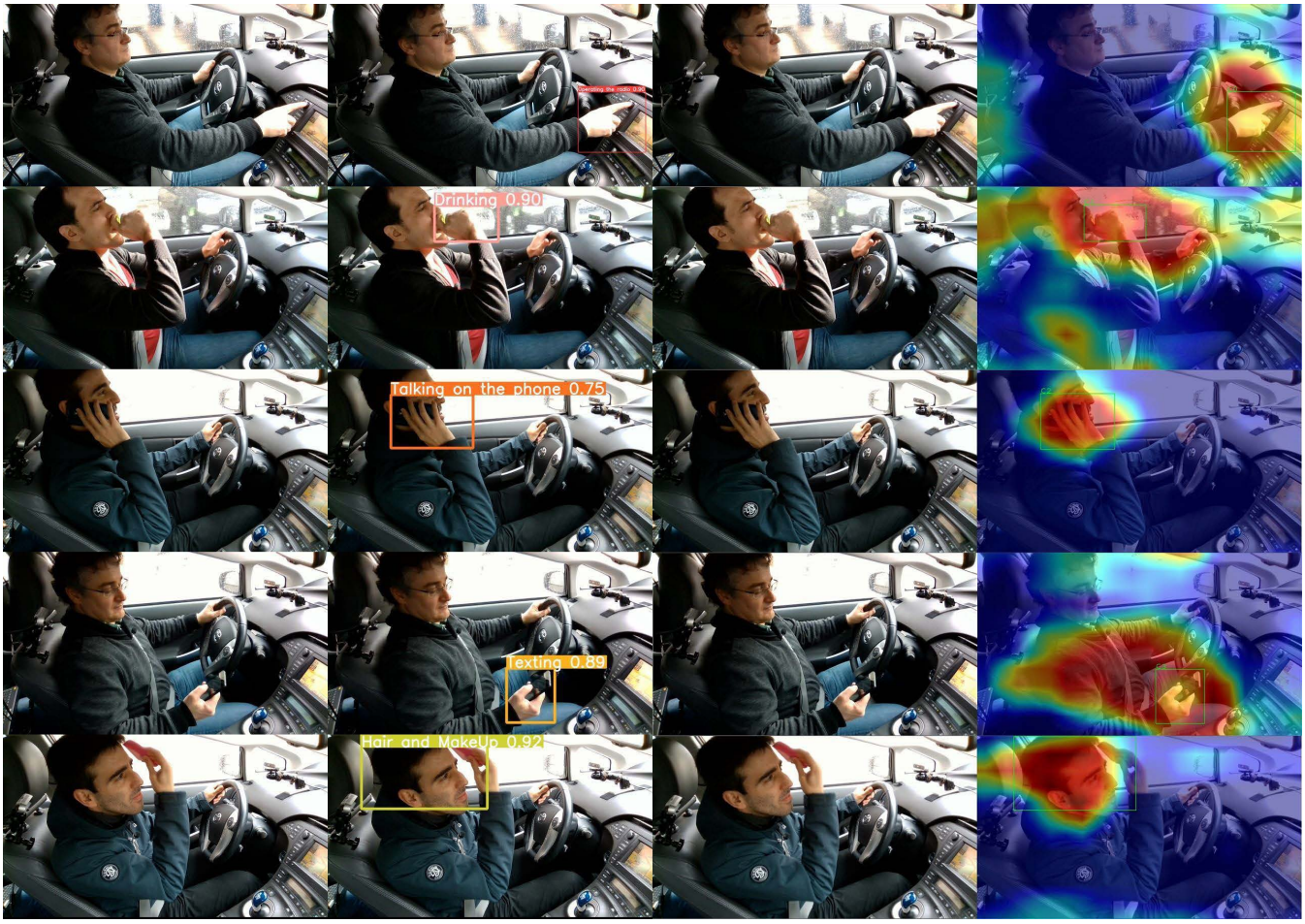


FIGURE 9. Visualization of driver distraction behavior detection results.

using classification networks. The algorithm proposed in this paper is based on object detection to detect distracted driving behavior, which is different from the current mainstream methods in principle. The distracted driving behavior of object detection can directly detect the distracted behavior of detailed categories, and at the same time, it can capture the action position of the subject. In addition, it is also seen in the table that the recognition accuracy of the model is the best. Compared with the traditional classifier method, the marking of the position will be more conducive to the research and development of safe driving.

V. CONCLUSION

Our approach to model improvement and optimization is to make the model as light as possible while delivering maximum performance gains. After research and study, the approach of structural optimization and new data augmentation strategies is proposed. The experimental results in Table 7 show that the AB-DLM model has significantly improved the detection accuracy, while the model simultaneously becomes larger and the detection speed decreases. This is because the SE layer and BiFPN modules are introduced in the overall structure. This improvement results in a larger

model in terms of structural complexity, as evidenced by the increase in WS and GFLOPs. It also leads to a slight decrease in detection speed, but this does not affect the real-time detection as it always remains at a high value. The 13.2% improvement in mAP is also an expected result and is in line with the original intention of the optimization and improvement.

In this paper, we apply the AB-DLM model and propose a driver distraction detection method based on an attention mechanism and a bi-directional feature pyramid network. It demonstrates better performance in terms of model lightness and detection accuracy compared to previous detection methods. In addition, with the use of the attention mechanism network, the AB-DLM model intensifies its focus on valid features when processing features upfront. This is an effective measure to ensure the model lightweight while improving detection performance. Enhanced feature fusion using BiFPN bi-directional feature pyramid network allows the model to achieve better results in classification and localization. For drivers, distracting actions are generated in a short and fast time, and the method provided in this paper has a high speed in detection speed. As well as being a lightweight model of only 15.9MB, this facilitates deployment to embedded or mobile devices.

In future work, the model will be trained using a more extensive and diverse dataset, and further optimization and lightweight of the current model will be considered. Another issue of interest is to propose in the future a driver distraction detection based on driver action behavior and eye attention tracking. The recognition of actions should be accompanied by the tracking of eye attention regions, and the two should be determined as coefficients. Ultimately, a calculation is to be made when a certain type of distraction is recognized. This will lead to more accurate detection of distracted driving.

REFERENCES

[1] (2018). *Global Status Report on Road Safety*. Accessed: Jan. 15, 2020. [Online]. Available: https://www.who.int/publications/i/item/global status-report-on-road-safety-2018

[2] Y. Yanbin, Z. Lijuan, L. Mengjun, and S. Ling, "Early warning of traffic accident in Shanghai based on large data set mining," in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Changsha, China, Dec. 2016, pp. 18–21.

[3] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. Paixão, F. Mutz, L. Veronese, T. Oliveira-Santos, and A. F. De Souza, "Self-driving cars: A survey," Oct. 2019, *arXiv:1901.04407*. Accessed: Aug. 4, 2022.

[4] C. Braunagel, W. Rosenstiel, and E. Kasneci, "Ready for take-over? A new driver assistance system for an automated classification of driver take-over readiness," *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 4, pp. 10–22, Winter 2017.

[5] *Uber Self-Driving Vehicles Involved in 37 Crashes Before Fatal Incident*. Accessed: Mar. 11, 2019. [Online]. Available: https://www.trtworld.com/business/uber-self-driving-vehicles-involved-in-37-crashesbefore-fatal-incident-31149

[6] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2339–2352, Jun. 2019.

[7] N. Neuhuber, G. Lechner, T. E. Kalayci, A. Stocker, and B. Kubicek, "Age-related differences in the interaction with advanced driver assistance systems—A field study," in *HCI in Mobility, Transport, and Automotive Systems. Automated Driving and In-Vehicle Experience Design* (Lecture Notes in Computer Science), vol. 12212. Cham, Switzerland: Springer, 2020.

[8] *Traffic Safety Facts: Distracted Driving 2013*, Nat. Highway Traffic Saf. Admin., Nat. Center Statist. Anal., U.S. Dept. Transp., Washington, DC, USA, 2015.

[9] CDC. (2016). *Distracted Driving*. Accessed: Jan. 24, 2020. [Online]. Available: https://www.cdc.gov/motorvehiclesafety/distracted.driving/

[10] S. G. Klauer, F. Guo, B. G. Simons-Morton, M. C. Ouimet, S. E. Lee, and T. A. Dingus, "Distracted driving and risk of road crashes among novice and experienced drivers," *New England J. Med.*, vol. 370, no. 1, pp. 54–59, Jan. 2014.

[11] N. Li, T. Misu, and A. Miranda, "Driver behavior event detection for manual annotation by clustering of the driver physiological signals," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 2583–2588.

[12] K. Lv, H. Sheng, Z. Xiong, W. Li, and L. Zheng, "Improving driver gaze prediction with reinforced attention," *IEEE Trans. Multimedia*, vol. 23, pp. 4198–4207, 2021.

[13] W. Li, J. Huang, G. Xie, F. Karray, and R. Li, "A survey on vision-based driver distraction analysis," *J. Syst. Archit.*, vol. 121, Dec. 2021, Art. no. 102319.

[14] X. Zuo, C. Zhang, F. Cong, J. Zhao, and T. Hamalainen, "Driver distraction detection using bidirectional long short-term network based on multiscale entropy of EEG," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 23, 2022, doi: 10.1109/TITS.2022.3159602.

[15] T. Čegovnik, K. Stojmenova, G. Jakus, and J. Sodnik, "An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers," *Appl. Ergonom.*, vol. 68, pp. 1–11, Apr. 2018.

[16] L. Alam and M. M. Hoque, "Real-time distraction detection based on Driver's visual features," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6.

[17] C. Craye, A. Rashwan, M. S. Kamel, and F. Karray, "A multi-modal driver fatigue and distraction assessment system," *Int. J. Intell. Transp. Syst. Res.*, vol. 14, no. 3, pp. 173–194, Sep. 2016.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[19] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[20] R. A. Berri, A. G. Silva, R. S. Parpinelli, E. Girardi, and R. Arthur, "A pattern recognition system for detecting use of mobile phones while driving," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, 2014, pp. 411–418.

[21] Y. Artan, O. Bulan, R. P. Loce, and P. Paul, "Driver cell phone usage detection from HOV/HOT NIR images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 225–230.

[22] C. Craye and F. Karray, "Driver distraction detection and recognition using RGB-D sensor," Feb. 2015, *arXiv:1502.00250*. Accessed: Aug. 4, 2022.

[23] X. Zhang, N. Zheng, F. Wang, and Y. He, "Visual recognition of driver hand-held cell phone use based on hidden CRF," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, Jul. 2011, pp. 248–251.

[24] C. H. Zhao, B. L. Zhang, X. Z. Zhang, S. Q. Zhao, and H. X. Li, "Recognition of driving postures by combined features and random subspace ensemble of multilayer perceptron classifiers," *Neural Comput. Appl.*, vol. 22, no. 1, pp. 175–184, 2013.

[25] C. Yan, F. Coenen, and B. Zhang, "Driving posture recognition by convolutional neural networks," *IET Comput. Vis.*, vol. 10, no. 2, pp. 103–114, Mar. 2016.

[26] S. Korkmaz, N. B. H. Ali, and I. F. C. Smith, "Configuration of control system for damage tolerance of a tensegrity bridge," *Adv. Eng. Informat.*, vol. 26, no. 1, pp. 145–155, Jan. 2012.

[27] N. Mofid, J. Bayrooti, and S. Ravi, "Keep your AI-es on the road: Tackling distracted driver detection with convolutional neural networks and targeted data augmentation," Jun. 2020, *arXiv:2006.10955*. Accessed: Aug. 4, 2022.

[28] B. Baheti, S. Gajre, and S. Talbar, "Detection of distracted driver using convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1032–1038.

[29] S. Masood, A. Rai, A. Aggarwal, M. N. Doja, and M. Ahmad, "Detecting distraction of drivers using convolutional neural network," *Pattern Recognit. Lett.*, vol. 139, pp. 79–85, Nov. 2020.

[30] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, "Driver distraction identification with an ensemble of convolutional neural networks," *J. Adv. Transp.*, vol. 2019, pp. 1–12, Feb. 2019.

[31] A. Rangesh and M. M. Trivedi, "HandyNet: A one-stop solution to detect, segment, localize & analyze driver hands," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1103–1110.
[32] N. Deo and M. M. Trivedi, "Looking at the driver/rider in autonomous vehicles to predict take-over readiness," *IEEE Trans. Intell. Vehicles*, vol. 5, no. 1, pp. 41–52, Mar. 2020.
[33] Y. Guo, Y. Lu, R. W. Liu, M. Yang, and K. T. Chui, "Low-light image enhancement with regularized illumination optimization and deep noise suppression," *IEEE Access*, vol. 8, pp. 145297–145315, 2020.
[34] L. Zhao, F. Yang, L. Bu, S. Han, G. Zhang, and Y. Luo, "Driver behavior detection via adaptive spatial attention mechanism," *Adv. Eng. Informat.*, vol. 48, Apr. 2021, Art. no. 101280.
[35] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018, pp. 1–13.
[36] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6687–6696.
[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[38] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12538, 2020, pp. 387–405.
[39] StateFarm. *State Farm Distracted Driver Detection*. Accessed: Jun. 15, 2017. [Online]. Available: https://www.kaggle.com/c/statefarm-distracted-driver detection
[40] J. M. Mase, P. Chapman, G. P. Figueredo, and M. T. Torres, "A hybrid deep learning approach for driver distraction detection," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2020, pp. 1–6.
[41] J. M. Mase, P. Chapman, G. P. Figueredo, and M. T. Torres, "Benchmarking deep learning models for driver distraction detection," in *Machine Learning, Optimization, and Data Science*, vol. 12566, G. Nicosia, V. Ojha, E. La Malfa, G. Jansen, V. Sciacca, P. Pardalos, G. Giuffrida, and R. Umeton, Eds. Cham, Switzerland: Springer, 2020, pp. 103–117.

TAIGUO LI received the B.Eng. degree in computer science and technology and the M.Eng. degree in computer software and theory from Wuhan University, China, in 2008 and 2010, respectively, and the Ph.D. degree in electronic science and technology from the China Academy of Space Technology (CAST), China, in 2017. Currently, he is a Lecturer at Lanzhou Jiaotong University, China. His research interests include electronic science and technology, computer vision, and driver behavior detection.

YINGZHI ZHANG was born in Laiwu, Shandong, China, in 1998. He received the B.Eng. degree in rail traffic signal and control from Jiangsu Normal University, China, in 2021. He is currently pursuing the master's degree in energy and power with Lanzhou Jiaotong University, China. His research interests include autonomous driving, computer vision, and driver behavior detection.

QUANQIN LI received the B.N. degree in rehabilitation therapeutics from the Shaanxi University of Chinese Medicine, China, in 2015. Currently, she is a Medical Care Staff at Shaanxi Kangfu Hospital, China. Her research interests include rehabilitation treatment and rehabilitation evaluation.

TIANCE ZHANG received the B.Eng. degree in electrical engineering and automation from the Henan University of Engineering, China, in 2018. He is currently pursuing the master's degree in energy and power with Lanzhou Jiaotong University, China. His research interests include machine learning and driver fatigue detection.