IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 7 November 2022, accepted 2 December 2022, date of publication 12 December 2022,
date of current version 15 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228331

RESEARCH ARTICLE

CEAM-YOLOv7: Improved YOLOv7 Based on Channel Expansion and Attention Mechanism for Driver Distraction Behavior Detection

SHUGANG LIU[1,3], YUJIE WANG[1,3], QIANGGUO YU[2], HONGLI LIU[4], (Member, IEEE), AND ZHAN PENG[1,3]

[1]School of Physics and Electronic Science, Hunan University of Science and Technology, Xiangtan 411201, China
[2]School of Electronic Information, Huzhou College, Huzhou 313000, China
[3]Key Laboratory of Intelligent Sensors and Advanced Sensing Materials of Hunan Province, Hunan University of Science and Technology, Xiangtan 411201, China
[4]College of Electrical and Information Engineering, Hunan University, Changsha 410082, China

Corresponding authors: Yujie Wang (wangyujie@mail.hnust.edu.cn) and Qiangguo Yu (yuqg@zjhzu.edu.cn)

This work was supported in part by the Foundation of Basic Strengthen Program in the Technical Field under Grant 2020-JCJQ-JJ-219, and in part by the National Nature Science Foundation of China under Grant 62173133.

ABSTRACT Driver distraction behavior is prone to induce traffic accidents. Therefore, it is necessary to detect it to caution drivers in time for traffic safety. In driver behavior recognition, the diversity of behaviors and driving environment can have a certain effect on detection accuracy, and most of the existing methods have serious information loss. These make it challenging to improve the real-time accuracy of driver distraction behavior. In this paper, we propose an improved YOLOv7 based on the channel expansion and attention mechanism for driver distraction behavior detection, named CEAM-YOLOv7. The global attention mechanism (GAM) module focuses on key information to improve accuracy. By inserting GAM into the Backbone and Head of YOLOv7, the global dimensional interaction features are scaled up, enabling the network to extract key features. Furthermore, In the CEAM-YOLOv7 architecture, the convolution computation has been significantly simplified, which is conducive to increasing the detection speed. Combined with the Inversion and contrast limited adaptive histogram equalization (CLAHE) image enhancement algorithm, a channel expansion (CE) algorithm for data augmentation is presented to further optimize the detection effect of infrared (IR) images. On the driver distraction IR dataset of Hunan University of Science and Technology (HNUST) and Hunan University (HNU), the verification results show that CEAM-YOLOv7 achieves a 20.26% higher *mAP* compared to the original YOLOv7 model and the *FPS* reaches 156, which illustrate that CEAM-YOLOv7 outperforms state-of-the-art methods in both accuracy and speed.

INDEX TERMS Deep learning, attention mechanism, YOLOv7, driver behavior recognition.

I. INTRODUCTION

Driver behavior detection is an essential component of Advanced Driver Assistance Systems (ADAS) [1]. To make this technology move towards practical applications, the key issue to be addressed is how to improve accuracy and real-time performance.

In recent years, most of the state-of-the-art target detection algorithms have achieved satisfactory results in autonomous driving tasks using Convolutional Neural Networks (CNNs)[2], such as the two-stage detectors Fast R-CNN [3], Faster R-CNN [4], and FPN [5], the single-stage detectors SSD [6], RetinaNet [7], and YOLO [8], [9], [10], [11], [12], [13], [14]. For the CNNs mentioned, the accuracy has been improved by increasing the network depth. In this way, the computation load, such as floating point operations (FLOPs), would significantly increase. So the

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

detection speed and data memory were sacrificed. At present, in-vehicle terminals with distraction behavior detection have become important technology products of assisted driving, which requires both high accuracy and fast speed. To accurately identify distracted behaviors, sufficient software and hardware resources are required to complete CNNs algorithms. Therefore, for the practical terminals of driving distracted behavior recognition, these CNNs algorithms are difficult to adopt directly because of their computational complexity.

As a single-stage detector, YOLOv7 has the advantages of high accuracy and speed [14]. In this study, we try applying YOLOv7 to driver distraction behavior detection. Meanwhile, we further optimized its accuracy and speed to better adapt to in-vehicle terminals. In this work, we propose the CEAM-YOLOv7 algorithm, an advanced version of YOLOv7 for detecting driver distraction behavior. The proposed algorithm can overcome the shortcomings of YOLOv7 in detection accuracy and speed. The main contributions of the work are as follows:

1) The global attention mechanism (GAM) module is inserted into the YOLOv7 network. Information is retained to amplify the global interactions across dimensions. By capturing important features in three dimensions, information loss is decreased, and the accuracy of behavior recognition is raised.
2) The network layers are pruned based on the idea of model lightweight. The computation load is greatly reduced, which improves the recognition speed. It is beneficial for the algorithm to be deployed on in-vehicle terminals.
3) A channel expansion (CE) algorithm is proposed to optimize YOLOv7 for infrared (IR) image recognition. Combined with the Inversion and contrast limited adaptive histogram equalization (CLAHE) image enhancement algorithms, the IR images' channels are expanded to three. This strategy more effectively improves the performance and robustness of the training model.
4) The proposed method is evaluated on the IR images dataset of Hunan University of Science and Technology (HNUST) and Hunan University (HNU) in Fig.1. The dataset is more suitable for real driving scenarios compared with the visible images whose grayscale and contrast are easily affected by lighting and flare.

II. RELATED WORK

A. DEEP LEARNING FOR DRIVER BEHAVIOR DETECTION

At present, deep learning has achieved great success in object detection. Driver behavior detectors based on deep learning have also been widely studied in the industrial and academic fields. Zhao et al. [15] proposed a driver behavior detection system based on an adaptive spatial attention mechanism. The discrimination region was extracted adaptively according to the driver's behavior classification. Then K-NN was used to classify multi-scale state vectors to identify specific driving behaviors. Masood et al. [16] used deep convolutional networks to detect distracted drivers. VGG16 and VGG19 models are employed to identify the distraction causes and effectively distinguish the driver's behavior. Shahverdy et al. [17] analyzed driver behavior with recursive graph technique, converting driving signals, such as acceleration, gravity and throttle, into images. Then CNNs recognized the images as different behaviors. Xing et al. [18] built a unified modeling system for multi-scale behavior recognition based on a deep encoder-decoder framework. The drivers' physical and mental states are recognized together, enhancing the unified model's inference ability.

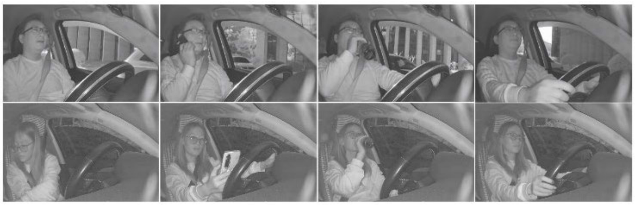


FIGURE 1. IR images dataset of HNUST and HNU for driver distraction behaviors.

Furthermore, Ghizlene et al. [19] presented a method to quickly detect the driver's eyes to identify the driver's drowsiness by combining the Haar cascade and YOLO algorithm. Based on YOLOv4-tiny, Zhao et al. [20] have integrated the Inception V3 architecture and RES-SEBlock module. The key feature information was extracted by adding attention module and squeeze-and-excitation module. As a result, the computation was reduced, and the average precision of mask-wearing detection reached 0.86. Qin et al. [21] built an enhanced eye-tracking object detection dataset for driving videos and proposed the increase-decrease YOLO network. The driver's selective attention mechanism was simulated to distinguish key objects in the driver's gaze area.

Most of the above networks use single-scale depth features, which are difficult to improve the detection performance in complex driving scenarios. Therefore, in the study, the GAM module is intended to optimize the YOLO network architecture for driver distraction behavior detection. In addition, most of the above algorithms have large models, which are difficult to deploy on in-vehicle terminals, and the *FPS* is too low to be applied in real driving scenarios. Therefore, we apply the idea of model lightweight to the YOLOv7 structure, which greatly reduces the calculation load and meets the application requirements.

B. IR IMAGES-BASED OBJECT DETECTION

IR images are not easily damaged by glare and lighting, so the object detection of IR images is widely concerned. Chen et al. [22] presented a novel R-Net based on IR image segmentation for human action recognition. The defined loss function comprehensively considered the shape, area and centroid of the images, which helps solve the impact of motion blur, low resolution and random noise on recognition accuracy. Yao et al. [23] used an effective single-stage algorithm for small IR targets based on FCOS and Spatio-temporal features, which enhanced the response to targets and suppressed the background response. Meanwhile, in order to eliminate the influence of static noise, time-domain features are added to the network as image sequences so that the network can learn the Spatio-temporal correlation features in the image sequences.

However, due to the low SNR and fuzzy edges of IR images, they are difficult to be used directly for recognition. To further improve the recognition of IR images, some studies have begun to focus on pre-processing image algorithms. Several mainstream algorithms are listed below.

1) INVERSION

It enables the processed IR image to be closer to the grayscale map of the visible image, which can significantly improve the recognition performance.

2) MEDIAN FILTERING + TOP-HAT AND BOTTOM-HAT TRANSFORM

The median filter can remove the Salt & Pepper Noise from the IR images. And Top-Hat and Bottom-Hat transform is used for image sharpening.

3) HISTOGRAM EQUALIZATION (HE)

Since the pixels of IR images are generally distributed in relatively concentrated intervals, histogram equalization is used for contrast enhancement;

4) CONTRAST LIMITED AHE (CLAHE)

Similar to HE, the processed pixel area becomes finer. In this way, noise can be suppressed while the contrast is enhanced.

Compared with visible images, IR images have fewer channels and thus contain less information. Therefore, choosing a suitable data augmentation method is crucial for IR image detection. In CEAM-YOLOv7, the CE algorithm consists of Inversion and CLAHE for image augmentation. In this way, we provide effective pre-processing of IR images.

III. CEAM-YOLOv7

A. DEEP LEARNING FOR DRIVER BEHAVIOR DETECTION

As the current state-of-the-art single-stage target detection algorithm, YOLO has been iterated to YOLOv7 since its release in 2016. In addition, there are many derivative algorithms based on the YOLO architecture, such as PP-YOLO [24], YOLOx [25], Scaled-yolov4 [26], YOLOR [27], and other optimized algorithms. The latest YOLOv7 [14] is optimized for deployment on edge terminals. It uses a composite scaling method to generate models at different scales to meet different inference speed requirements, such as YOLOv7-e6, YOLOv7-w6 and YOLOv7-x. The superior flexibility allows it to be easily deployed on in-vehicle terminals. The basic framework of YOLOv7 can be divided into three parts: Input, Backbone, and Head. The details are as follows:

Input: The Input part enriches the dataset by stitching data and requires only low computational cost.

Backbone: The Backbone part mainly consists of the E-ELAN module, which performs feature extraction through the CBS base convolution module.

Head: The Head part uses the SPPCSP and ELAN modules to aggregate image features. Then RepConv adjusts the channels of output features.1 × 1 convolution is used for prediction and output.

B. IMPROVED YOLOv7

The structure of the CEAM-YOLOv7 network is shown in Figure 2. Firstly, the GAM module[28] is introduced as our attention mechanism to extract key information. Secondly, based on the idea of model lightweight, we modify the network structure to improve the recognition speed. Moreover, some training tricks are used to enhance the performance of model.

1) NETWORK ARCHITECTURE

The original YOLOv7 network architecture is modified to make it specialized for the IR dataset. The CEAM-YOLOv7 network architecture can be divided into CBM, MP, SPPCSPC, and GAM modules. CBM is the basic convolution module, which consists of convolution blocks with different step sizes. As a multiple convolution module, Catconv uses the output of the other convolution layers for concat operation to improve the accuracy of the network. MP is a downsampling module that takes into account both the maximum and local value information of local regions. SPPCSPC is an improved spatial pyramid pooling structure (SPP) [29] that combines spatial pyramid pooling with the CSP structure.

The original network is designed for visible images, so detection accuracy cannot be guaranteed when directly used for IR images. Therefore, GAM modules are inserted at the output of the Backbone and Head parts of the architecture. Despite the increase in computation and memory overhead, object detection accuracy has improved.

2) GLOBAL ATTENTION MECHANISM

Different driver behavior is a fine-grained activity, and the attention should be directed to the region of interest. For example, drinking is mainly recognized by focusing on the shape and position of the hand and water bottle. GAM is an attention mechanism module that extracts relevant information by selectively focusing on the desired part of the channel and space to improve recognition accuracy. As shown in Fig.2, the sequential channel-spatial attention mechanism from CBAM [30] is used, of which submodules are redesigned. The channel attention submodule uses 3D permutation to preserve information across three dimensions. Multi-layer perceptron is used to amplify the cross-dimensional channel-spatial correlation. The spatial attention submodule uses two convolutional layers for spatial information fusion. The performance of the deep neural network is improved by reducing information loss and amplifying global interaction features. It provides an effective trade-off between recognition speed and accuracy, and

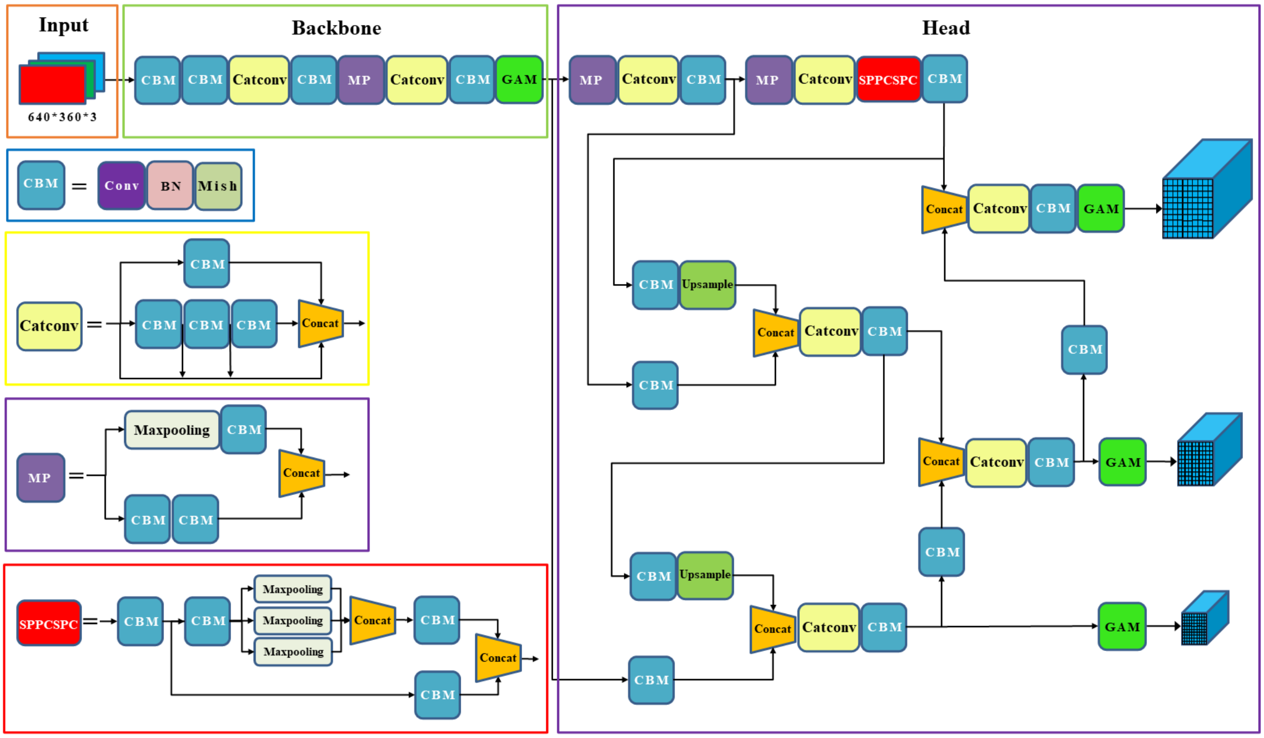


FIGURE 2. Architecture of CEAM-YOLOv7.

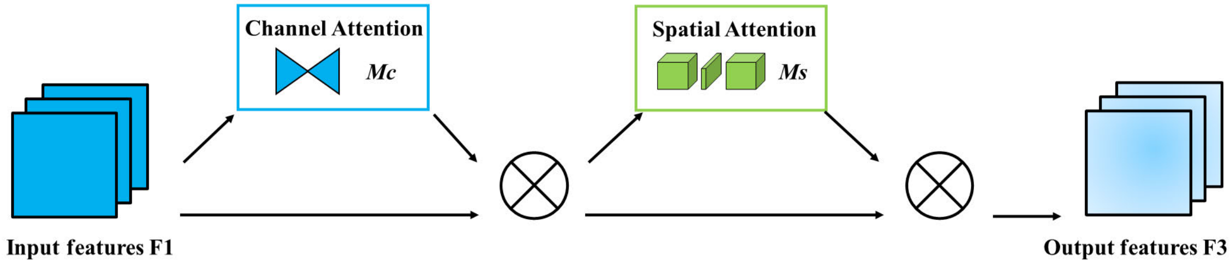


FIGURE 3. Global attention mechanism.

improves the recognition capability of targets in IR images. In addition, it corresponds to the CE algorithm in the data augment processing below.

The process is shown in Fig.2 and represented in equations (1) and (2). The given input feature mapping F_1, intermediate state F_2 and output F_3 are defined as:

$$F_2 = M_c(F_1) \otimes F_1 \tag{1}$$

$$F_3 = M_s(F_2) \otimes F_2 \tag{2}$$

where M_C and M_S represent channel and spatial attention maps, and $\otimes$ denotes element multiplication.

3) MODEL LIGHTWEIGHT

Deep neural networks are designed to extract deeper features. IR images have much fewer features than visible images, so a deep convolutional structure applied to IR images may bring about feature loss. Therefore, some convolution layers are removed from the original YOLOv7 network structure to reduce a large number of convolution operations. And the overall network structure has feature extraction capability while maintaining a moderate depth which is more suitable for object detection of IR images. As a result, we prune the original YOLOv7 layers from 306 to 235.

4) ACTIVATION FUNCTION

The SiLU activation function is replaced with Mish [31], whose upper-bound-free, smooth, and non-monotonic function properties allow better information deep into the network, thus contributing to training stability and final accuracy. Mathematically defined as:

$$f(x) = x\tanh(\mathrm{softplus}(x)) = x\tanh\left(\ln\left(1 + e^{x}\right)\right) \tag{3}$$

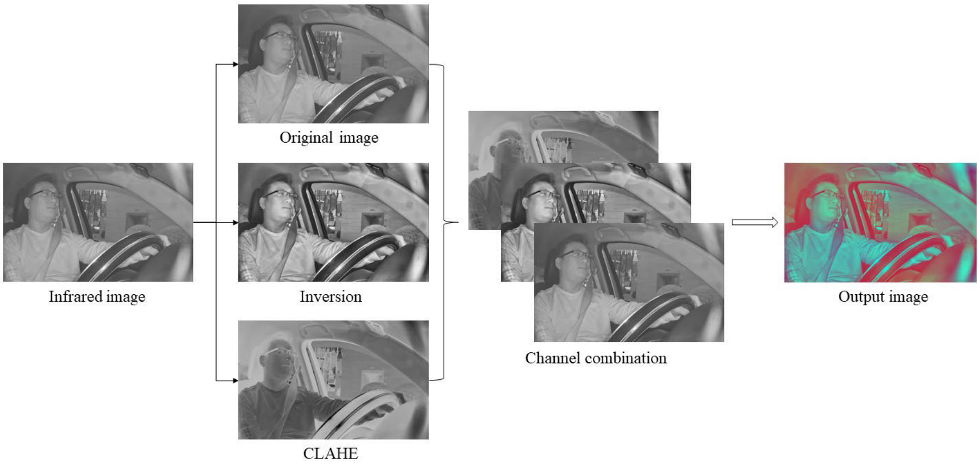


FIGURE 4. Channel expansion algorithm.

where softplus(·) represents the normalized exponential function which is a generalization of the binary classification function sigmoid on multi-classification, and x denotes input. In addition, we remove the Mosaic operation from the original YOLO to avoid reducing feature information due to image stitching.

C. CE ALGORITHM FOR IR IMAGE AUGMENTATION

Image enhancement algorithms can be broadly classified into color-oriented (e.g., luminance, contrast, and color projection) and geometry-oriented (e.g., scaling, flipping, panning, and zooming). The former enrich image information. And the latter artificially expands the size of the training dataset by data distortion or oversampling. Especially for IR images with low SNR, it is necessary to study a data enhancement method to enrich the image information and expand the dataset. So combined with Inversion and CLAHE, the CE algorithm is proposed in this study. The main functions of Inversion and CLAHE in the CE algorithm are as follows.

1) INVERSION

Inversion makes the network more adaptable to the processed IR image through the idea of domain migration. The Inversion operation can enhances the details of white and gray in dark areas of an image, facilitating the extraction of dark features.

2) CLAHE

CLAHE operation can make the grayscale distribution more uniform, enhance the contrast and suppress the noise simultaneously to increase the detail information of IR images.

CE algorithm generates images adapted to this work, increases the information content of the images, and improves the detection accuracy of the network. Besides, using data augmentation methods such as rotation and offset enrich the dataset for better training results. The application flow of the CE algorithm is shown in Fig.4.

D. LEARNING ALGORITHM

The task of driver behavior recognition is implemented using the well-trained CEAM-YOLOv7 model. The training procedure is summarized in Algorithm 1. The details are explained as follows.

1. In line 1, the structure of the CEAM-YOLOv7 model is constructed. The model consists of data augmentation, convolution, pooling, attention modules and activation functions.
2. In line 2, the parameters in the model are initialized. The parameters θ include weights w, bias b, reduction ratio r and learning rate α.
3. In lines 3-9, the CEAM-YOLOv7 model is trained using forward and backward propagation. In backward propagation, the optimization algorithm of SGD is used to update the parameters.
4. In line 9, the model training is completed when the end condition is satisfied to obtain the CEAM-YOLOv7 model with the best parameters for driver behavior recognition.

IV. EXPERIMENT AND ANALYSIS

A. EXPERIMENTAL SETTING

We implemented CEAM-YOLOv7 on PyTorch 1.10.1 and used NVIDIA GeForce RTX 2070 SUPER GPUs for training and testing. A partially pre-trained model of YOLOv7 was used in the training phase. Because CEAM-YOLOv7 and YOLOv7 share part of the network architecture, many weights can be transferred from YOLOv7 to CEAM-YOLOv7, and a lot of training time can be saved by using these weights. The model is trained on the dataset set for 300 epochs, using SGD optimizer for training, with 0.1 as the initial learning rate. The input image size is 640 × 320 pixels, and the batch size is 16. We use the evolve hyper-parameters method during the training process to optimize hyper-parameters continuously. Each baseline network architecture is trained with an identical optimization scheme.

Algorithm 1 Training Strategy of CEAM-YOLOv7

Input: Given training samples $X = \{x_1, \ldots, x_k\}$ and labels $Y = \{y_1, \ldots, y_k\}$, $k \in \mathbf{N}^+$.

Output: The well-trained model CEAM-YOLOv7

```
Construct the CEAM-YOLOv7 model shown in Fig. 2;
Initialize the parameters $\theta$ $(w, b, r, \alpha)$;
repeat
    Randomly select a batch of instances $X_b$ from $X$;
    Forward learn training samples through CEAM-YOLOv7 mode;
    Compute the training loss $L(\theta)$ by $L(\theta) = $ box__loss + object__loss + class__loss shown in Fig. 5;
    Propagate $L(\theta)$ back through CEAM-YOLOv7 and update the parameters with SGD;
    Find $\theta$ by minimizing $L(\theta)$ with $X_b$;
until End condition is satisfied.
```

The HNUST and HNU infrared images dataset is used for the experiment. The dataset was collected in a real driving situation, and the infrared camera was installed on the car center console to record the driver's behavior. The participants consisted of multiple male and female drivers in different driving environments to complete the dataset. The dataset contains four types of driver behaviors: normal(Safe), drinking (Drink), using a cell phone (Phone), and hands off the wheel (Danger). The drink and phone type are divided into left and right-handed, and the phone type is further subdivided into play phone and phone call. The numbers of safe, drink, phone, and danger images are 1000, 1200, 1500, and 1400, respectively. 3000 images were used for this experiment. They are randomly divided into training, validation and test set according to the 8:1:1 ratio. To avoid overfitting problem, there are different drivers in different sets. Fig. 1 shows the visual features of the original images. Based on the dataset, the Inversion and CLAHE data enhancement operations are used, with the proposed CE algorithm.

B. EVALUATION PARAMETERS

To demonstrate the advantages of the CEAM-YOLOv7, we use the following metrics: precision (P), recall (R), F1 score, average precision (AP), mean average precision (mAP), model size, parameters, $FLOPs$, and frames per second (FPS). The evaluation parameters equation is as follows:

$$\text{precision } (P) = \frac{TP}{TP + FP} \tag{4}$$

$$\text{recall } (R) = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 \text{ score} = 2 * \frac{P * R}{P + R} \tag{6}$$

$$AP_i = \int_0^1 P_i (R_i) \, dR_i \tag{7}$$

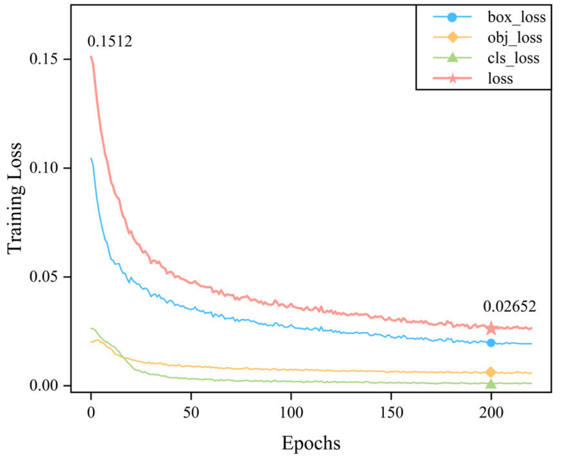


FIGURE 5. Training process of CEAM-YOLOv7.

TABLE 1. Model performance comparison.

Method	Model size (M)	*mAP*	*FPS*
Faster R-CNN	108	0.695	11
SSD	92.6	0.407	46
YOLOv3	117	0.671	62
YOLOv4	105.6	0.63	135
YOLOv5s	14.4	0.603	142
YOLOv7	74.9	0.612	66
CEAM-YOLOv7	10.6	0.736	156

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{8}$$

In the above equation, TP represents true positive samples, FP represents false positive samples, and FN represents false negative samples. In addition, P represents the number of true positive predictions in the overall prediction results, while R is the number of true positive predictions in all ground truths. *F1* score is the harmonic mean of P and R. A higher *F1* score indicates better target detection accuracy. AP evaluates the model's performance for each category by considering both P and R metrics. The mAP represents the average of AP and is used to measure the overall detection accuracy of the target detection algorithm. In summary, for the YOLO algorithm, the AP and mAP are the best metrics to measure the detection accuracy of the model.

C. EXPERIMENTAL ANALYSIS

1) CONVERGENCE ANALYSIS

To observe the convergence of CEAM-YOLOv7, we analyzed the training process. In the experiments, we set the initial parameters. The task is to identify four driver behaviors in the dataset. A mini-batch learning scheme of 16 per batch is used to speed up the training process. In an epoch, the model updates all parameters once after each mini-batch training is completed. Fig.5 illustrates the curve of training loss relative to the number of epochs. In this figure, the loss is a sum

TABLE 2. Effect of using different methods on the YOLOv7.

Method	Model size (M)	*mAP*@0.5	Parameters (M)	*FLOPs* (G)	*FPS*
YOLOv7	74.9	0.612	36.4	103.4	66
YOLOv7+CE	74.9	0.698	36.4	103.4	65
YOLOv7+AM	10.6	0.679	5.1	12.7	158
CEAM-YOLOv7	10.6	0.736	5.1	12.7	156

of box_loss, object_loss and class_loss. As the number of training epochs increases, the curve becomes flat. It indicates the convergence of CEAM-YOLOv7. Starting from epoch 200, the training loss of CEAM-YOLOv7 is basically stable.

2) ALGORITHM COMPARISON

We evaluated CEAM-YOLOv7 on NVIDIA GeForce RTX 2070 SUPER GPU and compared it with the two-stage detector Faster R-CNN and one-stage detectors SSD, YOLOv3, YOLOv4, YOLOv5s and the original YOLOv7. The detailed results are shown in Table 1.

First, it can be seen that the model size of CEAM-YOLOv7 is 10.6 M, which is easy to deploy on in-vehicle terminals and can be used for vehicle-side real-time detection. The parameters in the training process are 5.1 M, and the *FLOPs* are 12.7 G. Therefore, our model is trained faster and easy to deploy on hardware devices. Secondly, the *mAP* of CEAM-YOLOv7 reaches 0.736, which is significantly higher than other methods. As shown in Fig.6, the *AP* in all categories is higher than other methods, and the hard case 'Safe' is significantly improved, which proves the effectiveness of the new network structure. Finally, using *FPS* as an index to evaluate the object detection speed shows that our method can meet the real-time requirements for detection, especially faster than the two-stage detector Faster R-CNN by 14 times. Overall, our method has high accuracy for IR image detection and can achieve a balance between recognition accuracy and speed. The model size is suitable for deployment on in-vehicle terminals and has application meaning.

We visualize the detection results, as shown in Fig.7. Our method successfully identifies four types of driver behaviors, including Danger, Drink, Phone, and Safe, with high recognition accuracy and almost no missed and false detection.

D. ABLATION STUDY

In order to distinguish the respective features of the methods more visually, ablation experiments were conducted for all the proposed optimization methods, and the results are shown in Table 2.

1) EFFECT OF CE ALGORITHM

With data augmentation operation, the CE algorithm extends the number of channels of IR images from 1 to 3, which greatly increases the amount of information. In table 2, the *mAP* of YOLO+CE is increased to 0.698, which is 14.05% higher than that of YOLOv7. However, *FLOPs* are almost unchanged. Even compared with YOLOv7+AM, the *mAP* slightly improved from 0.679 to 0.698. Obviously, the addition of CE can improve recognition accuracy without increasing computation.

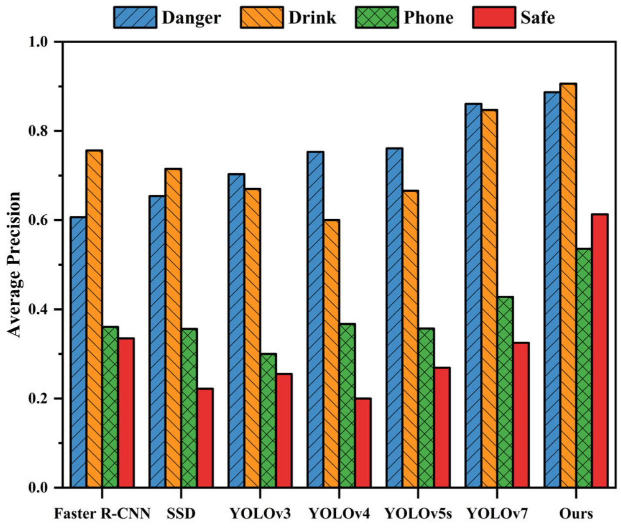


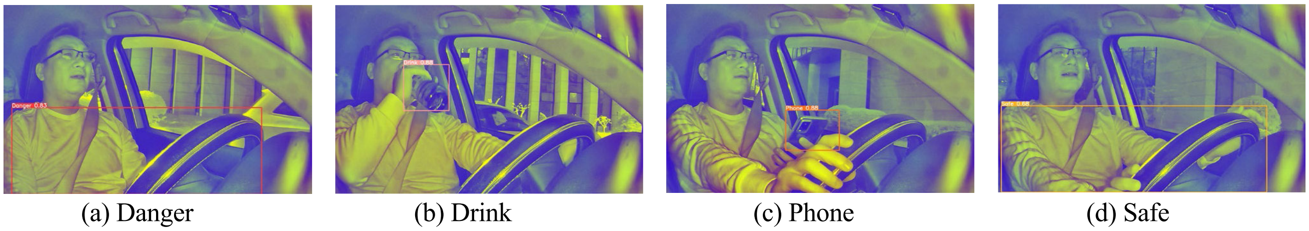
FIGURE 6. Comparison of the AP of different models.

2) EFFECT OF GAM

With the insertion of attention mechanism and layer prune, model size and parameters are significantly reduced by more than 80%. Moreover, compared with YOLO, the *FPS* of YOLO7+AM is nearly 2.5 times higher, and *FLOPs* rapidly drop to 12.7. These sufficiently indicate that the GAM module and layer prune can achieve an excellent balance between *FPS* and *mAP*.

3) EFFECT OF MODEL ENSEMBLE

Fig.8 shows the visualization results of the ablation experiments on the YOLOv7 model. It is observed from the results that the *mAP* of CEAM-YOLOv7 increased by 23.20%. According to the analysis of each recognition category, it can be found that the impact of algorithm optimization on the detection performance of each category is different, where the biggest improvement is in the 'Safe' category, with the *mAP* doubled. The 'Phone' category is raised to about 0.6. And the recognition ability of 'Danger' and 'Drink' remains better, with *mAP* staying above 0.875. The *mAP* indexes indicate that the method has achieved good results in target identification. Meanwhile, the small model size means faster network



(a) Danger (b) Drink (c) Phone (d) Safe

FIGURE 7. Detection results.

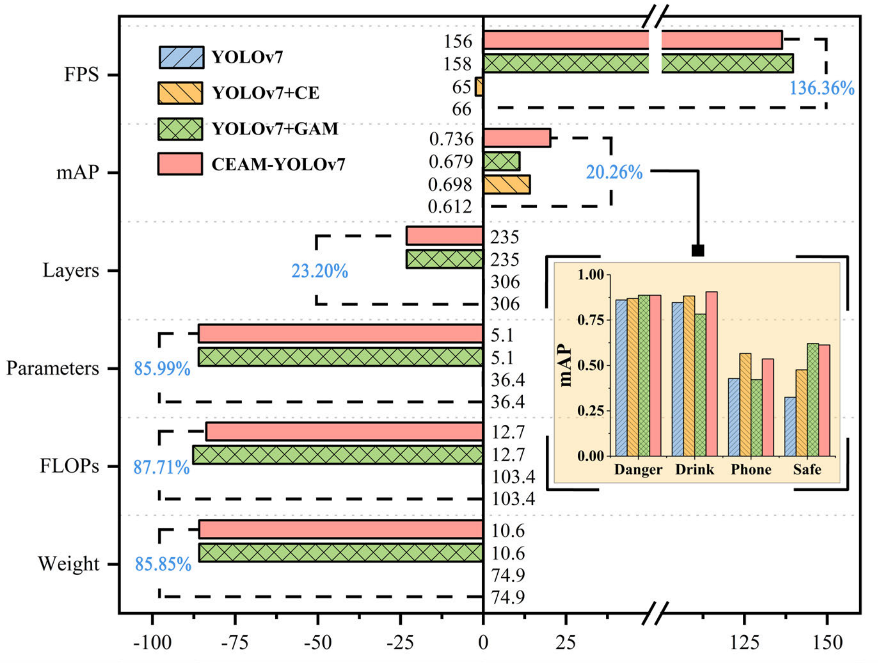


FIGURE 8. Comparison of indexes after YOLOv7 optimization. The black numbers are the evaluation index values, and the bolded blue numbers are the change rates of the index after optimization.

training speed and lower training equipment requirements. Our method can be easily deployed on in-vehicle terminals with an *FPS* of 156, which meets the requirements of real-time vehicle-side detection.

V. CONCLUSION

In this study, we propose the CEAM-YOLOv7, which outperforms the existing single-stage detections. The GAM module inserted into the network promotes the feature extraction ability of driver behavior. Layer prune operation makes IR image features easier to be extracted and models easier to be deployed. The data augmentation strategy optimizes the dataset through the CE algorithm. Based on the driver distraction IR images dataset of HNUST and HNU, the trained model can better adapt to the light changes of driving scenes. The experimental results show that the method has a fast detection speed of 156 *FPS*, and the mAP increases by 20.26% over the original YOLO7 network. The trained model is small in size and can be easily deployed on in-vehicle terminals for real-time driver behavior recognition.

There are many more distraction behaviors of drivers with different manifestations from person to person. We plan to explore further a more comprehensive object detection model and deploy it on in-vehicle terminals.

REFERENCES

[1] SAE On-Road Automated Vehicle Standards Committee, "Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems," *SAE Standard J.*, vol. 3016, pp. 1–16, Jan. 2014.
[2] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, and C. Gläser, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.
[3] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[5] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[8] 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.

[9] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[10] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[11] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[12] J. Glenn. *YOLOv5*. Accessed: Jun. 9, 2020. [Online]. Available: https://github.com/ultralytics/yolov5/releases/tag/v6.1.2022

[13] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[15] L. Zhao, F. Yang, L. Bu, S. Han, and G. Zhang, "Driver behavior detection via adaptive spatial attention mechanism," *Adv. Eng. Inf.*, vol. 48, Apr. 2021, Art. no. 101280.

[16] 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.

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

[18] Y. Xing, C. Lv, D. Cao, and E. Velenis, "Multi-scale driver behavior modeling based on deep spatial–temporal representation for intelligent vehicles," *Transp. Res. C, Emerg. Technol.*, vol. 130, Sep. 2021, Art. no. 103288.

[19] B. Ghizlene, M. Zoulikha, and H. Pomares, "An efficient framework to detect and avoid driver sleepiness based on YOLO with Haar cascades and an intelligent agent," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, 2019, pp. 699–708.

[20] Z. Zhao, K. Hao, X. Ma, X. Liu, T. Zheng, J. Xu, and S. Cui, "SAI-YOLO: A lightweight network for real-time detection of driver mask-wearing specification on resource-constrained devices," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–15, Nov. 2021.

[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] S. Chen, X. Xu, N. Yang, X. Chen, F. Du, S. Ding, and W. Gao, "R-Net: A novel fully convolutional network–based infrared image segmentation method for intelligent human behavior analysis," *Infr. Phys. Technol.*, vol. 123, Jun. 2022, Art. no. 104164.

[23] S. Yao, Q. Zhu, T. Zhang, W. Cui, and P. Yan, "Infrared image small-target detection based on improved FCOS and spatio-temporal features," *Electronics*, vol. 11, no. 6, p. 933, Mar. 2022.

[24] X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding, and S. Wen, "PP-YOLO: An effective and efficient implementation of object detector," 2020, *arXiv:2007.12099*.

[25] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[26] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13029–13038.

[27] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "You only learn one representation: Unified network for multiple tasks," 2021, *arXiv:2105.04206*.

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

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[30] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[31] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019, *arXiv:1908.08681*.

SHUGANG LIU received the B.Sc. degree in electronic information engineering from Hunan University, Changsha, China, in 2000, and the M.Sc. degree in software engineering from the Beijing University of Technology, Beijing, China, in 2005, and the Ph.D. degree in electronic science and technology from Hunan University, Changsha, in 2011. He is currently a Senior Lecturer with the Hunan University of Science and Technology. His research interests include deep learning, semantic communication, and embedded systems.

YUJIE WANG received the B.Eng. degree in measurement and control technology and instrument from the Hunan University of Science and Technology, Xiangtan, China, in 2021. He is currently pursuing the master's degree with the School of Physics and Electronic Science, Hunan University of Science and Technology. His research interests include deep learning and embedded systems.

QIANGGUO YU received the B.Sc. degree in electronic information engineering from Hunan University, Changsha, China, in 2000, and the M.Sc. degree in software engineering from Central South University, Changsha, in 2016. Since 2020, he has been with the Huzhou College, where he is currently a Senior Engineer. His current research interests include intelligent control and pattern recognition.

HONGLI LIU (Member, IEEE) received the B.Sc. degree in electrical engineering and the Ph.D. degree in control theory and engineering from Hunan University, Changsha, China, in 1985 and 2000, respectively. He is currently a Professor and the Department Head of the College of Electrical and Information Engineering, Hunan University. His research interests include intelligent information processing and transmission technology.

ZHAN PENG received the B.Eng. degree in electronic information science and technology from the Hunan University of Science and Technology, Xiangtan, China, in 2022. She is currently pursuing the master's degree with the School of Physics and Electronic Science, Hunan University of Science and Technology. Her research interest includes semantic communication.