electronics

MDPI

Article

HAR-Net: An Hourglass Attention ResNet Network for Dangerous Driving Behavior Detection

Zhe Qu [1], Lizhen Cui [1] and Xiaohui Yang [2,*]

[1] School of Software, Shandong University, Jinan 250100, China; quzhe@mail.sdu.edu.cn (Z.Q.); clz@sdu.edu.cn (L.C.)
[2] School of Information Science and Engineering, University of Jinan, Jinan 250022, China
* Correspondence: ise_xhyang@ujn.edu.cn

Abstract: Ensuring safety while driving relies heavily on normal driving behavior, making the timely detection of dangerous driving patterns crucial. In this paper, an Hourglass Attention ResNet Network (HAR-Net) is proposed to detect dangerous driving behavior. Uniquely, we separately input optical flow data, RGB data, and RGBD data into the network for spatial–temporal fusion. In the spatial fusion part, we combine ResNet-50 and the hourglass network as the backbone of CenterNet. To improve the accuracy, we add the attention mechanism to the network and integrate center loss into the original Softmax loss. Additionally, a dangerous driving behavior dataset is constructed to evaluate the proposed model. Through ablation and comparative studies, we demonstrate the efficacy of each HAR-Net component. Notably, HAR-Net achieves a mean average precision of 98.84% on our dataset, surpassing other state-of-the-art networks for detecting distracted driving behaviors.

Keywords: dangerous driving behavior detection; driving assistant; vehicle technology; gesture recognition; deep learning

Citation: Qu, Z.; Cui, L.; Yang, X. HAR-Net: An Hourglass Attention ResNet Network for Dangerous Driving Behavior Detection. *Electronics* **2024**, *13*, 1019. https://doi.org/10.3390/electronics13061019

Academic Editor: Daniel Gutiérrez Reina

Received: 14 December 2023
Revised: 22 February 2024
Accepted: 6 March 2024
Published: 8 March 2024

1. Introduction

Normal driving behavior is paramount for ensuring safe driving. Statistics and analysis have revealed that driver distraction accounts for over 70% of traffic accidents [1]. Engaging in dangerous driving behaviors like eating, drinking, using mobile phones, talking, smoking, and other distractions significantly elevates the risk of accidents. Consequently, detecting such behaviors has emerged as a pivotal research area in intelligent transportation systems.

Typically, a driver's state can be discerned through their hand movements during driving. However, this endeavor presents unique challenges. Firstly, distinguishing between various hand behaviors is inherently more challenging than recognizing broader body postures or distinct features. Additionally, collecting hand data often encounters issues like external occlusions. Secondly, the driving environment itself poses challenges, including varying backgrounds, light intensities, and image jitters caused by vehicle movement. Despite these obstacles, the importance of accurate hand-based driving behavior detection remains paramount for enhancing road safety.

Recently, the technology of deep learning has advanced significantly, sparking the interest of numerous researchers across various domains such as lane detection [2], distracted driver classification [3], object recognition [4], and data envelopment analysis [5]. In our study, we categorize dangerous driving behaviors into five distinct groups: eating, drinking, smoking, making phone calls, and playing with a phone. To address this challenge, we introduce a novel network named Hourglass Attention ResNet Network (HAR-Net) for dangerous driving behavior analysis.

In the HAR-Net architecture, we separately process optical flow data, RGB data, and RGBD data, enabling spatial–temporal fusion. For spatial fusion, we leverage the strengths of both ResNet-50 and the hourglass network, integrating them as the backbone of

CenterNet. Additionally, to enhance the model's performance, we incorporate an attention mechanism and integrate center loss with the traditional Softmax loss.

The main contributions of this study are as follows:

(1) We have proposed HAR-Net, a deep learning architecture designed specifically for identifying dangerous driving behaviors. Within this framework, optical flow data, RGB data, and RGBD data are individually inputted into the network, facilitating spatial–temporal fusion for enhanced analysis.
(2) We have gathered data on dangerous driving behaviors and labeled them according to five distinct categories to create a comprehensive dataset for our study.
(3) Comprehensive experiments have been conducted on three datasets, and the results of ablation and comparison experiments demonstrate that the proposed method significantly outperforms the baseline methods in terms of performance.

2. Related Works

2.1. Target Detection

Generally, the detection of dangerous driving behavior is a type of target detection. Existing target detection methodologies can be broadly categorized into two types: the two-stage target detection approach and the one-stage target detection approach.

The two-stage target detection method involves generating candidate regions using an algorithm and subsequently classifying these samples by using a convolutional neural network. For example, in the region-based convolutional neural network (R-CNN) algorithm proposed in [6], the candidate regions of all samples are extracted using a selective search algorithm and then classified using a convolutional neural network. However, a significant drawback of this selective search algorithm is that it often extracts a considerable amount of redundant information which can hinder the achievement of optimal results. To address this issue, the Fast RCNN algorithm proposed in [7] combines region of interest (ROI) pooling with a selective search algorithm to reduce redundant information. Additionally, Fast R-CNN incorporates a multitask loss function that integrates candidate region classification loss and location regression loss, directly incorporating boundary regression into CNN training. While this approach mitigates the issue of redundant information, it still faces challenges in terms of computational efficiency as it requires significant time to calculate the candidate regions. Subsequently, the Faster RCNN algorithm is proposed in [8]. Faster R-CNN integrates feature extraction, bounding box representation, and classification into a unified network, resulting in a substantial improvement in detection speed.

The one-stage target detection algorithm eliminates the need for generating candidate boxes by directly framing the target boundary location problem as a regression task. The You-Only-Look-Once (YOLO) algorithm proposed in [9] takes the entire image as input to the CNN and directly regresses the position and category of the bounding box in the output layer. It divides the images into S $\times$ S grid cells. Any grid cell containing the center of an object becomes responsible for predicting that object. Redmon et al. [10] propose the YOLOv3 algorithm, incorporating multiscale prediction and DarkNet53. In [11], the YOLOv4 algorithm is proposed, and its backbone is CSPDarknet53. CSPDarknet5 changes the leak ReLU activation function to the Mish activation function [12] and adds the idea of cross-stage concatenation. It also adds the spatial pyramid pooling algorithm [13] in the middle of the network to strengthen the feature information. It can be observed that the YOLOv4 algorithm shows advantages in accuracy and real-time computation, but it relies on a large number of anchors. The CenterNet used in this study does not need to distinguish whether an anchor is an object or background. Each target corresponds to only one anchor, which is extracted from the heatmap. In other words, it is an anchor-free target detection network with greater advantages in terms of speed and accuracy.

2.2. Dangerous Driving Behavior Detection

Previous studies examining dangerous driving behaviors, particularly those focused on drivers' hand movements, have predominantly employed traditional algorithm-based,

device-based, and deep learning-based methodologies. In terms of research based on traditional algorithms, Zhao et al. [14] propose a method to extract driver behavior features based on homomorphic filtering, skin-like region segmentation, and the contour wave transform. The driving posture dataset is used to extract the features. When comparing the performance of various classification algorithms, such as the random forest classifier, k-nearest neighbor (KNN), and multilayer perceptron (MLP), the random forest classifier emerges as the most effective. Within a driver-centered driving assistance system, the classification accuracy for the class "Eating" using a radio frequency classifier is greater than 88%, proving the effectiveness of the method. Furthermore, the study employs a pyramid gradient histogram to capture local features and establishes the SEU-DP driving posture dataset for training and testing purposes. This dataset contains "operate the steering wheel and gear lever", "eat cake", and "use mobile phone". Finally, multilayer perception classifiers are used to classify and predict different driving behaviors [15]. Although research is often based on traditional algorithms, research can also be conducted with multilevel and wide-field classification. It often demands significant computational resources and has difficulty achieving accurate recognition requirements.

Device-based research utilizes various instruments to gather depth and sensor information for analyzing hand movements. A popular device in this domain is the wearable data glove, equipped with numerous sensors. These gloves leverage magnetic positioning sensors to pinpoint the wearer's hand location in three-dimensional space, precisely detecting the global hand position, finger joint positions, and the extent of finger bending. This allows for the quantification of any hand action, enabling the identification of distinct hand behaviors. Fang et al. [16] proposes a new data glove for gesture capturing and recognition based on inertial and magnetic measurement units. This data glove is composed of a three-axis gyroscope, three-axis accelerometer, and three-axis magnetometer. Three-dimensional movements that include the arms, palms, and fingers are completely captured by data gloves. The captured data can be combined with extreme learning machines for hand detection and behavior recognition. This method accurately distinguishes subtle differences between different actions in scenarios that involve high speed and wide application. However, it should be noted that the data glove is an intrusive device, potentially affecting the user's natural behavior during use. Consequently, to preserve normal driving behavior, this method is not considered suitable for use in the driving environment.

In recent years, with the remarkable advancement of deep learning, numerous studies have focused on analyzing dangerous driving behaviors. Jha et al. [17] propose a formulation based on probabilistic models to determine salient regions for driver's visual attention description. A bidirectional posture–appearance interaction network (BPAI-Net) is proposed in [18], and in their method, RGB frames and skeleton data are adopted for driver behavior recognition. Ansari et al. [19] propose a driver mental fatigue and drowsiness detection method by monitoring drivers' head posture motions. Benjamin et al. [20] propose a driver posture classification system to detect whether the driver is using a mobile phone or eating food. Simultaneously, a new dataset is established to train and evaluate different learning models. The dataset is captured using two infrared cameras, achieving accuracies of 92.88% and 90.36% for the left and right-side camera data, respectively. In summary, the method based on deep learning is widely used and has a high fault tolerance, which is the most suitable method at present. However, we find that many networks lose a considerable amount of information in the process of feature extraction. To correct this defect, our study improves the entire network and proposes HAR-Net for the detection of dangerous driving behavior.

3. Related Improvements

To improve the precision of the network, this study adopts some existing improvement methods to improve the structures of the hourglass network and ResNet-50.

3.1. Hourglass Network

The hourglass network derives its name partly from its distinctive shape. During feature processing, the network initially down-samples the input, subsequently up-samples it, and finally incorporates skip connections to enhance its performance. In this study, we incorporate dilated convolution into the hourglass network, utilizing it as a fundamental unit. This modification aims to minimize down-sampling while preserving maximum information. Zhang et al. [21] suggest that the design goal of dilated hourglass models (DCM) is to make full use of different feature levels and reduce information loss. Traditional methods tend to use up/down-sampling to expand the perception domain and obtain high-level features. However, this up/down-sampling structure usually leads to the loss of information and resolution; thus, it has a significant impact on the accuracy of the determined target position. To address this limitation, dilated convolution is employed as the fundamental component, enabling the expansion of the receptive field while preserving the original resolution.

The DCM module consists of three layers, as shown in Figure 1. Skip connection connects two ordinary convolution layers, and the output of the final module is the addition of skip connection and the extended convolution layer output. The improved hourglass network uses DCM to replace the residual module; this replacement reduces the subsampling time and information loss. The relationship between the input size and output size of the network with dilated convolution is described in Equation (1).

$$W_2 = \frac{W_1 + 2p - d(k-1) - 1}{s} + 1 \tag{1}$$

where W_1 is the input feature map size, W_2 is the output feature map size, p is padding, d is dilation, and k is the kernel size. In the proposed model, the dilation coefficient is the cyclic structure of [1, 2, 5, 1, 2, 5]. The advantages of employing dilated convolution can be summarized as follows: Firstly, it efficiently enlarges the receptive fields of the network. Secondly, it mitigates gridding issues that arise from the stacking of multiple identical dilated convolutions.

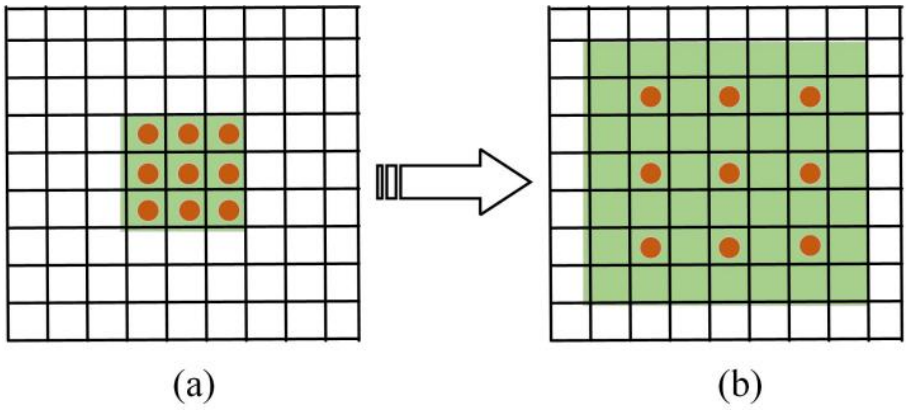

Figure 1. Dilated convolution analytic graph. (**a**) Ordinary convolution: 1-dilated convolution. (**b**) Dilated convolution: 2-dilated convolution. It is mainly composed of a 3×3 expanded convolution layer and two 3×3 ordinary convolution layers. The extended convolution layer is located between two ordinary convolution layers, and the expansion rate is 2.

3.2. ResNet-50 Network

In ResNet [22], the residual structure and cross-layer connection are added to the network to solve the problem of exploding gradient and degradation.

The down-sampling block of ResNet has two paths: path A and path B. Path A has two 1×1 convolution layers and one 3×3 convolution layer; we call this a bottleneck structure.

Path B has a 1 × 1 convolution layer in steps of 2. The output of the final down-sampling block is the sum of the two paths' outputs.

Through the ResNet structure described previously, it can be found that in the down-sampling part (as shown in Figure 2), path A has the first convolution with a stride of 2 whose convolutional kernel is 1 × 1, so three-quarters of the characteristic information is ignored. Similarly, the convolution layer in path B also ignores three-quarters of the characteristic information. In [23], the authors adjust the stride of the two convolution layers in path A to tackle these challenges. Specifically, they assign a stride of 2 to the 3 × 3 convolution layer while maintaining a stride of 1 for the remaining convolution layers. Additionally, they enhance path B by setting the stride of the convolution layer to 1 and introducing a 2 × 2 average pooling layer with a stride of 2 before the convolution layer. In our study, we apply this approach to refine ResNet-50. Our experimental results indicate that this modification has minimal impact on computational cost while significantly boosting accuracy.

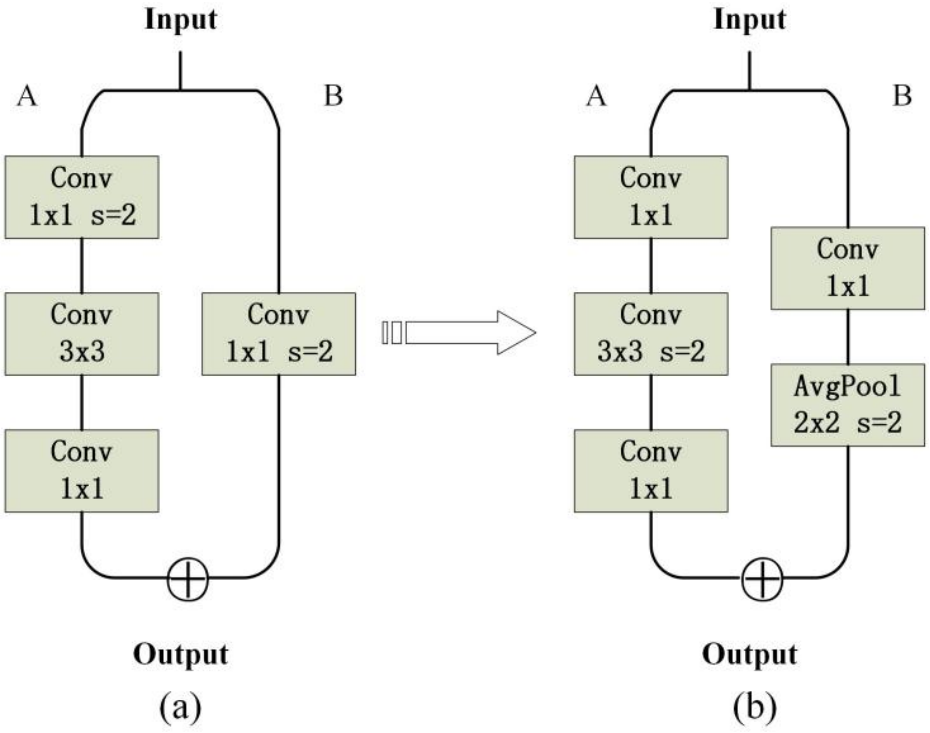

Figure 2. Down-sampling structure in ResNet. (**a**) The original network structure. (**b**) The improved network structure.

4. Proposed Method

In this study, we design an efficient and accurate target detection classifier based on CenterNet [24] that improves the accuracy index and accelerates the network processing speed. Compared with YOLO, SSD, and Fast R-CNN, the detection network relies on a relatively large number of anchors. CenterNet does not need to distinguish whether the anchor is an object or a background; each target corresponds to only one anchor, which is extracted from the heatmap. In other words, it is an anchor-free target detection network, which has advantages in terms of speed and accuracy.

4.1. Dataset

Before conducting the experiments, we gather and preprocess data to streamline the design, training, and assessment of our model. To investigate driving behavior, we compile information on various activities such as smoking, drinking, eating, talking on the phone, and playing with a phone. Subsequently, we label and categorize the dataset, using it to

evaluate our proposed network through both training and testing sets. In our study, we adopt the Philips CVR300 automobile data recorder, which captures high-definition color video streams at a rate of 30 frames per second, with a resolution of 1920 × 1080.

(1) Data Collection: As depicted in Figure 3, we position the automobile data recorder at the top right corner of the steering wheel and on the passenger side window to capture driving behavior videos of 14 drivers, 5 females and 9 males, ranging in age from 17 to 22 years old and with driving experience varying between 1 and 5 years. The driving behaviors observed in our study encompass smoking, drinking, eating, talking on the phone, and playing with a phone. Data are predominantly collected during daylight hours but also include sequences in dark and complex lighting conditions. After filtering the recorded video footage, we convert each frame into individual images and save them for further analysis. Figure 4 provides an illustrative example of one such saved image.

(2) Data Preprocessing: Incorporating optical flow into the model has been shown to be an effective way to improve accuracy [25,26]. Optical flow is the instantaneous speed of a spatially moving object moving in pixels, and it is a method used to calculate the motion information of an object between adjacent frames. In general, optical flow is the projection of the motion of an object in three-dimensional space on a two-dimensional pixel plane, which is generated by the relative velocities of the object and the camera. It reflects the moving direction and speed of the image pixels corresponding to the object in a very small time. Therefore, the optical flow map can unambiguously describe the short-term movements of the driver. In our study, the method of [27] is adopted to obtain the optical flow for every video frame in the dataset, as shown in Figure 5.

Figure 3. Dash camera placement. Dash cameras were placed in two locations to collect data from the front and from the side.

Figure 4. Our dataset. The upper part is collected from the side, and the lower part is collected from the front. From left to right are drinking, eating, phone playing, phone talking, and smoking.

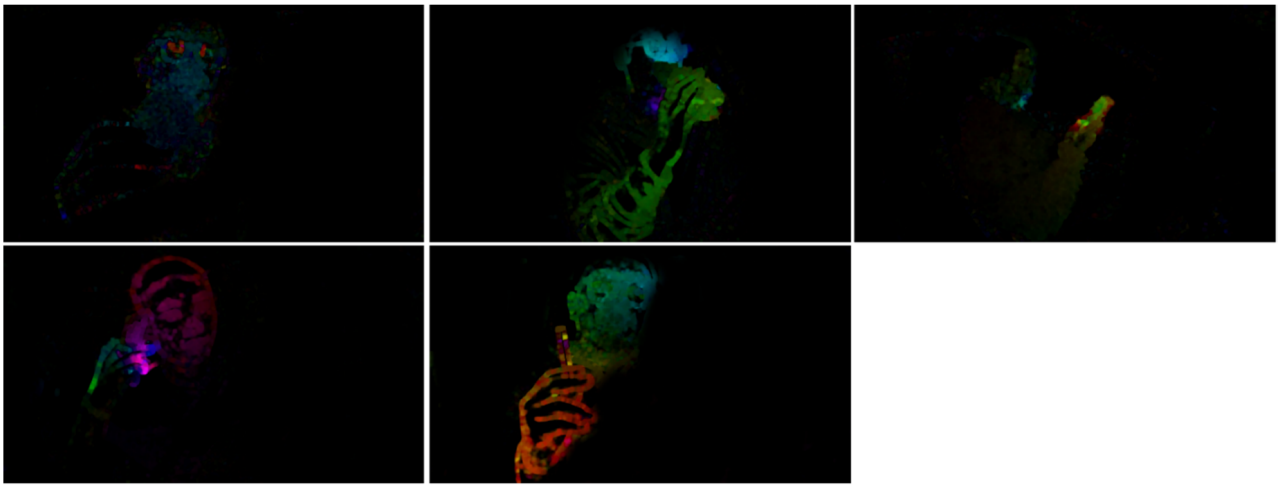

Figure 5. Optical flow data of five dangerous driving behaviors. The optical flow is calculated using the sequential frames in the driving video, and the color in the figure reflects the moving direction and speed of the image pixels corresponding to the object in a very small timeframe.

We use depth data in the model based on prior knowledge to make better use of the data. Depth images record the distance from the camera to points in the scene, reflecting the geometry of objects in the scene. The depth data are obtained using the method of [28], which is shown in Figure 6. Our model will jointly use optical flow information, depth information, and raw video frames as the input data.

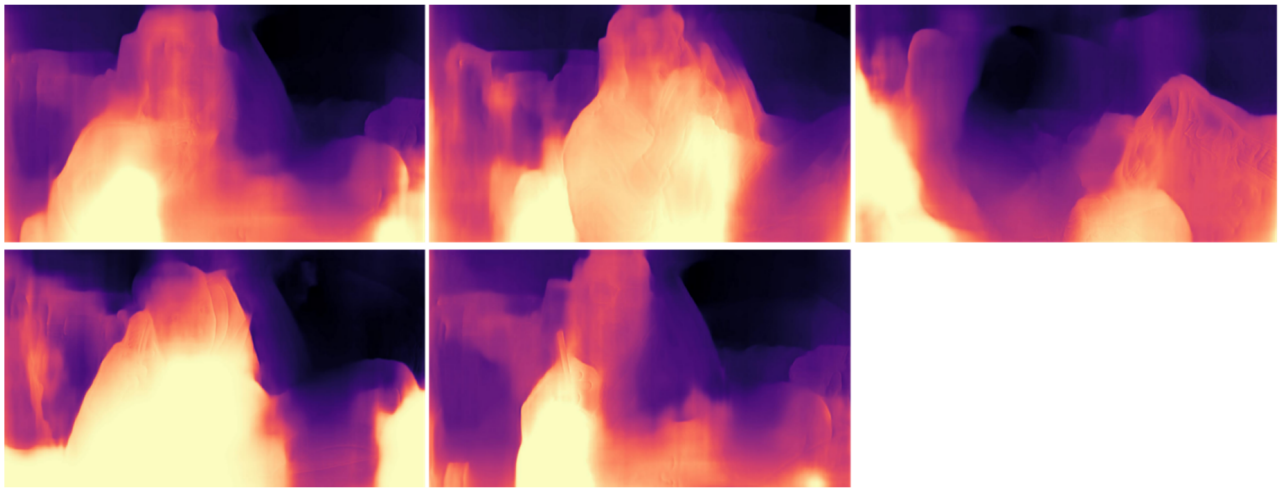

Figure 6. Depth data of five dangerous driving behaviors.

4.2. Overall Framework

We introduce a sophisticated deep neural network architecture, named HAR-Net, designed to detect dangerous driving behaviors. Our innovative approach involves leveraging multichannel input data of various types, including optical flow, RGB, and depth information. By integrating temporal and spatial details through fusion layers, our model enhances data utilization, leading to improved performance. The comprehensive structure of HAR-Net is depicted in Figure 7.

Considering the need for temporal fusion of RGB features and optical flow features, we choose to use the same ResNet50 in the RGB channel and the optical flow channel. After the features' temporal fusion, they are fused with the RGBD features extracted by the hourglass network in the spatial domain and output. The above is the overall framework

of our model, and the detailed structure of each component of the HAR-Net model will be described in subsequent chapters.

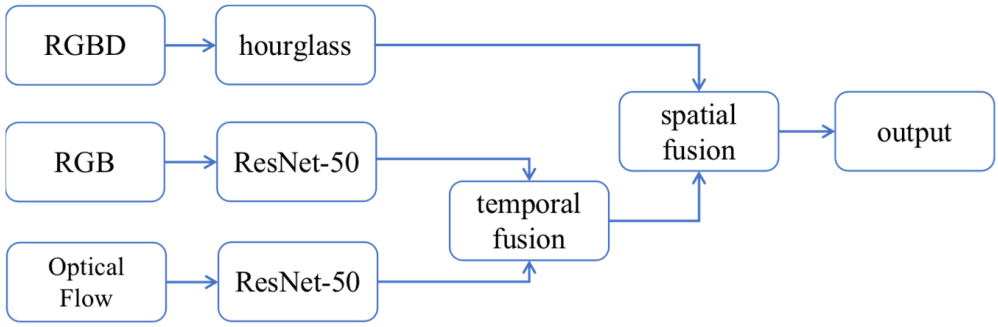

Figure 7. Overall process of the network.

4.3. Network Fusion

The original CenterNet network has three backbones for target detection: ResNet-18, Hourglass-104, and DLA-34. ResNet50 has been widely used in recent years. However, with an increase in the feature layer, the context information and global relationship gradually decreases, which may greatly reduce the performance of the model. On the other hand, the resolution of the feature map cannot meet the pixel-level requirements. We combine ResNet-50 and the hourglass network as the backbone of CenterNet because the hourglass network has a greater advantage in this regard. In this paper, we propose two network combination methods, as follows.

(1) Loss value combination: The loss value represents the distance between the model output and the real result. The calculation method for this distance is defined by the loss function. This combination method, based only on the loss value, is the most basic combination method. The data pass through ResNet-50 and the hourglass network in parallel, and features are extracted to obtain the corresponding loss value and combine it. The combination method is obtained as follows:

$$L = L_1 + L_2 \tag{2}$$

where L is the loss value after the combination, and L_1 and L_2 are the loss values through ResNet-50 and the hourglass network, respectively.

(2) Feature map combination: According to the above introduction, ResNet-50 loses a large amount of information as the number of network layers increase. Thus, the size of the feature map will be reduced. In this case, to combine the feature maps of the two networks, the first task is to populate the feature maps' output by ResNet-50. For example, the size of the data input to the model is 512 × 512, the size of the feature map A output by the hourglass network is also 512 × 512, and the feature map B output by ResNet-50 is only 16 × 16. Therefore, we should first pad feature map B to a size of 512 × 512 and then combine it with feature map A to obtain feature map C for the detection and recognition. Figure 8 shows the framework of the combination of the two networks.

The advantage of the feature graph combination method is that it does not simply combine the results of the two networks, but combines the extracted feature graphs of the two networks. This is equivalent to combining good features and removing bad features to achieve better recognition and classification results.

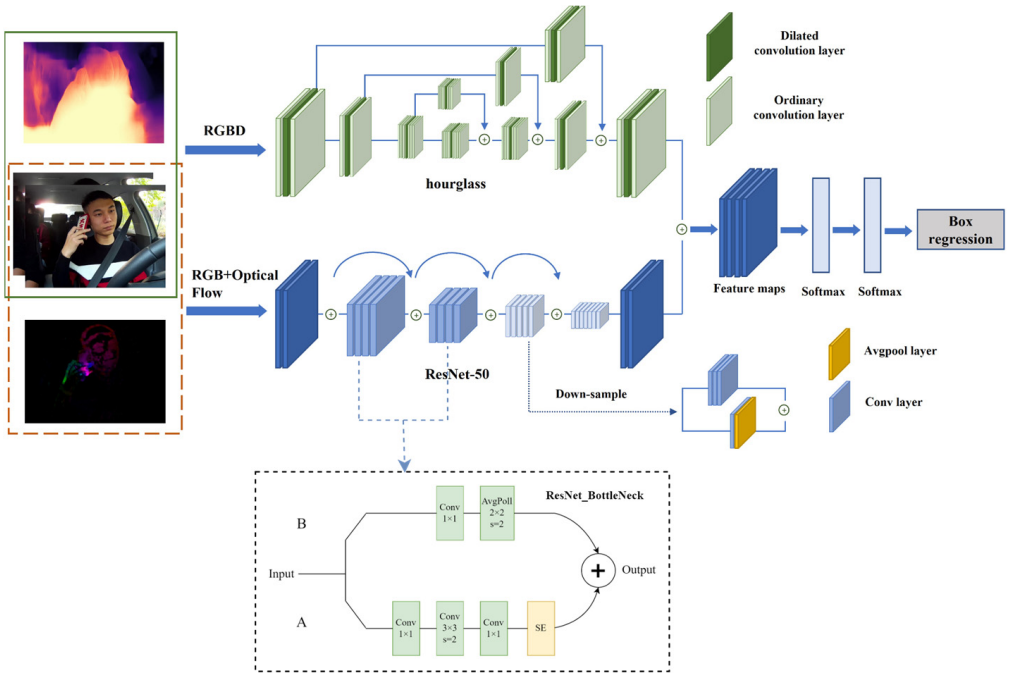

Figure 8. The framework of HAR-Net.

4.4. Attention Mechanism

The attention mechanism mentioned in this study is the squeeze-and-excitation network (SENet), which filters the attention of the channel by learning the correlation between channels. The increased computational workload can be negligible [29]. The core idea of SENet is to learn the feature weights through network loss. The weight of an effective feature map is large, and the weight of an invalid feature map is small; thus, good results can be achieved. Certainly, the SE block embedded in some original classification networks inevitably increases the number of parameters and the computation, but this increase is acceptable considering the obtained effect.

The general principle of the module is as follows: The SE block processes the feature map and obtains a one-dimensional vector as the evaluation score. The score is then applied to the corresponding channel, and only one block is added to the original basis. In the proposed model, an attention mechanism is added to the last layer of the ResNet-50 and hourglass networks.

4.5. Loss Function Reconstruction

The loss function serves as a metric to quantify the discrepancy between the model's predictions and the actual values. Its purpose is to establish a benchmark that facilitates the optimization of parameters during the training phase, ultimately aiming to attain the highest possible accuracy for the network. This concept is intuitively understandable and akin to many real-life scenarios. For instance, when parking a car, we rely on the rear-view mirror and adjust the steering wheel based on the parking lines visible. These lines serve as a reference, analogous to the loss function in machine learning, guiding the model toward convergence during training.

After the above improvements, the accuracy is significantly improved. However, in terms of model complexity, even if each improvement adds only a small amount of calculation, the impact on model speed is evident. Therefore, we improve the running speed of the model from the perspective of loss function reconstruction. In this study, we propose the use of both SoftMax loss and center loss. First, SoftMax loss is used to separate different categories, and then center loss is used to compress the same category and finally obtain discriminative features.

SoftMax loss is one of the most common loss functions. It is a combination of SoftMax loss and cross-entry loss. The formula is as follows:

$$L_s = \sum_{j=1}^{T} y_j log S_j \tag{3}$$

where L_s is the SoftMax loss. S_j is the j-th value of the output vector S of SoftMax, which represents the probability that this sample belongs to the j-th category. y is a $1 \times T$ vector (only the value of the position corresponding to the real label is 1, and the other $T - 1$ values are 0). Therefore, this formula has a simpler form:

$$L_s = -log S_j \tag{4}$$

The principle of center loss is to set several center points for classification so features of different categories are as close to their center points as possible. In other words, it is hoped that the within-class distance will become smaller and the between-class distance will become larger. The formula is as follows:

$$L_C = \frac{1}{2} \sum_{i=1}^{m} \parallel x_i - C(y_i) \parallel_2^2 \tag{5}$$

where x refers to the feature, and C is the category center corresponding to each sample in each batch. Similar to the dimension of feature x, it is updated with model training. It should be noted that the change in each category center is calculated only by the picture characteristics belonging to this category.

The SoftMax loss and center loss are used together in the model to obtain L_1 and L_2 in (2), i.e., k = 1 or 2 in (6). And λ is the weight, which is the best parameter obtained from multiple experiments and prior knowledge. It can be formulated as follows:

$$L_k = L_S + \lambda L_C \tag{6}$$

After the experiment, the effect of the improved loss function on the accuracy level is not obvious, but the speed is significantly affected.

5. Experiments

5.1. Evaluation Metrics

To evaluate the performance of our proposed method, we use evaluation metrics commonly used in classification tasks: macro-average, micro-average, and mAP. Macro-average is used to calculate the index value of each class first and then calculate the arithmetic mean of all classes. Equations (7)–(9) represent the calculations of macro-precision, macro-recall, and macro-average, respectively.

$$Macro\_P = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{7}$$

$$Macro\_R = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{8}$$

$$Macro\_F = \frac{2 \times Macro\_P \times Macro\_R}{Macro\_P + Macro\_R} \tag{9}$$

Micro-average is used to establish a global confusion matrix for each instance in the dataset regardless of category and then calculate the corresponding indicators.

Equations (10)–(12) represent the calculations of micro-precision, micro-recall, and micro-average, respectively.

$$Micro\_P = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{10}$$

$$Micro\_R = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{11}$$

$$Micro\_F = \frac{2 \times Micro\_P \times Micro\_R}{Micro\_P + Micro\_R} \tag{12}$$

The evaluation index in this study is the mean average precision (mAP) [30]. The principle is used to calculate the AP of each category and take the average value. AP is measured for a single category, and mAP is measured for all categories.

In our study, we conduct experiments based on Python 3.8 and PyTorch 1.8, using a NVIDIA RTX2080 with 12 GB memory machine. Our model is trained with 70 epochs, and the batch size is 8. It takes about 53 min to train an epoch.

5.2. Ablation Experiment

To test the performance of the HAR-Net, we first complete an overall experiment using the dataset as described. Figure 9 shows the results of the overall network experiment. Figure 10 shows the performance of various driving behaviors. As shown in the figure, our network performs well for smoking, drinking, and phone talking, but there are also some errors for the eating and phone-playing behaviors.

In this study, we introduce several enhancements to various network modules. While the overall network experiments demonstrate improved performance, the specific contributions of individual modules remain unclear. To address this, we undertake ablation studies on each module, aiming to assess its impact on the overall network performance. The primary modules under investigation include the hourglass network, ResNet-50, and attention modules, which are systematically arranged and combined for the ablation tests. Table 1 presents a comparative analysis of the mean average precision (mAP) results obtained from these ablation experiments.

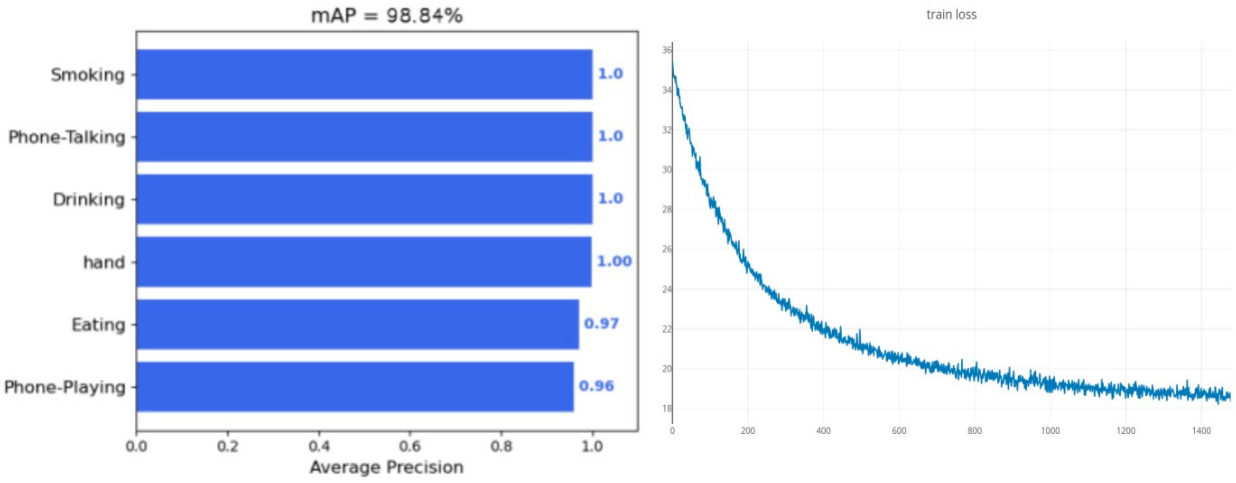

Figure 9. The results of the overall network experiment.

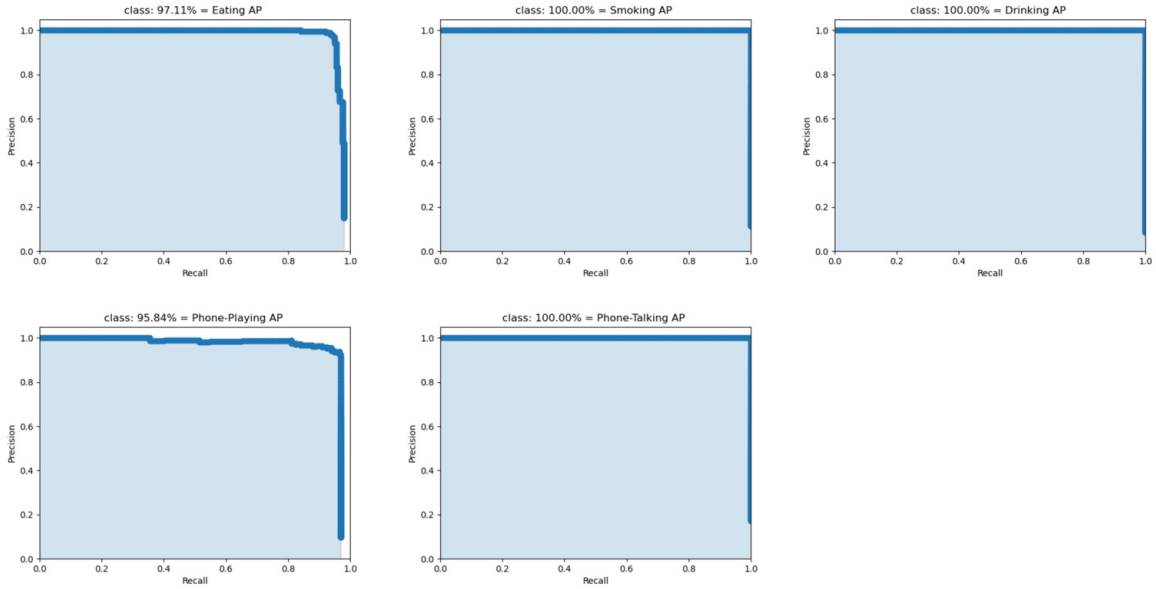

Figure 10. AP performance for various categories.

Table 1. Results of ablation experiments.

Networks	Modification Method	mAP (%)
Original hourglass network	—	88.17
Original ResNet-50	—	93.25
Modified hourglass network	Using extended convolution	92.14
Modified ResNet-50	Modified down-sampling	94.75
Attention mechanism	CenterNet + SE block	96.56
Loss combined network	Combine hourglass and ResNet-50 with loss value	98.72
HAR-Net	Combine hourglass and ResNet-50 at feature map	98.61

Lines 1 and 2 in Table 1 are the original hourglass network and ResNet-50 network, and lines 3 and 4 are the improved hourglass network and ResNet-50 network, respectively. It can be observed that the mAP is improved compared with the original networks. Line 5 represents the network with a separate attention mechanism [31], and lines 6 and 7 are the result of the two network combination methods. Specific results are shown in the tables. The ablation experiments demonstrate that the enhancements we introduced in each module result in improvements in mean average precision (mAP). The feature map combination method is better than the loss value combination method. This is because the feature map combination method does not simply combine the results of the two networks but combines the extracted feature maps of the two networks; this is equivalent to combining the useful features. Simultaneously, bad features are removed to achieve better recognition and classification results.

5.3. Comparative Experiment

After the ablation experiment, we compare HAR-Net with traditional target detection and several networks that have been reported in other papers. To compare the network comprehensively, we also process the CVPR-Hands 3D dataset [32] and StateFarm dataset, which are captured in dynamic driving environments, and utilize them for network training and testing with our dataset. During processing, the CVPR-Hands 3D dataset is classified into six types: normal driving, console manipulation, eating, texting, drinking, and reading. The StateFarm dataset is classified into five types: talking on the phone, playing with a phone, drinking, touching the face, and console manipulation.

The networks used in the comparative experiments are as follows: RetinaNet-101 [33], YOLOv3 [10], YOLOv4 [11], DDGNet-YOLO [34], Ensemble Inception V3 [35], and C-SLSTM [36]. Among them, RetinaNet-101, YOLOv3, and YOLOv4 are commonly used target detection networks. DDGNet-YOLO, C-SLSTM, and Ensemble Inception V3 are the networks mentioned in other papers related to dangerous driving behavior. Tables 2–4 show the experimental results on the three datasets.

Table 2. Results of comparative experiments on CVPR-Hands 3D dataset.

Network	CVPR-Hands 3D Dataset		
	Macro-Average	Micro-Average	mAP (%)
RetinaNet-101	0.76	0.77	75.7
YOLOv3	0.84	0.85	85.1
YOLOv4	0.85	0.87	87.3
DDGNet-YOLO	0.84	0.83	83.8
Ensemble Inception V3	0.81	0.82	81.22
C-SLSTM	0.82	0.83	82.59
HAR-Net	0.86	0.87	87.93

Table 3. Results of comparative experiments on StateFarm dataset.

Network	CVPR-Hands 3D Dataset		
	Macro-Average	Micro-Average	mAP (%)
RetinaNet-101	0.87	0.88	87.39
YOLOv3	0.95	0.95	96.10
YOLOv4	0.96	0.97	97.28
DDGNet-YOLO	0.96	0.96	96.52
Ensemble Inception V3	0.92	0.92	92.06
C-SLSTM	0.93	0.93	93.17
HAR-Net	0.97	0.97	97.23

Table 4. Results of comparative experiments on our dataset.

Network	CVPR-Hands 3D Dataset		
	Macro-Average	Micro-Average	mAP (%)
RetinaNet-101	0.88	0.88	88.15
YOLOv3	0.96	0.96	96.41
YOLOv4	0.97	0.97	97.79
DDGNet-YOLO	0.97	0.98	97.21
Ensemble Inception V3	0.91	0.92	92.32
C-SLSTM	0.92	0.93	93.06
HAR-Net	0.98	0.98	98.61

Upon analyzing Tables 2–4, it is evident that HAR-Net exhibits strong performance on our dataset, while YOLOv3 and YOLOv4 also achieve impressive results, surpassing 90% accuracy. However, when tested on the CVPR-Hands 3D dataset, all networks experience a decrease in mAP by less than 10%. We hypothesize that this discrepancy is primarily due to differences in data collection angles. Specifically, our dataset and the StateFarm dataset are captured from frontal and side perspectives, whereas the CVPR-Hands 3D dataset is predominantly collected from behind, making it more susceptible to sunlight interference. Additionally, the CVPR-Hands 3D dataset introduces a higher level of complexity with more classes, a wider range of motions, and more intricate actions, thus posing a greater challenge for target identification.

Despite these challenges, our HAR-Net maintains satisfactory performance, indicating the effectiveness of our implemented improvements in enhancing network performance. Nevertheless, when compared to the widely adopted YOLO networks, the

robustness of HAR-Net still has room for improvement, necessitating further research in this area. Figures 10–13 provide visualizations of HAR-Net's detection results across the three datasets.

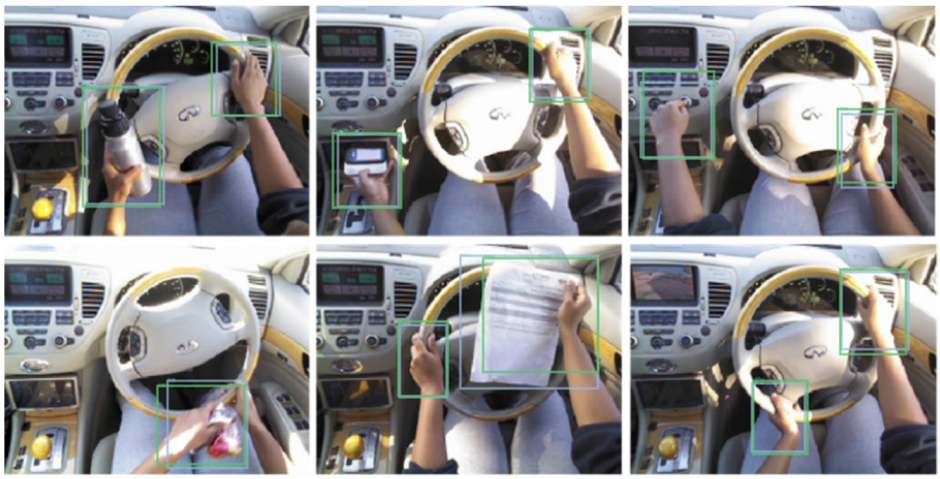

Figure 11. Detection results of HAR-Net on CVPR-Hands 3D dataset.

Figure 12. Detection results of HAR-Net on StateFarm dataset.

Figure 13. Detection results of HAR-Net on our dataset.

In terms of model speed, we utilize frames per second (FPS) as the evaluation metric. Following the refinement of the loss function, our proposed model achieves an FPS of 40, demonstrating its capability in real-time applications.

To more intuitively show the performance of the network, we draw the cumulative match characteristic (CMC) curve of these methods, as shown in Figure 14. The purpose of the CMC curve is to calculate a top-k hit probability, which is mainly used to evaluate the accuracy of the rank in the closed set. It can be observed from the curves that although all methods have similar detection results after rank 5, our method has a higher matching rate when the rank value is small.

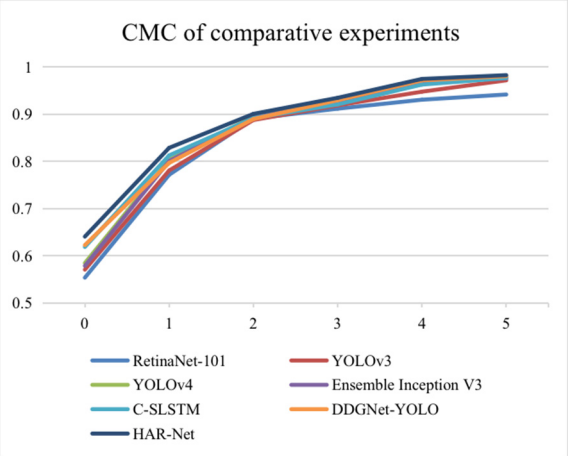

Figure 14. CMC of comparative experiments. CMC is utilized to calculate a top-k hit probability, which is mainly used to evaluate the accuracy of the rank in the closed set.

6. Conclusions

In this paper, we introduce HAR-Net, a deep learning-based network designed specifically for detecting dangerous driving behaviors. Our approach involves separately feeding optical flow, RGB, and depth data into the network for spatial–temporal fusion. For spatial fusion, we integrate ResNet-50 and the hourglass network as the foundation of CenterNet. The key findings from our research are summarized as follows:

1. To assess the network's performance, we construct a dataset featuring dangerous driving behaviors captured under natural conditions. The experimental results demonstrate that enhancements made to each module positively impact the model's overall performance.
2. To validate the efficacy of HAR-Net and mitigate any potential biases, we conduct a comparative analysis against traditional target detection methods and various networks mentioned in prior studies. We process and combine data from the CVPR-Hands 3D and StateFarm datasets with our own dataset for comprehensive network training and testing. HAR-Net achieves an impressive mAP of 98.84% on our dataset, surpassing the performance of other networks. However, we acknowledge some limitations in terms of robustness when applied to other datasets, an area we intend to focus on in future research.

Author Contributions: Conceptualization, Z.Q. and L.C.; methodology, Z.Q.; software, Z.Q.; validation, Z.Q., L.C. and X.Y.; formal analysis, Z.Q.; investigation, Z.Q.; resources, Z.Q.; data curation, Z.Q.; writing—original draft preparation, Z.Q.; writing—review and editing, Z.Q. and X.Y.; visualization, Z.Q.; supervision, L.C.; project administration, X.Y.; funding acquisition, L.C. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key R&D Program of China (No. 2021YFF0900800), the Shandong Provincial Key Research and Development Program (Major Scientific and Technological

Innovation Project) (No. 2021CXGC010108), the Shandong Provincial Natural Science Foundation (No. ZR202111180007), and the Fundamental Research Funds of Shandong University.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of University of Jinan (protocol code UJN-ISE-2024-001 and date of approval 5 March 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fitch, G.M.; Soccolich, S.A.; Guo, F.; McClafferty, J.; Fang, Y.; Olson, R.L.; Perez, M.A.; Hanowski, R.J.; Hankey, J.M.; Dingus, T.A. *The Impact of Hand-Held and Hands-Free Cell Phone Use on Driving Performance and Safety-Critical Event Risk*; DOT HS 811 757; NHTSA: Washington, DC, USA, 2013.
2. Liu, B.; Feng, L.; Zhao, Q.; Li, G.; Chen, Y. Improving the accuracy of lane detection by enhancing the long-range dependence. *Electronics* **2023**, *12*, 2518. [CrossRef]
3. Abbas, T.; Ali, S.F.; Mohammed, M.A.; Khan, A.Z.; Awan, M.J.; Majumdar, A.; Thinnukool, O. Deep learning approach based on residual neural network and SVM classifier for driver's distraction detection. *Appl. Sci.* **2022**, *12*, 6626. [CrossRef]
4. Yang, B.; Yang, S.; Zhu, X.; Qi, M.; Li, H.; Lv, Z.; Cheng, X.; Wang, F. Computer vision technology for monitoring of indoor and outdoor environments and HVAC equipment: A review. *Sensors* **2023**, *23*, 6186. [CrossRef] [PubMed]
5. Mirmozaffari, M.; Yazdani, M.; Boskabadi, A.; Ahady Dolatsara, H.; Kabirifar, K.; Amiri Golilarz, N. A novel machine learning approach combined with optimization models for eco-efficiency evaluation. *Appl. Sci.* **2020**, *10*, 5210. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
7. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
10. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2015**, arXiv:1804.02767.
11. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
12. Mish, M.D. A self regularized non-monotonic neural activation function. *arXiv* **2019**, arXiv:1908.08681.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
14. Zhao, C.H.; Zhang, B.L.; He, J.; Lian, J. Recognition of driving postures by contourlet transform and random forests. *IET Intell. Transp. Syst.* **2012**, *6*, 161–168. [CrossRef]
15. Zhao, C.H.; Zhang, B.L.; Zhang, X.Z.; Zhao, S.Q.; Li, H.X. Recognition of driving postures by combined features and random subspace ensemble of multilayer perceptron classifiers. *Neural Comput. Appl.* **2013**, *22*, 175–184. [CrossRef]
16. Fang, B.; Sun, F.; Liu, H.; Liu, C. 3D human gesture capturing and recognition by the IMMU-based data glove. *Neurocomputing* **2018**, *277*, 198–207. [CrossRef]
17. Jha, S.; Busso, C. Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions. *IEEE Trans. Intell. Veh.* **2023**, *8*, 59–72. [CrossRef]
18. Tan, M.; Ni, G.; Liu, X.; Zhang, S.; Wu, X.; Wang, Y.; Zeng, R. Bidirectional posture-appearance interaction network for driver behavior recognition. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 13242–13254. [CrossRef]
19. Ansari, S.; Naghdy, F.; Du, H.; Pahnwar, Y. Driver mental fatigue detection based on head posture using new modified reLU-BiLSTM deep neural network. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 10957–10969. [CrossRef]
20. Wagner, B.; Taffner, F.; Karaca, S.; Karge, L. Vision based detection of driver cell phone usage and food consumption. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 4257–4266. [CrossRef]
21. Zhang, Y.; Liu, J.; Huang, K. Dilated hourglass networks for human pose estimation. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 2597–2602.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 558–567.

24. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
25. Shi, Y.; Nie, X.; Liu, X.; Yang, L.; Yin, Y. Zero-shot hashing via asymmetric ratio similarity matrix. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 5426–5437. [CrossRef]
26. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
27. Sun, D.; Roth, S.; Lewis, J.P.; Black, M.J. Learning optical flow. In Proceedings of the 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 83–87.
28. Wang, C.; Buenaposada, J.M.; Zhu, R.; Lucey, S. Learning depth from monocular videos using direct methods. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2022–2030.
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
30. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
31. Ma, Z.; Yang, X.; Zhang, H. Dangerous driving behavior recognition using CA-CenterNet. In Proceedings of the 2nd IEEE ICBAIE, Nanchang, China, 26–28 March 2021; pp. 556–559.
32. Ohn-Bar, E.; Trivedi, M.M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2368–2377. [CrossRef]
33. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
34. Zhou, Y.; Lv, Z.; Zhou, Y. DDGNet-YOLO: A target detection network for dangerous driving gestures. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 3047–3052.
35. Eraqi, H.M.; Abouelnaga, Y.; Saad, M.H.; Moustafa, M.N. Driver distraction identification with an ensemble of convolutional neural networks. *J. Adv. Transp.* **2019**, *2019*, 4125865. [CrossRef]
36. Mafeni Mase, J.; Chapman, P.; Figueredo, G.P.; Torres Torres, M. Benchmarking deep learning models for driver distraction detection. In Proceedings of the International Conference on Machine Learning, Siena, Italy, 19–23 July 2020; pp. 103–117.