

Received 14 April 2024, accepted 6 May 2024, date of publication 13 May 2024, date of current version 20 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399789

## RESEARCH ARTICLE

# Recongnition of Distracted Driving Behavior Based on Improved Bi-LSTM Model and Attention Mechanism

ZHANFENG WANG<sup>1</sup> AND LISHA YAO<sup>2</sup>

<sup>1</sup>School of Computer Science and Artificial Intelligence, Chaohu University, Hefei, Anhui 238000, China

<sup>2</sup>School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui 230088, China

Corresponding author: Lisha Yao (jsjyaolisha@163.com)

This work was supported in part by the Key Scientific Research Project of Anhui Provincial Research Preparation Plan under Grant 2023AH051806, Grant 2023AH051807, Grant 2023AH052097, and Grant 2023AH052103; in part by Chaohu College Research Fund under Grant XLZ-202208; in part by the School Level Teaching and Research Projects under Grant ch21jxyj01; in part by the Special Support Plan for Innovation and Entrepreneurship Leaders in Anhui Province and the Teaching Innovation Team of Computer Basic Course Group under Grant 2022cxtid097; in part by the Key Research Project of Natural Science in Universities of Anhui Province under Grant KJ2020A0782; in part by Anhui Xinhua University Level Research Project under Grant 2022zr003 and Grant 2022zr010; in part by Anhui University Students Innovative Training Program Project under Grant S202312216026 and Grant S202212216044; and in part by Anhui Province Quality Engineering Project under Grant 2019mooc278 and Grant 2022xsxx089.

**ABSTRACT** Distracted driving, a leading cause of traffic accidents with severe consequences, still faces numerous technical challenges in practical implementation for recognizing unsafe driving behavior. These challenges include the complexity of feature extraction using traditional convolutional neural networks (CNNs) for driver behavior analysis and the lack of real-time perception during driving. To address these issues, this study proposes an improved method for distracted driving behavior recognition by combining the Bi-LSTM model with an attention mechanism based on Dilated Convolutional Neural Networks (ID-CNN). Firstly, we employ a dilated convolution model to extract features efficiently with fewer parameters while enhancing multi-scale feature extraction capabilities and widening the receptive field. Subsequently, we integrate the attention mechanism into the Bi-LSTM model to enhance its effectiveness in solving the driving behavior classification problem. The integrated Bi-LSTM model with attention mechanism calculates correlation between intermediate and final states to obtain a probability distribution of attention weights at each moment, thereby reducing information redundancy while preserving useful information effectively. Furthermore, image feature vectors are enhanced to further improve accuracy in image classification tasks. Compared to other methods, the proposed approach exhibits faster convergence rates and more stable model accuracy. Specifically, on both the StateFarm dataset and our own collected Drive&Act-Distracted data, we achieved accuracies of 95.8367% and 97.8911%, respectively. This indicates that incorporating dilated convolution and attention mechanisms strengthens sequence data learning and feature weighting within our network model, resulting in significantly improved accuracy for driving behavior recognition.

**INDEX TERMS** Distracted driving, Bi-LSTM, CNN, attention mechanisms.

## I. INTRODUCTION

Road traffic accidents currently surpass AIDS, tuberculosis, and diarrhea as the leading cause of mortality [1]. Among individuals aged 5 to 29, traffic accidents are the primary

The associate editor coordinating the review of this manuscript and approving it for publication was Wu-Shiung Feng.

contributor to fatalities. As per the 2018 World Traffic Situation Report published by the United Nations Global Conference on Sustainable Transport, road traffic accidents claim over 1.2 million lives annually and result in up to 51 million injuries [2]. The report highlights an increase in accidents attributed to high-risk distracted driving incidents. Distracted driving, defined by the International Organization

for Standardization as impaired driving ability due to engaging in non-driving activities while operating a vehicle under normal conditions, is acknowledged as a significant risk factor by the U.S. Department of Transportation. Data from Columbia University's Transportation Institute reveals that distracted driving poses three times higher likelihood of causing hazardous accidents compared to regular driving practices. According to statistics released by China's Ministry of Transport in 2019, approximately 57,000 motor vehicle accidents occurred on Chinese roads including around 43,000 involving cars and roughly 10,000 involving motorcycles [3].

To effectively identify unsafe distracted driving behavior exhibited by drivers during normal road operations and provide timely early warnings, this study focuses on investigating distracted driving behavior as the research subject. By employing deep learning techniques, a model is developed to detect abnormal drivers with the aim of minimizing traffic accidents. However, there are still several technological challenges associated with implementing the algorithm for recognizing unsafe driving behavior in practical applications. Considering its ability to simultaneously capture both forward and reverse information from input sequences while processing sequential data and exhibiting superior generalization capabilities, this paper selects the Bi-LSTM model as the fundamental framework.

The challenge in recognizing human behavior in existing video sequences lies in the fact that the target action occupies only a small area or portion of the sequence. Simultaneously, identifying the target face is susceptible to disturbances from surrounding background information, such as noise, lighting conditions, occlusions, and more. Consequently, extracting effective spatio-temporal information pertaining to facial or human behavior from video sequences has emerged as a pivotal concern for behavior recognition. The primary objective of this study is to effectively mitigate interference caused by surrounding background information and extract meaningful spatio-temporal cues related to human behavior. The attention mechanism (AM) serves as an internal resource allocation mechanism in deep learning models. Its integration enhances the extraction of key semantic information, thereby improving algorithmic recognition accuracy. Although there are numerous documents describing the application of Bi-LSTM models in human behavior recognition, face recognition, object recognition, etc., scarce literature exists on utilizing the organic fusion of Bi-LSTM model structure and attention mechanism for face detection and fatigue driving.

Consequently, this paper proposes a technique that combines the Bi-LSTM model with the attention mechanism to identify distracted driving behavior. The Bi-LSTM model is introduced based on feature extraction from cavity convolution, and the attention mechanism is utilized to calculate different weights between states in the Bi-LSTM model, thereby significantly enhancing its capacity for feature expression. Firstly, dilated convolution is employed to broaden receptive aspects and improve multi-scale representation ability

of feature information while extracting local fine-grained features of expressions and reducing computational costs. Secondly, the linkage relationship between information is fully considered in conjunction with LSTM.

The main contributions of this paper are as follows:

(1) Incorporating an attention mechanism into the Bi-LSTM model structure enhances the model's generalization ability by allowing it to focus on relevant information and ignore irrelevant information when processing sequence data;

(2) Combining the Bi-LSTM model with dilated convolution enables multi-scale feature extraction and perception, making it better suited for image recognition tasks with complex backgrounds and multi-scale targets;

(3) This model has achieved excellent performance on the StateFarm dataset. Simultaneously, the validity of this model is further substantiated through its application on our self-collected Drive&Act-Distracted dataset.

## II. RELATED RESEARCH

In foreign countries, drivers' level of distraction is initially assessed based on brainwave patterns and heart rate. For instance, the analysis of brainwave waveforms and heart rate data provides input to evaluate a driver's degree of distraction while driving. Monitoring these two parameters can accurately assess the driver's state under normal conditions. However, widespread implementation in real-life scenarios is challenging due to the requirement for drivers to constantly wear biometric sensors, which may interfere with their ability to drive safely.

In 2017, Craye and Karray [4] developed an efficient module that utilizes active sensors and deep learning to assess the driver's dangerous driving state. The module incorporates sensor-based color recognition of driving hazards, head position analysis, and facial expression evaluation to determine the driver's current driving state. To achieve accurate classification, a combination of hidden Markov model and AdaBoost classifier techniques is employed. The experimental results demonstrate the high recognition accuracy of this approach. However, it should be noted that the dataset used for identification is solely based on simulations, lacking real-world environment inspections. Li and Liu [5] made significant contributions to the initial exploration. In 2018, they proposed a method for identifying anomalous driving behavior using a multi-class LogitBoost classifier and a covariance manifold based on a dichotomous notion. The method achieved an impressive correct recognition rate of up to 81.08% for various recognition targets. In 2019, Yin [6] introduced a technique for detecting driver fatigue based on gated cyclic units and full convolutional networks. Firstly, an infrared camera acquisition system captured the driver's face image. Subsequently, a multi-task cascaded convolutional neural network was employed to detect the driver's face and locate feature points, enabling extraction of the driver's eye image through geometric position relationships

of these key points on the face. Finally, a convolutional neural network recognition algorithm was utilized to determine whether the extracted eye image represented open or closed states accurately. Overall, this approach exhibited excellent performance in detecting driver fatigue. In 2019, Tamas and Maties [7] proposed a real-time model for recognizing dangerous driving behaviors based on convolutional networks, enabling real-time assessment of the driver's current driving state. The recognition delays for texting and hands leaving the steering wheel were approximately 0.05s and 0.08s respectively. In 2020, Karri et al. [8] introduced a network structure grounded in support vector machine principles that accurately identifies various unsafe driving states under realistic conditions. Additionally, in 2020, Dong et al. [9] successfully achieved identification of dangerous driving states using a combination of convolutional neural networks and the Region Proposal method. However, in practical applications, the recognition accuracy is suboptimal, the model convergence speed is sluggish, and the real-time perception of driver behavior during driving is inadequate. In 2021, Liu et al. proposed a real-time system for detecting driver fatigue based on convolutional neural networks and short- and long-term memories (CNN-LSTM) [10]. Building upon an enhanced short- and long-duration memory network, Shi et al. suggested a method for detecting driving behavior by incorporating attention mechanisms to improve the network structure. They further developed a hybrid dual-flow convolutional neural network algorithm based on both CNNs and long- and short-term memory networks [11]. Feng et al. [12] utilized an integrated approach, incorporating eye, mouth, and head features for driver fatigue detection. This method involves face detection based on LBP features, extraction of drivers' facial feature points using a multi-cascade residual regression tree algorithm, determination of drivers' head posture through 3D face model matching, establishment of a fatigue detection model specific to drivers, and subsequent training. The optimization and acceleration of the face detection process were achieved by reducing the background difference-based detection area and video frame image size. Fu et al. [13] combined a neural network with distracted driving prediction methodology that comprehensively considers external uniform speed and driver state to provide more accurate predictions.

The memory and forgetting functions of LSTM contribute to the enhancement of recognition accuracy for temporal images, making LSTM models widely employed in video behavior analysis, facial recognition, as well as monitoring fields like attendance and security. Bi-LSTM model, a variant of LSTM that integrates forward LSTM and reverse LSTM, is capable of capturing more comprehensive contextual information. In this study, we enhance the Bi-LSTM model by incorporating ID-CNN to improve its multi-scale feature extraction capability. Additionally, we introduce an attention mechanism to facilitate the extraction of crucial semantic information, thereby further improving the algorithm's recognition rate.

### III. METHODS

#### A. ID-CNN

The Convolutional Neural Network (CNN) is a neural network architecture developed by Malini et al., inspired by the biological brain [14]. It consists of multiple hidden layers, similar to recurrent neural networks. To reduce network complexity, measures such as limited range perception are employed. Additionally, CNN exhibits adaptability to translation, rotation, and scale changes in data. While CNN excels in strong feature extraction ability and parameter efficiency, it may lack fine-grained features and their interrelationships. Therefore, this research employs the dilated convolution model with moderate parameters to widen the receptive field of features and enhance multi-scale feature extraction capabilities for expression analysis.

In contrast to regular convolution, dilated convolution introduces a new parameter known as the dilation rate, which determines the spacing between values during data processing by the convolution kernel. By skipping certain input elements while maintaining a constant kernel size, dilated convolution expands the receptive field of the kernel. This approach preserves data structure and avoids downsampling, offering distinct advantages.

The receptive field is the size of the area mapped by the pixels on the feature map output by each layer of the network on the original image. The calculation formula of receptive field  $r_{i+1}^2$  as shown in Eq. (1).

$$r_{i+1}^2 = [(r_i - 1) + (2d + 1)]^2, \quad (1)$$

where  $r_i$  represents the change of receptive field in  $i$ th layer, and  $d$  represents the expansion coefficient of dilated convolution.

Dilated convolution has the same size as the convolution kernel of ordinary convolution, and its parameters remain unchanged in neural networks; however, it possesses a larger receptive field. Dilated convolution is achieved by introducing zero-padding between each weight in the ordinary convolution kernel, thereby expanding the network's expansion coefficient. The operation of ordinary convolution can be expressed as shown in Eq. (2).

$$(s) = (x * k)(s) = \sum_{i=0}^{m-1} x(s-i) \cdot k(i). \quad (2)$$

In Eq. (2),  $x$  is the input sequence,  $*$  is the convolution operation,  $k$  is the convolution kernel, and  $m$  is the convolution kernel size. The dilated convolution is expressed as shown in Eq. (3).

$$F(s) = (x *_d k)(s) = \sum_{i=0}^{m-1} x(s-d \cdot i) \cdot k(i). \quad (3)$$

In Eq. (3),  $_d^*$  is the dilated convolution calculation of expansion rate  $d$ , and when  $d = 1$ , ordinary convolution is the same as dilated convolution.

The receptive field of dilated convolution increases exponentially. As depicted in Figure 1, all convolution kernels are of size  $3 \times 3$ . FIGURE 1(a) illustrates the characteristic diagram obtained from the original image using dilated

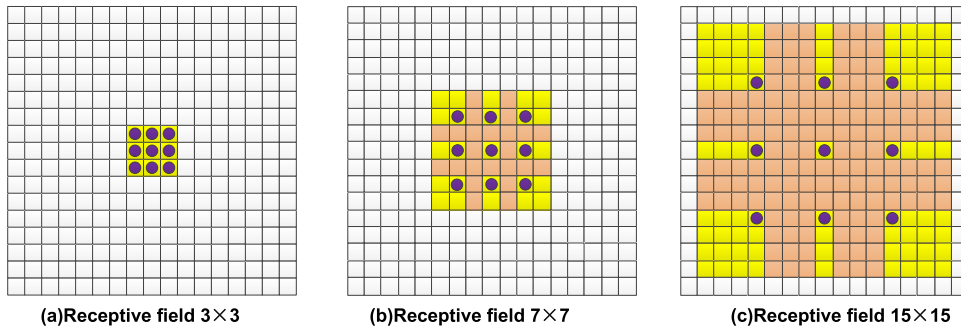


FIGURE 1. Receptive field of dilated convolution.

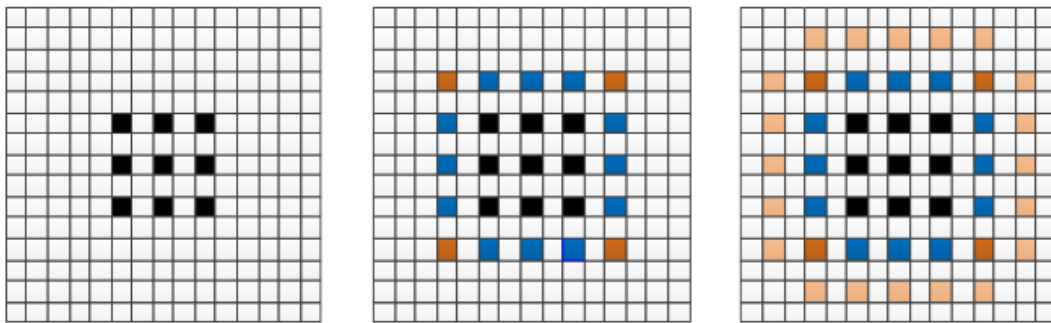


FIGURE 2. Convolution kernel result graph with expansion rate of 2.

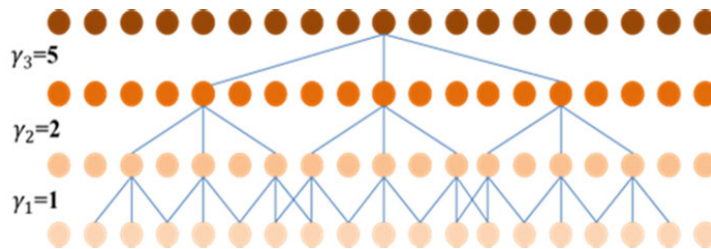


FIGURE 3. Stepped dilated convolution structure diagram.

convolution with  $d = 1$ , which is equivalent to ordinary convolution. Each element in the first layer represents a  $3 \times 3$  portion of the original image, while maintaining a receptive field size of  $3 \times 3$ . The feature map in FIGURE 1(b) is obtained through a dilated convolution operation with  $d = 1$  applied to the original map at layer 2. With an expansion rate of 2, the convolution kernel is effectively distributed across the positions indicated by dots in the map, resulting in a receptive field of each element in layer 2 being  $7 \times 7$  relative to the original map. In FIGURE 1(c), we observe the characteristic diagram of the third layer obtained by applying dilated convolution with  $d = 4$  on the second layer. Similarly, with an expansion rate of 4, each element in the third layer has a receptive field of size  $15 \times 15$ .

The conventional 3-layer  $3 \times 3$  convolutional kernels can only achieve a receptive field of  $7 \times 7$ . While the number of factors involved in the convolution remains unchanged and

so does the computational workload, increasing the size of the kernel allows for feature values in the feature map to correspond to larger regions, thereby expanding their range of perception.

Although dilated convolution effectively expands the receptive field and preserves feature map information, it introduces a limitation where not all image features contribute to the convolution calculation due to the presence of dilation in the convolution kernel. If continuous convolution layers have identical expansion rates, a gridding effect may occur. FIGURE 2 illustrates this phenomenon when stacking multiple  $3 \times 3$  convolution kernels with an expansion rate of 2. Furthermore, while dilated convolutions enhance receptive fields, they can hinder the extraction of detailed features during convolutional analysis. The severity of detail feature loss increases with higher expansion rates.

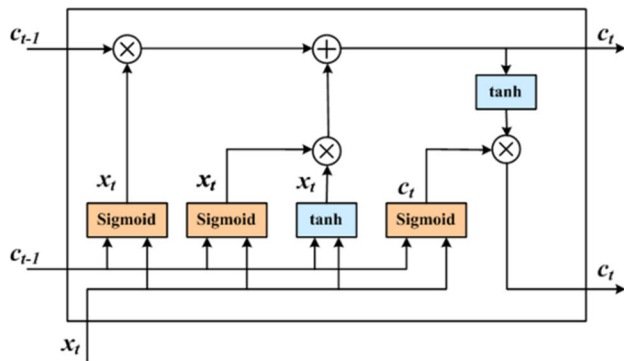


FIGURE 4. LSTM model structure diagram.

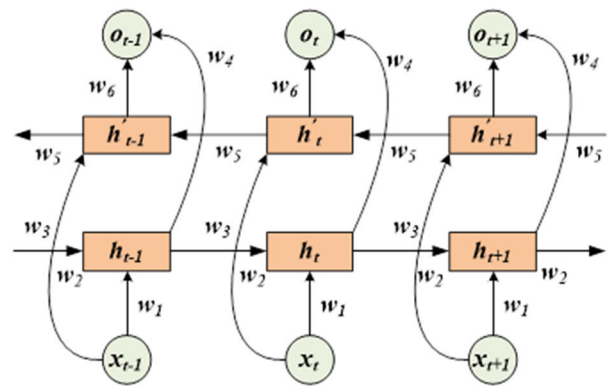


FIGURE 5. Flowchart of Bi-LSTM module.

Therefore, in this paper, we adopt a stepped expansion rate strategy by setting the expansion rates to 1, 2, and 5 in the three-layer convolutional network.

Revised sentence: In contrast to utilizing the same expansion rate, stepped dilated convolution employs convolution kernels with varying expansion rates. Stepped dilated convolution offers the following two advantages:

1. The stepped dilated convolution enables the acquisition of a multi-scale extended receptive field without introducing additional model parameters or computational complexity, thereby facilitating the extraction of long-range inter-feature information.

2. The utilization of stepped dilated convolution effectively mitigates the “gridding” phenomenon arising from local information loss caused by expansion of the convolution kernel.

The receptive fields of elements in each dilated convolution layer are 3, 7, and 19 respectively when the convolution kernel is set to  $3 \times 3$ , as depicted in FIGURE 3. By sharing feature mapping parameters across all dilated convolution layers, model parameters can be significantly reduced while computational overhead is saved.

The incremental expansion rate ensures the preservation of input information, while simultaneously accommodating the requirements for detailed features and inter-feature correlations.

By setting the stepped expansion rate, the receptive field can be adjusted to obtain multi-scale feature information. In our experiment, we adaptively determine the size of the feature map extracted by void convolution based on the original image size and then set the expansion rate accordingly. Convolution with different expansion rates enables us to capture behavior features at multiple scales, while assigning a weight of 0 to non-relevant points in the convolution process.

### B. BIDIRECTIONAL LONG SHORT MEMORY NETWORK

The Long short-term Memory (LSTM) [15] is a Recurrent Neural Network (RNN) that incorporates memory capacity [16], representing an advancement over traditional RNNs. When dealing with long-term dependent data, the generalization capabilities of RNNs are suboptimal. Specifically, when

handling distant nodes, issues such as gradient disappearance and explosion may arise. Numerous subsequent solutions have been explored, among which the threshold recurrent neural network stands out as particularly noteworthy. The LSTM model demonstrates superior generalization capabilities compared to other threshold recurrent neural networks. By incorporating three types of gating units and a state update mechanism within a single cell in the hidden layer, LSTM enables dynamic internal circulation weights. Consequently, it allows for variable integral scaling at different time points while keeping network parameters unchanged. FIGURE 4 illustrates the structural diagram of the LSTM model.

The LSTM model exhibits significant advantages over RNN in handling sequential data, owing to its distinctive long and short-term memory module that incorporates a memory unit with enhanced capacity.

The input gate possesses the capability to either accept or reject the input features, thereby generating memory cells. Subsequently, the input gate at the subsequent layer can transmit the acquired memory cell state to the neuron in that layer or discard it as output to the next layer. Furthermore, by regulating alterations in cell state, the forgetting gate determines whether current cell characteristics should be retained or past characteristics ought to be disregarded.

The Bi-LSTM model extracts input information and effectively captures the temporal dependencies between connected moments in the sequential data [17], [18]. This enables it to optimize the utilization of input model information by accurately capturing the flow of information.

The forward and backward modules in the model were independently constructed based on the bidirectional long short-term memory network with two distinct architectural layouts. All modules in the reciprocal backward positions shared an identical output layer. The temporal extension of the network enables all units in the output layer to access comprehensive information from both past and future tenses of the input layer at any given time point. In the task of recognizing distracted driving behavior, continuous driver state information can be stored in the hidden layer. Each operator within the model is associated with a corresponding weight  $W$ .

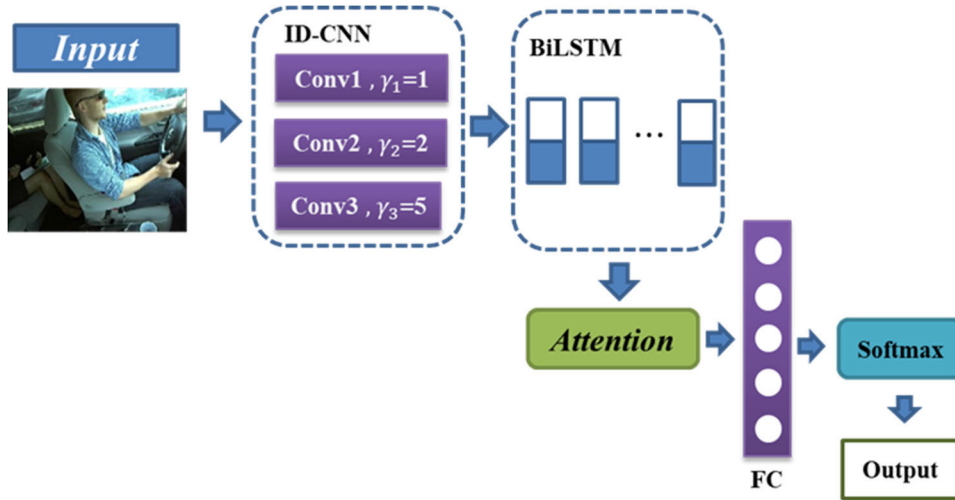


FIGURE 6. IDC-Bi-LSTM-Att model structure diagram.

The calculation process of Bi-LSTM is analogous to that of LSTM, albeit with distinctions in the output mode between the two. During Bi-LSTM operation, incoming data is propagated through both forward and backward hidden states, yielding a hidden layer output derived from bidirectional temporal operations.

As shown in FIGURE 5,  $x_{t-1}, x_t$  and  $x_{t+1}$  represent the input data corresponding to time  $t - 1, t$  and  $t + 1$  respectively.  $h_{t-1}, h_t$  and  $h_{t+1}$  represent the hidden state of the corresponding forward iteration of the LSTM, respectively.  $h'_{t-1}, h'_t$  and  $h'_{t+1}$  represent the hidden state of the corresponding backward iteration of the LSTM, respectively.  $o_{t-1}, o_t$  and  $o_{t+1}$  represent the corresponding output data.  $w_1, w_2, \dots, w_6$  indicates the weight of each layer.

The status update of the hidden layer of forward LSTM and backward LSTM and the final output process of BiLSTM are shown in Eqs. (4)-(6).

$$h_t = f_1(w_1x_t + w_2h_{t-1}), \tag{4}$$

$$h'_t = f_2(w_3x_t + w_5h'_{t+1}), \tag{5}$$

$$o_t = f_3(w_4h_t + w_6h'_t), \tag{6}$$

where,  $f_1, f_2$  and  $f_3$  are activation functions between different layers respectively.

### C. ATTENTION MODULE DESIGN

When utilizing CNN for image feature extraction, the computational complexity and training time are significantly increased due to its ability to perceive global image features. In the domain of image recognition, extracting information from photos often involves a substantial amount of irrelevant data, which not only hampers model training efficiency but also diminishes recognition accuracy. The incorporation of attention mechanism in deep learning aims to emulate the brain's functioning mechanism, thereby mitigating the adverse impact of superfluous information during model

training, enhancing recognition accuracy, and optimizing computer computing resources utilization [19].

The attention strategy employed in this essay is primarily focused on addressing the issue of multiple inputs with varying input vector sizes. It effectively enhances the ability to capture internal correlations within data or features, surpassing the performance of the original attention mechanism and reducing reliance on external information [20].

The weighted sum is the core of the attention mechanism. When the attention mechanism is applied in the practical application of deep learning, there exist  $n$  feature vectors to form the feature matrix  $X = [x_1, x_2, \dots, x_n]$ , assuming that its dimension is  $d$ , and the information of  $n$  features is integrated. Direct calculation of the relationship between features in attention can shorten the distance between remote features and calculate the attention feature vector  $g_i$  containing context information. The calculation process is as shown in Eq. (7).

$$g_i = \sum_{i \neq j} \alpha_{ij} \cdot x_j, \tag{7}$$

among them,  $\alpha_{ij} > 0$  is the attention weight, and  $\sum_j \alpha_{ij} = 1$ . Each feature vector's corresponding weight is determined using the tanh function. The following is the weight calculating process as shown in Eqs. (8), (9).

$$\alpha_{ij} = \frac{e^{s(x_i, x_j)}}{\sum_j e^{s(x_i, x_j)}}, \tag{8}$$

$$s(x_i, x_j) = v_a^T \tanh(w_a[x_i \oplus x_j]). \tag{9}$$

MLP is used to calculate  $s(x_i, x_j)$  to represent the correlation between  $x_i$  and  $x_j$  features. In Eqs. (8), (9),  $v$  and  $w$  are learnable parameter matrices.  $\oplus$  represents the addition operation.

The attention module dimension employed in this study is set to 512, while the batch size is fixed at 64. All samples

**TABLE 1.** Parameters of ID-CNN model.

Layer	Convolution Kernel	Kernel number	Expansion rate	Output
Input				$64 \times 64 \times 1$
Conv1	$3 \times 3$	16	1	$64 \times 64 \times 16$
Conv2	$3 \times 3$	32	2	$64 \times 64 \times 32$
Conv3	$3 \times 3$	64	5	$64 \times 64 \times 64$

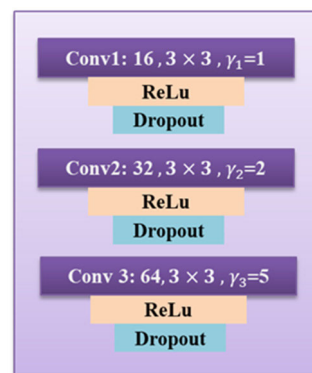
undergo a total of 30 training iterations. Weight calculation is based on the correlation degree between the feature vector and the recognized object. To obtain the probability distribution of attention and enhance recognition accuracy by optimizing image feature vectors, higher correlation degrees correspond to higher scores.

#### D. MODEL STRUCTURE DESIGN

The IDC-Bi-LSTM-Att network model developed in this study is illustrated in FIGURE 6. This model consists of three dilated convolution layers, one BiLSTM layer, an attention module, and a fully connected layer. In the experiment, two-dimensional images were used as input, and deep features were extracted from the data through convolutional operations. Subsequently, a Bi-LSTM network was constructed to capture temporal order information within the feature vectors. Furthermore, an attention mechanism was employed to enhance the focus on important regions of the image during training, thereby improving the network's ability to accurately identify hazardous driving conditions. Finally, a fully connected layer with Softmax classifier was utilized for image classification.

(1) The image is initially processed through the ID-CNN architecture, comprising three layers of cavity convolutional layers denoted as Conv1, Conv2, and Conv3 respectively, to extract structural features from the image. The first convolutional layer employs a  $3 \times 3$  parameterized convolution kernel with sixteen channels and applies the Rectified Linear Unit (ReLU) activation function. The second convolutional layer utilizes a  $3 \times 3$  parameterized convolution kernel with thirty-two channels and also applies the ReLU activation function. Similarly, the third layer of the convolutional network adopts a  $3 \times 3$  parameterized convolution kernel with sixty-four channels along with the ReLU activation function. In this three-layered configuration, expansion rates are set as 1, 2, and 5 correspondingly. The utilization of different expansion rates enables the capture of multi-scale context information. A smaller expansion rate facilitates the generation of fine-grained features, while a larger expansion rate allows for the inclusion of a broader range of features, thereby forming convolution nuclei with varying receptive fields and acquiring multi-scale information. The schematic diagram illustrating the structure of ID-CNN is presented in FIGURE 7. The detailed network structure configuration for ID-CNN is provided in Table 1.

(2) After passing through the ID-CNN module, the images are connected to the Bi-LSTM layer with 128 hidden units.

**FIGURE 7.** ID-CNN model structure diagram.

(3) The Attention module incorporates a dimension of 128 in its attention mechanism, reducing reliance on external data and enhancing its ability to capture internal correlations between information and features. This enables more effective extraction and retention of key information aspects.

(4) The wdilated connection layer has 10 output units, utilizing the softmax function for classification.

## IV. EXPERIMENTAL ANALYSIS

### A. DATA SET

The experiment utilized the StateFarm dataset, which was published by State Farm Insurance on Kaggle (<https://www.kaggle.com/datasets/rightway11/state-farm-distracted-driver-detection>). The organizers of the study aim to employ images captured by a compact dashboard camera in order to discern hazardous driving conditions, thereby notifying drivers and safeguarding their lives.

The data set is divided into 10 categories:

- c0: safe driving
- c1: texting - right
- c2: talking on the phone - right
- c3: texting - left
- c4: talking on the phone - left
- c5: operating the radio
- c6: drinking
- c7: reaching behind
- c8: hair and makeup
- c9: talking to passenger

Dataset example is shown in Table 2.

The dataset comprises a total of 10 states, denoted as c0-c9, and the data distribution for each state is illustrated

TABLE 2. Dataset example.

<p>c0: safe driving</p> 	<p>c1: texting - right</p> 
<p>c2: talking on the phone - right</p> 	<p>c3: texting - left</p> 
<p>c4: talking on the phone - left</p> 	<p>c5: operating the radio</p> 
<p>c6: drinking</p> 	<p>c7: reaching behind</p> 
<p>c8: hair and makeup</p> 	<p>c9: talking to passenger</p> 



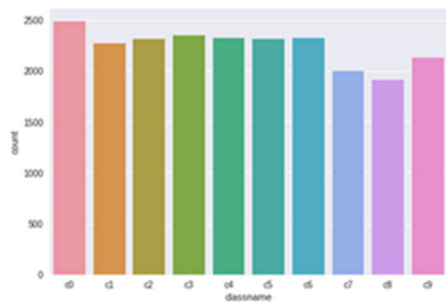


FIGURE 8. Dataset distribution.

in Figure 8, demonstrating a relatively uniform distribution pattern.

### B. DATA ENHANCEMENT

In the real driving environment, variations in personal height and habits result in differences in the distance and angle between the head and the operating wheel. Additionally, external factors such as driving turbulence and lens fouling further influence this relationship. Therefore, it is crucial for the model to acquire a diverse range of training data to enhance network robustness. Data augmentation techniques can effectively expand the original dataset by incorporating various operations.

- Rotation: The image is subjected to clockwise or counterclockwise rotation within a maximum angle of 20 degrees in order to preserve data integrity during the dangerous driving behavior recognition experiment.
- Zoom: Adjust the magnification level of the image, ensuring that excessive zooming does not result in information loss.
- Flip: Apply inward or outward tilting transformation to the image.
- Splicing: Combine images of different drivers belonging to the same category through cutting and proportioning techniques.

### C. EXPERIMENTAL ENVIRONMENT AND IMPLEMENTATION DETAILS

The ratio of the training set to the test set's image count is 4:1. The training set consists of approximately 30,840 photos, while the test set comprises roughly 7,720 images. The photo distribution between the training and test sets remains consistent. The comparison experiment described above directly utilizes the pre-drive training set.

During the data reading phase, maintain an input image size of  $64 \times 64$ . To achieve rapid convergence, this paper utilizes the Adam optimizer with an initial learning rate of 0.00001 and employs Relu as the activation function. The learning rate decay strategy is applied every 10 epochs with a decay factor of 0.9. The batch size is set to 4, and a total of 30 epochs are trained. Experimental parameter settings are presented in Table 4.

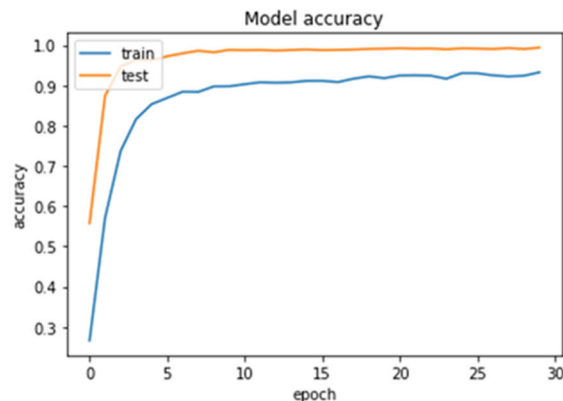


FIGURE 9. The training and testing accuracy of the model in this paper.

The experimental setup includes Windows10 operating system, TensorFlow-2.0 software framework, Nvidia GeForce GTX2080 GPU with video memory capacity of 24GB, and Python programming language (version 3.7).

The deep learning framework selected for this paper is TensorFlow, an artificial intelligence system developed by Google and globally available free of charge. Renowned as one of the most popular deep learning frameworks on Github, TensorFlow is an open-source software library that employs data flow diagrams for numerical computations.

### D. COMPARATIVE EXPERIMENTAL ANALYSIS

The following experiments examine the efficacy and superiority of the proposed model from various perspectives.

#### Experiment 1:

The IDC-Bi-LSTM-Att model is employed in this study to predict each type of action in the dataset of abnormal driving behavior for statistical analysis. The confusion matrix [21] can be utilized to represent the distribution of predicted actions compared to the actual data set. Table 5 presents the classification mixture matrix of data sets in IDC-Bi-LSTM-Att, while FIGURE 9 illustrates the test and training accuracies of the model using the StateFarm dataset from this research.

Through the analysis of the experimental data in Table 5, it can be observed that in the distracted driving dataset, there is a similarity between the hand movements and state changes associated with calling on hair (C8) and right (C2), which leads to potential confusion between these two types of distracted states. The lower recognition rate for certain driving states, as indicated by the confusion matrix, can be attributed to both similarities among objects and relatively limited data samples available for some categories. Consequently, achieving a high recognition rate becomes challenging. Notably, normal driving (C0) and tuned radio (C5) exhibit relatively high accuracy in recognition due to their distinguishable characteristics compared to other driving states and also owing to an abundance of corresponding images within the distracted driving dataset.

TABLE 3. Data enhancement contrast.

	Safe driving	Right hand texting	Right hand calling	left hand texting	left hand calling
label	C0	C1	C2	C3	C4
Raw data set	2486	2250	2310	2396	2387
Enhanced data set	4126	3712	3927	4073	3819
	Tune the radio	Have a drink	Find something	Fix your hair	Talk to the passengers
label	C5	C6	C7	C8	C9
Raw data set	2379	2385	2048	1874	2246
Enhanced data set	3806	3816	3481	3748	4042

TABLE 4. Table of experimental parameters.

Parameter	Parameter setting
Optimizer	Adam
Input size	64×64
Epoch	30
Batch Size	4
Learning rate	0.00001
Strategy for learning rate decay	Exponential decay
Coefficient of learning rate decay	0.9
Activation function	Relu

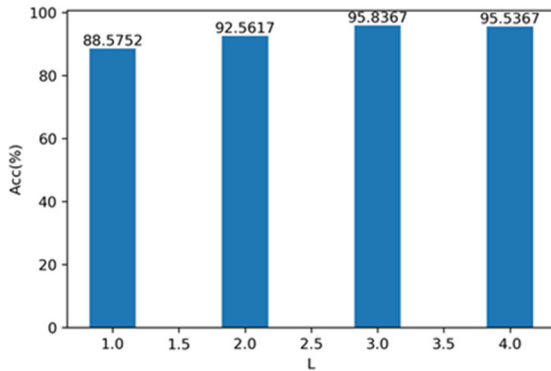


FIGURE 10. The training and testing accuracy of the model in this paper.

Experiment 2: Complexity analysis.

Based on this model, a comparison is made between the complexities of ordinary convolution and dilated convolution. Comparative experiments are conducted in the same experimental environment and dataset to evaluate their respective parameters and computational complexity, as presented in Table 6. From the table, it can be observed that the number of parameters and computations required by dilated convolution is approximately one third of those needed by ordinary convolution. This demonstrates that dilated convolution effectively reduces both parameter count and computational load in the model.

Experiment 3: Layer analysis.

The performance of the entire network is influenced by the number of layers in the feature extraction component,

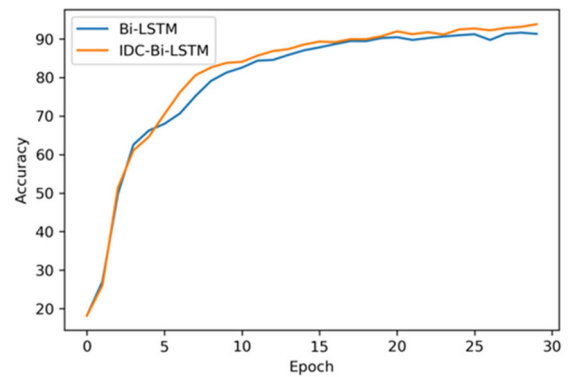


FIGURE 11. Accuracy comparison chart.

and this paper presents an improvement to the algorithm through adjustments made to the number of convolutional layers. Specifically, hyperparameter L denotes the quantity of cavity convolution modules utilized.

The performance of the ID-CNN module was analyzed with varying layers of dilated convolution, as depicted in FIGURE 10, which illustrates the model’s accuracy when adopting different layers for the ID-CNN convolution layer L. As evident from FIGURE 10, an increase in the number of layers can indeed enhance the model’s performance when initially few; for instance, a three-layer dilated convolution improves the model’s accuracy by 7.2615%. This improvement signifies that an expanded receptive field enables capturing longer distance dependencies and enhances feature extraction capabilities. However, surpassing three convolutional layers does not yield further improvements but rather escalates computational complexity. Thus, this paper concludes that L=3 is appropriate.

The experimental results demonstrate that the number of layers in the model significantly impacts its performance during the feature extraction stage. As the number of layers deepens, the model’s accuracy progressively improves until it reaches a certain threshold, beyond which it gradually diminishes. This phenomenon is primarily influenced by the dataset size, as an excessively large model with increased parameters can ultimately compromise its performance.

Experiment 4: Effect of expansion rate on model performance.

TABLE 5. Confusion matrix.

	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
C0	0.9300	0.0000	0.0000	0.0100	0.0000	0.0000	0.0300	0.0000	0.0000	0.0300
C1	0.0000	0.9160	0.0200	0.0100	0.0274	0.0300	0.0000	0.0100	0.0000	0.0100
C2	0.0000	0.0200	0.9400	0.0000	0.0000	0.0000	0.0300	0.0100	0.0000	0.0000
C3	0.0100	0.0000	0.0000	0.9500	0.0400	0.0000	0.0000	0.0000	0.0000	0.0000
C4	0.0000	0.0000	0.0000	0.0600	0.9400	0.0000	0.0000	0.0000	0.0000	0.0000
C5	0.0000	0.0000	0.0000	0.0000	0.0000	0.9600	0.0000	0.0300	0.0000	0.0100
C6	0.0000	0.0400	0.0300	0.0000	0.0000	0.0000	0.9100	0.0000	0.0200	0.0000
C7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9600	0.0200	0.0200
C8	0.0100	0.0000	0.0700	0.0000	0.0000	0.0000	0.0000	0.0000	0.9000	0.0200
C9	0.0220	0.0100	0.0000	0.0200	0.0000	0.0000	0.0100	0.0200	0.0000	0.9180

TABLE 6. Complexity analysis.

Convolution type	Parameter quantity(K)	computational complexity(K)
Ordinary convolution	83	245m
Dilated convolution	28	82m

TABLE 7. Effect of expansion rate on model performance in cavity convolution module.

L	T	Accuracy/%
2	1, 2	92.0354
2	2, 2	91.3411
3	2, 2, 2	92.5871
3	1, 2, 3	94.7894
<b>3</b>	<b>1, 2, 5</b>	<b>95.8367</b>
3	2, 3, 4	94.9578

After determining the number of layers  $L=3$ , the expansion rate  $T$  is adjusted to achieve an optimal receptive field. The performance of the model is influenced by adjusting the expansion rate  $T$  in the cavity convolution module due to its impact on filling within the module. The results are presented in Table 7, where bold numbers indicate superior values. Notably, parameter  $T$  is determined by parameter  $L$ . As shown in Table 7, optimal performance is observed when  $T$  takes values  $\{1,2,5\}$ , indicating that the model performs best with sequential expansion factors of 1, 2, and 5 for each of the three cavity convolution modules.

Experiment 5: Influence analysis of dilated convolution module.

In order to validate the efficacy and superiority of the proposed dilated convolution module, we conducted experiments based on a Bi-LSTM model. We compared and analyzed the IDC-Bi-LSTM baseline model with an added dilated

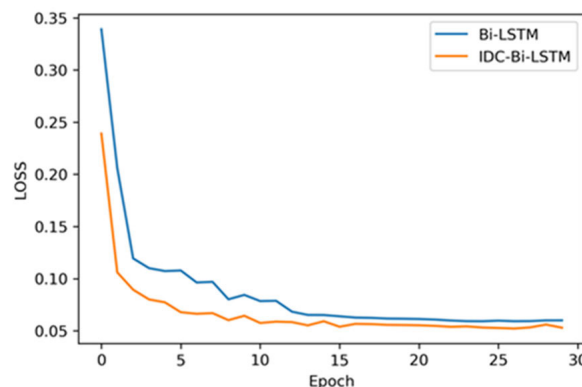


FIGURE 12. Loss comparison chart.

convolution module. The change curves of accuracy and loss values for both models are depicted in FIGURE 11 and FIGURE 12, respectively.

FIGURE 11 demonstrates a significant improvement in accuracy for the IDC-Bi-LSTM module compared to the basic model. Additionally, FIGURE 12 illustrates that the IDC-Bi-LSTM model with dilated convolution module exhibits faster convergence speed. Furthermore, FIGURE 13 presents a comparison of model accuracy, indicating that the Bi-LSTM model outperforms CNN, ID-CNN and LSTM models. This is why we have chosen it as our base model. Compared to the Bi-LSTM model alone, incorporating the dilated convolution module has resulted in an improved accuracy of 4.4895%. This suggests that multi-scale feature extraction ability has been enhanced by considering all contextual information during feature extraction; thus enabling more comprehensive capture of dependencies between features. These results confirm the effectiveness of using a dilated convolution module.

Experiment 6: Effect analysis of attention module.

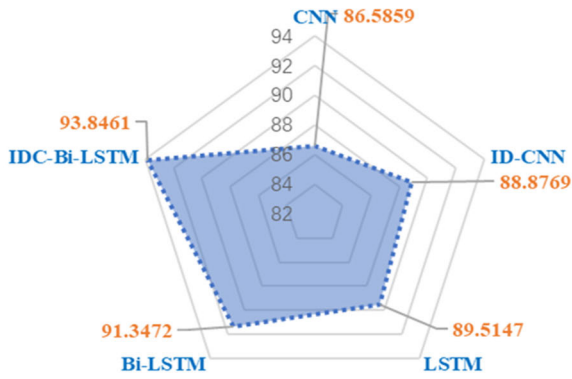


FIGURE 13. Accuracy comparison.

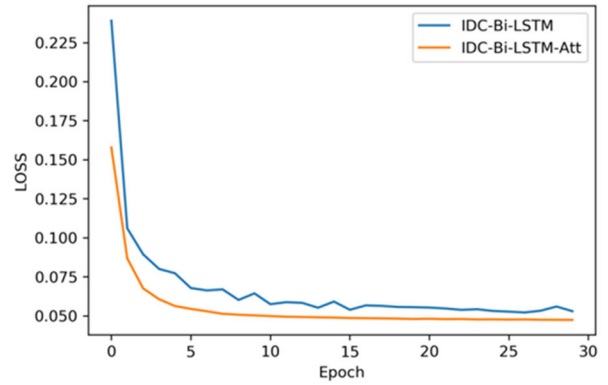


FIGURE 15. Loss comparison chart.

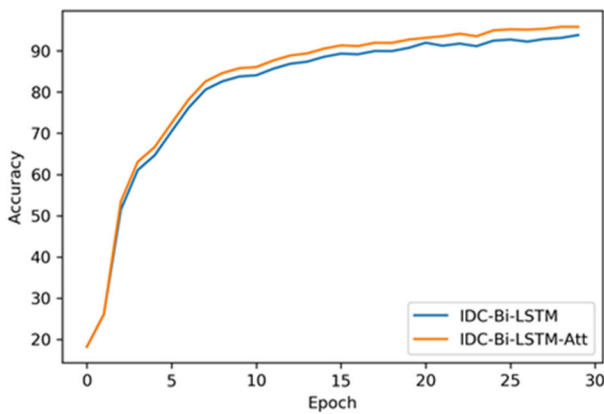


FIGURE 14. Accuracy comparison chart.

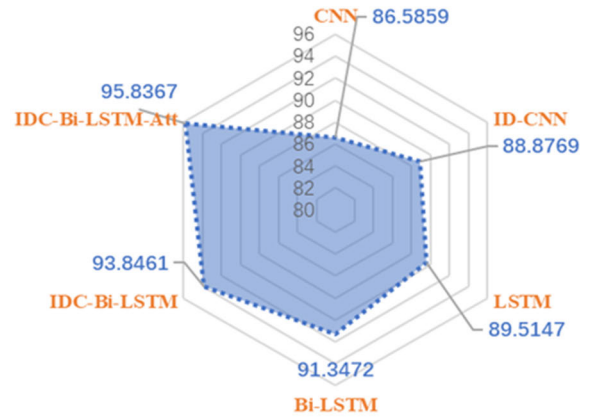


FIGURE 16. Accuracy comparison.

To validate the efficacy and superiority of the attention module proposed in this study, we conducted experiments based on IDC-Bi-LSTM. A comparative analysis was performed between the basic model and a modified version called IDC-Bi-LSTM-Att, which incorporates an attention module. The change curves of accuracy and loss value for both models are depicted in FIGURE 14 and FIGURE 15, respectively. As illustrated in FIGURE 14, it is evident that IDC-Bi-LSTM-Att outperforms IDC-Bi-LSTM in terms of accuracy, indicating that the attention mechanism compels the model to focus on discriminative features while enhancing its ability to extract directional features. Furthermore, as shown in FIGURE 15, the inclusion of an attention module accelerates model convergence rate significantly. Additionally, these results demonstrate that the hole convolution module effectively preserves positional information of features.

FIGURE 16 is a comparison of the accuracy of the models. The accuracy of the IDC-Bi-LSTM model is enhanced by 1.9906% when compared to the IDC-Bi-LSTM-Att model, indicating that the incorporation of an attention module enables optimal feature weighting, prioritization of crucial information, and extraction of superior features.

Experiment 7: Contrast experiment

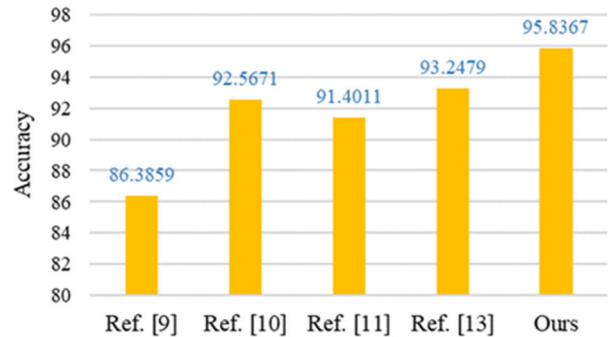
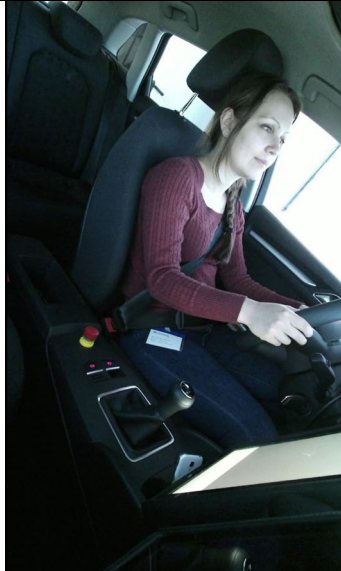
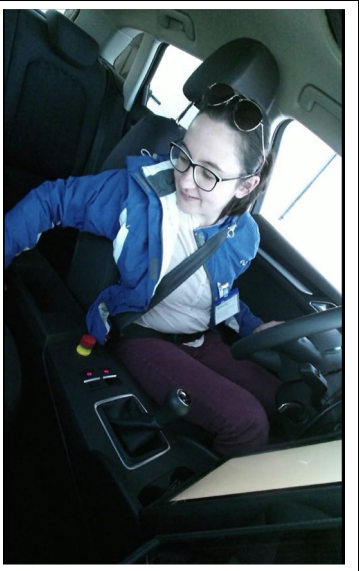
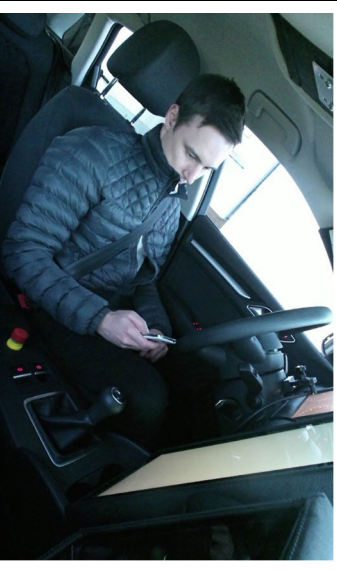
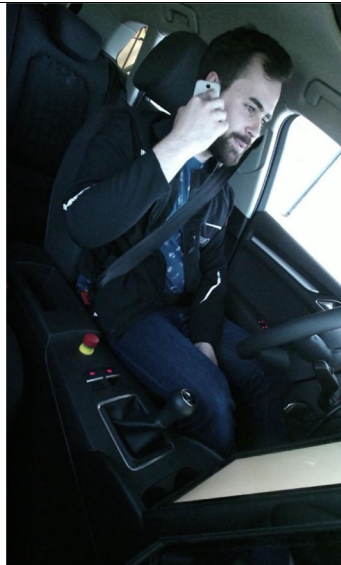
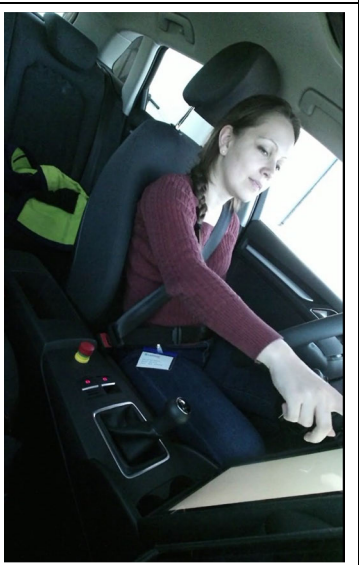



FIGURE 17. Accuracy comparison result.

To further validate the performance of IDC-Bi-LSTM-Att model in recognizing distracted driving behavior, a comparative model is established as follows:

- Model 1: Dong et al. [9] proposed the Fatigue Driving Detection Based on Involuntary Expert Neural Network, which aims to detect fatigue driving.
- Model 2: Liu et al. [10] introduced a real-time driver fatigue detection model based on CNN-LSTM.
- Model 3: Shi et al. [11] presented an enhanced long and short-term memory network for detecting driving behavior.

TABLE 8. Details of the drive&act-distracted dataset.

c0: safe driving	c1: reaching behind	c2: texting
		
c3: talking on the phone	c4: operating the radio	c5: eating
		

- Model 4: Fu et al. [13] Hybrid Neural Network predicts driving behavior risk by incorporating distracted driving behavior data.

The comparison of the accuracy of the aforementioned models is illustrated in FIGURE 17. The results demonstrate that the proposed strategy exhibits superior accuracy. CNN-based feature extraction excels at capturing intricate features, while the integration of Bi-LSTM module and attention mechanism significantly enhances recognition accuracy.

Experiment 8: Test of generalization ability

In order to further validate the model’s generalization ability, we utilized the Drive&Act dataset [22]

(<https://www.driveandact.com/>), which comprises a collection of Distracted drivers behavior data named Drive&Act - Distracted. This dataset consists of 1200 images obtained from 15 drivers across six categories of driving behavior, including safe driving, reaching behind, texting, talking on the phone, operating the radio, and eating. The distribution of images in the training set and test set follows a ratio of 4:1 as detailed in Table 8.

The parameter settings and comparison model remain consistent with those used in experiment 7. In this study, the IDC-Bi-LSTM-Att model is evaluated using the Drive&Act-Distracted dataset. Figure 18 presents a comparative analysis of the accuracy achieved by this model on the

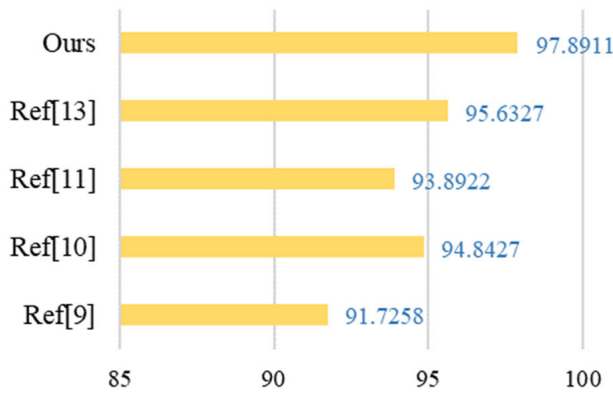


FIGURE 18. Accuracy comparison result.

Drive&Act-Distracted dataset. The results demonstrate that the proposed IDC-Bi-LSTM-Att model achieves an impressive accuracy rate of 97.8911%, surpassing other models under evaluation. These findings provide evidence for the robust generalization ability of our proposed IDC-Bi-LSTM-Att model.

## V. DISCUSSION

Building upon the advancements made in the Bi-LSTM model, this paper proposes a novel approach for recognizing distracted driving behavior by integrating the Bi-LSTM model with the attention mechanism of Dilated Convolutional Neural Networks (ID-CNN).

The key assumptions underlying this method are as follows:

1) Feature extraction is performed using the dilated convolutional model, which possesses a limited number of parameters, facilitates an expanded feature sensitivity field, and exhibits robust multi-scale feature extraction capabilities.

2) By incorporating the attention mechanism, distinct weights are computed between states within the Bi-LSTM model to further enhance its expressive power.

This methodology offers several advantages. It combines the Bi-LSTM model with dilated convolutions to enable multi-scale feature extraction and perception, making it particularly well-suited for complex background and multi-scale target image recognition tasks. Additionally, by introducing an attention mechanism into the structure of the Bi-LSTM model, relevant information can be selectively attended to while disregarding irrelevant information during sequence data processing, thereby augmenting its generalization ability.

The potential limitations and challenges of this approach are:

1) Computational complexity: When dealing with high-resolution or high-frame-rate video data, the computational requirements of the proposed model in this paper will further increase. Consequently, its implementation system may have limited applications.

2) Data requirements: Training the model necessitates a substantial amount of labeled distracted driving behavior data. However, labeling such data is time-consuming and costly.

3) Model generalization ability: Although the model performs well on the dataset used in this paper, its applicability to more complex scenarios or environments with fast driving speeds remains uncertain. The performance of the model can be influenced by various factors including individual driver differences, vehicle types, road conditions, etc.

4) Real-time performance: While this approach demonstrates good offline performance evaluations, maintaining high performance in real-time driving environments poses greater challenges. Further optimization is required to enhance real-time performance which will be addressed in future research.

## VI. CONCLUSION

This method employs dilated convolution to reduce the calculated parameters of the model, thereby expanding the receptive field and focusing more on details while disregarding irrelevant information. By incorporating bidirectional long-term memory networks and an Attention mechanism to capture relationships between key features, this approach mitigates gradient disappearance issues and enables the network to pay greater attention to driving state details. The proposed model achieves an accuracy of 95.8367% on the StateFarm dataset, surpassing both the Bi-LSTM network model by 4.4895% and the IDC-Bi-LSTM network model by 1.9906%. At the same time, we get 97.8911% on Drive&Act-Distracted data set, which is better than other models. This study presents a novel method for recognizing distracted driving behaviors, which can enhance recognition accuracy in intelligent driving systems and provide safer guarantees for future implementations. As behavior occurrences often involve time information, accurately identifying their start and end from data streams holds significant implications for final behavior identification. Online behavior detection has emerged as a promising direction for future research efforts. Furthermore, existing approaches primarily focus on detecting events that have already occurred; however, predicting anomalies and issuing alarms before such events transpire would greatly expand the application scope of this technology—a challenge that warrants further investigation.

## REFERENCES

- [1] J. Jun, J. Ogle, and R. Guensler, "Relationships between crash involvement and temporal spatial driving behavior activity patterns: Use of data for vehicles with global positioning systems," *Transp. Res. Rec.*, vol. 2019, no. 1, pp. 246–255, 2018.
- [2] Y. Y. Zhang, S. Zhang, and Y. Zhang, "Multi-modality fusion perception and computing in autonomous driving," *J. Comput. Res. Develop.*, vol. 57, no. 9, pp. 1781–1799, 2020.
- [3] J. K. Li, Z. W. Li, and W. W. Hua, "Statistical analysis of urban road traffic accidents," *Technol. Innov. Appl.*, vol. 11, no. 21, pp. 74–76, 2021.
- [4] C. Craye and F. Karray, "Driver distraction detection and recognition using RGB-D sensor," 2015, *arXiv:1502.00250*.
- [5] C. J. Li and Y. P. Liu, "Abnormal driving behavior detection based on covariance manifold and LogitBoost," *Laser Optoelectron. Prog.*, vol. 55, no. 11, pp. 338–345, 2018.

- [6] H. B. Yin, "Fatigue driving detection method and system design based on gate control recurrent neural network," Tianjin Polytech. Univ., Tianjin, China, Tech. Rep., 2019.
- [7] V. Tamas and V. Maties, "Real-time distracted drivers detection using deep learning," *Amer. J. Artif. Intell.*, vol. 3, no. 1, pp. 1–8, 2019.
- [8] S. L. Karri, L. C. De Silva, D. T. C. Lai, and S. Y. Yong, "Identification and classification of driving behaviour at signalized intersections using support vector machine," *Int. J. Autom. Comput.*, vol. 18, no. 3, pp. 480–491, Jun. 2021.
- [9] C. J. Dong, G. H. Lin, and C. X. Wu, "Fatigue driving detection based on convolutional expert neural network," *Comput. Eng. Design*, vol. 41, no. 10, pp. 2812–2817, 2020.
- [10] M. Liu, X. Xu, J. Hu, and Q. Jiang, "Real time detection of driver fatigue based on CNN-LSTM," *IET Image Process.*, vol. 16, no. 2, pp. 576–595, Feb. 2022.
- [11] D. M. Shi, F. Xiao, and J. Hu, "Study on driving behavior detection method based on improved long and short term memory network," *Automot. Eng.*, vol. 43, no. 8, pp. 1203–1209, 2021.
- [12] X. F. Feng, X. Xu, and J. Hu, "Fatigue driving detection method based on fusing eye mouth head feature," *J. Saf. Environ.*, vol. 22, no. 1, pp. 263–270, 2022.
- [13] X. Fu, H. Meng, X. Wang, H. Yang, and J. Wang, "A hybrid neural network for driving behavior risk prediction based on distracted driving behavior data," *PLoS ONE*, vol. 17, no. 1, Jan. 2022, Art. no. e0263030.
- [14] A. Malini, P. Priyadarshini, and S. Sabeena, "An automatic assessment of road condition from aerial imagery using modified VGG architecture in faster-RCNN framework," *J. Intell. Fuzzy Syst.*, vol. 40, no. 6, pp. 11411–11422, Jun. 2021.
- [15] R. Zhu, Y. Lv, Z. Wang, and X. Chen, "Prediction of the hypertension risk of the elderly in built environments based on the LSTM deep learning and Bayesian fitting method," *Sustainability*, vol. 13, no. 10, p. 5724, May 2021.
- [16] S. Gupta, S. Pawar, N. Ramrakhiyani, G. K. Palshikar, and V. Varma, "Semi-supervised recurrent neural network for adverse drug reaction mention extraction," *BMC Bioinf.*, vol. 19, no. S8, pp. 212–226, 2018.
- [17] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.
- [18] N. K. Karnam, S. R. Dubey, A. C. Turlapaty, and B. Gokaraju, "EMGHandNet: A hybrid CNN and bi-LSTM architecture for hand activity classification using surface EMG signals," *Biocybern. Biomed. Eng.*, vol. 42, no. 1, pp. 325–340, Jan. 2022.
- [19] J. Yu, "Short-term airline passenger flow prediction based on the attention mechanism and gated recurrent unit model," *Cogn. Comput.*, vol. 14, no. 2, pp. 693–701, Mar. 2022.
- [20] L. Wang, L. Wang, and S. Chen, "ESA-CycleGAN: Edge feature and self-attention based cycle-consistent generative adversarial network for style transfer," *IET Image Process.*, vol. 16, no. 1, pp. 176–190, Jan. 2022.
- [21] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020.
- [22] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelwagen, "Drive&Act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2801–2810.



**ZHANFENG WANG** was born in Shou County, Lu'An, Anhui, China, in 1984. She received the degree from Huaibei University, in 2007, and the master's degree from Anhui University, in 2010.

Since 2010, she has been a Lecturer with the School of Computer and Artificial Intelligence, Chaohu University. She has written more than ten articles. Her research interests include deep learning, machine learning, and image processing.



**LISHA YAO** was born in Anhui, China, in 1986. She received the master's degree in applied computer technology from Anhui University, in 2011, and the Ph.D. degree in computer science from the National University of the Philippines, in 2020.

She has been a Teacher with Anhui Xinhua University, since 2011. She is currently an Associate Professor. She presided more than six scientific research projects, published more than 20 papers in domestic and foreign academic journals and international conferences, and obtained one national invention patent.

...