frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 18 April 2024
DOI 10.3389/fphy.2024.1387338

Check for updates

OPEN ACCESS

EDITED BY
Yuanyuan Huang,
Chengdu University of Information Technology, China

REVIEWED BY
Li Chaorong,
Yibin University, China
Yufei Wang,
Jilin Engineering Normal University, China
Fei Wang,
Nanjing University of Aeronautics and Astronautics, China

*CORRESPONDENCE
Quan Wen,
✉ wenquan@jlu.edu.cn

RECEIVED 17 February 2024
ACCEPTED 02 April 2024
PUBLISHED 18 April 2024

CITATION
Luan X, Wen Q and Hang B (2024), Driver emotion recognition based on attentional convolutional network.
*Front. Phys.* 12:1387338.
doi: 10.3389/fphy.2024.1387338

COPYRIGHT
© 2024 Luan, Wen and Hang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Driver emotion recognition based on attentional convolutional network

Xing Luan[1], Quan Wen[1]* and Bo Hang[2]

[1]College of Communication Engineering, Jilin University, Changchun, China, [2]Hubei University of Arts and Science, Xiangyang, China

Unstable emotions, particularly anger, have been identified as significant contributors to traffic accidents. To address this issue, driver emotion recognition emerges as a promising solution within the realm of cyber-physical-social systems (CPSS). In this paper, we introduce SVGG, an emotion recognition model that leverages the attention mechanism. We validate our approach through comprehensive experiments on two distinct datasets, assessing the model's performance using a range of evaluation metrics. The results suggest that the proposed model exhibits improved performance across both datasets.

KEYWORDS

**road rage detection, driver emotion recognition, facial expression recognition, attention mechanism, deep learning**

## 1 Introduction

The driver emotion recognition has garnered substantial scholarly attention as a consequential application within cyber-physical-social systems. A driver emotion recognition system is structured into three pivotal layers: perception, cognition and decision, and interaction [1]. Within this framework, perception involves the deployment of sensors in the cockpit to meticulously acquire data pertaining to the driver's emotional state. Cognition and decision denote the integration of emotion recognition models with real-time data to analyze the driver's emotions. The interaction layer includes a vehicle warning system to detect and alert on the driver's emotional instability or fatigue. This paper endeavors to explore an approach to driver emotion recognition with a specific focus on the cognitive and decision layers. The emphasis lies in the sophisticated integration of emotion recognition models with dynamic data streams, facilitating a nuanced real-time analysis of the driver's emotional states.

A number of methods for driver emotion recognition by facial expression have emerged due to the low price of vision sensors and their easy installation and realization in the driving environment [2]. After acquiring the data, it is also critical to extract the emotional characteristics from the data. The attention mechanism [3] in deep learning is a way to mimic human vision, allowing neural networks to focus more on top of important information and improve the effectiveness of their models. Jiyong Xue proposed a deep convolutional model based on multi-head self-attention that fuses utterance-level acoustic features and frame-level acoustic features [4]. Wei Tao proposes an attention-based convolutional recurrent neural network (ACRNN) which extracted more discriminative features in EEG (electroencephalogram) signals through the attention mechanism [5]. These methods have yielded excellent results.

In the field of emotion recognition, researchers often choose the emotion model (sad, happy, fear, disgust, surprise, and angry) proposed by Ekman [6] as the starting point of research. From the perspective of preventing traffic accidents, some emotions are somewhat redundant. Including an excessive variety of emotions as the subject of study amplifies model redundancy and diminishes the recognition rate. Therefore, it is necessary to consider several emotions that are most relevant to drivers for the study. Research with drivers of varying ages has found that anger is associated with speeding, fear with stronger braking, lower speeds, and poorer lateral vehicle control [7]. Additionally, anger and happiness were found to be associated with more driving errors than fear or a neutral emotional state [8]. Henceforth, within this dissertation, we elect to investigate the emotional domains of happiness, anger, fear, and sadness as our primary research subjects. Employing the attention convolutional network, our objective is to discern and analyze drivers' facial expressions with precision and relevance.

# 2 Related work

## 2.1 Contact method and contactless method

According to the different information obtained, the types of driver emotion recognition can be divided into two types: contact method and contactless method [9]. The contact method uses special equipment to measure the drivers' physiological signals [10, 11], such as body temperature [12], electrocardiographic signals [13], skin electrical signal [14] et al. While the contactless method analyzes the drivers facial [15, 16] or voice information [17] through cameras or microphones. Using the contact method for emotion recognition has high accuracy and high real-time performance. However, the effect in actual use is often not satisfactory. This is not only because the drivers' physiological signals are inconvenient to obtain and the identification device is difficult to wear, but also because the device will cause psychological stress to the driver, which makes it impossible for drivers to drive vehicles in a relaxed environment [18].

Within the realm of contactless methods investigations, facial expressions predominantly serve as indicators of the driver's emotional state [19]. Through the scrutiny of the driver's facial image information, it becomes feasible to intuitively ascertain the driver's ongoing emotional state. Employing this methodology not only avoids causing any disturbance to the driver but also enables the continuous monitoring of the driver.

## 2.2 Face emotion recognition

Miyajia [20] uses Kohonen neural network as a classification algorithm to recognize the drivers' facial emotion and proposes an early warning method of the driver's angry state. However, the KNN solely concentrates on the emotion of anger and overlooks other emotions in drivers that might possibly lead to a collision. Alessandro [21] used the VGG (Visual Geometry Group) model to detect the drivers anger and recognized the continuous image by sliding the window. Both frontal and non-frontal facial expressions have been explored in literature, however, their precision falls short when compared to contemporary methods. Geesung [22] used a variety of CNN (Convolutional Neural Networks) models to test the driver's facial expression and acquired the driver's skin electric signal to judge the driver's emotion synthetically. The DRER model proposed in the literature has an accuracy of 88.6%, but it only identifies emotional states for a short period. However, emotions are continuous, and the model requires improvement to identify emotional states in real-time. H. Varun Chand [16] presented a multi-layer drowsiness detection system based on CNN and emotion analysis, which achieved an accuracy rate of 93%. The spatial transformer network [23] adjusts the image by learning spatial transformations and applying them to the original image. This approach significantly reduces the interference of environmental factors in the extraction of emotional features. However, real-time monitoring of the driver's mood has not yet been attained. In this work, benefiting from attention convolutional neural networks, we obtain higher recognition accuracy while maintaining recognition speed.

# 3 Methods

## 3.1 Preprocessing

The preprocessing stage mainly includes face detection and segmentation. When the camera captures an image of the driver, the first step is to extract the facial information in the image. In this paper, we use OpenCV (Open-Source Computer Vision Library) to detect face Haar features [24] by loading the pre-trained classifiers. The advantage of this method is that the recognition speed is fast, it can be used for real-time detection, and it has a high recognition rate. At the same time, the model is small and can be run on an embedded platform, which is more suitable for the scene of driver face detection.

## 3.2 The proposed framework

Figure 1 illustrates the structural framework of SVGG, a driver emotion recognition model based on attention mechanisms. When the driver's face image information is acquired, the face image is first input into the STN model for face alignment to reduce the interference of environmental factors. Then the processed image is inputted into the improved VGG network model for feature extraction of the image in depth, in which the convolution adopts the Ghost Module to reduce the parameters of the model and speed up the inference speed of the network, and the activation function adopts the Mish function to speed up the convergence speed of the model. After obtaining the feature maps with complete feature extraction, the channel attention model ECA-Net will redistribute the weights of the feature maps in the channel, and finally the feature maps of multiple channels will be passed through a fully connected layer for emotion classification and recognition.

### 3.2.1 Spatial Transformer Networks

Spatial Transformer Networks (STNs) [23] is a neural network model as well as a spatial attention mechanism. Among the ways of spatial transformations are translation, rotation or scaling. The

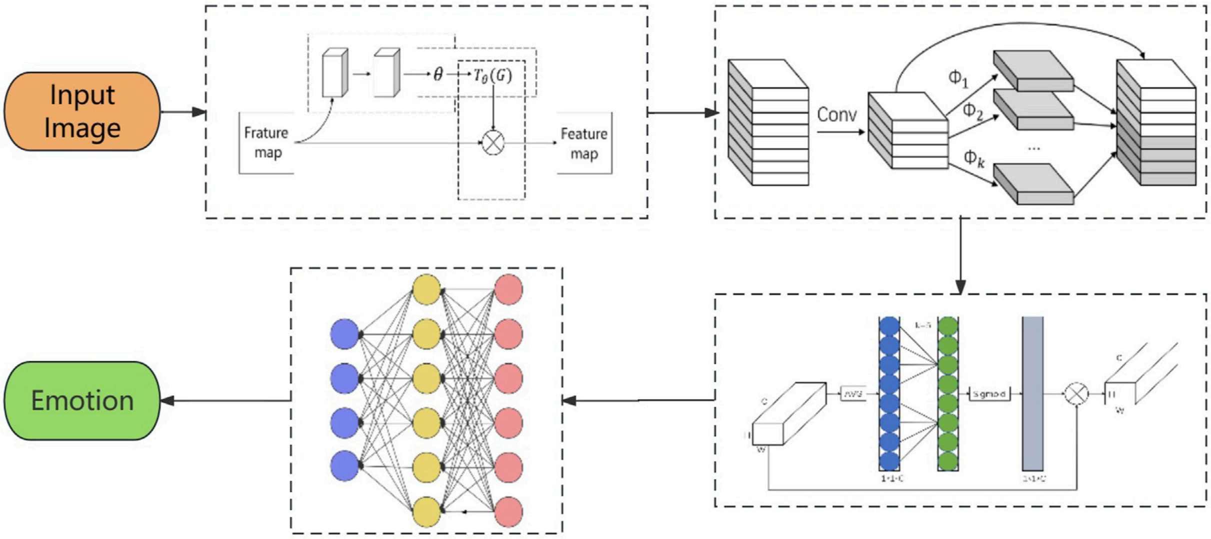

FIGURE 1
Framework of the model.

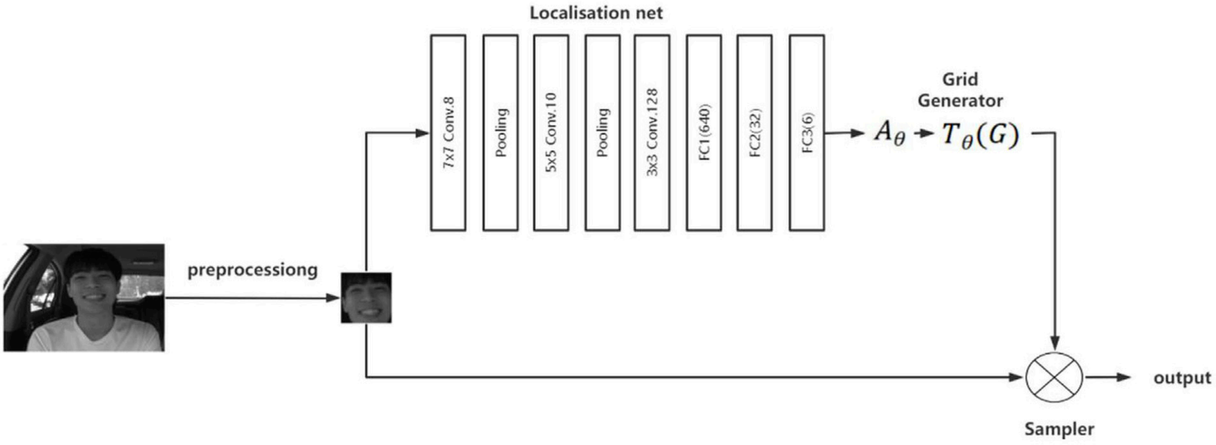

FIGURE 2
Spatial transform network. Reprinted with permission from "Driver's Facial Expression Recognition in Real-Time for Safe Driving" by Mira Jeong, ByoungChul Ko. The license for this content can be found at https://cvpr.kmu.ac.kr/KMU-FED.htm.

TABLE 1 Comparison of model parameter sizes.

| Model | No. of parameters (in million) |
|---|---|
| VGG19 | 20.1 |
| MobileNet | 3.2 |
| LiveEmoNet [30] | 1.3 |
| CNN [31] | 1.3 |
| SVGG | 10 |

advantage of the spatial transformation network is that the network can autonomously learn certain key changes in the natural image without human labeling, and thus adjust the image so that the network focuses on these changes. Theoretically, the spatial transform network can be added to any layer of the convolutional neural network, but in practice, most researchers add the network to the convolutional neural network before preprocessing the original image, and then input the image to the convolutional layer for feature extraction, so as to ensure the integrity of the original input data. The spatial transform network only adjusts the pixel positions of the original image, and does not adjust the size of the original image, i.e., the input and output images of the spatial transform network have the same size.

The spatial transformation network model was introduced in 2015, which contains three models: Localization net, Grid generator, and Sampler. The model structure diagram is shown in Figure 2. When we get the driver's face image after preprocessing, the image

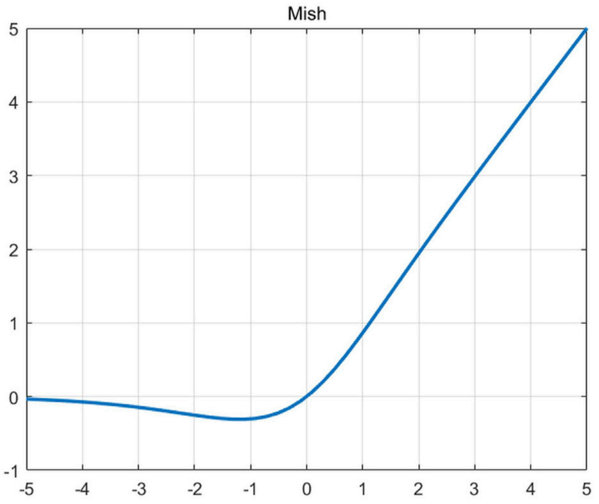

FIGURE 3
Mish activation function.

TABLE 2 Comparison of experimental results on the KMU-FED dataset.

| Model | Accuracy (%) | Recognition speed (ms/frame) |
|---|---|---|
| VGG19 | 94.3 | 140 |
| SVGG | 96.6 | 80 |

will go through the spatial transformation network to extract the key regions of the face. First through the Localization net to generate affine transformation $A_\theta$. The Grid generator used affine transformation $A_\theta$ to create a sampling grid. At last, the drivers face image and the sampling grid are taken as inputs to the Sampler. And we can finally get the most relevant parts of a face image from Sampler.

Localization net is a small convolutional neural network used to generate affine transformation parameters. The input image shape is 48*48*1 and the output is $A_\theta$, the affine transformation $A_\theta$ has 6 parameters, as Eq. 1. These parameters are constantly optimized during the training process to identify the most relevant face regions in an image.

$$\boldsymbol{A_\theta} = \begin{bmatrix} \boldsymbol{\theta_{11}, \theta_{12}, \theta_{13}} \\ \boldsymbol{\theta_{21}, \theta_{22}, \theta_{23}} \end{bmatrix} \tag{1}$$

Grid generator uses the affine transformation matrix $A_\theta$ to create. Assume that pixel in the input image is $(x_i^s, y_i^s)$, the pixel in the output image is $(x_t^i, y_t^i)$, and the corresponding relationship between an input image pixel and output image pixel is as shown in Eq. 2. It can obtain the values of the output image pixel values by taking the inverse.

$$\begin{pmatrix} \boldsymbol{x_i^s} \\ \boldsymbol{y_i^s} \end{pmatrix} = \boldsymbol{T_\theta(G_i)} = \boldsymbol{A_\theta} \begin{pmatrix} \boldsymbol{x_i^t} \\ \boldsymbol{y_i^t} \\ \boldsymbol{1} \end{pmatrix} = \begin{bmatrix} \boldsymbol{\theta_{11}} & \boldsymbol{\theta_{12}} & \boldsymbol{\theta_{13}} \\ \boldsymbol{\theta_{21}} & \boldsymbol{\theta_{22}} & \boldsymbol{\theta_{23}} \end{bmatrix} \begin{pmatrix} \boldsymbol{x_i^t} \\ \boldsymbol{y_i^t} \\ \boldsymbol{1} \end{pmatrix} \tag{2}$$

TABLE 3 Performance in KMU-FED by 5-fold cross-validation.

| Model | Accuracy (%) |
|---|---|
| SqueezeNet | 89.7 |
| Modified SqueezeNet | 95.8 |
| MobileNetV2 | 93.8 |
| MobileNetV3 | 94.9 |
| LMRF | 95.1 |
| SVGG | 96.6 |

The sampler module is employed for the execution of spatial transformations. It applies the previous sampling grid to the input image, to produce the final image which is the most relevant parts of a face image.

### 3.2.2 VGG network

The VGG network [23] is employed for recognizing emotions in driver face images that are processed by the Spatial Transform Network. The VGG19 [23] network contains five convolutional groups and 16 convolutional layers, each of which contains a large number of redundant computations. For the special environment of driving, in order to make the VGG network have a more efficient recognition effect, this paper is influenced by the idea of "Cheap Operations" in Ghost Net, and adopts the convolutional method of Ghost Module to improve the convolutional layers in the VGG network. In an ordinary convolution operation, assuming the input feature map

$$\boldsymbol{X \in R^{h^* w^* c}} \tag{3}$$

where $h$ is the height of the input image, $w$ is the width of the feature map, and c is the number of channels of the feature map, and the output feature map

$$\boldsymbol{Y \in R^{h'^* w'^* n}} \tag{4}$$

where, $h'$ and $w'$ are the height and width of the feature map, respectively, and n is the number of channels of the feature map, and the convolution kernel is, and k is the size of the kernel, and n is the number of the kernels of the convolution, the computation of this convolution operation can be expressed in Eq. 5:

$$\boldsymbol{F = h'^* w'^* n^* c^* k^* k} \tag{5}$$

In convolutional operations, the number of n and c is high, so the feature maps generated by ordinary convolution consume a high amount of computation and have a large redundancy. Ghost Module, on the other hand, utilizes the redundancy of convolutional operations and uses simple convolutional and linear operations to obtain the same feature maps as normal convolution, which is called "Cheap Operations". Specifically, Ghost Module first uses the regular convolution to generate the eigenfeature map $Y'$, which contains m feature maps, the number of m is less than n. The amount of computation $F_1$ needed after omitting the bias term in the convolution operation can be expressed as Eq. 6:

$$\boldsymbol{F_1 h'^* w'^* m^* c^* k^* k} \tag{6}$$

After that, the Ghost Module performs linear operations on the obtained eigenfeature maps $Y'$ to generate phantom feature maps, and the process can be expressed as Eq. 7:

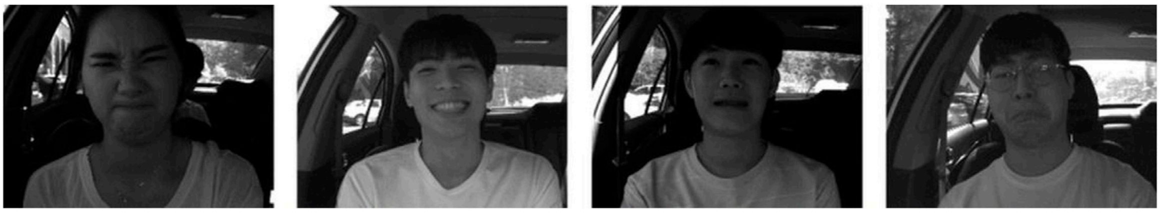

FIGURE 4
The four expression images contained in the KMU-FED dataset. Reprinted with permission from "Driver's Facial Expression Recognition in Real-Time for Safe Driving" by Mira Jeong, ByoungChul Ko. The license for this content can be found at https://cvpr.kmu.ac.kr/KMU-FED.htm.

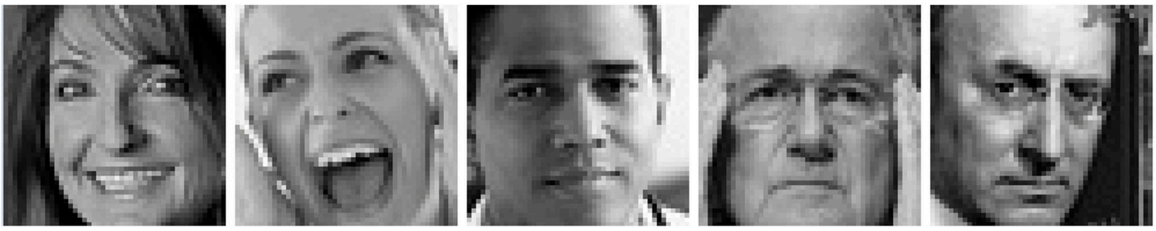

FIGURE 5
The five expression images contained in the FER-2013 dataset. Reprinted with permission from "Challenges in Representation Learning: Facial Expression Recognition Challenge" by Dumitru, Ian Goodfellow, Will Cukierski, Yoshua Bengio. https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge.

$$y_{ij} = \varnothing_{i,j}\left(y_i^{'}\right), \forall i = 1, 2, ..., m, j = 1, 2, ..., s \tag{7}$$

Comparing the convolution formed by Ghost Module with ordinary convolution, it can be seen that the computation of Ghost Module is only 1/s of ordinary convolution, as shown in Eq. 8. In this paper, s is fixed to 2, i.e., compared with the ordinary convolution in the VGG network, the computation of the convolution layer in the improved VGG will be reduced to 1/2. The total number of parameters for SVGG is 10,067,914 compared to the literature as shown in Table 1.

$$r_s = \frac{n * h' * w' * c * k * k}{\frac{n}{s} * h' * w' * c * k * k + (s-1)\frac{n}{s} * h' * w' * d * d} \approx s \tag{8}$$

The activation function in VGG networks is the ReLU function, which is a simple and effective nonlinear activation function and has been widely used in many neural network models. However, it has an obvious drawback: the "dead neuron" problem. In the ReLU function, when the value of the input is less than 0, the output value will always be 0, which means the neuron is "dead". As the number of neurons increases, the number of dead neurons will also increase, resulting in some neurons cannot be effectively used in the backpropagation. Also, this function does not solve the problem of vanishing gradients. This problem can be solved by using the Mish activation function, whose functional formula is shown in Eq. 9 and the graph is shown in Figure 3.

$$f(x) = x^{*} \tanh(\ln(1 + e^{x})) \tag{9}$$

### 3.2.3 ECA-Net

ECA-Net [25] represents a lightweight and efficient channel attention model proficient in capturing inter-channel feature map information with the introduction of a minimal number of parameters. Unlike the SENet (Squeeze-and-Excitation Networks) structure, ECA-Net does not use the fully connected layer in SENet, but chooses to use one-dimensional convolution to dynamically adjust the size of the kernel, and establishes the relationship between the feature channels and the size of the convolution kernel using an approximate linear mapping to achieve the ability of focusing on channel convolution information exchange while avoiding dimensionality degradation.

# 4 Experiment

## 4.1 Datasets

The KMU-FED dataset [26] was selected as the primary dataset to assess the proposed approach. In order to evaluate the performance of the model more thoroughly, the FER-2013 dataset was also selected for testing (Figure 4).

KMU-FED dataset is a real-world driver's facial image dataset collected by the CVPR laboratory of Keimyung University, contains 1,106 emotional images about drivers, the picture pixel is 1,600*1,200. It is a dataset of driver emotions in real environments captured by infrared cameras placed on the steering wheel or dashboard and contains emotional pictures of multiple drivers under different lighting conditions. It contains six emotions: anger, surprise, happiness, fear, disgust, and sad.

The FER-2013 dataset [32], compiled by Google in 2013, comprises 35,887 grayscale images stored in a CSV file with dimensions of 48 × 48 pixels. These images are categorized into seven emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. Despite being

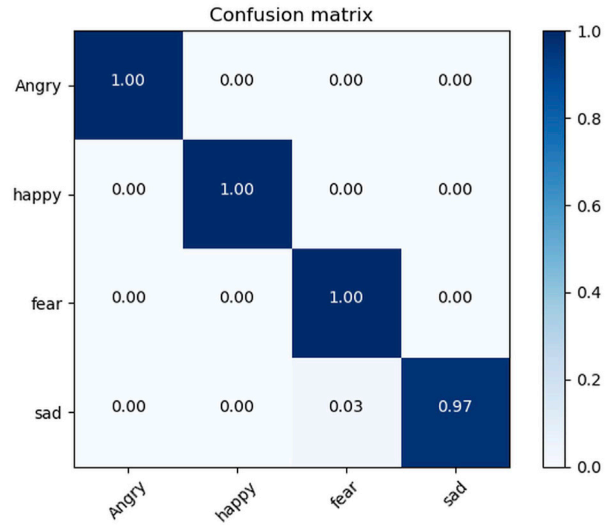

FIGURE 6
Confusion matrix in KMU-FED.

TABLE 4 Effect of different modules in the model on experimental results.

| STN | Convolution | Activation | ECA-net | Accuracy (%) |
|---|---|---|---|---|
| - | - | - | - | 70.4 |
| √ | - | - | - | 72.0 |
| √ | √ | - | - | 71.5 |
| √ | √ | √ | - | 71.8 |
| √ | √ | √ | √ | 72.4 |

TABLE 5 Comparison of results of different models.

| Model | Accuracy (%) |
|---|---|
| MobileNetV2 | 68.3 |
| SqueezeNet | 64.5 |
| LiveEmoNet [30] | 69.0 |
| CNN(31) | 65.0 |
| SVGG | 72.4 |

originally designed for general in-the-wild conditions and including animated characters displaying diverse emotions, this dataset is not explicitly tailored for driver-centric scenarios (Figure 5).

For our analysis, we specifically focus on four emotions, resulting in a subset of 26,217 images. This subset includes 4,953 images depicting anger, 8,989 images displaying happiness, 5,121 images conveying fear, and 6,077 images depicting sadness. Notably, since all images in this dataset represent facial expressions, no additional image preprocessing steps were applied.

## 4.2 Training procedures and evaluation criteria

Prior to discussing model performance, we will provide a brief overview of the training procedures and evaluation criteria used in this study. Training epochs of the model is set to 300, and an early stopping strategy is used to avoid overfitting. When the validation accuracy did not improve in 30 iterations, the training would stop. The batch size is set to 128, and an Adam algorithm [27], a useful optimizer, is set to optimize the model parameters. Adam optimizers learning rate is set to 0.001.

In order to comprehensively validate the effectiveness of the Ghost Net-SSD algorithm, it is necessary to comprehensively evaluate both the average precision and the recognition speed of the algorithm recognition. Among them, the average precision (AP) is calculated by combining the recognition accuracy (Precision) and the recall rate (Recall).

Recognition Precision is the probability that a face is correctly detected among all detected samples, assuming that the number of samples in which a face is detected is TP and the number of samples in which a non-face is detected is FP, then the formula for the Recognition Precision (Precision) can be expressed as Eq. 10.

Recall is the probability of correctly detected faces among all faces that should be detected. Assuming that the number of incorrectly detected non-face samples is FN, the formula for recall (Recall) can be expressed as Eq. 11.

The P-R curve can be established with the recognition precision rate as the vertical coordinate and the recall rate as the horizontal coordinate, then the area of the curve combined with the axes is the Average Precision (AP, Average Precision), which is calculated as shown in Eq. 12.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{10}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{11}$$

$$\boldsymbol{AP} = \int_{0}^{1} \boldsymbol{P}(\boldsymbol{R})\boldsymbol{dR} \tag{12}$$

The recognition speed is calculated by synthesizing the recognition speed of multiple images by the model on the experimental platform to determine whether the algorithm meets the real-time requirements.

## 4.3 Results and analysis

### 4.3.1 KMU-FED dataset

In order to verify the effectiveness of emotion recognition in real driving environments, the SVGG model and the VGG19 model are compared and experimented on the KMU-FED dataset, and the experimental results are shown in Table 2. As evidenced by the data presented in Table 2, the SVGG model achieves a recognition accuracy of 96.6% on the KMU-FED dataset, surpassing the accuracy of the VGG19 model at 94.3%. This signifies an improvement of 2.3% relative to the VGG19 model. Regarding recognition speed, the SVGG model operates at 80 m/frame, while the VGG19 model exhibits a recognition speed of 140 m/frame. This

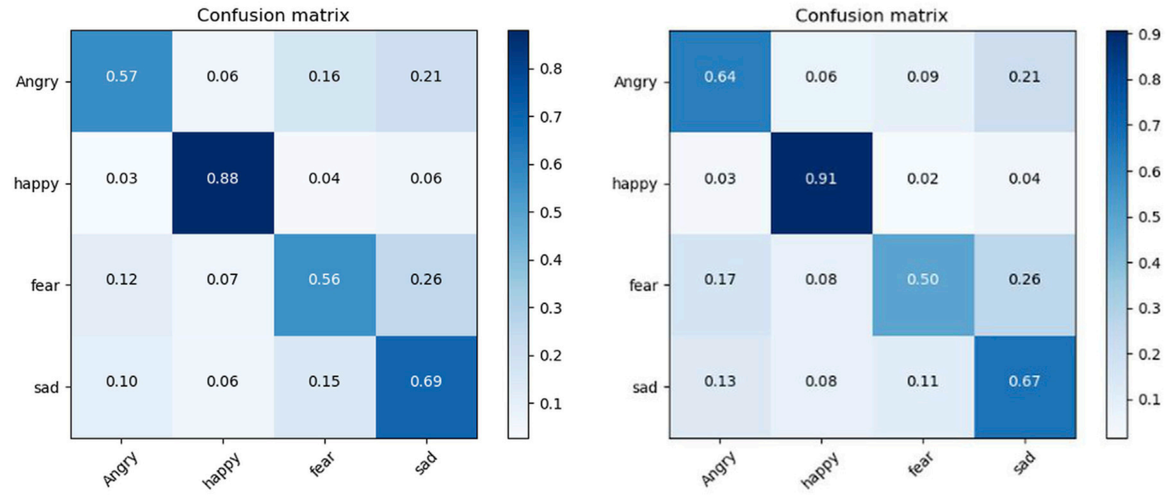

FIGURE 7
Confusion matrix in FER-2013 used VGG left and our method right.

represents a significant improvement of 43% in comparison to the VGG19 network model.

Table 3 demonstrates the results of comparing the recognition accuracy of different methods on the KMU-FED dataset, from the data in the table, it can be seen that the SVGG model is higher than the classical network model in terms of recognition accuracy. Specifically, compared to the Squeeze Net and MobileNetV3 models, the recognition accuracy of the SVGG model has improved by 6.9% and 1.7%. Meanwhile, comparing the new network models proposed in recent years, such as LMRF [28], Modified Squeeze Net [29], *etc.*, the SVGG model also has a more obvious advantage in terms of recognition accuracy, in which the recognition accuracy of the SVGG model is improved by 1.5% compared with the LMRF model, and compared with Modified Squeeze Net. SVGG model has improved the recognition accuracy by 0.8%.

Figure 6 shows the confusion matrix of the model in the case of using the KMU-FED dataset. The horizontal axis of the confusion matrix represents the predicted emotion categorization, the vertical axis represents the true emotion categorization, and the diagonal of the matrix represents the correct recognition rate of emotions. In the figure, the model demonstrates precise recognition of anger, happiness, and fear. Notably, only a minimal number of instances portraying sad expressions are misclassified as fear.

### 4.3.2 FER-2013 dataset

The experiments conducted in the preceding subsection confirmed the effectiveness of the SVGG model in real-world driving scenarios. In this subsection, additional experiments will be carried out on the FER-2013 dataset to assess the model's generalization capabilities First of all, this paper does is experiments on the improvement effect of different modules on the overall model. The experimental results are shown in Table 4. Where, the mark "√" indicates that the module is used in the model, and the mark "-" indicates that the module is not used in the model.

From the results of the experiment in Table 4, it can be seen that the recognition accuracy of the original model is 70.4%, and the face alignment implemented by the STN module has the most obvious effect on the model recognition accuracy improvement, which can improve the recognition accuracy by 1.6%. The improvement to the convolution module will make the number of parameters of the model decrease significantly and improve the recognition speed of the model, but from the results of this experiment, the improvement to the convolution module will make the model recognition accuracy decrease slightly. The activation function module and the ECA-Net module both have positive enhancement effects for the model. The recognition accuracy of the SVGG model on the FER-2013 dataset is 72.4%, which is an improvement of 2.0% compared to the original model, which verifies the validity of the model proposed in this paper.

To more thoroughly evaluate the efficacy of the proposed SVGG model, it is essential to conduct comparative experiments with other emotion recognition models. In this paper, experimental comparisons are conducted on the FER-2013 dataset. The experimental results are shown in Table 5.

From the results of the comparison experiments in Table 5, it can be seen that in terms of recognition accuracy, the SVGG model has a significantly higher recognition accuracy than some lightweight network models. For example, compared with MobileNetV2, the average accuracy of SVGG model is improved by 4.1%, and compared with Squeeze Net, the average accuracy of SVGG model is improved by 7.9%. Also, the recognition accuracy of SVGG model is better than network models optimized for emotion recognition in recent years, e.g., compared with LiveEmoNet, the recognition accuracy of SVGG model is improved by 3.4%. Compared with CNN, the recognition accuracy of SVGG model is improved by 7.4%.

Figure 7 shows the accuracy of the two models under the confusion matrix model. Contrasting the anger that causes road rage, it can be seen from the figure that our model's recognition rate of anger is 7% higher than that of the VGG model. At the same time, it can be seen that the recognition rates of fear and sadness are not high for both two models, and it is easy to misjudge these two emotions. This is due to the relatively small amount of data for these two emotions compared to the other emotions in the dataset.

# 5 Conclusion

In this paper, an emotion recognition model SVGG based on the attention mechanism is proposed. The SVGG model addresses picture jitter in driving environments through a spatial transformation network for facial alignment. To enhance recognition speed, it employs the Ghost Module's convolution method and the Mish activation function for accelerated convergence. Addressing low accuracy, the SVGG model utilizes ECA-Net to redistribute output channel weights. The experimental results demonstrate that our model achieves a 43% improvement in processing speed, with a rate of 80 milliseconds per frame, compared to the VGG model. Furthermore, it attains an accuracy of 96.6% on the KMU-FED dataset and 72.4% on the FER-2013 dataset, suggesting its high potential for practical applications.

This paper focuses on analyzing frontal or partially frontal face images of drivers, emphasizing specific camera placement requirements within the driving environment. Future research aims to explore emotion recognition from the driver's side face, potentially overcoming camera placement limitations. Additionally, we consider integrating the perception and interaction layers to form a more holistic Cyber-Physical Social System, enhancing the system's overall functionality in driver-assistance technologies.

# Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data https://cvpr.kmu.ac.kr/KMU-FED.htm.

# Author contributions

XL: Data curation, Methodology, Software, Writing–original draft. QW: Conceptualization, Funding acquisition, Methodology, Resources, Writing–review and editing. BH: Formal Analysis, Investigation, Supervision, Writing–review and editing.

# Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Jilin Science and Technology Development Program Project: Research on Intelligent Recognition and Prevention of Driver's Emotions under Multimodality, Project No. 20230203032SF.

# Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# References

1. Li W, Wu L, Wang C, Xue J, Hu W, Li S, et al. Intelligent cockpit for intelligent vehicle in metaverse: a case study of empathetic auditory regulation of human emotion. *IEEE Trans Syst Man, Cybernetics: Syst* (2023) 53(4):2173–87. doi:10.1109/tsmc.2022.3229021

2. Yang L, Yang H, Hu BB, Wang Y, Lv C. A Robust driver emotion recognition method based on high-purity feature separation. *IEEE Trans Intell Transportation Syst* (2023) 24(12):15092–104. doi:10.1109/tits.2023.3304128

3. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv* (2016). Available from: http://arxiv.org/abs/1409.0473. doi:10.48550/arXiv.1409.0473

4. Xue J, Li W, Zhang Y, Xiao H, Tan R, Xing Y, et al. Driver's speech emotion recognition for smart cockpit based on a self-attention deep learning framework. In: *2021 5th CAA international conference on vehicular control and intelligence (CVCI)* (2021). p. 1–5. Available from: https://ieeexplore.ieee.org/document/9661268.

5. Tao W, Li C, Song R, Cheng J, Liu Y, Wan F, et al. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Trans Affective Comput* (2023) 14(1):382–93. doi:10.1109/taffc.2020.3025777

6. Ekman P. Facial expression and emotion (1993). Available from: https://www.semanticscholar.org/paper/Facial-Expression-and-Emotion-Ekman/b0153a91c7124644f8515625e3a0e41193b2fc23.

7. Roidl E, Frehse B, Höger R. Emotional states of drivers and the impact on speed, acceleration and traffic violations—a simulator study. *Accid Anal Prev* (2014) 70:282–92. doi:10.1016/j.aap.2014.04.010

8. Jeon M, Walker BN, Yim JB. Effects of specific emotions on subjective judgment, driving performance, and perceived workload. *Transportation Res F: Traffic Psychol Behav* (2014) 24:197–209. doi:10.1016/j.trf.2014.04.003

9. Oh G, Jeong E, Kim RC, Yang JH, Hwang S, Lee S, et al. Multimodal data collection system for driver emotion recognition based on self-reporting in real-world driving. *Sensors* (2022) 22(12):4402. doi:10.3390/s22124402

10. Singh RR, Conjeti S, Banerjee R. Biosignal based on-road stress monitoring for automotive drivers. In: 2012 National Conference on Communications (NCC); 03-05 February 2012; Kharagpur, India (2012). p. 1–5. Available from: https://ieeexplore.ieee.org/document/6176845.

11. Singh RR, Conjeti S, Banerjee R. An approach for real-time stress-trend detection using physiological signals in wearable computing systems for automotive drivers. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC); 05-07 October 2011; Washington, DC, USA (2011). p. 1477–82. Available from: https://ieeexplore.ieee.org/document/6082900.

12. Muhammad G, Hossain MS. Light deep models for cognitive computing in intelligent transportation systems. *IEEE Trans Intell Transportation Syst* (2023) 24(1):1144–52. doi:10.1109/tits.2022.3171913

13. Prasolenko O, Lobashov O, Bugayov I, Gyulyev N, Filina-Dawidowicz L. Designing the conditions of road traffic in the cities taking into account the human factor. In: 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS); 05-07 June 2019; Cracow, Poland (2019). p. 1–8. Available from: https://ieeexplore.ieee.org/document/8883381.

14. Lingelbach K, Bui M, Diederichs F, Vukelić M. Exploring conventional, automated and deep machine learning for electrodermal activity-based drivers' stress recognition. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 17-20 October 2021; Melbourne, Australia (2021). p. 1339–44. Available from: https://ieeexplore.ieee.org/document/9658662.

15. Ujir H, Jee EM, Farhaan Iqbal M, Mun QK, Hipiny I. Real-time driver's monitoring mobile application through head pose, drowsiness and angry detection. In: 2021 8th International Conference on Computer and Communication Engineering (ICCCE); 22-23 June 2021; Kuala Lumpur, Malaysia (2021). p. 1–6. Available from: https://ieeexplore.ieee.org/document/9467232.

16. Chand V, Karthikeyan J. CNN based driver drowsiness detection system using emotion analysis. *Intell Automation Soft Comput* (2022) 31:717–28. doi:10.32604/IASC.2022.020008

17. Du G, Wang Z, Gao B, Mumtaz S, Abualnaja KM, Du C. A convolution bidirectional long short-term memory neural network for driver emotion recognition. *IEEE Trans Intell Transportation Syst* (2021) 22(7):4570–8. doi:10.1109/tits.2020.3007357

18. Li W, Cui Y, Ma Y, Chen X, Li G, Zeng G, et al. A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: emotions triggered by video-audio clips in driving scenarios. *IEEE Trans Affective Comput* (2023) 14(1):747–60. doi:10.1109/taffc.2021.3063387

19. Luo J, Yoshimoto H, Okaniwa Y, Hiramatsu Y, Ito A, Hasegawa M. Emotion monitoring sensor network using a drive recorder. In: 2023 IEEE 15th International Symposium on Autonomous Decentralized System (ISADS); 15-17 March 2023; Mexico City, Mexico (2023). p. 1–8. Available from: https://ieeexplore.ieee.org/document/10092139

20. Miyajia M. Driver's anger state identification by using facial expression in cooperation with artificial intelligence. *J Fundam Appl Sci* (2017) 9(7S):87–97. doi:10.4314/JFAS.V9I7S.9

21. Leone A, Caroppo A, Manni A, Siciliano P. Vision-based road rage detection framework in automotive safety applications. *Sensors* (2021) 21(9):2942. doi:10.3390/s21092942

22. Oh G, Ryu J, Jeong E, Yang JH, Hwang S, Lee S, et al. DRER: deep learning–based driver's real emotion recognizer. *Sensors* (2021) 21(6):2166. doi:10.3390/s21062166

23. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. *arXiv* (2016). Available from: http://arxiv.org/abs/1506.02025. doi:10.48550/arXiv.1506.02025

24. Viola P, Jones MJ. *Robust real-time face detection*.

25. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-net: efficient Channel Attention for deep convolutional neural networks. In: *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2020). p. 11531–9. Available from: https://ieeexplore.ieee.org/document/9156697.

26. Jeong M, Ko BC. Driver's facial expression recognition in real-time for safe driving. *Sensors* (2018) 18(12):4270. doi:10.3390/s18124270

27. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv* (2017). Available from: http://arxiv.org/abs/1412.6980. doi:10.48550/arXiv.1412.6980

28. Jeong M, Nam J, Ko BC. Lightweight multilayer random forests for monitoring driver emotional status. *IEEE Access* (2020) 8:60344–54. doi:10.1109/access.2020.2983202

29. Sahoo GK, Das SK, Singh P. Deep learning-based facial emotion recognition for driver healthcare. In: 2022 National Conference on Communications (NCC).; 24-27 May 2022; Mumbai, India (2022). p. 154–9. Available from: https://ieeexplore.ieee.org/document/9806751.

30. Podder T, Bhattacharya D, Majumdar A. Time efficient real time facial expression recognition with CNN and transfer learning. *Sādhanā* (2022) 47(3):177. doi:10.1007/s12046-022-01943-x

31. Kaviya P, Arumugaprakash T. Group facial emotion analysis system using convolutional neural network. In: 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184); 15-17 June 2020 (2020). p. 643–7. Available from: https://ieeexplore.ieee.org/document/9143037.

32. Dumitru Goodfellow L, Cukierski W, Bengio Y. Challenges in Representation Learning: Facial Expression Recognition Challenge. *Kaggle* (2013). Available online at: https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge.