MDEmoNet: A Multimodal Driver Emotion Recognition Network for Smart Cockpit

Chenhao Hu
School of Electronics and Information
Hangzhou Dianzi University
Equipment Electronics Key Lab
Hangzhou, China
chenhaohu@hdu.edu.cn

Shenyu Gu
School of Electronic and Information
Hangzhou Dianzi University
Equipment Electronics Key Lab
Hangzhou, China
shenyugu@hdu.edu.cn

Mengjie Yang
School of Electronic and Information
Hangzhou Dianzi University
Equipment Electronics Key Lab
Hangzhou, China
mengjieyang@hdu.edu.cn

Gang Han
National Engineering Laboratory for Wireless Security
Xi'an University of Posts and Telecommunications
Xi'an, China
hangang668866@xupt.edu.cn

Chun Sing Lai*
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
chunsing.lai@brunel.ac.uk

Mingyu Gao
School of Electronics and Information
Hangzhou Dianzi University
Equipment Electronics Key Lab
Hangzhou, China
mackgao@hdu.edu.cn

Zhexun Yang
Faculty of Materials Science and Chemistry
China University of Geosciences Wu han
Wuhan, China
2030059839@qq.com

Guojin Ma*
School of Electronics and Information
Hangzhou Dianzi University
Equipment Electronics Key Lab
Hangzhou, China
magj@hdu.edu.cn

***Abstract*—The automotive smart cockpit is an intelligent and connected in-vehicle consumer electronics product. It can provide a safe, efficient, comfortable, and enjoyable human-machine interaction experience. Emotion recognition technology can help the smart cockpit better understand the driver's needs and state, improve the driving experience, and enhance safety. Currently, driver emotion recognition faces some challenges, such as low accuracy and high latency. In this paper, we propose a multimodal driver emotion recognition model. To our best knowledge, it is the first time to improve the accuracy of driver emotion recognition by using facial video and driving behavior (including brake pedal force, vehicle Y-axis position and Z-axis position) as inputs and employing a multi-task training approach. For verification, the proposed scheme is compared with some mainstream state-of-the-art methods on the publicly available multimodal driver emotion dataset PPB-Emo.**

Keywords—smart cockpit, driver emotion recognition, deep learning， multimodal fusion

I. INTRODUCTION

Relying on the development of artificial intelligence [1-4] and computing systems [5-8], smart cars are rapidly gaining popularity and becoming widespread worldwide. The popularization of smart cars has brought about dramatic changes in smart cockpits, an in-vehicle consumer electronics product. Driver emotion recognition technology infers the driver's emotional state by analyzing a variety of data such as the driver's facial expression, voice, physiological signal and behavior, helping the smart cockpit better understand the driver's state and needs. Thus, the smart cockpit can provide appropriate services and support. Driver's negative emotions are one of the main causes of traffic accidents. Emotions can affect a driver's decision-making, alertness and driving behavior, and even lead to road rage and aggressive behavior [9, 10]. Therefore, the development of automatic driver emotion recognition and corresponding human-computer interaction systems to alleviate drivers' negative emotions and avoid traffic accidents is of great value in improving driving safety.

Driver emotion recognition technology can be divided into invasive and non-invasive methods. Invasive methods require drivers to wear physiological sensors or devices, which can provide more accurate recognition results but may affect driving safety and comfort. Non-invasive methods infer emotions by analyzing drivers' facial expressions, voice features, and behavioral actions, without contact with the driver directly, making drivers safer and more comfortable. However, currently, the accuracy of non-intrusive emotion recognition is relatively low, and improving its accuracy is a crucial research direction. Studies [11] have shown significant differences in facial action units between driving scenarios and other life situations, highlighting the importance of using driving scenario datasets in driver emotion recognition research. Research [12, 13] has also demonstrated that fusing multimodal data leads to higher emotion recognition performance compared to using single modal alone, as the latter may lack robustness in complex situations. Moreover, the potential role of driving behavior data in driver emotion recognition remains largely unexplored. Furthermore, multi-task learning, which leverages shared information from different tasks, enhances model performance and generalization, making it particularly suitable for training multimodal networks. Therefore, in our research, we utilize facial video and behavior data (including brake pedal force, vehicle Y-axis, and Z-axis positions) as inputs, conducting non-invasive emotion recognition studies in driving scenarios, and adopting a multi-task learning approach for training. The main contributions of our work are summarized as follows:

- We first propose and design a multimodal driver emotion recognition network based on facial video and driving behavior.

This work was supported by the National Natural Science Foundation of China (62206062), the Zhejiang Provincial Major Research and Development Project of China (2023C01132), the Hangzhou Major Science and Technology Innovation Project of China (2022AIZD0009), the Central Guiding Local Science and T
and the Zhejiang Provincial Key Lab of Equipment Electronics.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICCE59016.2024.10444365, 2024 IEEE International Conference on Consumer Electronics (ICCE).

- We design the driving behavior feature extractor that can improve emotion recognition accuracy and speed up computation.
- We adopt multi-task learning, which can more effectively integrate the two modalities compared to single-task learning.

II. Related Work

A. Emotion Classification

Typically, researchers classify emotions based on two models. One is the discrete emotion model [14], which categorizes emotions into different classes. In the 20th century, Ekman and Friesen [15] defined six cross-cultural basic emotions: anger, disgust, fear, happiness, sadness, and surprise. Later, contempt was added as one of the basic emotions [16]. The other model is the dimensional emotion model, which uses multiple dimensions to label emotions. Most dimensional emotion models include valence and arousal [17] as two dimensions. In the research of driver emotion recognition, the discrete emotion model is more commonly used. This is because the discrete emotion model is more intuitive and easier to interpret, providing concise emotion classification results that are easier for drivers and researchers to understand and apply. This paper also adopts the discrete emotion model.

B. Driver Emotion Recognition

Psychologists such as Albert Mehrabian proposed that emotional expression is 55% through facial expression [18]. Facial expressions are considered the most powerful way of expressing emotions. Researchers have proposed different models for facial video-based emotion recognition [19]. For example, Meng D et al. introduced the Emotion-FAN model [20], which automatically focuses on frames with distinctive features. Zhao Z et al. proposed the Former-DFER model [21], which learns facial features in both spatial and temporal dimensions. Previous studies [11] have shown that human emotional expressions in driving scenarios differ from those in other life situations, making it essential to conduct driver emotion recognition research using datasets specifically collected in driving scenarios. Li et al. [22] publicly released the first and currently the only multimodal dataset for driving tasks, named PPB-Emo, in 2022. Additionally, Oh et al. [23] proposed an onboard system for collecting multimodal data in real driving environments. In recent years, research on driver emotion recognition has been trending towards the direction of multimodality. Li et al. [24] proposed the multimodal model CogEmoNet, which takes the driver's facial video and cognitive information (including age, gender, and driving age) as inputs. Mou et al. [25] proposed a multimodal model using eye movements, driving behavior, and environmental information as inputs, adopting multitask learning based on a convolutional long short-term memory network and a hybrid attentional mechanism. Their research demonstrated the effectiveness of driving behavior as an auxiliary modality for driver emotion recognition. To the best of our knowledge, the fusion network combining facial video and driving behavior modalities has not been explored yet. Therefore, this paper aims to investigate the fusion of these two modalities in our research.

III. Model Description

A. Overview

MDEmoNet is a multimodal driver emotion recognition model with inputs modal to facial video and driving behavior data (including brake pedal force, vehicle Y-axis position, and Z-axis position), and outputs discrete emotion classification. As shown in Fig. 1, the model contains a facial video modal processing module, a driving behavior modal processing module, and a decision module. The network employs multi-task learning to fuse the two modalities more effectively.

B. Facial Video Modal Processing Module

We first segment the video samples into 8 segments with 2 randomly selected frames in each segment. In this way, each video sample is transformed into a sequence of 16 frames of 112×112 RGB facial images, denoted as $X \in \mathbb{R}^{16\times3\times112\times112}$. Next, this facial image sequence will be input into the convolution module, and for each frame, the features of the facial image are first extracted through four convolutional layers to obtain the feature map $M \in \mathbb{R}^{C\times H'\times W'}$, where C denotes the number of channels of the feature map, and H' and W' denote the height and width of the feature map respectively. Flatten the feature map into a one-dimensional sequence $M^f \in \mathbb{R}^{Q\times C}$ where $Q=H'W'$. We then encode spatial positions by adding a visual word embedding m^f_p of length C to a learnable position embedding e_p. The calculation process is given as follows:

$$z_p = m^f_p + e_p \quad p \in \{1,2,\mathrm{L}\ ,Q\} \tag{1}$$

The next part is the Spatial Transformer, which is used to extract spatial features from facial images. It consists of a spatial encoder that includes multi-head self-attention and feed-forward networks. Self-attention calculations are performed by computing query(q), key(k), and value(v) vectors. The calculation process is given as follows:

$$q^k_p = W^k_Q LN(z_p) \tag{2}$$
$$k^k_p = W^k_K LN(z_p) \tag{3}$$
$$v^k_p = W^k_V LN(z_p) \tag{4}$$

where $LN(\cdot)$ represents the layer normalization operation, W represents the weight matrix of the k-th multi-head attention head, where $k \in \{1, \cdots, K\}$, K=8 is the total number of attention heads. The self-attention weight λ^k_p of each query p can be obtained through the dot product. The calculation process is given as follows:

$$\lambda^k_p = \mathrm{softmax}\left(\frac{q^{kT}_p}{\sqrt{C'}} \cdot \left\{k^k_{p'}\right\}_{p'=1,\mathrm{L}\ ,Q}\right) \tag{5}$$

C' is the potential dimension of each attention head. To compute the encoding z within the block, first, the weighted sum of value vectors is calculated using the self-attention coefficients of the self-attention heads. The calculation formula is as follows (6).

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICCE59016.2024.10444365, 2024 IEEE International Conference on Consumer Electronics (ICCE).

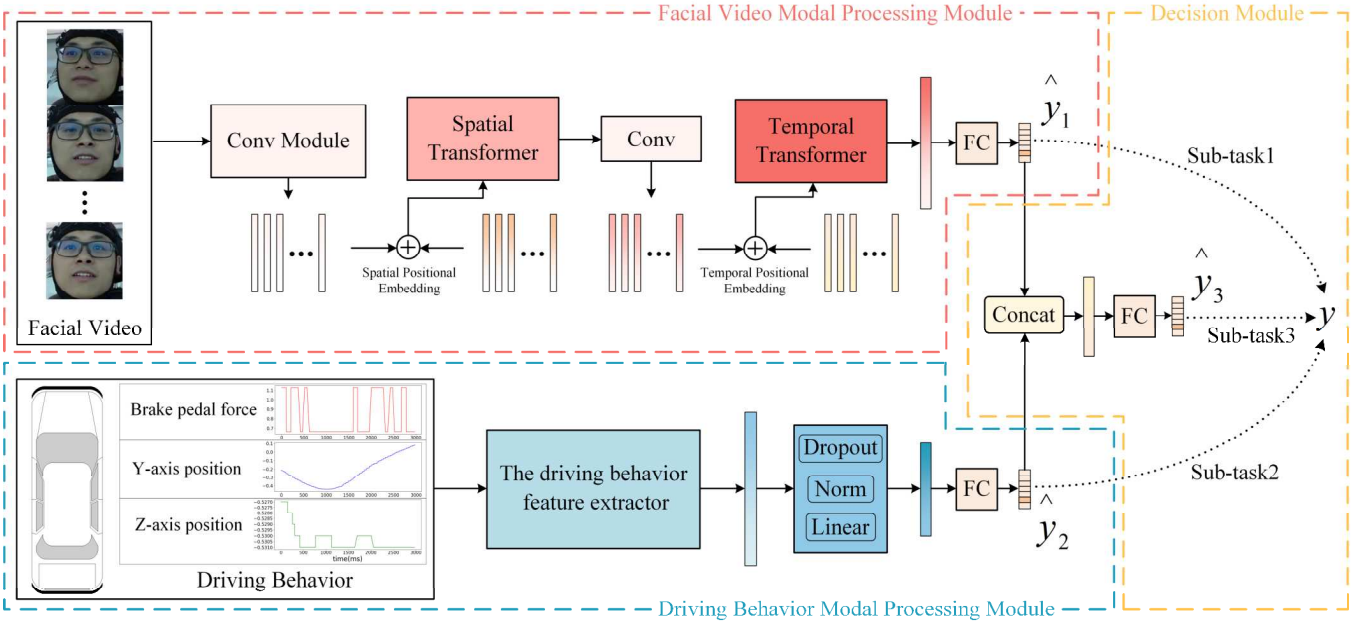

Fig. 1. The structure of the proposed method.

$$s_p^k = \sum_{p'=1}^{Q} \lambda_{p,p'}^k v_{p'}^k \tag{6}$$

Then, all the outputs from the self-attention heads are concatenated and passed through an *MLP* projection with a residual connection. The calculation can be represented as:

$$z'_p = W \begin{bmatrix} s_p^1 \\ \mathrm{M} \\ s_p^k \end{bmatrix} + z_p \tag{7}$$

$$z_p = MLP(LN(z'_p)) + z'_p \tag{8}$$

Finally, the Q encoded z^p are concatenated together to generate the refined feature map $Mr \in \mathbb{R}^{C \times H' \times W'}$. Then, it goes through a convolutional layer, and the feature embedding $x'_t \in \mathbb{R}^F$ for each frame can be calculated as follows:

$$x'_t = GAP(g(Mr)) \quad t \in \{1,2,\mathrm{L}\ 16\} \tag{9}$$

where $g(\cdot)$ represents the convolution block, and $GAP(\cdot)$ stands for global average pooling. All frames share the same Spatial Transformer module, given input $X \in \mathbb{R}^{16\times3\times112\times112}$, which yields output $X' \in \mathbb{R}^{16\times F}$, where F represents the dimensionality of the features. Next, we input the feature sequence of the facial images into the Temporal Transformer. It consists of 3 temporal encoders, each comprising a multi-head self-attention and a feed-forward network. Each temporal encoder is augmented with temporal encoding embeddings. For example, the formula for the temporal encoding embedding in the first temporal encoder is as follows:

$$z_{t'}^1 = x'_{t'} + e_{t'} \quad t' \in \{0,1,\mathrm{L}\ ,16\} \tag{10}$$

where $e_{t'}$ represents the learned position embedding, used to encode the temporal position. Unlike Spatial Transformer, we add a special learnable vector x'_0 at the first position of the sequence to represent the embedding of the class token. Next, in each temporal encoder, we calculate the query(q), key(k), and value(v) vectors, following the same process as in Spatial Transformer. The input of the next temporal encoder is the output of the previous temporal encoder. Then, we use the output $z^3{}_0$ of the class token from the final layer of the Temporal Transformer to represent the video's features. Finally, we perform classification using a fully connected network to obtain the emotion recognition results for the video modality. The calculation process is given as follows:

$$\hat{y}_1 = FC(z_0^3) \in \mathrm{i}^{\ 7} \tag{11}$$

where $FC(\cdot)$ represents a fully connected network, and 7 indicates the number of emotion categories.

C. Driving Behavior Modal Processing Module

The PPB-Emo dataset records driving behavior modal data, including acceleration, lateral acceleration, gas pedal position, brake pedal force, gear, steering wheel position, velocity, lateral velocity, x position, y position, and z position. We initially extracted a total of 9396 features for each sample from the driving behavior modal data, such as time reversal asymmetry statistic, fft coefficient, fft aggregated, etc. To identify features that have significant impacts on different emotion categories, we performed a two-sided Mann-Whitney U test [26] to determine whether each feature exhibits significant differences between different emotion categories. We calculated the p-value corresponding to the U statistic, which represents the

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICCE59016.2024.10444365, 2024 IEEE International Conference on Consumer Electronics (ICCE).

probability of observing the U statistic or more extreme values under the null hypothesis. If the p-value is small, it indicates that the feature has a significant impact on the binary emotion target. Conversely, if the p-value is large, it means that the feature does not have a significant impact. We applied the Benjamini-Hochberg method to control the false discovery rate and finally identified which features are significant. After conducting the significance test for all features, we selected a total of 95 relevant features related to Brake pedal force, Y-axis position, and Z-axis position. Due to space limitations, we present key relevant features in Table I. We designed a driving behavior feature extractor to extract these features. The raw features extracted by the driving behavior feature extractor are further processed through a linear layer, a normalization layer, and a dropout layer to extract more meaningful features. Subsequently, a fully connected network is employed for classification, producing the emotional recognition results for the driving behavior modality. The calculation formula is as follows:

$$Fdb' = Dropout(LN(Linear(Fdb))) \tag{12}$$

$$\hat{y}_2 = FC(Fdb') \in ¡^{7} \tag{13}$$

where F_{db} represents the raw features extracted by the driving behavior feature extractor. *Linear* denotes the linear layer, and *Dropout*(·) represents the dropout operation.

TABLE I. KEY DRIVING BEHAVIORAL FEATURES

Feature	Description
cwt coefficients	A continuous wavelet transform for the Ricker wavelet
fft coefficient	The fourier coefficients of the one-dimensional discrete Fourier Transform
c3	Measure non linearity in the time series
large standard deviation	Whether the time series has a large standard deviation?
variation coefficient	The variation coefficient
quantile	The q quantile of the time series
kurtosis	The kurtosis of the time series
number cwt peaks	Number of different peaks in the time series
median	The median of the time series
range count	Count observed values within the specific interval
agg linear trend	A linear least-squares regression for values of the time series that were aggregated over chunks versus the sequence from 0 up to the number of chunks minus one

D. Decision Module

The decision layers of the video modal processing part and the driving behavior modal processing part are concatenated, and then the classification is performed through a fully connected neural network. The calculation formulas are as follows:

$$F = cat(\hat{y}_1, \hat{y}_2) \in ¡^{14} \tag{14}$$

$$\hat{y}_3 = FC(F) \tag{15}$$

where the $cat(\cdot)$ represents the concatenation operation and $\hat{y}_3$ represents the final classification result.

E. Multi-task Learning

During the training process, we divided the overall task into three sub-tasks. Sub-task 1 solely used facial video modality features for classification, sub-task 2 solely used driving behavior modality features for classification, while sub-task 3 integrated information from both facial video modality and driving behavior modality for classification. During network validation, we only retained sub-task 3. The designed loss functions are as follows:

$$Loss = -\frac{1}{N}\sum_{m}\left[\sum_{n=1}^{N}\log\left(\frac{\exp\left(\hat{y}_{m_{n,target_n}}\right)}{\sum_{j}\exp\left(\hat{y}_{m_{n,j}}\right)}\right)\right] \tag{16}$$

where N represents the number of samples and $m \in \{1, 2, 3\}$ represents the different sub-tasks.

IV. EXPERIMENT

A. Dataset

The PPB-Emo dataset [22] is currently the only publicly available multimodal driver emotion dataset. The dataset collected psychological, physiological, and behavioral data (including driving behavior and facial videos) from 40 participants in 240 driving tasks.

B. Implementation Details

1) Data Pre-processing: For the facial video modality, we performed face alignment on each frame image to ensure that the face is positioned at a standard location and resized to 112×112 pixels. In the original PPB-Emo dataset, there are a total of 240 samples with a duration of 30 seconds. In order to obtain more training data, we cut each original sample into 10 subsamples of 3 seconds, resulting in 2400 samples.

2) Training Setting: We implemented this model using the platform PyTorch and optimized the parameters using the SGD optimizer with an initial learning rate of 0.01, which was divided by 10 every 40 epochs. The batch size was set to 64, and training was stopped after 130 epochs. We randomly split the dataset into training and validation sets in an 8:2 ratio. The final experimental results were obtained by averaging the scores from five experiments.

3) Validation Metrics: Driver emotion recognition is a multi-classification task, and the primary evaluation metric is accuracy [27, 28]. To ensure a fair comparison, we also include the Macro F1 score [29] as a supplementary evaluation metric. The complexity of the model determines the computing speed of the system. Considering the storage resources of the in-vehicle system are limited, we take the size and complexity of the model as additional evaluation metrics.

C. The Impact of Driving Behavior Feature Selection on Model Performance

We input all the features and the filtered features into the network separately, and modify the input layer dimension of the driving behavior modal processing module accordingly. The

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICCE59016.2024.10444365, 2024 IEEE International Conference on Consumer Electronics (ICCE).

experimental results are shown in Table II. The accuracy and F1 score of the model with filtered features input are significantly improved compared to the model with all features input, increasing by 6.04% and 0.071 respectively. This indicates that the filtered features can better capture the emotion-related information in driving behavior and better distinguish different emotion categories. In contrast, the full set of features may contain some redundant information, which reduces the performance of the model. The experimental results show that feature selection can significantly improve the model performance.

TABLE II. RECOGNITION RESULTS OF ALL FEATURES AND FILTERED FEATURES INPUT INTO THE NETWORK

Input Driving Behavior Feature	Acc(%)	Macro F1 score
All the features	61.88	0.6070
Filtered features	67.92	0.6780

D. *The Impact of Different Driving Behavior Modality Feature Extraction Methods on Model Performance*

Since the BiLSTM neural network is widely used for handling time series data, we compared this RNN model with the proposed driving behavior feature extractor for extracting driving behavior modality features in terms of accuracy, F1 score, model size, and computational speed. The experimental results (shown in Table III) demonstrate that the model using the driving behavior feature extractor performs the best in accuracy and F1 score, achieving 67.92% and 0.6780, respectively. The model using 2-layer BiLSTM for feature extraction has slightly lower accuracy compared to the driving behavior feature extractor, with a larger model size and slower computational speed. The models using 1-layer or 3-layer BiLSTM for feature extraction have lower accuracy. Therefore, the results suggest that our designed driving behavior feature extractor is a superior choice, providing higher accuracy, faster computational speed, and relatively smaller model size.

TABLE III. RECOGNITION RESULTS OF DIFFERENT DRIVING BEHAVIOR MODAL FEATURE EXTRACTION METHODS

Driving behavior modal feature extraction method	Acc (%)	Macro F1 score	Model Size (MB)	Complexity (GFLOPs)
1-layer BiLSTM	60.62	0.5980	18.57	9.32
2-layer BiLSTM	66.67	0.6613	20.04	12.17
3-layer BiLSTM	58.96	0.5857	21.72	15.01
the driving behavior feature extractor	67.92	0.6780	18.02	8.32

E. *Ablation Analysis*

To validate the effectiveness of each component in MDEmoNet, the necessary ablation analysis is conducted in this part. We separately studied the effectiveness of multimodality, multi-task learning, and decision layer fusion methods.

1) Evaluation of multimodality: We compared the classification results between single-modal and multimodal networks. We separately retained the network with only the driving behavior modality and the network with only the facial video modality. The experimental results, as shown in Table IV, indicate that the single-modal network using driving behavior has lower accuracy and F1 score, at 21.88% and 0.1756. The single-modal network using facial video shows a significant improvement in accuracy and F1 score compared to the driving behavior modality, reaching 58.96% and 0.5854. The multimodal network demonstrates even greater improvement in accuracy and F1 score achieving 67.92% and 0.6780. The fusion of multiple modalities effectively utilizes information from both driving behavior and facial video modalities, leading to a significant enhancement in model performance. Overall, the experimental results validate the effectiveness of multimodal fusion.

2) Evaluation of multi-task learning: We removed subtask 1 and subtask 2, keeping only task 3, and compared it with MDEmoNet, using multi-task learning. The experimental results, as shown in Table IV, indicate that the model using single-task learning achieved an accuracy of 35.63% and an F1 score of 0.3314. On the other hand, MDEmoNet with multi-task learning achieved a higher accuracy of 67.92% and an F1 score of 0.6780. The multi-task learning strategy significantly improved the model's classification performance. Therefore, multi-task learning is effective in this model.

3) Evaluation of decision layer fusion: We modified the network to perform feature-level fusion and compared it with MDEmoNet using decision-level fusion. The results, as shown in Table IV, indicate that the model with feature-level fusion achieved an accuracy of 63.83% and an F1 score of 0.6348, while the model with decision-level fusion achieved a higher accuracy of 67.92% and an F1 score of 0.6780. The use of decision-level fusion improved the performance of the model compared to the feature-level fusion model. Therefore, using decision-level fusion in this model is reasonable.

TABLE IV. EVALUATION OF COMPONENTS IN MDEMONET

Model	Acc(%)	Macro F1 score
Driving behavior unimodal	21.88	0.1756
Facial video unimodal	58.96	0.5854
Using single-task learning	35.63	0.3314
Using feature layer fusion	63.83	0.6348
MDEmoNet	67.92	0.6780

F. *Comparison with State-of-the-Arts*

We compared our MDEmoNet model with several state-of-the-art models on the PPB-Emo dataset from recent years. These advanced methods include the Emotion-FAN model [20] (2019) for facial video emotion recognition, the Former-DFER model [21] (2021) for the in-the-wild scenario facial emotion recognition, and the CogEmoNet model [24] (2022) for driver emotion recognition using both facial video and cognitive features (age, gender, and driving experience) as inputs. The results in Table V show that our MDEmoNet achieved excellent performance on the PPB-Emo dataset. It outperformed other models in most sentiment categories, overall accuracy and F1 scores. Additionally, our model has relatively smaller model size and computational complexity. By effectively leveraging both video and driving behavior modalities, MDEmoNet captures the driver's emotional features better. These results demonstrate the superiority of our approach in driver emotion recognition and highlight the advantages of using both facial video and driving behavior modalities.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICCE59016.2024.10444365, 2024 IEEE International Conference on Consumer Electronics (ICCE).

TABLE V. COMPARISON WITH STATE-OF-THE-ART METHODS ON PPB-EMO. BOLD DENOTES THE BEST. V DENOTES VIDEO MODAL. COG DENOTES COGNITIVE MODAL. DB DENOTES DRIVING BEHAVIOR MODAL. H DENOTES HAPPINESS. SAD DENOTES SADNESS. N DENOTES NEURAL. A DENOTES ANGER. S DENOTES SURPRISE. D DENOTES DISGUEST. F DENOTES FEAR.

Model	Modality	Accuracy of Each Emotion (%)							Acc (%)	Macro F1	Model Size (MB)	Complexity (GFLOPs)
		H	*Sad*	*N*	*A*	*S*	*D*	*F*				
Emotion-FAN [20] (2019)	V	49	30	69	31	33	36	22	40.21	0.3833	11.18	1.46
Former-DFER [21] (2021)	V	68	51	84	49	53	**58**	46	58.96	0.5854	19.01	8.32
CogEmoNet [24] (2022)	V-Cog	64	51	79	66	63	55	54	62.29	0.6228	21.36	15.51
MDEmoNet (Ours)	V-DB	**75**	**58**	**86**	**69**	**69**	44	**64**	**67.92**	**0.6780**	18.02	8.32

V. CONCLUSION

In this paper, we propose a multi-modal driver emotion recognition network for automotive intelligent cockpits, which combines facial video and driving behavior data to classify driver emotions. By employing multi-task training and decision-level fusion methods, the accuracy of driver emotion recognition is improved. The effectiveness of the entire scheme is validated through comparisons with advanced methods on the publicly available multi-modal driver dataset PPB-Emo. Experimental results demonstrate superior performance in accuracy, F1 score, model size, and computational complexity compared to state-of-the-art methods. In conclusion, the proposed multi-modal driver emotion recognition network for automotive intelligent cockpits holds significant practical significance. In intelligent cockpits, this model can enhance the driving experience, improve driving safety, and contribute to more innovations and possibilities in the development of intelligent vehicles.

REFERENCES

[1] X. Ji, et al., "EMSN: An Energy-Efficient Memristive Sequencer Network for Human Emotion Classification in Mental Health Monitoring," IEEE Trans. Consum. Electron., pp. 1–1, 2023.

[2] M. Wang, Y. He, et al., "A Review of AC and DC Electric Springs,". IEEE Access, vol. 9, pp. 14398-14408, Jan. 2021.

[3] Z. Dong, et al., "Memristor-Based Hierarchical Attention Network for Multimodal Affective Computing in Mental Health Monitoring," IEEE Consum. Electron. Mag., vol. 12, pp. 94–106, Jul. 2023.

[4] J. Wang, X. Zhang, et al. "MDGN: Circuit Design of Memristor-based Denoising Autoencoder and Gated Recurrent Unit Network for Lithium-ion Battery State of Charge Estimation," IET Renew. Power Gener., vol. 2023, pp. 1-12, Jul. 2023.

[5] Z. Dong, et al., "Design and Implementation of a Flexible Neuromorphic Computing System for Affective Communication via Memristive Circuits," IEEE Commun. Mag., vol. 61, pp. 74–80, Jan. 2023.

[6] Z. Dong, et al, "Neuromorphic Extreme Learning Machines with Bimodal Memristive Synapses," Neurocomputing, vol. 453, pp. 38-49, Apr. 2021

[7] X. Ji, et al., "A Flexible Memristor Model With Electronic Resistive Switching Memory Behavior and Its Application in Spiking Neural Network," IEEE Trans. NanoBioscience, vol. 22, pp. 52–62, Jan. 2023.

[8] Z. Dong, X. Ji, J. Wang, Y. Gu, J. Wang, and D. Qi, "ICNCS: Internal Cascaded Neuromorphic Computing System for Fast Electric Vehicle State of Charge Estimation," IEEE Trans. Consum. Electron., pp. 1–1, 2023.

[9] M. Jeon, "Don't Cry while you're Driving: Sad Driving is as Bad as Angry Driving," Int. J. Hum.-Comput. Interact., vol. 32, pp. 777-790, 2016.

[10] G. Underwood, P. Chapman, S. Wright, and D. Crundall, "Anger while driving," Transp. Res. Part F: Traff. Psychol. Behav., vol. 2, pp. 55-68, 1999.

[11] W. Li et al., "A Spontaneous Driver Emotion Facial Expression (DEFE) Dataset for Intelligent Vehicles: Emotions Triggered by Video-Audio Clips in Driving Scenarios," IEEE Trans. Affect. Comput., vol. 14, pp. 747–760, Jan. 2023.

[12] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework," IEEE Access, vol. 6, pp. 49325–49338, 2018.

[13] Z. Dong, X. Ji, G. Zhou, M. Gao, and D. Qi, "Multimodal Neuromorphic Sensory-Processing System With Memristor Circuits for Smart Home Applications," IEEE Trans. Ind. Appl., vol. 59, pp. 47–58, Jan. 2023.

[14] P. Ekman, "An Argument for Basic Emotions," Cognition and Emotion, vol. 6, pp. 169–200, 1992.

[15] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," J. Pers. Soc. Psychol., vol. 17, pp. 124–129, 1971.

[16] D. Matsumoto, "More Evidence for the Universality of a Contempt Expression," Motivation and Emotion, vol. 16, pp. 363–368, Dec. 1992.

[17] P. J. Lang, "The Emotion Probe: Studies of Motivation and Attention," American Psychologist, vol. 50 , pp. 372–385, 1995.

[18] A. Mehrabian and S.R. Ferris, "Inference of attitudes from nonverbal communication in two channels," J. Consult. Psychol., vol. 31, pp. 248-252, 1967

[19] X. Ji, Z. Dong, Y. Han, C. S. Lai, and D. Qi, "A Brain-inspired Hierarchical Interactive In-memory Computing System and its Application in Video Sentiment Analysis," IEEE Trans. Circuits Syst. Video Technol., pp. 1–1, 2023.

[20] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame Attention Networks for Facial Expression Recognition in Videos," ICIP 2019, pp. 3866–3870, Sep. 2019.

[21] Z. Zhao and Q. Liu, "Former-DFER: Dynamic Facial Expression Recognition Transformer," MM '21, pp. 1553–1561, Oct. 2021.

[22] W. Li et al., "A Multimodal Psychological, Physiological and Behavioural Dataset for Human Emotions in Driving Tasks," Scientific Data, vol. 9, p. 481, Aug. 2022.

[23] G. Oh et al., "Multimodal Data Collection System for Driver Emotion Recognition Based on Self-Reporting in Real-World Driving," Sensors, vol. 22, 2022.

[24] W. Li et al., "CogEmoNet: A Cognitive-Feature-Augmented Driver Emotion Recognition Model for Smart Cockpit," IEEE Trans. Comput. Social Syst., vol. 9, pp. 667–678, Jun. 2022.

[25] L. Mou et al., "Driver Emotion Recognition with a Hybrid Attentional Multimodal Fusion Framework," IEEE Trans. Affect. Comput., pp. 1–12, 2023.

[26] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," The Ann. Math. Stat., vol. 18, pp. 50–60, 1947.

[27] X. Ji , et al., "A Physics-Oriented Memristor Model with the Coexistence of NDR Effect and RS Memory Behavior for Bio-Inspired Computing," Mater. Today Adv., vol. 16, pp. 1-14, Sep. 2022.

[28] L. Cai, H. Wang et al., "A Multi-Fault Diagnostic Method Based on Category-Reinforced Domain Adaptation Network for Series-Connected Battery Packs," J. Energy Storage, vol. 60, pp. 1-11, Jan. 2023.

[29] J. Wang, Y. Chen et al., "Improved YOLOv5 Network for Real-Time Multi-Scale Traffic Sign Detection," Neural Comput. Appl., vol. 35, pp. 7853-7865, Dec. 2023.