This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/JIOT.2025.3527206, IEEE Internet of Things Journal

A Dual-Pathway Driver Emotion Classification Network Using Multi-Task Learning Strategy: A Joint Verification

Zhekang Dong, *Senior Member*, *IEEE*, Chenhao Hu, Liyan Zhu, Xiaoyue Ji, *Member*, *IEEE*, Chun Sing Lai, *Senior Member*, *IEEE*

***Abstract*—Negative emotion (e.g., anger, fear) may influence normal driver behavior, resulting in serious traffic accidents. Thus, developing an automatic driver emotion classification method is necessary and urgent. Most of the existing methods are performed in realistic indoor environment and always lack effective utilization of heterogeneous information, resulting in low accuracy and reliability. In this paper, a novel dual-pathway driver emotion classification network using multi-task learning strategy is proposed. To illustrate the design of the proposed driver emotion classification network, three modules are constructed: 1) visual-facial data processing module; 2) driving behavioral data processing module; 3) fusion output module. Meanwhile, considering the influence of emotional states on driving behavior, a comprehensive analysis is conducted to distinguish the positive, neutral, and negative influence on driving behavior. Furthermore, a joint verification in both realistic indoor environment (i.e., laboratory simulation on the PPB-Emo dataset) and real-world outdoor scenario is performed. The experimental results illustrate that the proposed network exhibits superior performance in terms of classification accuracy and response time, achieving good balance between classification accuracy and running speed in internet of things scenarios.**

***Index Terms*—Dual-pathway, driver emotion classification, driving behavior, joint verification**

I. Introduction

In the process of vehicle driving, a driver's emotion is affected by different factors (e.g., the traffic conditions, psychophysiological states), which may result in hazardous driving behaviours and potentially severe traffic accidents, particularly during significant emotional fluctuations [1-3]. Based on this, timely and accurate recognition of emotional state using internet of thing (IoT) embedded technologies is beneficial to implement healthcare and safety interventions, as well as for the development of a friendly and well-organized driving environment in smart city.

In general, driver emotion classification technology mainly relies on analysing available invasive physiological signals and non-invasive multi-modal signals through different machine/deep learning methods [4, 5]. Accordingly, driver emotion classification methods can be divided into two categories, i.e., the invasive methods [4-8] and non-invasive methods [9-17]. For the invasive methods, the physiological signals, including electroencephalography (EEG) signal, electrocardiography (ECG) signal, and electromyography (EMG) signal can be directly measured to achieve accurate classification performance, while having negative impact on driving behavior, especially when an emergency occurs.

Different with the invasive methods, non-invasive methods implement the driver emotion classification using non-invasive multi-modal signals (e.g., facial expression and voice signal), which has almost no effect on driving performance, thereby developing a secure and comfortable driving environment [9-17]. For example, a deep learning approach is proposed to monitor various drivers' expressions in different pose variations, illuminations, and occlusions using facial images [9]. A driver emotion classification network based on voice modality was proposed, which achieves accurate classification results by fusing the global acoustic features and local spectrogram features [10]. In [11], a driver emotion classification network using facial expressions and cognitive process features (age, gender, driving experience) was proposed. [12] proposed a hybrid network (i.e., Emotion-FAN) using frame attention and deep convolutional neural network (CNN) for emotion classification. Then, a dynamic driver emotion classification network (i.e., Former-DFER) was introduced to tackle the challenges posed by occlusion and non-frontal poses during driving [13]. After that, [14] proposed a clip-ware emotion-rich feature learning network (i.e., CEFLNET) for robust driver emotion classification. A facial expression-based driver emotion classification network with intensity-aware loss (IAL) function was designed [15]. In [16], a self-supervised autoencoder (i.e., MARLIN) that can learn universal facial representations from non-annotated web-crawled videos was

This work was supported by the Postdoctoral Fellowship Program of CPSF under Grant GZB20230356, China Postdoctoral Science Foundation under Grants 2024T170463 and 2024M751676, National Natural Science Foundation of China under Grant 62206062, Ministry of Science and Technology - Yangtze River Delta Science and Technology Innovation Program under Grant YDZX20233100004028, and Zhejiang Xinmiao Talents Program under Grant 2024R407C065. *(Corresponding author: Xiaoyue Ji).*

Z. Dong, C. Hu, and L. Zhu are with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China, and also with the Zhejiang Provincial Key Lab of Equipment Electronics, Hangzhou 310018, China (e-mail: englishp@hdu.edu.com; chenhaohu@hdu.edu.cn; 232040154@hdu.edu.cn).

X. Ji is with the Center for Brain-Inspired Computing Research (CBICR), Beijing Innovation Center for Future Chip, Optical Memory National Engineering Research Center, Department of Precision Instrument, Tsinghua University, Beijing 100084, China. (e-mail: jixiaoyue@mail.tsinghua.edu.cn)

C. S. Lai is with Department of Electronic and Computer Engineering, Brunel University London, London, UB8 3PH, UK, (e-mail: chunsing.lai@brunel.ac.uk).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/JIOT.2025.3527206, IEEE Internet of Things Journal

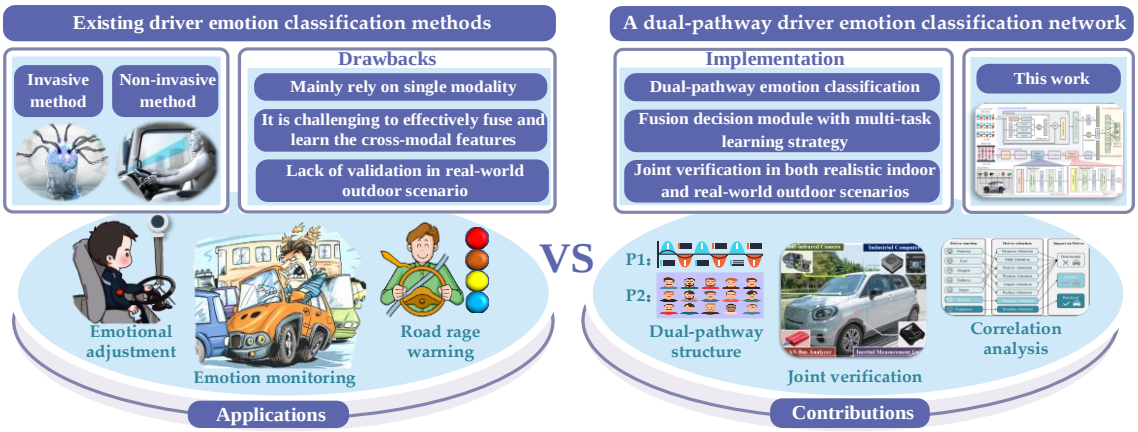

Fig. 1. Systemic comparison of the existing driver emotion classification methods versus the proposed method

proposed to perform the accurate driver emotion classification. [17] developed a multi-3D dynamic facial emotion classification network (i.e., M3DFEL), which can improve the driver emotion classification accuracy.

Non-invasive methods have been proved effective in driver emotion classification task, enabling a safer and more comfortable driving environment. However, these methods still suffer from three limitations [18-20]:

Limitation 1: Recent studies almost focus on single modality, e.g., visual-facial data (VFD), which may degrade the accuracy and reliability of driver emotion classification. The complementary modality, e.g., driving behavioral data (DBD), and coupling relationship between cross-modal information have not been fully considered.

Limitation 2: Since different modalities have their own characteristics (e.g., data structure, spatial-temporal complexity), it is challenging to effectively fuse and learn cross-modal features.

Limitation 3: The existing methods are almost conducted in the realistic indoor environment (e.g., laboratory simulation). It is hard to balance the trade-off between classification accuracy and running speed in the real-world outdoor scenario.

Based on this, a novel non-intrusive driver emotion classification network is proposed in this work, which aims to address the above-mentioned three limitations. For clarity, the systemic comparison of the existing driver emotion classification methods versus the proposed method is illustrated in **Fig. 1**. The main contributions can be summarized below:

1) A dual-pathway driver emotion classification network (DDECNet) is proposed, which can effectively capture high-level facial features from spatial-temporal perspectives and time-series features from different receptive fields, enabling efficient utilization of heterogeneous information to improve the classification performance.

2) A fusion output module with multi-task learning strategy is designed that can sufficiently integrate cross-modal features and provide a reliable analysis to distinguish the positive, neutral, and negative influence on driving behavior.

3) The proposed emotion classification network is performed in both the realistic indoor environment and the real-world outdoor scenarios. The experimental results demonstrate that the entire scheme has advantages in balancing the trade-off between running speed and classification performance.

The remaining of this work is organized as follows: Section II elaborates on the design of the entire driver emotion classification network from the perspectives of VFD processing module, DBD processing module, and fusion output module. In Section III, the dataset and the corresponding pre-processing method is provided detailed. In Section IV, a joint verification is conducted to demonstrate the correctness and effectiveness of the entire scheme. Section V provides the conclusion and future direction of the entire work.

II. Design of The Proposed DDECNet

A. Overall Network Architecture

In this work, we propose a novel driver emotion classification network (i.e., DDECNet) which can effectively capture and fuse heterogeneous information from in-vehicle environment, enabling automatic driver emotion classification. The specific architecture is illustrated in **Fig. 2**. To facilitate understanding of the DDECNet design, we describe it using three modules, i.e., the VFD processing module, the DBD processing module, and the fusion output module.

B. Visual-Facial Data Processing Module

In VFD processing module, each video sample is uniformly partitioned into 8 segments. We then randomly select 2 frames per segment, transforming each video sample into a 16-frame facial image sequence with dimensions 112×112, denoted as $X_{FV} \in \mathbb{R}^{16\times3\times112\times112}$. For each facial image frame, initial features are extracted using a two-dimensional (2D) convolutional layer and three residual convolutional blocks. The generated feature map is denoted as $M \in \mathbb{R}^{H\times W\times C}$, where H, W, and C are the height, width, and channel number of feature map, respectively. Then, the feature map is converted into Q one-dimensional (1D) vectors with the length of C (denoted as $M^f \in \mathbb{R}^{Q\times C}$, where $Q=H\cdot W$). Furthermore, the feature map is subsequently injected into the spatial transformer that consists of spatial positional embedding and S-layer spatial encoders. In spatial positional embedding, the spatial positions can be encoded by adding visual word embeddings m_p^f to a learnable position embedding e_p, which can be mathematically expressed by:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/JIOT.2025.3527206, IEEE Internet of Things Journal

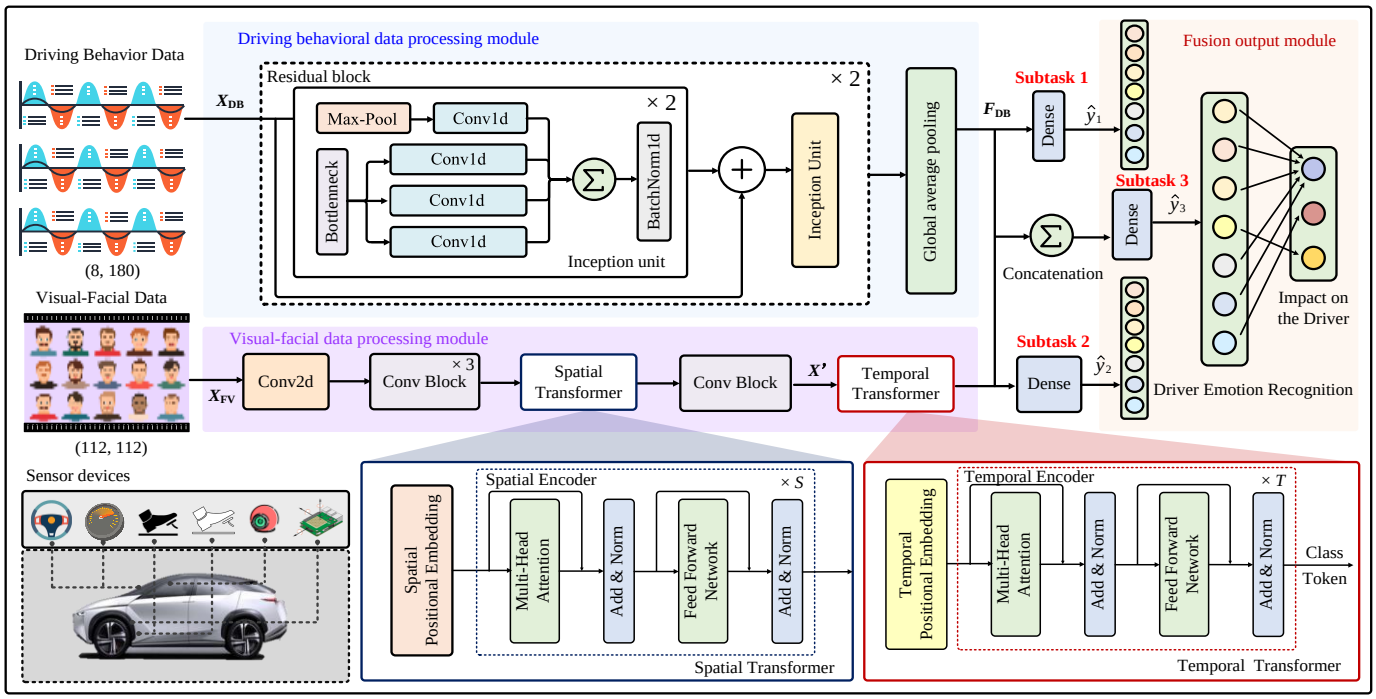

Fig. 2. Architecture of the proposed DDECNet.

$$z_p^0 = m_p^f + e_p \quad p \in \{1, 2, \cdots, Q\} \tag{1}$$

where encoded result z_p^0 is then input to the S-layer spatial encoders (mainly comprising multi-head self-attention and feedforward network). In the l-th layer of the spatial encoder, the self-attention computation can be achieved by:

$$q_p^{(l,k)} = W_Q^{(l,k)} LN(z_p^{l-1}) \tag{2}$$

$$k_p^{(l,k)} = W_K^{(l,k)} LN(z_p^{l-1}) \tag{3}$$

$$v_p^{(l,k)} = W_V^{(l,k)} LN(z_p^{l-1}) \tag{4}$$

where $q_p^{(l,k)}$, $k_p^{(l,k)}$, and $v_p^{(l,k)}$ denote the query, key, and value vectors. $LN(\cdot)$ represents the layer normalization. $W_Q^{(l,k)}$, $W_K^{(l,k)}$, and $W_V^{(l,k)}$ are all weight matrices for the k-th head in the l-th layer, where $k \in \{1, \cdots, K\}$, and K is the total number of attention heads. For the k-th attention head, the self-attention weight $\lambda_p^{(l,k)}$ can be calculated by:

$$\lambda_p^{(l,k)} = soft\max\left(\frac{q_p^{(l,k)\top}}{\sqrt{C'}} \cdot \left\{k_{p'}^{(l,k)}\right\}_{p'=1,\cdots,Q}\right) \tag{5}$$

where C' denotes the latent dimensionality of each attention head.

Then, the output of the l-layer spatial encoder z_p^l can be mathematically expressed by:

$$z_p^l = MLP(LN(\tilde{z}_p^l)) + \tilde{z}_p^l \tag{6}$$

$$\tilde{z}_p^l = W^l \begin{bmatrix} s_p^{(l,1)} \\ \vdots \\ s_p^{(l,k)} \end{bmatrix} + z_p^{l-1} \tag{7}$$

$$s_p^{(l,k)} = \sum_{p'=1}^{Q} \lambda_{p,p'}^{(l,k)} v_{p'}^{(l,k)} \tag{8}$$

where W, $MLP(\cdot)$, $\lambda_{p,p'}^{(l,k)}$ are the projection matrix, MLP mapping operation, and self-attention weight, respectively.

Furthermore, the Q encodings z_p^S are concatenated at the spatial level to generate the refined feature map $Mr \in \mathbb{R}^{H \times W \times C}$. The feature embedding $x'_t \in \mathbb{R}^F$ for each frame is computed by:

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

where $g(\cdot)$ and $GAP(\cdot)$ represent the convolution and global average pooling operations, respectively.

Next, the temporal transformer consisting of temporal positional embedding and T-layer temporal encoder is introduced. The input to the temporal encoder can be given by:

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

where $e_{t'}$ denotes the learned temporal positional embedding.

The output of the temporal encoder z_0^T represents the high-level facial features. The emotion classification results $\hat{y}_1$ can be written by:

$$\hat{y}_1 = FC(z_0^T) \tag{11}$$

where $FC(\cdot)$ denotes a fully connected network.

C. Driving Behavioral Data Processing Module

In the DBD processing module, eight driving behavior data, including steering wheel position, gas pedal position, brake pedal force, forward direction acceleration, lateral acceleration, forward direction velocity, lateral velocity, and vertical velocity are selected as inputs X_{DB}. The DBD processing module consists of two residual blocks and a global average pooling layer. Notably, a max-pooling operation cascaded with 1D convolution is introduced to alleviate the sensitivity to minor noise. An average pooling layer is incorporated to average the output features over the entire temporal dimension.

After a fully connected network, the time-series features F_{DB} from different receptive fields can be captured. The emotion classification results $\hat{y}_2$ can be expressed by:

$$\hat{y}_2 = FC(F_{DB}) \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/JIOT.2025.3527206, IEEE Internet of Things Journal

D. Fusion Output Module

In fusion output module, the features derived from dual-pathway processing modules (i.e., the VFD processing module and DBD processing module) are directly fused through a concatenation unit. The final emotion classification result $\hat{y}$ can be obtained by:

$$\hat{y} = FC(F) \tag{13}$$

$$F = cat(z_0^T, F_{DB}) \tag{14}$$

where F is the fused features and $cat(\cdot)$ denotes the concatenation operation.

E. Multi-task Learning Strategy and Loss Functions

During the training process, the driver emotion classification task can be decomposed into three classification subtasks. Specifically, the classification subtask 1 independently employs the high-level facial features with loss function $Loss1$; The classification subtask 2 independently utilizes the time-series features with loss function $Loss2$; The classification subtask 3 used the fused features with loss function $Loss3$.

$$Loss1 = CrossEntropyLoss(\hat{y}_1, y) \tag{15}$$

$$Loss2 = CrossEntropyLoss(\hat{y}_2, y) \tag{16}$$

$$Loss3 = CrossEntropyLoss(\hat{y}_3, y) \tag{17}$$

where y is the ground-truth label, $CrossEntropyLoss(\cdot)$ denotes the cross-entropy loss function.

Although each subtask focuses on different features, there are synergies between subtasks. Namely, these three subtasks are learned jointly. The overall loss function of DDECNet can be given by weighted summation of the individual subtask loss functions:

$$Loss = \alpha Loss1 + \beta Loss2 + \gamma Loss3 \tag{18}$$

where parameters α, β, and γ denote the weight assignments.

F. Influence of Driver Emotion on Driving Behavior

Different emotional states always have different influences on the driving behavior. Inspired by [21-24], seven basic emotions (i.e., surprise, fear, disgust, sadness, anger, neutral, and happiness) are mapped into three categories (negative, neutral, and positive) according to their different impacts on driving behavior. The specific mapping relationship between the driver emotion and driving behavior is illustrated in **Fig. 3**.

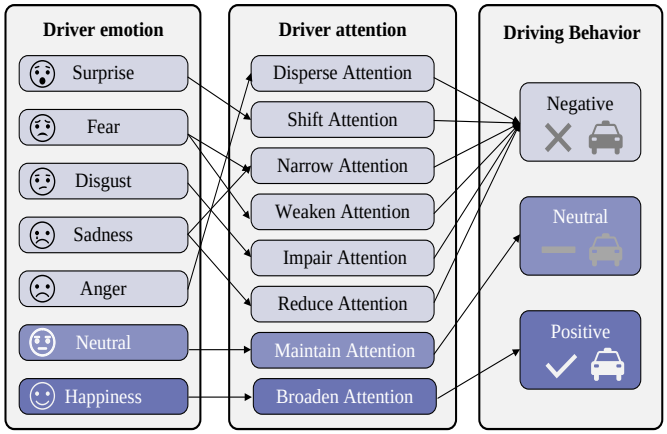

Fig. 3. The correspondence between the driver emotion and driving behavior

From **Fig. 3**, the feeling of anger is one of the negative emotions during driving, it may disperse the driver's attention and even lead to aggressive driving behavior, making the entire driving process challengeable and unpredictable. Similarly, the feeling of sadness may narrow or reduce attention, which may also trigger aggressive driving behavior. The feeling of fear may have a broad impact on driver attention, e.g., narrowing/weakening attentional focus, increasing attentional difficulty, and excessive focusing on the threatening stimuli. The feeling of disgust weakens or blocks the perception ability, which may impair driver's attention. The feeling of surprise may delay the initiation of discrete actions and interrupt continuous actions, which may shift driver's attention. The feeling of happiness is associated with an assimilative processing style, which may broaden attentional focus. The feeling of neutral is able to maintain attention, which has negligible effect on driving behavior.

III. Dataset and Pre-processing

A. Dataset Description

In this work, the PPB-Emo dataset [25], a widely used multimodal dataset, is employed to verify the effectiveness and feasibility of the proposed DDECNet in realistic indoor environment. Specifically, the PPB-Emo dataset records VFD and DBD from 40 participants in 240 valid driving tasks. Each sample has a certain driver emotion label, including seven categories, i.e., surprise, fear, disgust, sadness, anger, neutral, and happiness. Each driver emotion in the PPB-dataset is evenly distributed. During the training and testing phases, the 5-fold cross-validation [13] is employed to evaluate the proposed DDECNet.

To demonstrate the generalizability and transferability, the well-trained DDECNet is also evaluated on a real-world outdoor dataset collected by an electric vehicle (Leapmotor T03). Specifically, this dataset includes 50 participants (25 males and 25 females) with five different age groups (18 ~ 27 years old, 28 ~ 37 years old, 38 ~ 47 years old, 48 ~ 57 years old, more than 58 years old), five different driving experiences (less than 10,000km, 10,000 ~ 30,000km, 30,000 ~ 50,000km, 50,000 ~ 100,000km, more than 100,000km), and five different education backgrounds (primary school, junior middle school, senior middle school, university, and higher education). Meanwhile, more than 10 scenarios (urban road, highway, rural road, tunnel under different weather and light conditions) are included in this dataset. Similar with PPB-Emo dataset, seven emotion categories are also evenly-distributed, and each emotion category contains 50 samples. The necessary sensor devices and the corresponding data samples are provided in **Table I**. Specifically, the signal acquisition devices include the near-infrared camera, CAN bus analyzer, and inertial measurement unit. Notably, considering these devices are all mature, reliable, and commercially available, they can be easily installed and run in the vehicle. Different with the invasive sensors, these non- invasive devices have minimal impact on driving performance and help create a secure and comfortable driving environment.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/JIOT.2025.3527206, IEEE Internet of Things Journal

TABLE I
THE SENSOR DEVICES AND THE CORRESPONDING DATA

Sensor devices	Collected Data	Unit
Near-infrared camera (ZWAK 3306)	Visual-facial data	Frame
CAN bus analyzer (GCAN USBCAN-I Pro)	Steering wheel position	Rad
	Gas pedal position	Degree
	Brake pedal force	N
	Forward direction velocity	m/s
	Lateral velocity	m/s
	Vertical velocity	m/s
Inertial measurement unit (Wit-motion HWT906)	Forward acceleration	m/s^2
	Lateral acceleration	m/s^2

B. Dataset Pre-processing

The dataset pre-processing contains the linear interpolation & normalization for DBD and the face alignment for FVD. Notably, the face alignment is performed to ensure the face is in a standard position, as illustrated in **Fig. 4.** The specific process can be divided to three steps:

Step 1 (Landmarks detection): We first employ a facial landmark predictor to detect facial landmarks.

Step 2 (Transformation matrix calculation): we calculate the centre coordinates of both eyes using the coordinates of landmarks. The horizontal distance dX, the vertical distance dY, and the angle θ between the two eyes can be obtained. After calculating the rotation angle, scaling factor, and translation position, a transformation matrix is developed for translating, rotating, and scaling.

Step 3 (Face alignment): According to the obtained transformation matrix, the aligned facial image can be achieved by translating, rotating, and scaling operations.

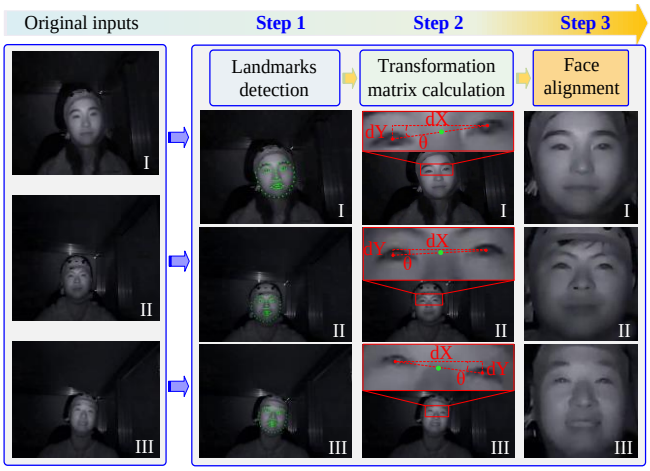

Fig. 4. The preprocessing procedure for near-infrared facial images

IV. EXPERIMENT AND ANALYSIS

A joint verification in both realistic indoor environment (i.e., laboratory simulation on the PPB-Emo dataset) and real-world outdoor scenario is carried out. The specific experimental description is provided below:

A. Experimental Setup and Evaluation Metrics

The proposed DDECNet is trained on an open-source PyTorch platform (dual NVIDIA GeForce RTX 4090 GPUs). During the training phase, the total training epochs are set as 300. The dynamic learning rate (with the initial value of 0.01) is divided by 10 every 100 epochs, and the batch size is set to 128. The stochastic gradient descent (SGD) optimizer (momentum and weight decay are set as 0.9 and 0.0001, respectively) is utilized for parameter optimization.

Notably, the random cropping and horizontal flipping operations are performed on VFD to improve the robustness of the model. Meanwhile, the Gaussian random noise is introduced to DBD to better simulate the real signal acquisition process in real-world scenario.

To evaluate the performance of the proposed DDECNet, a series of common evaluation metrics [26, 27] including seven-class classification accuracy (Acc-7), Macro F1 score (F1-7), three-class classification accuracy (Acc-3), weighted F1 score (F1-3), average accuracy (Acc), F1 score (F1), and computational complexity are introduced in this work.

B. Model Hyperparameter Selection

The success of most deep learning models in classification tasks is largely attributed to the depth of their architectures [28]. Based on this, the influence of depth of spatial transformer, temporal transformer, and inception unit on the performance of DDECNet are explored. In our model, the initial depths of spatial encoder, temporal encoder, and inception unit are set as 1, 1, 6 respectively. Then, a series of experiments comparing different depth combinations (S, T, I) on PPB-Emo dataset are conducted. The specific experimental and visualization results are collected in **Table II** and **Fig. 5**, respectively.

TABLE II
THE INFLUENCE OF DEPTH COMBINATION ON MODEL PERFORMANCE

Setting			Metrics				
S	T	I	Acc-7 (%)	F1-7 (%)	Acc-3 (%)	F1-3 (%)	Complexity (GFLOPs)
1	1	6	77.87	77.90	90.37	90.07	8.33
3	1	6	79.00	79.18	90.72	90.59	9.15
1	**3**	**6**	**81.63**	**82.49**	**94.00**	**93.95**	**8.40**
3	3	6	80.37	80.56	91.89	91.77	9.23
6	3	6	78.16	78.21	90.62	90.38	10.46
3	6	6	80.67	80.70	91.80	91.75	9.33
6	6	6	80.29	80.34	91.83	91.71	10.57
1	3	3	77.58	77.53	90.90	90.78	10.53
1	3	9	80.21	80.43	91.27	91.18	10.61

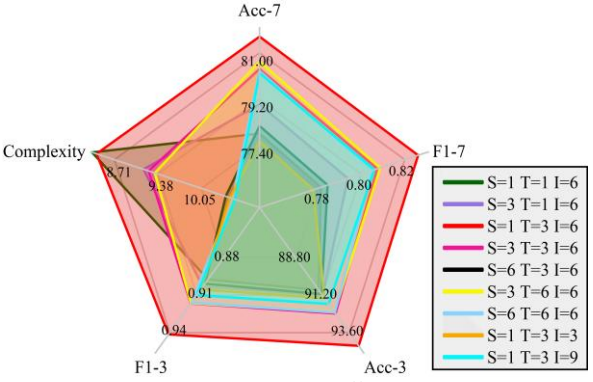

Fig. 5. The visualization results with different depth combinations

From **Table II** and **Fig. 5**, when the neural network depth combination (S, T, I) is set as (1, 3, 6), the proposed DDECNet achieves the best classification performance.

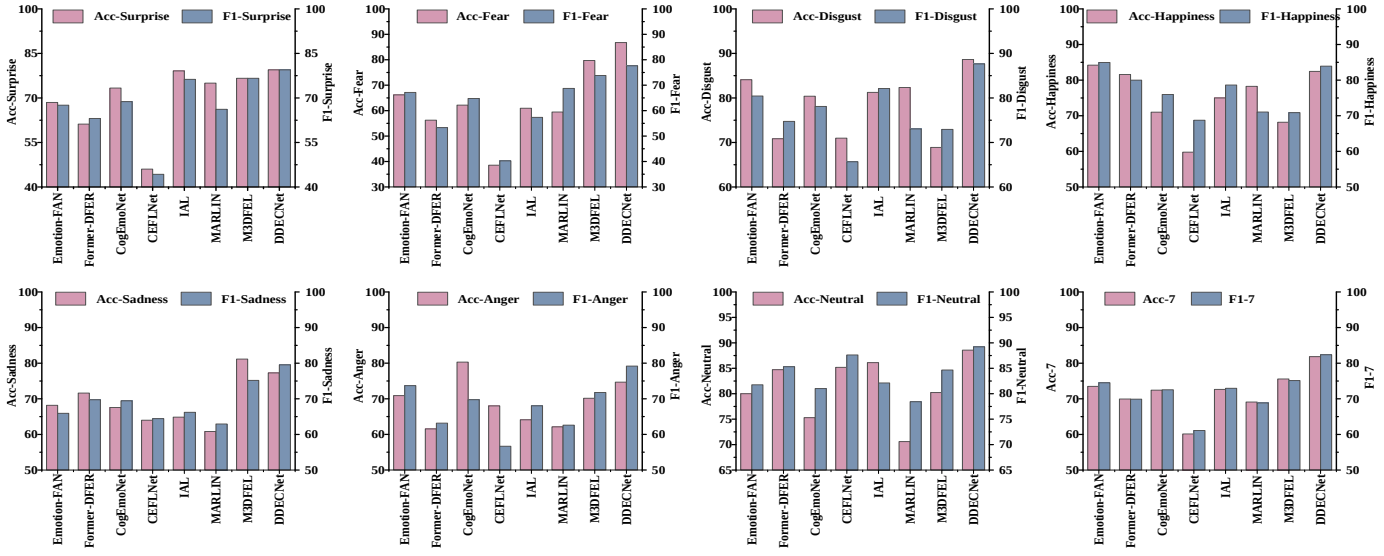

Fig. 7. The comparative results for driver emotion classification task

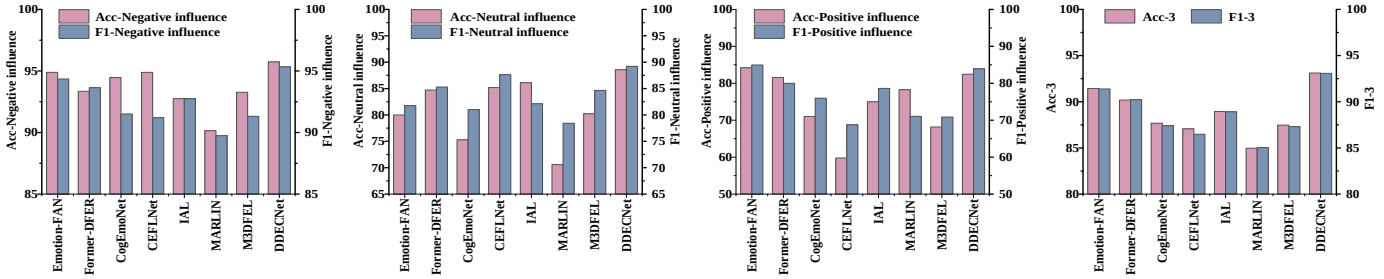

Fig. 8. The comparative results for the exploration of influence of driver emotion on driving behavior

C. Comparison with State-of-the-Art Models

To evaluate the overall performance of the proposed DDECNet, a fair comparison between the proposed DDECNet and state-of-the-art (SOTA) methods (including CogEmoNet [11], Emotion-FAN [12], Former-DFER [13], CEFLNET [14], IAL [15], MARLIN [16], M3DFEL [17]) is conducted. The convergence curves of accuracy and loss during training and testing phases are illustrated in **Fig. 6**.

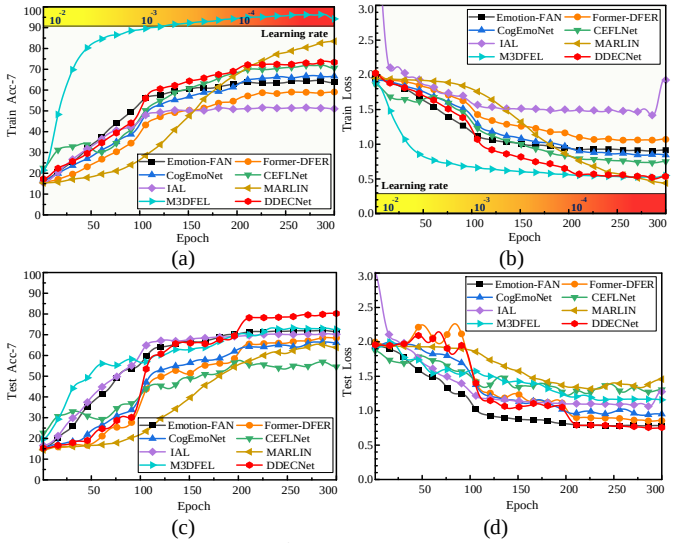

Fig. 6. The convergence curves of accuracy and loss during training and testing phases. (a) Acc-7 curve during training phase (b) Loss curve during training phase (c) Acc-7 curve during testing phase (d) Loss curve during testing phase

From **Fig. 6**, the proposed DDECNet model converges rapidly with variable learning rate, demonstrating that the proposed DDECNet has better convergence performance, compared to SOTA methods. Meanwhile, the similar results can also be observed in the accuracy curves during training and testing phase.

Then, the specific experimental results are recorded in **Table III** and **Table IV.** The visualization results are shown in **Fig. 7** and **Fig. 8**, respectively.

For the driver emotion classification task (as illustrated in **Table III** and **Fig. 7**), the proposed DDECNet achieves the highest classification Acc and F1 in surprise, fear, disgust, and neutral emotion over all the SOTA methods. Meanwhile, the classification performance of happiness, sadness, and anger emotions also wins the second place over currently advanced approaches in terms of Acc and F1. Notably, both the Acc-7 and F1-7 achieve the first place, outperforming other competitors.

For the exploration of influence of driver emotion on driving behavior (as illustrated in **Table IV** and **Fig. 8**), the proposed DDECNet outperforms other competitors, especially for the negative influence on driving behavior (+0.85% Acc, +0.98% F1) and neutral influence on driving behavior (+2.46% Acc, +1.59% F1). Meanwhile, the proposed method is superior to the other competitors in terms of Acc-3 and F1-3.

D. Ablation Analysis

To validate the role of VFD and DBD, as well as the effectiveness of the multi-task learning strategy, a series of ablation experiments are carried out.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/JIOT.2025.3527206, IEEE Internet of Things Journal

TABLE III
THE COMPARATIVE RESULTS OF DRIVER EMOTION CLASSIFICATION TASK

Ref.	Surprise		Fear		Disgust		Happiness		Sadness		Anger		Neutral		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc-7	F1-7
[11]	73.33	68.75	62.16	64.79	80.39	78.10	71.01	75.97	67.57	69.44	**80.30$_1$**	69.74	75.29	81.01	72.44	72.54
[12]	68.49	67.57	66.18	67.16	84.09	80.43	**84.21$_1$**	**84.96$_1$**	68.18	65.93	70.89	**73.68$_2$**	80.00	81.75	73.49	74.50
[13]	61.19	63.08	56.25	53.33	70.83	74.73	81.58	80.00	71.62	69.74	61.54	63.16	84.72	85.31	69.94	69.91
[14]	46.03	44.27	38.55	40.25	70.97	65.67	59.78	68.75	64.00	64.43	68.00	56.67	85.19	**87.62$_2$**	60.13	61.09
[15]	**79.10$_2$**	76.26	60.94	57.35	81.25	**82.11$_2$**	75.00	78.62	64.86	66.21	64.10	68.03	**86.11$_2$**	82.12	72.65	72.96
[16]	75.00	66.18	59.46	68.75	**82.35$_2$**	73.04	78.26	71.05	60.81	62.94	62.12	62.60	70.59	78.43	69.10	68.84
[17]	76.62	**76.62$_2$**	**79.71$_2$**	**73.83$_2$**	68.89	72.94	68.18	70.87	**81.16$_1$**	**75.17$_2$**	70.15	71.76	80.23	84.66	**75.57$_2$**	**75.12$_2$**
This Work	**79.45$_1$**	**79.45$_1$**	**86.76$_1$**	**77.63$_1$**	**88.64$_1$**	**87.64$_1$**	**82.46$_2$**	**83.93$_2$**	**77.27$_2$**	**79.53$_1$**	**74.68$_2$**	**79.19$_1$**	**88.57$_1$**	**89.21$_1$**	**81.84$_1$**	**82.37$_1$**

Note: the subscript **1** and **2** represent the specific ranking results.

TABLE IV
THE COMPARATIVE RESULTS FOR THE EXPLORATION OF INFLUENCE OF DRIVER EMOTION ON DRIVING BEHAVIOR

Ref.	Negative influence		Neutral influence		Positive influence		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc-3	F1-3
[11]	94.46	91.51	75.29	81.01	71.01	75.97	87.68	87.41
[12]	**94.89$_2$**	**94.35$_2$**	80.00	81.75	**84.21$_1$**	**84.96$_1$**	**91.44$_2$**	**91.39$_2$**
[13]	93.35	93.64	84.72	85.31	81.58	80.00	90.19	90.22
[14]	**94.89$_2$**	91.20	85.19	**87.62$_2$**	59.78	68.75	87.06	86.48
[15]	92.75	92.75	**86.11$_2$**	82.12	75.00	78.62	88.94	88.91
[16]	90.15	89.74	70.59	78.43	78.26	71.05	84.97	85.04
[17]	93.27	91.32	80.23	84.66	68.18	70.87	87.47	87.30
This Work	**95.74$_1$**	**95.33$_1$**	**88.57$_1$**	**89.21$_1$**	**82.46$_2$**	**83.93$_2$**	**93.11$_1$**	**93.08$_1$**

TABLE V
ABLATION ANALYSIS ON DRIVER EMOTION CLASSIFICATION TASK

Setting		Surprise		Fear		Disgust		Happiness		Sadness		Anger		Neutral		Average	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc-7	F1-7
Modality ablation	FVD	70.13	70.13	55.07	56.30	62.22	55.45	63.64	66.14	60.87	63.64	73.13	64.05	66.28	73.08	64.72	64.10
	DBD	59.74	58.23	53.62	49.66	42.22	42.22	46.97	46.97	42.03	41.13	44.78	46.88	43.02	46.25	47.81	47.33
	FVD+DBD	**79.45**	**79.45**	**86.76**	**77.63**	**88.64**	**87.64**	**82.46**	**83.93**	**77.27**	**79.53**	**74.68**	**79.19**	**88.57**	**89.21**	**81.84**	**82.37**
Learning strategy ablation	STL	58.44	53.25	40.58	38.10	31.11	32.18	34.85	34.07	40.58	40.29	41.79	43.75	43.02	48.37	42.38	41.43
	MT-IL	71.64	71.64	71.19	68.29	74.47	73.68	84.72	82.43	63.53	67.92	73.49	75.78	86.36	82.61	74.74	74.62
	MT-JL	**79.45**	**79.45**	**86.76**	**77.63**	**88.64**	**87.64**	**82.46**	**83.93**	**77.27**	**79.53**	**74.68**	**79.19**	**88.57**	**89.21**	**81.84**	**82.37**

Note: STL denotes the single-task learning strategy, MT-IL denotes multi-task independent learning, MT-JL denotes multi-task joint learning.

TABLE VI
ABLATION ANALYSIS ON THE EXPLORATION OF INFLUENCE OF DRIVER EMOTION ON DRIVING BEHAVIOR

Setting		Negative influence		Neutral influence		Positive influence		Average	
		Acc	F1	Acc	F1	Acc	F1	Acc-3	F1-3
Modality ablation	FVD	92.05	89.19	66.28	73.08	63.64	66.14	83.51	83.12
	DBD	82.57	81.08	43.02	46.25	46.97	46.97	70.56	70.13
	FVD+DBD	**95.74**	**95.33**	**88.57**	**89.21**	**82.46**	**83.93**	**93.11**	**93.08**
Learning strategy ablation	STL	80.43	78.51	43.02	48.37	34.85	34.07	67.43	66.97
	MT-IL	92.08	93.45	86.36	82.61	84.72	82.43	90.19	90.07
	MT-JL	**95.74**	**95.33**	**88.57**	**89.21**	**82.46**	**83.93**	**93.11**	**93.08**

Modality ablation: the proposed DDECNet is performed under single modality (i.e., pure VFD or pure DBD) and dual modalities (both VFD and DBD), respectively. The specific experimental results are illustrated in **Table V** and **Table VI**.

Specifically, the dual-modality setting achieves best classification performance in terms of Acc and F1, compared to single-modality setting. For diver emotion classification (as shown in **Table V**), the dual-modality setting shows a 17.12% and a 18.27% increase in Acc-7 and F1-7 respectively, compared to the pure VFD setting. Meanwhile, compared to the pure DBD setting, a 34.03% and a 35.04% improvement in Acc-7 and F1-7 are observed in the dual-modality setting. For the exploration of influence of driver emotion on driving behavior (as shown in **Table VI**), the dual-modality setting shows a 9.60% and a 9.96% increase in Acc-3 and F1-3 respectively, compared to the pure VFD setting. Meanwhile, compared to the pure DBD setting, a 22.55% and a 22.95% improvement in Acc-3 and F1-3 are observed in the dual-modality setting. Correspondingly, the confusion matrix (as shown in **Fig. 9)** also illustrates that the proposed DDECNet using dual modalities can achieve better classification performance. Because dual-pathway construction enables the effective fusion of data from both VFD and DBD.

Learning strategy ablation: the proposed DDECNet is performed using single-task learning (STL) strategy, multi-task independent learning (MT-IL) strategy, and multi-task joint

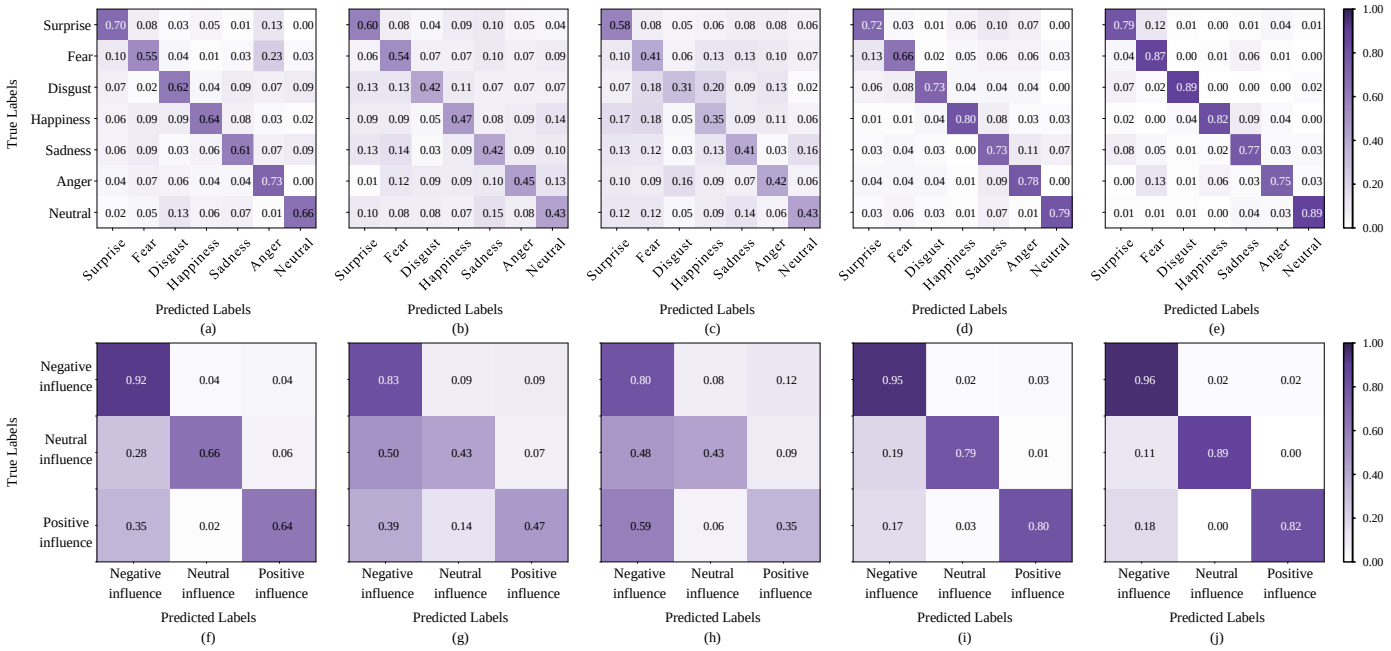

Fig. 9. Confusion matrix. (a) Driver emotion classification with pure FVD. (b) Driver emotion classification with pure DBD. (c) Dual-modality driver emotion classification with STL. (d) Dual-modality driver emotion classification with MT-IL. (e) Dual-modality driver emotion classification with MT-JL. (f) Exploration of influence of driver emotion on driving behavior with FVD. (g) Exploration of influence of driver emotion on driving behavior with pure DBD. (h) Exploration of influence of driver emotion on driving behavior with STL and dual-modality inputs. (i) Exploration of influence of driver emotion on driving behavior with MT-IL and dual-modality inputs. (j) Exploration of influence of driver emotion on driving behavior with MT-JL and dual-modality inputs.

TABLE VII
EVALUATION OF CORE COMPONENTS IN DDECNET

Setting				Acc-7 (%)	F1-7 (%)	Acc-3 (%)	F1-3 (%)
S1	**S2**	**S3**	**S4**				
CBAM [29]	Temporal Transformer	Inception Unit	Concatenation Unit	78.29	78.28	91.02	91.00
Spatial Transformer	GRU [30]	Inception Unit	Concatenation Unit	75.37	75.14	87.89	87.71
Spatial Transformer	BiLSTM [31]	Inception Unit	Concatenation Unit	56.58	56.00	74.32	74.14
Spatial Transformer	Temporal Transformer	Transformer [32]	Concatenation Unit	77.04	77.12	90.61	90.48
Spatial Transformer	Temporal Transformer	GRU [30]	Concatenation Unit	75.16	74.88	89.56	89.47
Spatial Transformer	Temporal Transformer	LSTM [33]	Concatenation Unit	74.32	74.58	89.14	89.20
Spatial Transformer	Temporal Transformer	Inception Unit	Cross Attention Fusion [34]	78.29	78.52	88.52	88.36
Spatial Transformer	Temporal Transformer	Inception Unit	Transformer-based Fusion [36]	73.70	73.09	84.55	84.35
Spatial Transformer	Temporal Transformer	Inception Unit	MISA Fusion [35]	61.38	61.07	80.79	81.55
Spatial Transformer	Temporal Transformer	Inception Unit	Learned Weights Fusion [37]	78.29	78.26	92.28	92.20
Spatial Transformer	Temporal Transformer	Inception Unit	Concatenation Unit	**81.84**	**82.37**	**93.11**	**93.08**

Note: S1→Extract spatial features from FVD. S2→Extract temporal features from FVD. S3→Extract time-series features from DBD.S4→Concatenate operation.

learning (MT-JL) strategy, respectively. The experimental results are presented in **Table V** and **Table VI**.

From these two tables, it can be seen that the MT-JL strategy achieves best classification performance in terms of Acc and F1, compared to STL strategy and MT-IL strategy. Specifically, the MT-JL strategy shows obvious improvement (+39.46% Acc-7, +40.94% F1-7 compared to the STL strategy; +7.10% Acc-7, +7.75% F1-7 compared to the MT-IL strategy) on driver emotion classification task (as shown in **Table V**). For the exploration of influence of driver emotion on driving behavior (as shown in **Table VI**), the MT-JL strategy exhibits obvious improvement (+25.68% Acc-3, +26.11% F1-3 compared to the STL strategy; +2.92% Acc-3, +3.01% F1-3 compared to the MT-IL strategy). Correspondingly, the confusion matrix in **Fig. 9** also illustrates that the proposed DDECNet using MT-JL strategy can achieve best classification performance. The main reason may be that the MT-JL strategy aims to learn multiple related tasks jointly, so that the knowledge contained in a task can be leveraged by other tasks.

E. Effectiveness Analysis

To study the necessity and effectiveness of core components each module (i.e., VFD processing module, the DBD processing module, and the fusion output module), the effectiveness analysis is carried out. The specific comparison results are collected in **Table VII**.

1) Evaluation of the Spatial Transformer

The convolutional block attention module (CBAM) [29] is applied to replace the spatial transformer, an obvious decrease can be observed in terms of Acc-7 (−3.55%), F1-7 (−4.09%), Acc-3 (−2.09%), and F1-3 (−2.08%). Because the spatial transformer can guide the proposed DDECNet to capture spatial

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/JIOT.2025.3527206, IEEE Internet of Things Journal

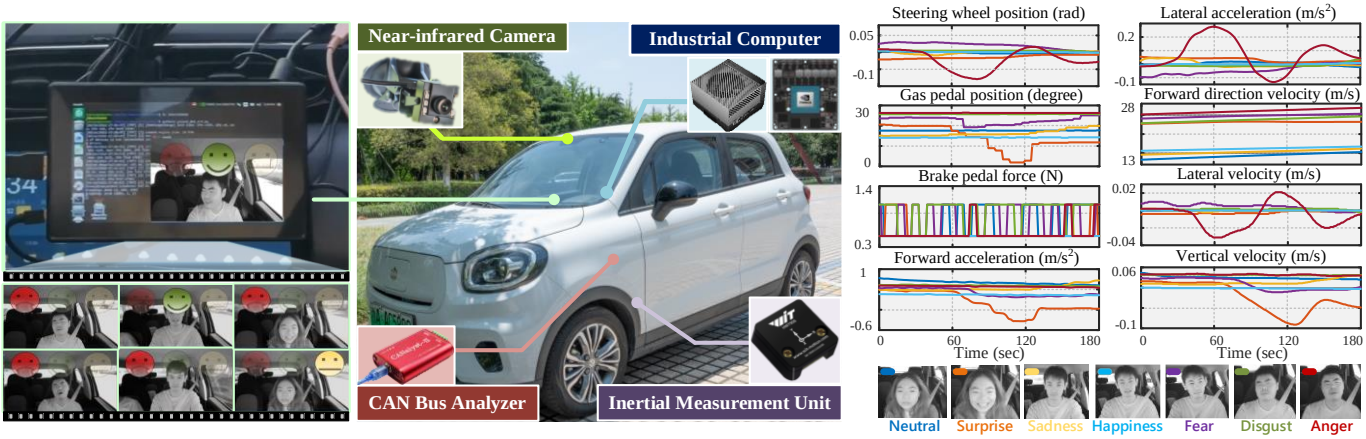

Fig. 10. Validation of the proposed DDECNet in real-world scenario

features, compared to CBAM. Meanwhile, the self-attention mechanism in spatial transformer enables capturing facial features with long-range dependencies.

2) Evaluation of the Temporal Transformer

The gated recurrent unit (GRU) [30] and bidirectional long short-term memory (BiLSTM) [31] are applied to replace the temporal transformer, the obvious decrease can be observed in classification performance (GRU: −6.47% Acc-7, −7.23% F1-7, −5.22% Acc-3, −5.37% F1-3; BiLSTM: −25.26% Acc-7, −26.37% F1-7, −18.79% Acc-3, −18.94% F1-3). Because the contextual facial features can be effectively extracted from temporal perspective in the temporal transformer.

3) Evaluation of the Inception Unit

The GRU [30], Transformer [32], and LSTM [33] are applied to replace the inception unit, the obvious decrease can be witnessed in classification performance (GRU: −6.68% Acc-7, −7.49% F1-7, −3.55% Acc-3, −3.61% F1-3; Transformer: −4.80% Acc-7, −5.25% F1-7, −2.50% Acc-3, −2.60% F1-3; LSTM: −7.52% Acc-7, −7.79% F1-7, −3.97% Acc-3, −3.88% F1-3). Because the inception unit using parallel convolutional blocks can speed up the learning of time-series features from different receptive fields.

4) Evaluation of Concatenation Unit

The cross-modal attention (CMA) fusion [34], MISA fusion [35], transformer-based fusion [36], and learned weights fusion [37] are applied to replace the concatenation unit, the obvious reduction can be witnessed in classification performance (CMA fusion: −3.55% Acc-7, −3.85% F1-7, −4.59% Acc-3, −4.72% F1-3; MISA fusion: −20.46% Acc-7, −21.30% F1-7, −12.32% Acc-3, −11.53% F1-3; transformer-based fusion: −8.14% Acc-7, −9.28% F1-7, −8.56% Acc-3, −8.73% F1-3; learned weight fusion: −3.55% Acc-7, −4.11% F1-7, −0.83% Acc-3, −0.88% F1-3). Because the correlation between VFD and DBD is difficult to explore by self-attention mechanism.

F. Validation in Real-World Scenarios

In this work, all the classification methods are further performed in an electric vehicle (Leapmotor T03) to validate the effectiveness in different real-world scenarios. The necessary sensor device deployment is shown in the middle of **Fig. 10**, the corresponding collected data samples are provided in the right of **Fig. 10**, and the negative (represented by red face), neutral (represented by yellow face), and positive (represented by green face) output examples can be seen in the screen embedded in the electric vehicle, as shown in the left of **Fig. 10**. Notably, all the classification methods are well-trained and deployed on the NVIDIA Jetson AGX Orin.

The experimental results of different classification methods are provided in **Fig. 11** (i.e., Latency, processing time and response time) and **Table VIII** respectively.

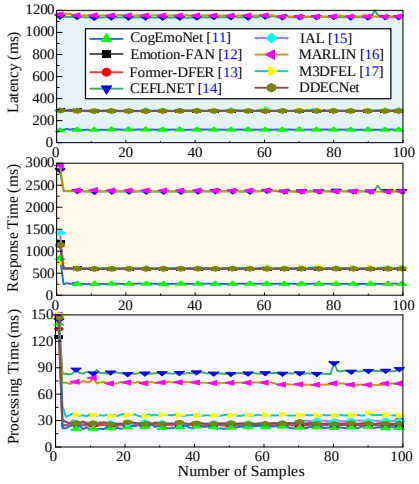

Fig. 11. The letancy, processing time and response time of different methods

TABLE VIII
THE COMPARATIVE RESULTS ON REAL-WORLD SCENARIO

Ref.	Negative influence		Neutral influence		Positive influence		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc-3	F1-3
[11]	85.97	83.28	69.94	75.72	65.82	69.85	73.91	76.28
[12]	86.55$_2$	86.23$_2$	74.62	75.25	78.19$_1$	78.33$_1$	79.79$_2$	79.94$_2$
[13]	85.63	85.41	78.35	79.43	75.34	74.16	79.77	79.67
[14]	86.33	83.32	79.03	81.54$_2$	53.97	62.29	73.11	75.72
[15]	84.95	84.15	80.06$_2$	76.36	69.45	72.47	78.15	77.66
[16]	83.84	81.84	64.78	72.61	72.23	65.62	73.61	73.36
[17]	85.50	83.67	74.11	78.88	62.58	64.54	74.06	75.70
Ours	**87.46$_1$**	**87.70$_1$**	**82.29$_1$**	**83.17$_1$**	**76.66$_2$**	**77.91$_2$**	**82.13$_1$**	**82.93$_1$**

Note: the subscript 1 and 2 represent the specific ranking results.

It can be seen that the latency, processing time, and response time of the proposed DDECNet (approximately 0.277s, 0.031s, and 0.585s) are smaller than almost all the other competitors (except for [11]), achieving the real-time requirements in the IoT scenario. Meanwhile, the comparative results of Acc and F1 demonstrate that the proposed DDECNet is superior in classification performance (ranking 1st in negative and neutral,

ranking 2nd in positive). Notably, the average results (Acc-3: 82.13%, F1-3: 82.93%) win the first place over currently advanced approaches. Namely, the proposed DDECNet is able to achieve a better trade-off between running speed and classification performance, compared with other competitors.

V. Conclusions

This paper focuses on the investigation of driver emotion classification network based on VFD and DBD (i.e., DDECNet). Specifically, the proposed DDECNet consists of three modules: the VFD processing module, the DBD processing module, and the fusion output module. The VFD processing module efficiently extracts high-level facial features from both spatial and temporal perspectives. The DBD processing module captures time-series features from different receptive fields. The fusion output module effectively integrates dual-modality features. Meanwhile, a multi-task learning strategy with a combined loss function is developed to oversee feature extraction across different modalities, enabling a reliable analysis to distinguish the positive, neutral, and negative influence on driving behavior. To verify the effectiveness of the proposed DDECNet, a joint verification in both realistic indoor environment (i.e., laboratory simulation on the PPB-Emo dataset) and real-world outdoor scenarios is carried out. The experimental results demonstrate that the proposed network is able to achieve a good balance between classification accuracy and running speed in the IoT scenario.

References

[1] N. Ying, Y. Jiang, C. Guo, D. Zhou and J. Zhao, "A multimodal driver emotion recognition algorithm based on the audio and video signals in internet of vehicles platform," *IEEE Internet Things*, Early Access 2024.

[2] 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.*, vol. 33, no. 12, pp. 7928-7942, Dec. 2023.

[3] W. Yue, C. Li, P. Duan and F. R. Yu, "Revolution on wheels: A survey on the positive and negative impacts of connected and automated vehicles in era of mixed autonomy," *IEEE Internet Things*, vol. 10, no. 24, pp. 21820-21835, 15 Dec.15, 2023.

[4] J. W. Li et al., "Single-channel selection for EEG-based emotion recognition using brain rhythm sequencing," *IEEE J. Biomed. Health. Inf.*, vol. 26, no. 6, pp. 2493-2503, June 2022.

[5] S. H. Kim, H. J. Yang, N. A. T. Nguyen, S. K. Prabhakar and S. W. Lee, "Wedea: A new EEG-based framework for emotion recognition," *IEEE J. Biomed. Health. Inf.*, vol. 26, no. 1, pp. 264-275, Jan. 2022.

[6] P. Wan, C. Wu, Y. Lin, and X. Ma, "On-road experimental study on driving anger identification model based on physiological features by ROC curve analysis," *IET Tntell. Transp. Syst*, vol. 11, no. 5, pp. 290-298, Jun. 2017.

[7] Z. Gao et al., "EEG-based spatio–temporal convolutional neural network for driver fatigue evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2755-2763, Sep. 2019.

[8] D. Lopez-Martinez, N. El-Haouij, and R. Picard, "Detection of real-world driving-induced affective state using physiological signals and multi-view multi-task machine learning," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Cambridge, UK, Sep. 2019, pp. 356–361.

[9] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. El-Sayed, and A. Elleithy, "Deep neural network approach for pose, illumination, and occlusion invariant driver emotion detection," *Int. J. Environ. Res. Public Health*, vol. 19, no. 4, p. 2352, Feb. 2022.

[10] W. Li et al., "Global-local-feature-fused driver speech emotion detection for intelligent cockpit in automated driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 2684-2697, Apr. 2023.

[11] W. Li et al., "Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit," *IEEE Trans. Comput. Soc. Syst.*, vol. 9, no. 3, pp. 667-678, Jun. 2022.

[12] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 3866-3870.

[13] Z. Zhao and Q. Liu, "Former-dfer: Dynamic facial expression recognition transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*, Chengdu, China, Oct. 2021, pp. 1553-1561.

[14] Y. Liu et al., "Clip-aware expressive feature learning for video-based facial expression recognition," *Inf. Sci.*, vol. 598, pp. 182-195, Jun. 2022.

[15] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington DC, USA, Jun. 2023, pp. 67-75.

[16] Z. Cai et al., "Marlin: Masked autoencoder for facial video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 1493-1504.

[17] H. Wang et al., "Rethinking the learning paradigm for dynamic facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 17958-17968.

[18] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, "A convolution bidirectional long short-term memory neural network for driver emotion recognition," *IEEE Trans. Intell. Transport. Syst.*, vol. 22, no. 7, pp. 4570-4578, Jul. 2021.

[19] M. Patil and S. Veni, "Driver emotion recognition for enhancement of human machine interface in vehicles," in *International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, Apr. 2019, pp. 0420-0424.

[20] L. Mou et al., "Driver emotion recognition with a hybrid attentional multimodal fusion framework," *IEEE Trans. Affective Comput.*, vol. 14, no. 4, pp. 2970-2981, Oct. 2023.

[21] W. Li et al., "Visual-attribute-based emotion regulation of angry driving behaviors," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 3, pp. 10-28, May 2022.

[22] 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. Affective Comput.*, vol. 14, no. 1, pp. 747-760, Jan. 2023.

[23] F. Eyben et al., "Emotion on the road: necessity, acceptance, and feasibility of affective computing in the car," *Adv. Hum. Comput. Interact.*, vol. 2010, p. 5:1-5:17, Jan. 2010.

[24] P. D. Paikrao, A. Mukherjee, D. K. Jain, P. Chatterjee and W. Alnumay, "Smart emotion recognition framework: A secured IoVT perspective," *IEEE Consum. Electron. Mag.*, vol. 12, no. 1, pp. 80-86, Jan. 2023

[25] W. Li et al., "A multimodal psychological, physiological and behavioral dataset for human emotions in driving tasks," *Sci. Data*, vol. 9, no. 1, Aug. 2022.

[26] X. Ji, Z. Dong, Y. Han, C. S. Lai, G. Zhou, and D. Qi, "EMSN: An energy-efficient memristive sequencer network for human emotion classification in mental health monitoring," *IEEE Trans. Consum. Electron.*, 2023, Early Access.

[27] Z. Dong, X. Ji, C. S. Lai, and D. Qi, "Design and implementation of a flexible neuromorphic computing system for affective communication via memristive circuits," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 74-80, Jan. 2023.

[28] H. Chen, H. Zhao, Z. Zhang and K. Li, "Discriminative feature learning-based federated lightweight distillation against multiple attacks," *IEEE Internet Things*, vol. 11, no. 10, pp. 17663-17677, 15 May15, 2024.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, Sep. 2018, pp. 3-19.

[30] 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*. 2023, Early Access.

[31] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631-2641, July 2019.

[32] Z. Dong, X. Ji, C. S. Lai, D. Qi, G. Zhou and L. L. Lai, "Memristor-based hierarchical attention network for multimodal affective computing in mental health monitoring," *IEEE Consum. Electron. Mag.*, vol. 12, no. 4, pp. 94-106, July 2023.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural*

Comput., vol. 9, no. 8, pp. 1735-1780, Nov. 1997.
[34] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019, pp. 603-612.
[35] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA, USA: ACM, Oct. 2020, pp. 1122-1131.
[36] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878-12895, Nov. 2023.
[37] E. Lim, H. J. Yang, H. J. Yang, S. H. Kim, S. Kim, J. E. Shin, and A. Kim, "Modality weights based fusion model for social perception prediction in video, audio, and text," in *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor*, Melbourne VIC Australia: ACM, Oct. 2024, pp. 12–19.