DECNet: A Non-Contacting Dual-Modality Emotion Classification Network for Driver Health Monitoring

Zhekang Dong[1,4], Chenhao Hu[1], Shiqi Zhou[1], Liyan Zhu[1], Junfan Wang[1], Yi Chen[2], Xudong Lv[1], Xiaoyue Ji[3*]
[1]Hangzhou Dianzi University [2]Zhejiang University [3]Tsinghua University
[4]Zhejiang Provincial Key Laboratory of Equipment Electronics
{englishp, chenhaohu, shiqizhou, 232040154, wangjunfan}@hdu.edu.cn
morningone@126.com 15B901019@hit.edu.cn jixiaoyue@mail.tsinghua.edu.cn

Abstract

Negative emotions have been identified as significant factors influencing driver behavior, easily leading to extremely serious traffic accidents. Hence, there is a pressing need to develop an automatic emotion classification method for driver health monitoring and road safety improvement. Most of the existing methods predominantly focus on single modalities, resulting in suboptimal classification performance due to the underutilization of heterogeneous information. In this work, we propose a novel non-contacting dual-modality driver emotion classification network (DECNet) to address these limitations. DECNet consists of three key modules: 1) facial video modality processing module; 2) driving behavior modality processing module; 3) fusion decision module. Meanwhile, we introduce a combined multi-task learning strategy within DECNet to improve the efficacy in the driver emotion classification task. To evaluate the effectiveness of the proposed DECNet, we conducted experiments on the PPB-Emo dataset, the experimental results showcase the superiority in terms of accuracy (⩾ 6.12% Acc-7) and F1-score (⩾ 7.25% F1-7) compared to existing state-of-the art methods. The model and code will be available at https://github.com/fqfqngxhs/DECNet.git

1. Introduction

During vehicular operation, driver's emotion is inevitably influenced by multiple factors (e.g., the surrounding environment, psychophysiological states, traffic conditions etc.), which may lead to risky driving behavior and even serious traffic accident especially when the significant emotional fluctuation occurs [1, 2, 3]. Timely and accurate recognition of emotional state is beneficial for the execution of healthcare and safety measures, as well as for the establishment of a congenial and structured driving ambiance within the framework of a smart city [37, 38].

Emotion classification technology typically monitors drivers' emotions by analyzing physiological signals such as facial expressions [9-13, 15-20], voice [14, 39], electrocardiography (ECG) [40, 41], electroencephalography (EEG) [4, 5], and electromyography (EMG) [42, 43] using various deep learning approaches. According to the signal acquisition methods, driver emotion classification technology can be roughly divided into two categories, i.e., the contact methods [4-8, 40-43] and non-contact methods [9-20, 39]. For the contact methods, drivers are always required to wear some contact sensors during driving. Although these methods may appear to perform well due to the accurate measurement, the process of signal acquisition can adversely affect driving behavior, particularly in emergency situations.

Different with the contact methods, non-contact methods employ sensors that do not require physical contact (e.g., near-infrared cameras) to classify emotions, thereby minimizing impact on driving performance and fostering a safer, more comfortable driving environment. Xiao et al. [9] proposed a transfer learning model to realize emotion classification based on facial expressions. Du et al. [10] combined facial skin information with RGB component variations for driver emotion classification. In [11], a driver emotion classification method by fusing local binary patterns and facial features was proposed, which is able to address the problem of varying illumination conditions. Li et al. [12] proposed a driver emotion recognition model considering both facial expressions and cognitive process features for driver emotion classification. Mou et al. [13] proposed a multimodal fusion framework that incorporates a hybrid attention mechanism to fuse non-invasive multimodal data from eyes, vehicles, and surrounding environment for driver emotion classification. A speech-based emotion classification network was

*Corresponding author. jixiaoyue@mal.tsinghua.edu.cn
This work was supported in part by the National Postdoctoral Researcher Support Program under Grant GZB20230356, Ministry of Science and Technology - Yangtze River Delta Science and Technology Innovation Program under Grant YDZX20233100004028, Fundamental Research Funds for the Provincial University of Zhejiang under Grant GK229909299001-06, and Shuimu Tsinghua Scholar Program under Grant 2023SM035.

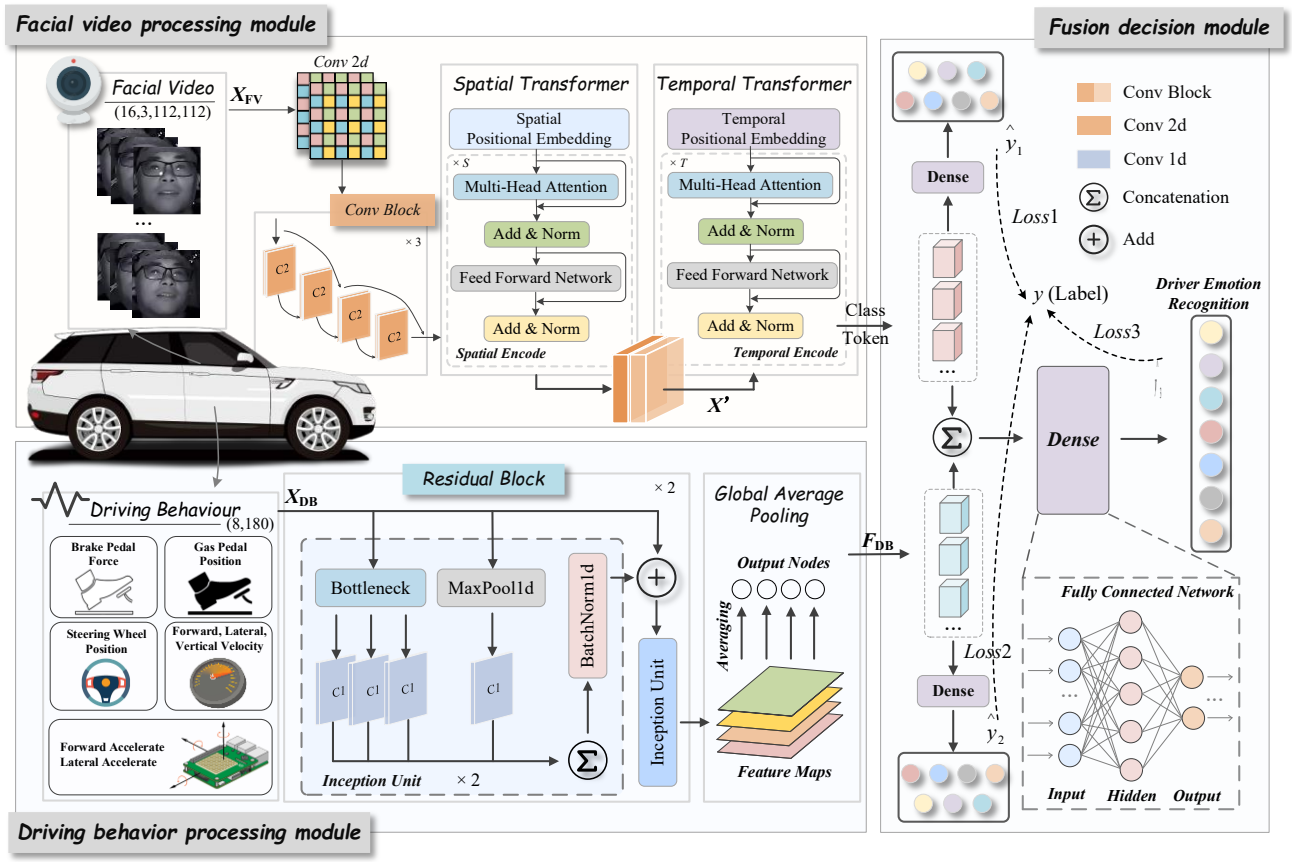

Figure. 1. Overview of the proposed DECNet

proposed [14], leveraging both global acoustic and local spectrogram features for accurate emotion detection. Meng et al. [15] developed Emotion-FAN, a hybrid network combining deep Convolutional Neural Networks (CNN) with frame attention for emotion classification. Similarly, Zhao et al. [16] proposed a dynamic facial expression recognition method (Former-DFER) that can address the occlusion and non-frontal poses issues during driving. In [17], a clip-ware emotion-rich feature learning network (CEFLNET) for robust video-based facial emotion expression classification was proposed. A dynamic facial expression classification network using intensity-aware loss (IAL) was developed [18], which can address the problem of large intra-class and small inter-class differences. In [19], a self-supervised facial video masked autoencoder (MARLIN) was proposed for accurate facial expression recognition, learning universal facial representations from non-annotated videos. Wang et al. [20] created a multi-3D dynamic facial expression learning network (M3DFEL) to address inexact labeling issues and enhance driver emotion classification accuracy.

Non-contact driver emotion classification methods are safer and more comfortable. However, they still suffer from two main limitations. Specifically, (1) Most of existing driver emotion classification methods rely on single modality (e.g., facial video). Some related modality data (e.g., driving behavior modality) and coupling relation between different modalities have not been fully explored; (2) The feature extraction in existing emotion classification methods is more suitable for single-modality data, how to supervise feature extraction from different modalities is a critical problem.

Based on this, a dual-modality non-contact driver emotion classification network (DECNet) is proposed in this work, which aims to deal with the two outlined research gaps in the field of driver emotion classification. The main contributions are three-folds: (1) Different with the most of exiting driver emotion classification methods, a dual-modality driver emotion classification network based on facial video and driver behavior (i.e., DECNet) is proposed, enabling efficient utilization of heterogeneous information to improve the classification performance. (2) Compared with single-task learning strategy, a multi-task learning strategy with combined loss function is designed, which is beneficial to supervise feature extraction from different modalities. (3) A series of comparative experiments and analysis (including ablation analysis and effectiveness analysis) are carried out. The experimental

results demonstrate that the proposed DECNet has advantages in terms of classification accuracy and F1-score.

2. Method

The overview of the proposed DECNet is depicted in Fig. 1. It consists of three main modules, i.e., the facial video modality processing module (Sec. 2.1), the driving behavior modality processing module (Sec. 2.2), and the fusion decision module (Sec. 2.3). The corresponding dual-modality input signals injected to the DECNet include the video signals (i.e., facial videos) and driving behavior signals (i.e., steering wheel position, gas pedal position, brake pedal force, forward direction acceleration, lateral acceleration, forward direction velocity, lateral velocity, and vertical velocity). The output of the DECNet is the discrete emotion classification result. Fig. 1 illustrates the signal acquisition devices (e.g., near-infrared camera, the position sensor, the acceleration sensor, etc.) and the corresponding signal examples for DECNet. Notably, different with the contact sensors, these non-contact devices have almost no effect on driving performance, beneficial to develop a safe and comfortable driving environment.

2.1. Facial Video Modality Processing Module

In facial video modality processing module, each input video sample is converted into a 16-frame facial image sequence with the size of 112×112, denoted as $X_{FV}\in\mathbb{R}^{16\times3\times112\times112}$. For each frame of the facial image, features are initially extracted using a two-dimensional convolutional layer followed by three residual convolutional blocks. The feature map is denoted as $M\in\mathbb{R}^{C\times H'\times W'}$, where C, H', and W' represent the channel number, height, and width of the feature map, respectively. Subsequently, the feature map is flattened into a one-dimensional sequence denoted as $M^f\in\mathbb{R}^{Q\times C}$, where $Q=H'\cdot W'$. The spatial transformer comprises spatial positional embedding and S-layer spatial encoders. In the spatial positional embedding process, the spatial positions are encoded by:

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

where m_p^f and e_p are visual word embedding and learnable position embedding, respectively.

The encoded result z_p^0 is then fed into the S-layer spatial encoders. 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=8 denotes the total number of attention heads. For the k-th attention head, the self-attention weight $\lambda_p^{(l,k)}$ is 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' represents the latent dimensionality of each attention head.

Then, the output of the l-layer spatial encoder z_p^l can be obtained 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 represents the projection matrix, $MLP(\cdot)$ means the MLP mapping, and $\lambda_{p,p'}^{(l,k)}$ denotes the self-attention weight.

The feature embedding $x'_t\in\mathbb{R}^F$ for each frame is computed by:

$$x'_t = GAP(g(Mr)) \quad t\in\{1,2,\cdots16\} \tag{9}$$

where $Mr\in\mathbb{R}^{C\times H'\times W'}$ denotes the refined feature map, $g(\cdot)$ represents the convolution operation, and $GAP(\cdot)$ denotes global average pooling.

The temporal transformer module consists of temporal positional embedding and T-layer temporal encoder. The input for the temporal encoder can be expressed by:

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

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

Notably, when t'=0, the learnable vector x'_0 represents the embedding of the class token.

Within each temporal encoder, the calculation of query, key, and value vectors follows the same procedure as in the spatial encoder. The output z_0^T of the class token from the terminal layer of the temporal encoder represents the facial video features. The emotion classification results $\hat{y}_1$ can be yielded by.

$$\hat{y}_1 = FC(z_0^T)\in\mathbb{R}^7 \tag{11}$$

where $FC(\cdot)$ denotes a fully connected network, 7 represents the number of categories for facial expressions

2.2. Driving Behavior Modality Processing Module

Eight types of driving behavior data are selected, namely, steering wheel position, gas pedal position, brake pedal force, forward direction acceleration, lateral acceleration,

forward direction velocity, lateral velocity, and vertical velocity. The data is sampled at a frequency of 60 times per second, with each sample spanning a duration of 3 seconds. The input for the behavior modality can be represented as $X_{DB} \in \mathbb{R}^{8\times180}$. The driving behavior modality processing module comprises two residual blocks, designed to mitigate the vanishing gradient problem. Each residual block contains three inception units. An average pooling layer is employed to average the output features across the temporal dimension. Within the inception unit, the initial component is the bottleneck layer, a one-dimensional convolutional layer with a kernel size of 1 and a stride of 1, aiming at reducing the number of feature channels to improve computational efficiency. The subsequent component is one-dimensional convolution with different kernel sizes (i.e., 39, 19, and 9) and strides (i.e., 19, 9, and 4) applied to the same input, enabling the capture of features across different spatial extents.

To mitigate the model's sensitivity to minor noise, a parallel Max-Pooling operation followed by a one-dimensional convolution is applied in inception unit. The output features from each convolution are concatenated and processed through a BatchNorm layer. Then, the output of the BatchNorm layer is combined with the original input X_{DB} and then subjected to an additional inception layer. After a global average pooling and a fully connected network, the classification results are yielded by:

$$\hat{y}_2 = FC(F_{DB}) \in \mathbb{R}^7 \tag{12}$$

where F_{DB} represents the features extracted from the driving behavior modality, and $\hat{y}_2$ is the classification result associated with the driving behavior modality.

2.3. Fusion Decision Module

In fusion decision module, the features derived from both the facial video modality processing module and driving behavior modality processing module are merged through a concatenation unit. Then, the emotional classification result $\hat{y}$ can be obtained by:

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

$$F = cat(z_0^S, F_{db}) \in \mathbb{R}^{14} \tag{14}$$

where $cat(\cdot)$ denotes the concatenation operation, F represents the concatenated feature vector.

2.4. Multi-task Learning Strategy

During the training process, the dual-modality driver emotion classification task is decomposed into three subtasks. Specifically, the subtask 1 solely employs the features from the facial video modality; The subtask 2 exclusively utilizes the features from the driving behavior modality; The subtask 3 integrates the features from both the facial video modality processing module and driving behavior modality processing module. The overall loss function for training of DECNet *Loss* can be obtained by:

$$Loss = \alpha Loss1 + \beta Loss2 + \gamma Loss3 \tag{15}$$

$$Loss1 = CrossEntropyLoss(\hat{y}_1, y) \tag{16}$$

$$Loss2 = CrossEntropyLoss(\hat{y}_2, y) \tag{17}$$

$$Loss3 = CrossEntropyLoss(\hat{y}_3, y) \tag{18}$$

where y represents the target labels, $CrossEntropyLoss(\cdot)$ denotes the cross-entropy loss function. $Loss1$, $Loss2$, $Loss3$ represent the loss functions for subtask 1, subtask 2, and subtask 3, respectively. Parameters α, β, γ represent the weight assignments for $Loss1$, $Loss2$, $Loss3$.

3. Experiments

3.1. Experimental Setup

The proposed DECNet is trained and tested on a server equipped with dual NVIDIA GeForce RTX 4090 GPUs using the open-source PyTorch platform. The batch size was set to 128, and the learning rate was initially set to 0.01, with a reduction by a factor of 10 every 100 epochs. Training was terminated at the 300th epoch. The Stochastic Gradient Descent (SGD) optimizer, with a momentum of 0.9 and weight decay of 0.0001, was utilized for parameter optimization. To enhance the robustness of the model, the random cropping and horizontal flipping operations were applied to facial video frames. Meanwhile, Gaussian random noise was introduced to the driving behavior data to more accurately mimic the real signal acquisition process in a driving environment.

In this study, the driver emotions were categorized into seven classes, i.e., surprise, fear, disgust, happiness, sadness, anger, and neutral. The driver emotion classification accuracy (Acc-7), Macro F1 score (F1-7), averaged accuracy (Acc), and F1-score (F1) [26] were applied to evaluate the driver emotion classification performance. Meanwhile, considering the efficiency requirement of in-vehicle systems, the computational complexity [27] was also used as an evaluation metric.

3.2. Dataset

The PPB-Emo dataset [25] is currently the only publicly available multimodal dataset for driver emotion classification. It comprises physiological data, facial videos, and driving behavior data from 40 participants across 240 valid driving tasks. The samples for each emotion in the dataset are evenly distributed. For time-series driving behavior data, the linear interpolation and normalization are conducted. For time-series driving behavior data, the linear interpolation and normalization are conducted. For near-infrared facial video, face alignment is performed to standardize the position of the

face. In the PPB-Emo dataset, the 5-fold cross-validation is employed to evaluate the proposed DECNet, in which one-fold of the samples is used for testing, while the remaining samples are used for training.

3.3. Model Hyperparameter Optimization

We investigate the impact of hyperparameters on the classification performance, focusing on the effects of varying the depths of the spatial transformer, temporal transformer, and inception unit. The initial number of layers for spatial encoder, temporal encoder, and inception unit is set as 1, 1, 6 respectively. Three evaluation metrics including Acc-7, Macro F1, and computational complexity are collected in Tab. 1.

Setting			Metrics		
S	T	I	Acc-7 (%)	F1-7 (%)	Complexity (GFLOPs)
1	1	6	77.87	77.90	8.33
3	1	6	79.00	79.18	9.15
1	3	6	81.63	82.49	8.40
3	3	6	80.37	80.56	9.23
6	3	6	78.16	78.21	10.46
3	6	6	80.67	80.70	9.33
6	6	6	80.29	80.34	10.57
1	3	3	77.58	77.53	10.53
1	3	9	80.21	80.43	10.61

Table 1. The impact hyperparameters on model performance

From Tab. 1 and Fig. 2, it is observed that when the number of layers for spatial encoder, temporal encoder, and inception unit is set to 1, 3, and 6, DECNet achieves optimal classification performance.

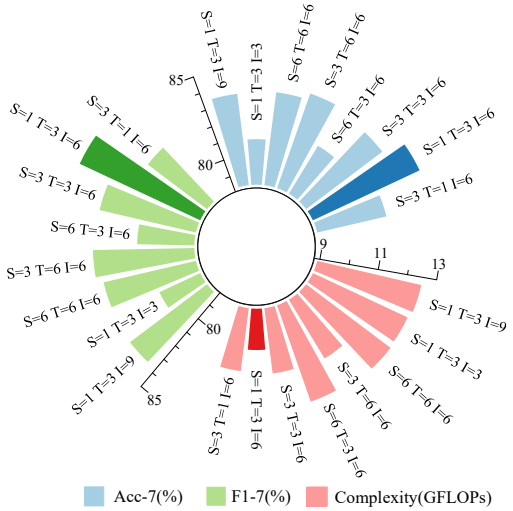

Figure 2. The impact of hyperparameters on classification performance

3.4. Experimental Results

In this part, a comparative analysis between the proposed DECNet and contemporary state-of-the-art models (including Emotion-FAN [15], Former-DFER [16], CogEmoNet [12], CEFLNET [17], IAL [18], MARLIN [19], M3DFEL [20]) is conducted. The corresponding visualized results are demonstrated in Fig. 2.

Method	Surprise		Fear		Disgust		Happiness		Sadness		Anger		Neutral		Average		Complexity (GFLOPs)
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc-7	F1-7	
[15]	68.49	67.57	**66.18**	**67.16**	**84.09**	**80.43**	**84.21**	**84.96**	**68.18**	65.93	**70.89**	**73.68**	80.00	81.75	**73.49**	**74.50**	**1.16**
[16]	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	10.49
[12]	73.33	68.75	62.16	64.79	80.39	78.10	71.01	75.97	67.57	69.44	**80.30**	69.74	75.29	81.01	72.44	72.54	15.51
[17]	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**	60.13	61.09	25.53
[18]	**79.10**	**76.26**	60.94	57.35	81.25	**82.11**	75.00	78.62	64.86	66.21	64.10	68.03	**86.11**	82.12	72.65	72.96	15.26
[19]	75.00	66.18	59.46	68.75	**82.35**	73.04	78.26	71.05	60.81	62.94	62.12	62.60	70.59	78.43	69.10	68.84	101.85
[20]	**76.62**	**76.62**	**79.71**	**73.83**	68.89	72.94	68.18	70.87	**81.16**	**75.17**	70.15	**71.76**	80.23	84.66	**75.57**	**75.12**	**9.18**
Ours	**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**	**8.40**

Note: Best results are highlighted as first, second, and third.

Table 2. The comparative results of DECNet in the driver emotion classification against state-of-the-art methods

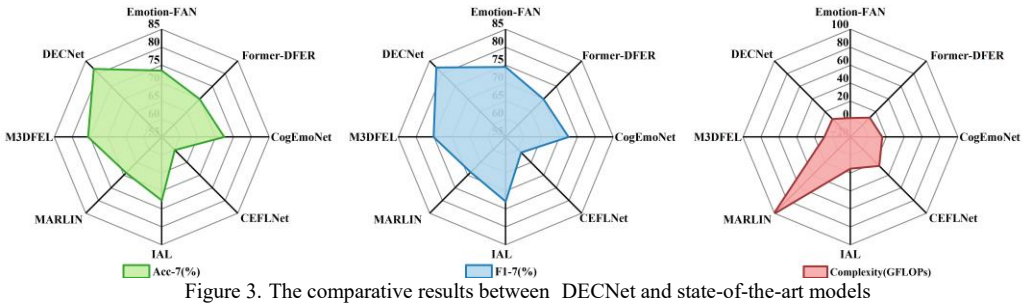

Figure 3. The comparative results between DECNet and state-of-the-art models

[19], M3DFEL [20]) is conducted on the PPB-Emo dataset. The corresponding results are summarized in Tab. 2, with corresponding visualizations displayed in Fig. 3. DECNet exhibits superior performance, securing top 2 positions in terms of accuracy, F1-score, and complexity. Specifically, for the driver emotion classification task, DECNet achieves the improvements on accuracy and F1-score in surprise, fear, disgust, and neutral emotion classification tasks over state-of-the-art methods. Meanwhile, the classification performance of happiness, sadness, and anger emotions also achieves the top two rankings, slightly outperforming other competitors. Notably, the average F1-score and accuracy win the first place over currently advanced approaches.

Fig. 4 illustrates the convergence curves of accuracy and loss on training and testing sets. Except for the M3DFEL model and the MARLIN model, the proposed DECNet outperforms other competitors on the training set. The M3DFEL model and the MARLIN model exhibit inferior performance on the testing set, and the proposed DECNet achieves the improvement over all the state-of-the-art methods on the testing set. The main reason for this may be the occurrence of overfitting in the M3DFEL and MARLIN models. The proposed DECNet model converges rapidly with a variable learning rate, confirming the efficacy of our learning rate adjustment strategy.

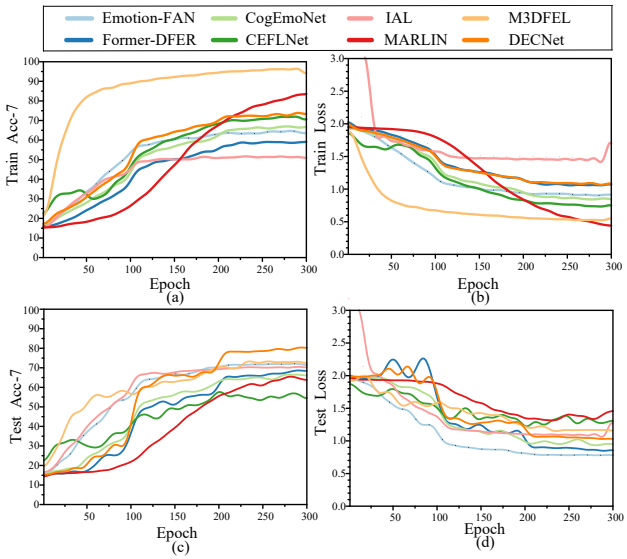

Figure 4: The convergence curves of accuracy and loss on training and testing sets. (a) Acc-7 in training phase (b) Loss in training phase (c) Acc-7 in testing phase (d) Loss in testing phase

3.5. Ablation Analysis

To assess the significance of the facial video modality and the driving behavior modality, along with the efficacy of the multi-task learning strategy, the modality ablation experiment and multi-task learning ablation experiment are conducted in this part.

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	FV	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	64.718	64.10
	DB	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	47.808	47.33
	FV+DB	**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**	**81.837**	**82.37**
Multi-task ablation	FV+DB	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	42.380	41.43
	FV+DB	**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: FV→facial video modality; DB→driver behavior modality.

Table 3. Experimental results of ablation analysis on driver emotion classification

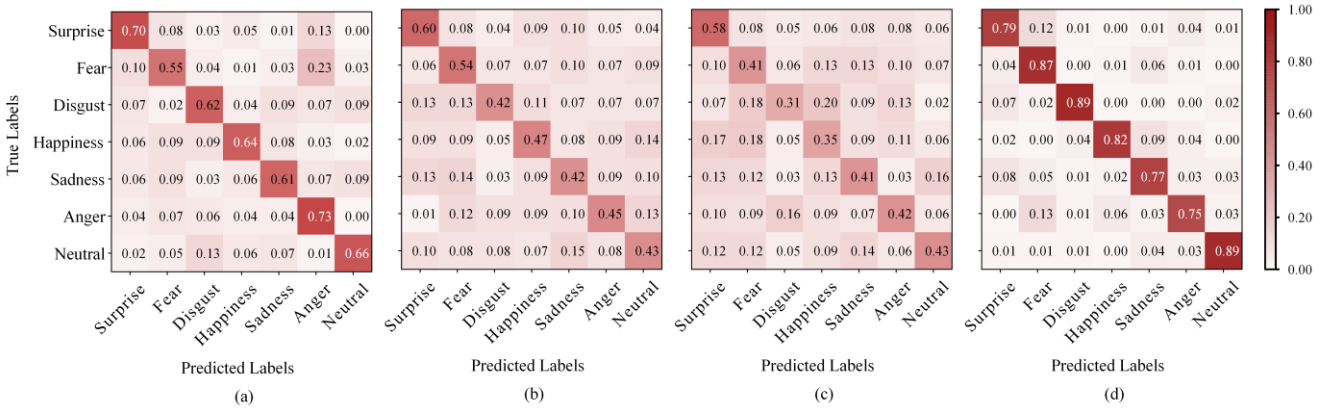

Figure 5. Confusion matrix. (a) Driver emotion classification with facial video modality. (b) Driver emotion classification with driving behavior modality. (c) Dual-modality driver emotion classification with single-task learning. (d) Dual-modality driver emotion classification with multi-task learning.

Modality Ablation

In the modality ablation experiment, DECNet is performed under single modality (either facial video modality or driving behavior modality) and dual-modality (both facial video modality and driving behavior modality), respectively. The results for driver emotion classification are shown in Tab. 3.

For driver emotion classification, the dual-modality configuration demonstrates a significant improvement in both accuracy and F1-score, compared to single modality setting. Specifically, the dual-modality setting exhibits a 17.12% increase in Acc-7 and an 18.27% improvement in F1-7 compared to the facial video modality alone; It shows a 47.81% increase in Acc-7 and a 35.04% improvement in F1-7 compared to the driving behavior modality alone. The confusion matrix in Fig. 5(a) and (b) further illustrate that dual-modality setting achieves better classification results. This is attributed to the effective integration of data from both facial video modality and driving behavior modality, significantly improving the classification performance.

Multi-task Learning Ablation

In the multi-task learning ablation experiment, DECNet is performed with single-task learning strategy and multi-task learning strategy, respectively. The corresponding experimental results are shown in Tab. 3. In driver emotion classification, DECNet with the multi-task learning strategy shows a significant improvement in accuracy and F1-score compared to single-task learning strategy. Specifically, the multi-task learning strategy exhibits a 39.46% increase in Acc-7 and a 40.94% improvement in F1-7 compared to the single-task learning strategy. Correspondingly, the confusion matrix in Fig. 5(c) and (d) further illustrate that the multi-task learning strategy achieves better classification results. This is attributed to the multi-task learning strategy can effectively supervise feature learning, significantly improving the classification performance.

3.6. Effectiveness Analysis

We examine the significance of core components within the facial video modality processing module, the driving behavior modality processing module, and the fusion decision module (including spatial transformer, temporal transformer, inception unit, and concatenation unit) in the proposed DECNet. To investigate the effectiveness of these core components in each module, a series of experiments are conducted. The detailed comparison results are summarized in Tab. 4.

Evaluation of the Spatial Transformer

To validate the effectiveness of the spatial transformer in extracting spatial features from facial video modality, the Convolutional Block Attention Module (CBAM) [29] is used to replace spatial transformer. After that, a discernible decline in classification performance (−3.55% Acc-7, −4.09% F1-7) can be observed. The main reason may be that the spatial transformer is able to guide DECNet to capture spatial features that are robust to occlusion and pose variations. Meanwhile, different with CBAM, the self-attention mechanism of spatial transformer facilitates the learning of correlations between facial features with long-range dependencies.

Setting	Acc-7	F1-7
CBAM-T-I-C	78.29	78.28
S-GRU-I-C	75.37	75.14
S-BiLSTM-I-C	56.58	56.00
S-T-Transformer-C	77.04	77.12
S-T-GRU-C	75.16	74.88
S-T-LSTM-C	74.32	74.58
S-T-I-Cross Attention Fusion	78.29	78.52
S-T-I-Transformer-based Fusion	73.70	73.09
S-T-I-MISA Fusion	61.38	61.07
The proposed: S-T-I-C	**81.84**	**82.37**

Note: S→Spatial Transformer. T→Temporal Transformer. S3→Inception Unit. S4→Concatenation Unit.

Table 4. Evaluation of core components in DECNet

Evaluation of the Temporal Transformer

To demonstrate the effectiveness of the temporal transformer in extracting temporal features from facial video modality, the Gated Recurrent Unit (GRU) [30] and Bidirectional Long Short-Term Memory (BiLSTM) [31] are used to replace temporal transformer. After that, an obvious reduction in classification performance (GRU: −6.47% Acc-7, −7.23% F1-7; BiLSTM: −25.26% Acc-7, −26.37% F1-7) can be observed. The reason may be that the temporal transformer can effectively learn contextual facial features from a temporal perspective.

Evaluation of the Inception Unit

To validate the effectiveness of the inception unit in capturing time-series features from driving behavior modality, GRU [30], Transformer [32], and LSTM [33] are used to replace the inception unit. The classification performance witnesses a significantly decrease (GRU: −6.47% Acc-7, −6.68% F1-7; Transformer: −4.80% Acc-7, −5.25% F1-7; LSTM: −7.52% Acc-7, −7.79% F1-7). The main reason may be that the parallel convolutional blocks in the inception unit enable the learning of features across different receptive field sizes.

Evaluation of Concatenation Unit

Finally, the effectiveness of concatenation unit in integrating facial video and driving behavior modalities is investigated. Cross-modal attention fusion [34], MISA fusion [35], and transformer-based fusion [36], are employed to replace the concatenation unit. The experimental results demonstrate that there is a significant

reduction in classification performance (cross-modal attention fusion: −3.55% Acc-7, −3.85% F1-7; MISA fusion: −20.46% Acc-7, −21.30% F1-7; transformer-based fusion: −8.14% Acc-7, −9.28% F1-7). This decline may be attributed to the difficulty of exploring the correlation between facial video and driving behavior modalities using the self-attention mechanism.

4. Conclusion

This paper focuses on the investigation of a driver emotion classification network (i.e., DECNet). Specifically, DECNet comprises three modules: the facial video modality processing module, the driving behavior modality processing module, and the fusion decision module. The facial video modality processing module enables the extraction of high-level facial features from both spatial and temporal perspectives; Through the driving behavior modality processing module, time-series features from different-sized receptive fields can be captured well; The fusion decision module effectively integrates features from both modalities. Meanwhile, a multi-task learning strategy with combined loss function is developed to supervise the feature extraction across different modalities, yielding reliable driver emotion classification results. To demonstrate the superiority of DECNet, a series of comparative experiments and analysis (including ablation analysis and effectiveness analysis) are conducted with state-of-the-art methods on the PPB-Emo dataset. The results demonstrate that DECNet achieves the best classification performance, offering enhanced precision in driver health monitoring and driving safety in smart city.

References

[1] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W. Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Comput. Surv.*, 53(3):1-30, 2021.

[2] Xiaoyue Ji, Zhekang Dong, Yifeng Han, Chun Sing Lai and Donglian Qi. A brain-inspired hierarchical interactive in-memory computing system and its application in video sentiment analysis. *IEEE Trans. Circ. Syst. Vid. Technol.*, 33(12):7928-794, 2023.

[3] Zhekang Dong, Xiaoyue Ji, Chun Sing Lai, Donglian Qi, Guangdong Zhou, and Loi Lei Lai. Memristor-based hierarchical attention network for multimodal affective computing in mental health monitoring. *IEEE Consum. Electron. Mag.*, 12(4):94-106, 2023.

[4] Jia Wen Li, Shovan Barma, Peng Un Mak, Fei Chen, Ming Tao Li, Mang I Vai, Sio Hang Pun. Single-channel selection for EEG-based emotion recognition using brain rhythm sequencing, *IEEE J. Biomed. Health. Inf.*, 26(6):2493-2503, 2022.

[5] Sun-Hee Kim, Hyung-Jeong, Ngoc Anh Thi Nguyen, Sunil Kumar Prabhakar and Seong-Whan Lee. WeDea: A new EEG-based framework for emotion recognition. *IEEE J. Biomed. Health. Inf.*, 26(1):264-275, 2022.

[6] Ping Wan, Chaozhong Wu, Yingzi Lin, and Xiaofe Ma. On-road experimental study on driving anger identification model based on physiological features by ROC curve analysis. *IET Tntell. Transp. Syst*, 11(5):290-298, 2017.

[7] Zhongke Gao, Xinmin Wang, Yuxuan Yang, Chaoxu Mu, Qing Cai, Weidong Dang, and Siyang Zuo. EEG-based patio–temporal convolutional neural network for driver fatigue evaluation. *IEEE Trans. Neural Networks Learn. Syst.*, 30(9):2755-2763, 2019.

[8] Boon Giin Lee, Teak Wei Chong, Boon Leng Lee, Hee Joon Park, Yoon Nyun Kim, and Beomjoon Kim. Wearable mobile-based emotional response-monitoring system for drivers. *IEEE Trans. Hum. Mach. Syst.* 47(5):636-649, 2017.

[9] Huafei Xiao, Wenbo Li, Guanzhoug Zeng, Yingzhang Wu, Jiyong Xue, Juncheng Zhang, Chengmou Li, and Gang Guo. On-road driver emotion recognition using facial expression. *Appl. Sci.*, 12(2):807, 2022.

[10] Guanglong Du, Zhiyao Wang, Boyu Gao, Shahid Mumtaz, Khamael M. Abualnaja, and Cuifeng Du. A convolution bidirectional long short-term memory neural network for driver emotion recognition. *IEEE Trans. Intell. Transp. Syst.*, 22(7):4570-4578, 2021.

[11] Mrinalini Patil and S. Veni. Driver emotion recognition for enhancement of human machine interface in vehicles. In *International Conference on Communication and Signal Processing*, pages 420-424, 2019.

[12] Wenbo Li, Guanzhou Zeng, Juncheng Zhang, Yan Xu, Yang Xing, Rui Zhou, Gang Guo, Yu Shen, and Fei-Yue Wang. Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit. *IEEE Trans. Computat. Social Syst.*, 9(3):667-678, 2022.

[13] Luntian Mou, Yiyuan Zhao, Chao Zhou, Bahareh Nakisa, Mohammad Naim Rastgoo, Lei Ma, Tiejun Huang, Baocai Yin, and Wen Gao. Driver emotion recognition with a hybrid attentional multimodal fusion framework. *IEEE Trans. Affect. Comput.*, 14(4): 2970-2981, 2023.

[14] Wenbo Li, Jiyong Xue, Ruichen Tan, Cong Wang, Zejian Deng, Shen Li, Gang Guo, and Dongpu Cao. Global-local-feature-fused driver speech emotion detection for intelligent cockpit in automated driving. *IEEE Trans. Intell. Veh.*, 8(4):2684-2697, 2023.

[15] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *IEEE International Conference on Image Processing*, pages 3866-3870, 2019.

[16] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553-1561, 2021.

[17] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhouwen Luo. Clip-aware expressive feature learning for video-based facial expression recognition. *Inf. Sci.*, 598:182-195, 2022.

[18] Hangting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 67-75, 2023.

[19] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1493-1504, 2023.

[20] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17958-17968, 2023.

[21] Wenbo Li, Bingbing Zhang, Peizhi Wang, Chen Sun, Guanzhong Zeng, Qiuyang Tang, Gang Guo, and Dongpu Cao. Visual-attribute-based emotion regulation of angry driving behaviors. *IEEE Intell. Transp. Syst. Mag.*, 14(3):10-28, 2022.

[22] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao. A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: emotions triggered by video-audio clips in driving scenarios. *IEEE Trans. Affect. Comput.*, 14(1):747 - 760, 2023.

[23] Florian Eyben, Martin Wöllmer, Tony Poitschke, Björn Schuller, Christoph Blaschke, Berthold Färber, and Nhu Nguyen-Thien. Emotion on the road: necessity, acceptance, and feasibility of affective computing in the car. *Adv. Hum. Comput. Interact.*, pages 1-17, 2010.

[24] Pavan D. Paikrao, Amrit Mukherjee, Deepak Kumar Jain, Pushpita Chatterjee, and Waleed Alnumay. Smart emotion recognition framework: A secured IoVT perspective. *IEEE Consum. Electron. Mag.*, 12(1):80-86, 2023.

[25] Wenbo Li, Ruichen Tan, Yang Xing, Guofa Li, Shen Li, Guanzhong Zeng, Peizhi Wang, Bingbing Zhang, Xinyu Su, Dawei Pi, Gang Guo, and Dongpu Cao. A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks. *Sci. Data*, 9(1):481, 2022.

[26] Xiaoyue Ji, Zhekang Dong, Yifeng Han, Chun Sing Lai, Guangdong Zhou, and Donglian Qi. EMSN: An energy-efficient memristive sequencer network for human emotion classification in mental health monitoring. *IEEE Trans. Consum. Electron.*, 69(4):1005-1016, 2023.

[27] Zhekang Dong, Xiaoyue Ji, Chun Sing Lai, and Donglian Qi. Design and implementation of a flexible neuromorphic computing system for affective communication via memristive circuits. *IEEE Commun. Mag*, 61(1): 74-80, 2023.

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436-444, 2015.

[29] Sanghyun Woo, Jongchan Park, and Joon-Young Lee. CBAM: Convolutional block attention module. In *European Conference on Computer Vision*, pages 3-19, 2018.

[30] Zhekang Dong, Xiaoyue Ji, Jiayang Wang, Yeting Gu, Junfan Wang and Donglian Qi. ICNCS: Internal cascaded neuromorphic computing system for fast electric vehicle state of charge estimation. *IEEE Trans. Consum. Electron.*, 2023.

[31] Yi Bin, Yang Yang, Fumin Shen, Ning Xie, Heng Tao Shen, and Xuelong Li. Describing video with attention-based bidirectional LSTM. *IEEE Trans. Cybernetics*, 49(7):2631-2641, 2019.

[32] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. TransFuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11): 12878-12895, 2023.

[33] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735-1780, 2010.

[34] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation, In *International Conference on Computer Vision*, pages 603-612, 2019.

[35] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122-1131, 2020.

[36] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger, Transfuser: Imitation with transformer-based sensor fusion for autonomous driving, *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11): 12878-12895, 2023.

[37] Zhekang Dong, Xiaoyue Ji, Guangdong Zhou, Mingyu Gao, and Donglian Qi. Multimodal neuromorphic sensory-processing system with memristor circuits for smart home applications. *IEEE Trans. Ind. Applicat.*, 59(1): 47-58, 2022.

[38] Xiaoyue Ji, Zhekang Dong, Chun Sing Lai, Guangdong Zhou, and Donglian Qi. A physics-oriented memristor model with the coexistence of NDR effect and RS memory behavior for bio-inspired computing. *Mater. Today Adv.*, 16: 100293, 2022.

[39] Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, and Rajesh Kumar Muthu. Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*, 2020.

[40] Foteini Agrafioti, Dimitris Hatzinakos, and Adam K. Anderson. ECG pattern analysis for emotion detection. *IEEE Trans. Affect. Comput.*, 3(1):102-115, 2011.

[41] Pritam Sarkar, and Ali Etemad. Self-supervised ECG representation learning for emotion recognition. *IEEE Trans. Affect. Comput.*, 13(3):1541-1554, 2020.

[42] Yumiao Chen, Zhongliang Yang, and Jiangping Wang. Eyebrow emotional expression recognition using surface EMG signals. *Neurocomputing*, 168: 871-879, 2015.

[43] Wei Chang Zhi. Stress emotion recognition based on RSP and EMG signals. *Adv. Mater. Research*, 709: 827-831, 2013.