

TFI²F-Net: Traffic-Facial Intra-Inter Frame Fusion Network for Driver Emotion Recognition under Naturalistic Driving

Chengmou Li, Zongxin Xiao, Yutian Zhang, Gang Guo, Zejian Deng,
Yang Xing, *Senior Member, IEEE*, and Wenbo Li*

Abstract—Driver emotion is a critical factor affecting both road traffic safety and human-machine interaction experience. Existing recognition methods mainly rely on individual affective cues, such as facial expressions, while rarely considering their collaborative modeling with traffic context. To address this limitation, this paper proposes a Traffic-Facial Intra-Inter Frame Fusion Network (TFI²F-Net), which explicitly models the stagewise interaction between traffic and facial frame sequence. The proposed framework consists of three key stages. First, an intra-frame cross-modal encoder based on dual-path attention mechanism is designed to adaptively model fine-grained interactions between traffic and facial features at the frame level. Second, an inter-frame reweighted semantic fusion module is developed to emphasize key frames and integrate temporal representations under disentangled semantic guidance. Third, auxiliary unimodal supervision and modality separation losses are incorporated to enhance single modality discriminability and promote effective disentanglement. In addition, we construct a naturalistic scenario driver emotion dataset (Scenario-Emo) comprising outside-view and inside-view video clips, where a participant-expert joint annotation protocol is employed to ensure the reliability. Extensive experiments on the Scenario-Emo and AIDE datasets demonstrate that the proposed TFI²F-Net consistently outperforms state-of-the-art methods. Overall, our model and dataset are expected to facilitate the development of affective interaction in intelligent cockpits.

Index Terms—Driver emotion recognition, multimodal learning, naturalistic driving, intelligent cockpit

I. INTRODUCTION

EMOTION plays a crucial role in human cognitive activities, and driving tasks are no exception. One major reason for focusing on driver emotion is its

strong relevance to driving safety, as both negative and positive emotions can influence how drivers focus their attention and behave [1], [2]. Over 90% of road traffic accidents are related to human factors [3], and the risk of accidents associated with extreme emotional states, such as anger or sadness, is approximately 9.8 times higher than baseline [4]. At the same time, intelligent cockpit systems that emphasize emotional interaction have become an important area of research in human-machine interaction [5]. Given that driving environments frequently alternate between high-stress and relaxed conditions, cockpit systems that continuously monitor emotion states and offer affective interaction responses can improve the driving experience and comfort [6]. Therefore, developing driver emotion recognition methods that perform effectively under naturalistic driving is essential for enabling timely emotional regulation and intervention, which is crucial for improving road traffic safety and human-machine interaction experience, as shown in Fig. 1.

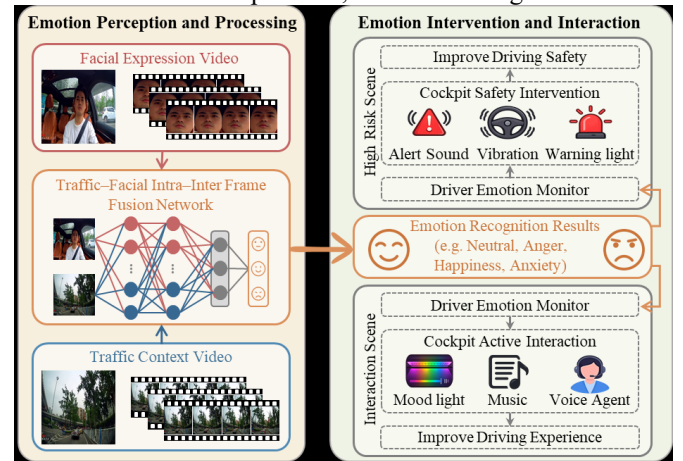


Fig. 1. Application schematic of proposed TFI²F-Net in automotive intelligent cockpits.

Accurate recognition of driver emotions is a prerequisite for affective interaction in intelligent cockpits. Existing studies have explored a variety of methods based on multimodal signals such as facial expressions [7], [8], [9], speech [10], physiological responses [11], [12], and driving behaviors [13]. These methods follow the dominant paradigm in affective computing, which assumes that emotions are associated with relatively stable expressive patterns and can therefore be inferred from observable signals [14]. However, this strategy

This work was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China (grant number 52302497); the Project of Chongqing Municipal Commission of Economics and Information Technology Science and Technology under grant number SEKJP001ZJY.

*Wenbo Li is the corresponding author for this paper.

Chengmou Li, Zongxin Xiao, Yutian Zhang, Gang Guo and Wenbo Li are with the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China (e-mail: chengmou_li@stu.cqu.edu.cn; zongxin_xiao@stu.cqu.edu.cn; yutian_zhang@stu.cqu.edu.cn; guogang@cqu.edu.cn; wenbo_li@cqu.edu.cn).

Zejian Deng is with the Department of Data and Systems Engineering, The University of Hong Kong, Hong Kong SAR, 999077, China (e-mail: z49deng@hku.hk).

Yang Xing is with the Centre for Autonomous and Cyber-Physical Systems, Cranfield University, UK (email: yang.x@cranfield.ac.uk)

may have limitations under dynamic and complex driving scenarios. Specifically, high cognitive load during driving may suppress salient facial muscle movements, while variations in lighting and head pose further degrade the reliability of facial cues. Physiological signals, such as EEG and ECG, are generally sensitive to emotional variations, but their reliance on multiple sensors increases deployment complexity and makes them vulnerable to motion-induced noise. Approaches based on speech and behavior are less intrusive, yet the discontinuous nature of speech signals and the strong correlation between driving behaviors and road condition limit their generalizability. Consequently, these challenges make emotion recognition methods that rely solely on individual signals difficult to generalize to naturalistic driving.

According to cognitive appraisal theory [15], emotional changes arise from an individual's subjective evaluation of the significance of situational events. In the context of driving, the complex and dynamic traffic environment is the primary situation that drivers constantly encounter. Increasing research has revealed specific associations between emotional states and traffic context. For example, Roidl et al. [16] studied how emotions like anger and anxiety are triggered by context factors including goal relevance, responsibility attribution, and goal congruence in laboratory. Additionally, Dittrich et al. [17] explored the connections between common emotions and traffic factors across temporal, spatial, and event-related aspects. Even under autonomous driving, Alsaied et al. [18] found that variables like intersection type and traffic flow significantly affect driver emotions. These results indicate that including traffic context as an input modality can provide emotion-related situational cues, which holds considerable promise for enhancing emotion recognition accuracy.

Nevertheless, traffic context alone cannot fully determine driver emotions, as similar events may elicit different affective appraisals across individuals. Therefore, we introduce facial expressions as individual-level cues to enable multimodal emotion recognition. However, effectively integrating these two modalities still involves three key challenges:

Challenge 1: Recent studies still mainly rely on individual affective cues, with limited exploration of the collaborative modeling between facial expressions and traffic context, thereby restricting naturalistic driver emotion recognition.

Challenge 2: Since facial and traffic modalities exhibit substantial semantic heterogeneity and temporal inconsistency in emotion-relevant cues, it is challenging to learn an effective and robust multimodal representation.

Challenge 3: Existing emotion datasets collected under naturalistic driving conditions remain limited in data scale, annotation quality, and multimodal completeness, restricting comprehensive evaluation of model generalizability.

To address these challenges, we propose the Traffic-Facial Intra-Inter Frame Fusion Network (TFI²F-Net) for naturalistic driver emotion recognition. Meanwhile, an on-road emotion dataset is constructed to provide a more comprehensive evaluation benchmark. The main contributions of this paper are summarized as follows:

- We propose TFI²F-Net, a progressive multimodal fusion

framework that jointly models facial and traffic video signals through intra-frame and inter-frame interaction, thereby achieving the collaborative modeling of internal affective cues and external inducing context.

- We design two dedicated modules, namely Intra-Frame Cross-modal Encoder and Inter-Frame Reweighted Semantic Fusion, to enhance semantic collaboration across modalities and discriminative temporal representation. Modality auxiliary supervision and semantic separation loss are also introduced to facilitate optimization.
- We introduce Scenario-Emo, a naturalistic scenario driver emotion dataset with 6,900 synchronized outside-view and inside-view clips annotated with four common emotion categories. Extensive experiments on Scenario-Emo and the public AIDE dataset showed that TFI²F-Net significantly outperforms state-of-the-art methods.

The structure of this paper is as follows: Section II reviews related work on driver emotion recognition. Section III presents the proposed framework and details of its main modules. Section IV explains the creation of the Scenario-Emo dataset and its sample distribution. Section V details the training configurations and evaluation metrics. Section VI reports the experimental results and analysis. Finally, Section VII concludes the paper and discusses future directions.

II. RELATED WORK

A. Driver Emotion Recognition

Accurate emotion recognition is essential for improving driving safety and interaction experience. Table 1 summarizes recent advances over the past three years in terms of methods, signals, datasets, and recognition performance. Existing studies have primarily followed two research directions:

1) *Algorithm optimization based on global and local facial features*: To better utilize facial information, recent studies have shifted from full face modeling to explicitly separating key local areas to capture detailed emotional cues [7], [8], [9], [19], [20]. For example, Hsu et al. [9] divided the facial feature into global, local, and salient regions based on landmarks, and used three parallel CNN branches for robust emotion recognition under varying lighting conditions. Li et al. [8] further introduced global facial key points, combined with a multi-class joint spatial filter and spatio-temporal feature extraction module, to enhance the global-local fusion of facial representation. Moreover, Xiang et al. [19] combined multi-scale facial features, remote photoplethysmography signals, and facial action units for joint emotion classification and dimensional prediction. These methods improve the discriminability of facial cues, but their effectiveness under naturalistic driving is still limited, since real-world facial expression changes are often weak and easily affected by head motion, pose variation, and occlusion.

2) *Algorithm optimization based on multimodal signals*: Multimodal signals have become a central focus in driver emotion recognition, with driving behavior widely used due to its strong connection to driving tasks [11], [13], [21], [22], [23]. For instance, Dong et al. [13] modeled facial expressions and driving behaviors with a dual-path network, while Wu et

al. [23] introduced a mid-level fusion strategy to facilitate facial-behavior alignment. Others like speech [10], [24], EEG [11], and eye movements [22] have also been found to improve prediction performance. Notably, since driver emotions are heavily influenced by external environment, traffic context have attracted growing interest. Liu et al. [25] used VGG and GRU to encode traffic context and driving behavior, respectively, and demonstrated the effectiveness in

driver emotion recognition. Similarly, Liu et al. [26], [27] explored facial-traffic features fusion using horizontal-vertical attention, but mainly relied on separate encoding followed by simple feature concatenation. Overall, these studies confirm the value of multimodal signals, especially traffic context, yet most existing methods rely on shallow fusion strategies, with limited modeling of sufficient and effective cross-modal interaction and temporal dynamics under naturalistic driving.

TABLE I
RELATED WORKS ABOUT DRIVER EMOTION RECOGNITION IN THE PAST THREE YEARS

Cite	Year	Method	Emotion	Signals	Dataset	Category	Metrics
[9]	2025	CaFpNet	Happy, Normal, Angry	1. Facial image	MLI-DER	Simulated	92.42% Acc
[8]	2025	GLF-STAF	Happy, Normal, Angry	1. Facial video	MLI-DER	Simulated	90.50% Acc
[19]	2025	MER-MFVA	Neutral, Happy, Angry, Sad, Fear	1. Facial video	MMDE	Simulated	86.98% Acc
[11]	2024	Stacking with Reinforcement	Anger, Disgust, Happiness, Sadness, Surprise, Fear, Neutral	1. EEG-Driving sequence 2. Psychological survey	PPB-Emo	Simulated	85.42% Acc
[13]	2024	DDECNet	Anger, Disgust, Happiness, Sadness, Surprise, Fear, Neutral	1. Facial video 2. Driving sequence	PPB-Emo	Simulated	81.84% Acc
[21]	2024	MDEmoNet	Anger, Disgust, Happiness, Sadness, Surprise, Fear, Neutral	1. Facial video 2. Driving sequence	PPB-Emo	Simulated	67.92% Acc
[28]	2024	DERNet	Neutral, Happy, Angry	1. Facial video	DEFE	Simulated	81.20% Acc
[20]	2023	HPFS	Happy, Normal, Angry	1. Facial image	MLI-DER	Simulated	88.24% Acc
[22]	2023	ConvLSTM + Hybrid Attention	Low-high valence, Low-high Arousal, Low-high dominance	1. Eye-Driving sequence 2. Environment info	Self-experiment Data	Simulated	97.12% Acc
[24]	2024	CNN+GRU+ Attention	Angry, Calm, Disgusted, Fearful, Happy, Neutral; Sad, Surprised.	1. Facial video 2. Audio sequence	RAVDESS	Generic	93.00% Acc
[23]	2024	MDERNet	Anger, Disgust, Sadness, Fear, Happiness, Surprise, Contempt	1. Facial video 2. Driving behavior	CK+	Generic	89.51% Acc
[10]	2023	GL-DSED	Anger, Disgust, Happiness, Sadness, Surprise, Fear, Neutral	1. Audio sequence	EmoDB	Generic	87.61% Acc
[27]	2025	MMTL-UniAD	Anxiety, Peace, Weariness, Happiness, Anger	1. Facial-Traffic-Body video 2. Gesture-Posture sequence	AIDE	Naturalistic	76.67% Acc
[26]	2025	UMD-Net	Anxiety, Peace, Weariness, Happiness, Anger	1. Facial-Traffic-Body video 2. Gesture-Posture sequence	AIDE	Naturalistic	76.56% Acc
[7]	2025	ILAB-CNN	Anger, Disgust, Happiness, Sadness, Surprise, Fear	1. Facial image	KMU-FED	Naturalistic	98.80% Acc
[29]	2025	TSN-w	Low stress, Medium stress, High stress	1. Traffic video	AffectiveROAD	Naturalistic	77.00% Acc
[30]	2024	BI-DED	Anger, Disgust, Happiness, Sadness, Surprise, Fear, Neutral	1. Facial image	RoadEmo	Naturalistic	93.20% Acc

Note: Some studies involve evaluations across multiple datasets; here, we briefly summarize a subset of the results. Acc denotes recognition accuracy.

B. Multimodal Emotion Recognition

Multimodal emotion recognition (MER) aims to combine data from various modalities like visual, speech, and text to make accurate and robust emotion predictions. In terms of representation learning, MER methods are typically classified into three types:

1) *Feature-level fusion*: Each modality is first encoded independently, and the resulting features are fused before final prediction, most commonly via concatenation or summation [13]. To account for the unequal contributions of different modalities, some studies further introduce gating networks [31] or attention mechanisms [32] for adaptive weighting. Although feature-level fusion is flexible and compatible with various encoder architectures, its interaction is typically confined to high-level representations, which may overlook useful cross-modal correspondences at earlier stages.

2) *Decision-level fusion*: Each modality is processed by its

own prediction network, and the final emotion label is made by combining the posterior probabilities from each modality network at the decision stage. In addition to basic voting or weighted averaging, recent studies have explored more advanced strategies, such as multi-statistical feature weighting [33], meta-learner-based prediction [34], and Bayesian optimal cue integration [35]. Since fusion is performed only at the output level, this strategy offers strong compatibility with heterogeneous networks, but lacks modality interaction and therefore may limit recognition performance.

3) *Model-level fusion*: This strategy focuses on optimizing the network architecture to enable cross-modal interaction at multiple stages. Owing to their dynamic weighting capability and structural flexibility, attention mechanisms have become a main approach for model-level fusion. For instance, Cheng et al. [36] proposed a dynamic interactive network with self-distillation for MER, in which facial and EEG features are fused through intra- and inter-modal Transformers. Zhang et

al. [37] applied cross-modal attention to facilitate interaction between visual and audio features, and further designed a token-based similarity adjustment module. Despite their effectiveness, these methods typically update mid features through addition or concatenation, which may insufficiently account for distributional differences across modalities and negatively impact subsequent interactions and prediction.

Disentangled representation learning [38] also has attracted growing interest in multimodal tasks. The main concept is to separate features into distinct semantic subspaces, thereby generating structured and interpretable representations for each modality. A key contribution is MISA by Hazarika et al. [39], which projects features from visual, acoustic, and textual modalities into modality-invariant and modality-specific subspaces, guided by similarity, difference, and reconstruction losses. Building on this idea, Yang et al. [40] introduced modality discriminators to guide the learning of invariant and specific encoders in an adversarial manner, while Zhou et al. [41] further decomposed modality-specific features into effective and ineffective components, discarding the latter to improve prediction. However, existing methods mainly emphasize global semantic separation across modalities, while overlooking the temporal distribution differences of emotion-relevant cues across frames. Treating all frames equally may reduce the effectiveness of disentangled representations in sequential multi-frame tasks.

C. Driver Emotion Dataset

Rich and reliable public datasets have established standardized benchmarks for driver emotion recognition. From the perspective of data sources, current driver emotion datasets can generally be divided into three categories:

1) *Generic facial expression datasets*: Considering the important role of facial expression in driver emotion cues,

some studies adopt generic facial datasets recorded in daily life as one of their benchmarks. This group is the most abundant, including static datasets such as FER-2013 [42], AffectNet [43], RAF-DB [44], as well as dynamic datasets like CK+ [45], DEAP [46], DFEW [47]. However, due to the lack of task-related conditions, the expressions contained in these datasets differ markedly from those arising in driving contexts [48], limiting their use to supplementary evaluation.

2) *Simulated driver emotion datasets*: Simulators offer controllable conditions and high reproducibility, making them become a popular platform for driver emotion research. They also allow safe deployment of various sensors to capture multimodal signals. For example, datasets like PPB-Emo [49], manD 1.0 [50], MMDE [51] include facial, physiological, and behavioral signals, whereas others such as DEFE [48] and MLI-DER [20] focus primarily on facial data. Nevertheless, simulators lack the real risk and lighting variation present in actual driving, which restricts their ecological validity.

3) *Naturalistic driver emotion datasets*: Unlike simulators, naturalistic driving emotions arise organically from real interactions inside and outside the cockpit, without relying on stimulus materials like videos or texts. This setup provides the highest ecological validity and is widely regarded as the gold standard for emotion recognition method. Early datasets such as MA-Dataset [52] and KMU-FED [53] focused mainly on facial expressions. Given the importance of driving context, recent datasets like AIDE [54] and RoadEmo [30] have recorded both driver facial and road traffic data. However, existing datasets still face some issues, such as poor in-cabin views and a lack of diverse traffic events. Accordingly, we introduce the Scenario-Emo dataset, a new resource collected under naturalistic driving conditions, to enable more accurate and generalizable assessment of emotion recognition methods.

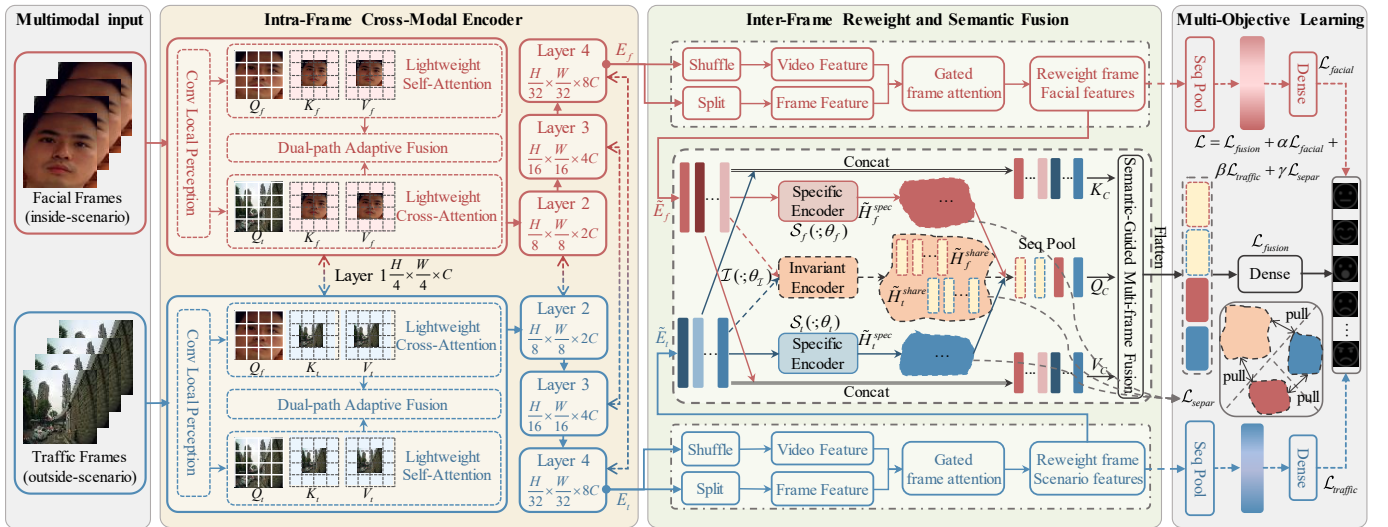


Fig. 2. Overview of Traffic-Facial Intra-Inter Frame Fusion Network (TFI²F-Net). Given synchronized facial and traffic input frames, TFI²F-Net first performs single-frame encoding through four hybrid layers including local convolution and dual-path attention modules. The resulting frame-level features are then temporally reweighted and further disentangled into modality-invariant and modality-specific representations to guide multi-frame aggregation. The network is optimized by a joint loss consisting of fusion prediction, unimodal supervision, and semantic separation regularization. The convolutional stem for low-level feature extraction is omitted here, and the base channel number C is 64.

III. METHODOLOGY

A. Network Overview

TFI²F-Net is designed to progressively capture multi-level interactions between facial and traffic modalities through two phases: single-frame encoding and multi-frame aggregation. The overall architecture is illustrated in Fig. 2. To achieve this, we introduce three key stages: 1) *Intra-Frame Cross-Modal Encoder (IFCME)*. This stage focuses on hierarchical interactions between the two modalities within a single frame. Built upon local feature extraction, it employs dual-path lightweight attention to capture intra-modal and cross-modal interactions, followed by an adaptive fusion module to dynamically integrate the two paths and enhance intra-frame consistency. 2) *Inter-Frame Reweighted Semantic Fusion (IFRSF)*. This stage aims to aggregate multi-frame features from IFCME through informative frame reweighting and semantic-guided cross-modal fusion. Specifically, frame shuffling and gating are used to emphasize salient frames, while modality-invariant and modality-specific semantics are further exploited to construct a structured semantic query for multi-frame aggregation. 3) *Multi-Objective Learning (MOL)*. This stage introduces a multi-objective joint loss combining fusion prediction loss, modality auxiliary loss, and semantic separation loss. This design promotes cross-modal consistency while preserving modality-specific information. The following sections describe the proposed TFI²F-Net in detail.

B. Task Definition

Given a multi-modal driver emotion dataset of N samples, denoted as $\{X_m \in \mathbb{R}^{N_m \times 3 \times T \times H \times W}\}$, where m denoted the numbers of modalities, each sample in the m -th modality contains T RGB frames of size $H \times W$. The target set Y is defined as $Y = \{Y_i\}_{i=1}^N, Y_i \in \{0, 1, \dots, C\}$, where Y_i represent the i -th emotion label and $|C|$ is the number of emotion categories. The goal of TFI²F-Net is to predict the driver emotional categories using facial data X_f and traffic data X_t .

C. Intra-Frame Cross-Modal Encoder

The IFCME serves as a cross-modal frame encoder, designed to perform hierarchical modeling and information interaction of frame features from two different modalities. For each input frame sample $x^m \in \mathbb{R}^{3 \times H \times W}$ from modality m , we first apply a ResNet-style convolutional stem [55] as early feature extractor to capture low-level visual patterns:

$$x_m^{(0)} = \text{Maxpool}(\text{ReLU}(\text{BN}(\text{Conv2d}(x^m)))) \quad (1)$$

Where Conv2d denotes 2D convolution with 7×7 kernel size, BN refers to batch normalization, ReLU is the activation function, and Maxpool represents the max-pooling layer.

Next, the early feature $x_m^{(0)}$ are processed through four layers L_i , each consisting of three modules: Convolutional Local Perception (CLP), Dual-Path Lightweight Attention (DPLA), and Dual-Path Adaptive Fusion (DPAF).

1) *Convolutional Local Perception*: Convolutional structure is well suited for image modeling due to its local connectivity

and translation-invariant property [55], [56]. Based on this, we design a Convolutional Local Perception (CLP) module composed of two stacked convolutional blocks. Given the input $x_m^{(i-1)}$ of layer L_{i-1} , the module produces the initial feature representation $c_m^{(i)}$ as:

$$c_m^{(i)} = \mathcal{B}^{(2)}(\mathcal{B}^{(1)}(x_m^{(i-1)})) \quad (2)$$

$$\mathcal{B}^{(k)}(x) = x + \mathcal{F}^{(k)}(x), k \in \{1, 2\} \quad (3)$$

Where $\mathcal{F}^{(k)}$ denotes the convolutional block. Each block follows the design of ResNet18 [55], consisting of two 3×3 Conv2d, two BN layers, and one ReLU activation. Except for the first layer, the CLP module doubles the channel dimension and halves the spatial resolution.

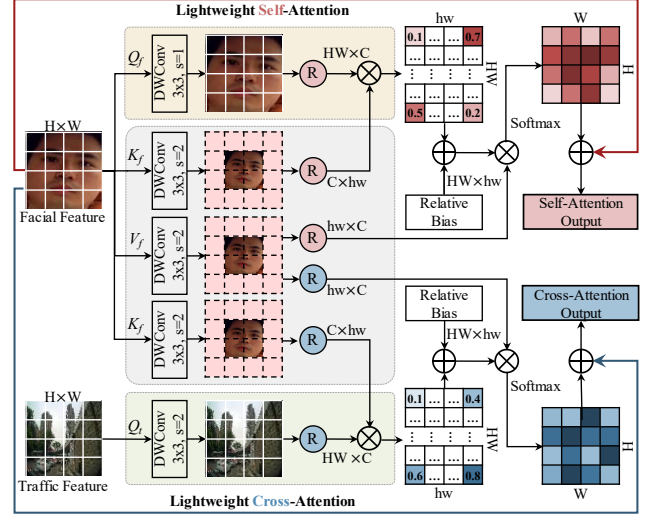


Fig. 3. DPLA module in the facial modality encoder ($s_f = 2$), where $\otimes$ denotes matrix multiplication, $\oplus$ denotes element-wise addition, and $\textcircled{R}$ denotes the reshape operation.

2) *Dual-Path Lightweight Attention*: Attention methods have been widely used in multimodal tasks [22], [36], [37], [39], primarily due to their dot-product similarity-based matching mechanism. To enhance cross-modal semantic collaboration during frame encoding, we design a DPLA module composed of self-attention and cross-attention. The self-attention branch captures intra-modal dependencies, while the cross-attention branch models inter-modal correlations, as illustrated in Fig. 3. Taking the facial modality encoder as an example, this process can be described as follows:

$$A_{f \rightarrow f}^{(i)} = \text{LightAttn}(Q_f^{(i)}, K_f^{(i)}, V_f^{(i)}) + c_f^{(i)} \quad (4)$$

$$A_{t \rightarrow f}^{(i)} = \text{LightAttn}(Q_t^{(i)}, K_f^{(i)}, V_f^{(i)}) + c_f^t \quad (5)$$

Where $A_{f \rightarrow f}^{(i)}$ and $A_{t \rightarrow f}^{(i)}$ represent the lightweight attention outputs in layer L_i for intra-facial and traffic-to-facial interactions, respectively. $Q_m^{(i)}$, $K_m^{(i)}$ and $V_m^{(i)}$ denote the query, key, and value vectors of modality m .

The original self-attention has quadratic complexity [57], which limits its scalability in multi-path attention framework. To address this, we adopt depthwise separable convolutions with a 3×3 kernel and stride s_i as the projection functions for K and V , reducing the attention cost by a factor of s^2 . Moreover, relative positional bias is added to retain spatial

information. Given the CLP output c_i^m , the lightweight attention computation is formulated as:

$$Q_m^{(i)} = \text{DWConv}_{3 \times 3, 1}^Q(\text{BN}(c_m^{(i)})) \quad (6)$$

$$K_m^{(i)} = \text{DWConv}_{3 \times 3, 5}^K(\text{BN}(c_m^{(i)})) \quad (7)$$

$$V_m^{(i)} = \text{DWConv}_{3 \times 3, 5}^V(\text{BN}(c_m^{(i)})) \quad (8)$$

$$\text{LightAttn}(Q_m^{(i)}, K_m^{(i)}, V_m^{(i)}) = \text{Softmax}\left(\frac{Q_m^{(i)} K_m^{(i)T}}{\sqrt{d_k}} + B^{(i)}\right) V_m^{(i)} \quad (9)$$

Where $B^{(i)} \in \mathbb{R}^{N_Q \times N_K}$ denotes a learnable relative positional bias, which is randomly initialized. The stride factors s_1 to s_4 are set to 8, 4, 2, and 1, respectively.

To enhance attention diversity, we implement both self-attention and cross-attention path in a multi-head form. Specifically, $Q_m^{(i)}$, $K_m^{(i)}$ and $V_m^{(i)}$ are split along the channel dimension into h subspaces, and each head computes lightweight attention independently. The head outputs are then concatenated and projected by a linear layer W_o to form the final output. This process can be outlined as:

$$\text{MHLA}(c_m^{(i)}) = \text{Concat}(LA^{(i,1)}, \dots, LA^{(i,h)}) W_o^{(i)} \quad (10)$$

3) *Dual-Path Adaptive Fusion*: Within the facial encoder, $A_{f \rightarrow f}^{(i)}$ and $A_{t \rightarrow f}^{(i)}$ share the same keys and values but use different queries, resulting in a distributional discrepancy between them. To balance these path differences, we propose an input-guided dual-path adaptive fusion module (DPAF), inspired by SKNet [56], as illustrated in Fig. 4. Specifically, we first integrate $A_{f \rightarrow f}^{(i)}$ and $A_{t \rightarrow f}^{(i)}$ into a global feature:

$$A_f^{(i)} = A_{f \rightarrow f}^{(i)} + A_{t \rightarrow f}^{(i)} \quad (11)$$

Next, $A_f^{(i)}$ is compressed by global average pooling GAP into a compact attention vector. To further improve nonlinear representation while maintaining efficiency, we design a bottleneck projection network $\mathcal{P}_{FC}$ composed of two linear layers to generate a doubled channel weighting feature $Z_f^{(i)} \in \mathbb{R}^{2C}$. This process is formulated as:

$$Z_f^{(i)} = \mathcal{P}_{FC}(\text{GAP}(A_f^{(i)})) \quad (12)$$

$$\mathcal{P}_{FC}(x) = W_{\text{upgrade}}(\text{ReLU}(\text{BN}(W_{\text{reduce}}x))) \quad (13)$$

Here, $W_{\text{reduce}} \in \mathbb{R}^{d \times C}$ and $W_{\text{upgrade}} \in \mathbb{R}^{2C \times d}$ are linear layers with output dimensions d and $2C$, respectively. In the proposed TFI²F-Net, d is fixed at 32 to matching the attention head.

Subsequently, we reshape the feature vector $Z_f^{(i)}$ into two sets of channel-wise weights $Z_f = \{Z_{f \rightarrow f}^{(i)}, Z_{t \rightarrow f}^{(i)}\} \in \mathbb{R}^{2 \times C \times 1 \times 1}$, and softmax operation is used to ensure that the weights of each channel satisfy the normalization constraint across two paths:

$$P_{j,c}^{(i)} = \frac{\exp(Z_{j,c}^{(i)})}{\sum_{k=1}^2 \exp(Z_{k,c}^{(i)})} \quad (14)$$

Where j denotes the two attention paths of $f \rightarrow f$ and $t \rightarrow f$, and $P_{j,c}^{(i)}$ represents the weight of channel c in path j . The final output is computed by applying channel-wise weighted summation of two attention maps:

$$x_f^{(i)} = P_{f \rightarrow f}^{(i)} \odot A_{f \rightarrow f}^{(i)} + P_{t \rightarrow f}^{(i)} \odot A_{t \rightarrow f}^{(i)} \quad (15)$$

$$x_t^{(i)} = P_{t \rightarrow t}^{(i)} \odot A_{t \rightarrow t}^{(i)} + P_{f \rightarrow t}^{(i)} \odot A_{f \rightarrow t}^{(i)} \quad (16)$$

Here, $x_f^{(i)}$ represents the output of the i -th layer of the facial encoder, $\odot$ denotes element-wise multiplication. Similarly, the output of the traffic encoder is denoted as $x_t^{(i)}$.

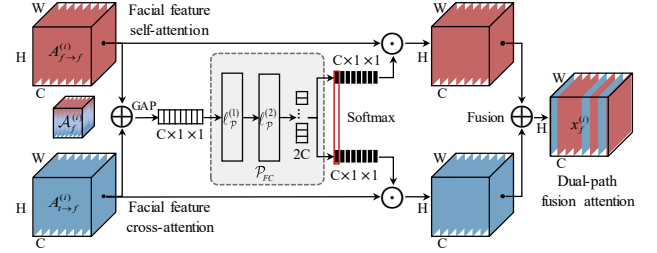


Fig. 4. DPAF module in the facial modality encoder, where $\odot$ denotes element-wise multiplication, $\oplus$ denotes element-wise addition.

D. Inter-Frame Reweighted Semantic Fusion

The IFRSF stage focuses on effectively integrating multi-frame features from both modalities and consists of three modules: Intra-Modal Multi-Frame Reweighting (IMMFR), Inter-Modal Semantic-Guided Fusion (IMSGF), and Multi-Modal Fusion Decision (MMFD).

1) *Intra-Modal Multi-Frame Reweight*: In dynamic driving scenarios, different frames contribute unequally to emotion recognition, with some frames containing more salient cues such as a vehicle suddenly cutting in or stopping. Therefore, IMMFR is introduced to adaptively highlight key frames while suppressing redundant frames based on global video context, as outlined in Fig. 5. This strategy helps generate more stable and discriminative temporal representations to support subsequent multimodal semantic disentanglement.

In detail, we first arrange the frame encoding features E_m^l produced by IFCME in temporal order to form a sequence $E_m = [E_m^1, \dots, E_m^L] \in \mathbb{R}^{L \times C}$, where m indicates the modality. Then, a shuffle operation is applied to rearrange the frame-wise features into a channel-wise format, so that the responses of each channel across all frames are grouped into a cross-frame vector. A channel-wise grouped projection is subsequently used to aggregate these vectors and generate a global video-level feature $G_m \in \mathbb{R}^C$. This process can be formulated as:

$$G_m = \mathcal{W}(\mathcal{S}(E_m)) \quad (17)$$

$$\mathcal{W}(x_c) = (\omega^{(c)})^T x_c + b^{(c)} \quad (18)$$

Here, $\mathcal{S}$ denotes the channel shuffle operation, and $\mathcal{W}$ represents the channel-wise grouped projection. Each channel $x_c \in \mathbb{R}^L$ corresponds to a cross-frame vector formed by the c -th channel across all L frames.

Subsequently, each frame feature E_m^l is concatenated with the global video feature G_m and fed into a frame gating $\mathcal{G}_{FC}$, which consists of two linear layers and a sigmoid activation, to generate the frame weight W_m^l . Finally, element-wise multiplication is performed between the frame weight and the corresponding frame vector, formulated as:

$$W_m^l = \mathcal{G}_{FC}(\text{Concat}(E_m^l, G_m)) \quad (19)$$

$$\mathcal{G}_{FC}(x) = \sigma(W_2(\text{ReLU}(\text{BN}(W_1x)))) \quad (20)$$

$$\tilde{E}_m^l = W_m^l \odot X_m^l, l \in \{1, \dots, L\} \quad (21)$$

Where $W_1 \in \mathbb{R}^{d \times 2C}$ and $W_2 \in \mathbb{R}^{1 \times d}$ are linear layers with output dimensions d and 1, respectively. $\tilde{E}_m^l \in \mathbb{R}^C$ represents the updated l -th frame feature vectors. The reweighted temporal vector is finally formed as $\tilde{E}_m = \{\tilde{E}_m^1, \dots, \tilde{E}_m^L\} \in \mathbb{R}^{L \times C}$.

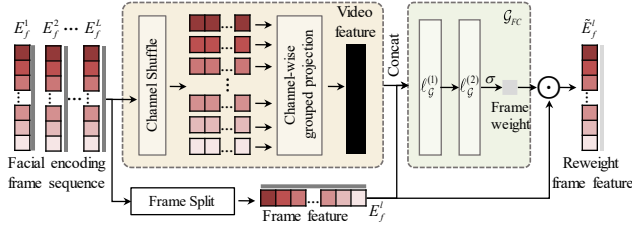


Fig. 5. IMMFR module for the facial encoding feature sequence, where $\odot$ denotes element-wise multiplication. Each frame is reweighted using this process.

2) *Inter-Modal Semantic-Guided Fusion*: Although facial and traffic features have interacted in the encoding stage, they still differ significantly in information source and semantic role. Facial features mainly characterize internal affective expressions, whereas traffic features provide external contextual cues for emotion elicitation. Therefore, directly concatenating multi-frame features as the fused representation may lead to information confusion and conflict. Inspired by MISA [39], we introduce the MSGF module to disentangle modality-invariant and modality-specific semantics, thereby guiding the structured fusion of multi-frame reweighted features of the two modalities, as shown in Figure 1.

Specifically, for the reweighted facial vector $\tilde{E}_f$ and traffic vector $\tilde{E}_t$, we first apply three parallel linear encoders to disentangle modality-invariant and modality-specific semantic vectors, formulated as:

$$\begin{aligned} \tilde{H}_f^{invar} &= \mathcal{I}(\tilde{E}_f; \theta_x) \quad , \quad \tilde{H}_t^{invar} = \mathcal{I}(\tilde{E}_t; \theta_x) \\ \tilde{H}_f^{spec} &= \mathcal{S}_f(\tilde{E}_f; \theta_f) \quad , \quad \tilde{H}_t^{spec} = \mathcal{S}_t(\tilde{E}_t; \theta_t) \end{aligned} \quad (22)$$

Here, $\tilde{H}_{f/t}^{invar/spec}$ represents the modality semantic vector from four different perspectives. The invariant encoder $\mathcal{I}(\cdot; \theta_x)$ shares the parameters θ_x across all modalities, and the specific encoder $\mathcal{S}_m(\cdot; \theta_m)$ learn the specific parameters θ_m for each modality. Each encoder contains a linear projection and a sigmoid activation.

Next, the four semantic vectors are averaged along the temporal dimension and concatenated to form a cross-modal query $Q_C \in \mathbb{R}^{4 \times d}$. Compared with temporal mean queries, disentangled semantic queries provide a clearer and more structured retrieval condition for attention. The reweighted facial and scenario temporal features are concatenated as the unified key $K_C \in \mathbb{R}^{2L \times d}$ and value $V_C \in \mathbb{R}^{2L \times d}$, and multi-head attention is then applied to perform semantic-guided multi-frame fusion, formulated as:

$$Q_C = [\mathcal{AP}(\tilde{H}_f^{invar}), \mathcal{AP}(\tilde{H}_t^{invar}), \mathcal{AP}(\tilde{H}_f^{spec}), \mathcal{AP}(\tilde{H}_t^{spec})] \quad (23)$$

$$K_C = V_C = [\tilde{E}_f, \tilde{E}_t] \quad (24)$$

$$O_{fusion} = \text{MHA}(Q_C, K_C, V_C) \quad (25)$$

Here, $\mathcal{AP}$ denotes average pooling along the temporal dimension, and $\text{MHA}(\cdot)$ refers to the multi-head attention, $O_{fusion} \in \mathbb{R}^{4 \times d}$ is the multi-frame fusion representation.

3) *Multi Modal Fusion Decision*: In the final decision module, we first flatten the fused semantic representation $O_{fusion} \in \mathbb{R}^{4 \times d}$ into a $4d$ dimensional vector. The vector is then linearly projected into the target label space for final discrete emotion classification, formulated as:

$$\hat{y}_{fusion} = \text{FC}(\text{Flatten}(O_{fusion})) \quad (26)$$

Where $\hat{y}_{fusion}$ is the predicted emotion label, FC denotes the fully connected layer.

E. Multi-Objective Learning

The MOL stage is designed to establish a joint loss function that effectively supports the continuous optimization of TFI²F-Net. In addition to the primary loss based on the fused representation, we further introduce an Auxiliary Unimodal Loss and a Modality Separation Loss. The following sections provide details of these loss functions.

1) *Auxiliary Unimodal Loss*: For the driver emotion recognition task, the dependency on specific modalities varies across emotions. For example, happiness is often visibly reflected in facial expressions, whereas anxiety is more closely associated with constrained environments. Based on this observation, we design two auxiliary unimodal loss that leverages both facial and traffic encoders to enhance the discriminative ability of each modality.

Specifically, we apply temporal average pooling to the reweighted features $\tilde{E}_f$ and $\tilde{E}_t$, followed by two separate linear layers for discrete emotion prediction. The process is formulated as follows:

$$\hat{y}_{facial} = \text{FC}_f(\mathcal{AP}(\tilde{E}_f)), \hat{y}_{traffic} = \text{FC}_t(\mathcal{AP}(\tilde{E}_t)) \quad (27)$$

$$\mathcal{L}_{facial} = \text{CE}(\hat{y}_{facial}, y), \mathcal{L}_{traffic} = \text{CE}(\hat{y}_{traffic}, y) \quad (28)$$

Where $\hat{y}_{facial}$ and $\hat{y}_{traffic}$ are emotion prediction from the facial and traffic modality, with $\mathcal{L}_{facial}$ and $\mathcal{L}_{traffic}$ denoting their corresponding auxiliary losses computed using the cross-entropy function (CE).

2) *Modality Separation Loss*: The semantic-guided multi-frame fusion serves as a core component for generating multimodal representations. This process relies on the effective separation between modality-invariant and modality-specific semantic vectors in the representation space. To achieve this, we design a modality separation loss based on a soft orthogonality constraint.

In detail, the loss minimizes the squared Frobenius norms between invariant and specific semantics within each modality, as well as between specific semantics across modalities. This encourages the four types of semantic representations to be mapped into mutually independent subspaces, which is formulated as:

$$\mathcal{L}_{separ} = \|\tilde{H}_f^{invarT} \tilde{H}_f^{spec}\|_F^2 + \|\tilde{H}_t^{invarT} \tilde{H}_t^{spec}\|_F^2 + \|\tilde{H}_f^{specT} \tilde{H}_t^{spec}\|_F^2 \quad (29)$$

Where $\|\cdot\|_F^2$ represents the squared Frobenius norms, and $\mathcal{L}_{separ}$ denotes the modality separation loss.

3) *Overall Loss of the TFF-Net*: The overall training objective of the model is achieved by minimizing the following joint loss function:

$$\mathcal{L} = \mathcal{L}_{fusion} + \alpha\mathcal{L}_{facial} + \beta\mathcal{L}_{traffic} + \gamma\mathcal{L}_{separ} \quad (30)$$

Here, the multimodal fusion loss $\mathcal{L}_{fusion} = \text{CE}(\hat{y}_{fusion}, y)$ serves as the primary optimization target with its weight fixed at 1. The coefficients α , β , and γ control the relative contributions of each regularization component to the total loss $\mathcal{L}$.

IV. SCENARIO-EMO DATASET COLLECTION

A. Participants and Ethics Statement

Seven drivers with distinct styles participated in naturalistic emotion data collection, including five males and two females. Due to equipment failure, one male participant's data was incomplete. As a result, the final Scenario-Emo dataset contains outside-view and inside-view video data from six valid participants. All participants had verified driving experience to ensure data quality. Specifically, their ages ranged from 24 to 30 years (Mean = 27.5 years), and their driving experience spanned 5 to 9 years (Mean = 6.7 years). The experimental procedures were approved by the ethics review board of the College of Mechanical and Vehicle Engineering, Chongqing University. Each participant signed a GDPR informed consent which allows the dataset to be publicly available for research purposes.

TABLE II

SCENARIO-EMO DATASET CONTENT SUMMARY

Item	Content
No. of participants	6 (4 males, 2 females)
Age of participants	24-30 years (mean= 27.5 years)
Emotion Elicitation	In- and Out-of-Cockpit Interaction under Naturalistic Driving
Number of samples	6900
Emotion clips	Neutral (3000), Happiness (1500), Anxiety (900), Anger (1500)
Clips content	Outside-view and inside-view clips
Clips duration	3 s
Clips frame rate	30 fps
Clips resolution	2560×1440
Annotation strategy	Three-stage collaborative labeling protocol
Stage 1 sample ratio	Happiness (0.85), Anxiety (0.75), Anger (0.82)
Stage 2 sample ratio	Happiness (1.03), Anxiety (1.12), Anger (1.07)

Note: The Stage 1/2 sample ratio denotes the proportion of emotion samples retained at each annotation stage relative to the final sample set.

B. Data Collection Platform

The data collection platform comprised an SUV equipped with an automatic transmission and two Hikvision dash cameras with 2560×1440 resolution. In particular, the cameras were mounted at the front side of the interior rearview mirror and above the dashboard, respectively, to capture the outside-view and inside-view video during driving. Compared to other locations, the dashboard-mounted camera offered better coverage of facial expressions and aligned more closely with the camera placement in commercial driver monitoring systems. Both cameras recorded at 30 frames per second and were time-synchronized via simultaneous power-up. The setup

of the recording equipment is illustrated in Fig. 6.

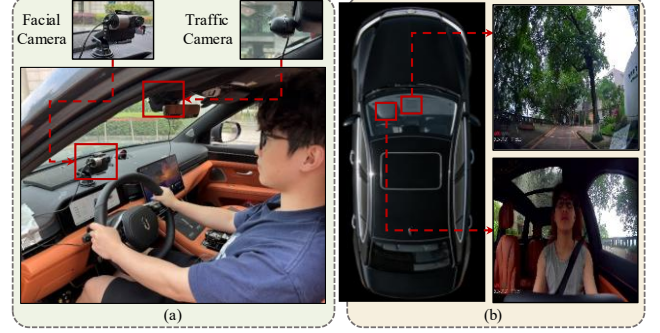


Fig. 6. Data collection platform of the naturalistic scenario driver emotion dataset (Scenario-Emo). (a) The data recording setup, and (b) outside-view and inside-view video acquisition.

C. Naturalistic Driving Data Collection

The experiment was carried out on an almost 60-minute route from Shapingba District to Yuzhong District in Chongqing City. Benefiting from Chongqing's unique topography, the route covered a variety of road scenarios (intersections, pedestrian-dense areas, tunnels of varying lengths, and so on), which helps to present a comprehensive driving context. Additionally, driving tasks were scheduled during two periods: the morning rush (8–9 a.m.) and the evening rush (6–7 p.m.), to increase the density of complex traffic events. Each participant completed the drive independently along a predefined route, while researchers were seated in the rear to monitor data collection. Unlike studies that induce emotions using specific stimuli material, our dataset was derived from participants' natural driving behavior without prior notice.

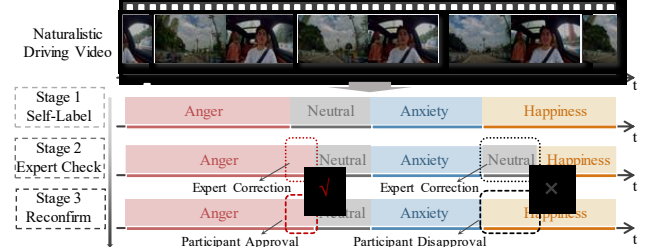


Fig. 7. Three-stage collaborative labeling protocol.

D. Emotion Annotation Process

To enhance the accuracy and consistency of emotion annotations, a three-stage collaborative labeling protocol was adopted, involving both participants and expert researchers. The annotation process is illustrated in Fig. 7. First, each participant reviewed the stitched video combining the outside-view and inside-view from their two driving tasks, and provided preliminary categorical labels by identifying segments where specific emotions occurred. Second, three research experts independently examined the participants' videos and annotations, identifying potential mislabels or omissions through a majority voting strategy. Finally, participants reconfirmed the flagged segments based on expert suggestions to validate or revise the emotion annotations,

thereby forming reliable consensus annotation results.

E. Scenario-Emo Dataset Content

After data quality checks and three-stage collaborative annotation, each emotional segment was divided into 3-second clips, with each clip assigned the label of its corresponding annotated segment. Based on the final annotations, we retained 6,900 primary common road emotion samples from complete data sample. A summary of the dataset is provided in Table II. Anxiety samples exhibit the highest revision ratio (0.37), mainly due to differences in driving style or skill among participants when appraising traffic. Emotion with extremely low ratio (e.g., surprise, sadness) were excluded. Each sample includes an outside-view clip, an inside-view clip, and the corresponding emotion category. Example frames from the Scenario-Emo dataset are shown in Fig. 8.



Fig. 8. Sample images of the Scenario-Emo dataset.

V. EXPERIMENTS

A. Experimental Dataset

1) AIDE: AIDE [54] is a multimodal, multi-task dataset collected under naturalistic driving conditions. It provides four video views from inside and outside the cockpit, along with annotations for driver emotion, behavior, traffic context, and vehicle state. The dataset includes 2,898 samples, randomly split into training (65%), validation (15%), and testing (20%) sets. In this study, we utilize only front-view and inside-view data, along with five emotion labels (Peace, Anxiety, Weariness, Happiness, Anger). Notably, the inside-view captures the driver's face only from a side perspective, and the emotional annotations were provided by external experts rather than the participants themselves, which may limit its suitability for driver emotion recognition tasks.

2) Scenario-Emo: Scenario-Emo is a two-view naturalistic driving dataset specifically designed for driver emotion research. Details of the data collection and annotation process are provided in Section IV. Unlike AIDE, Scenario-Emo captures inside-view using a dashboard-mounted camera, providing clearer facial recordings better suited for emotion analysis. A visual comparison of the two datasets is shown in Fig. 9. In addition, Scenario-Emo offers a larger number of samples and a higher proportion of non-neutral emotions, enabling more robust evaluation of model generalization. To

ensure consistency across benchmarks, Scenario-Emo is also randomly divided into training (65%), validation (15%), and testing (20%) sets.



Fig. 9. Visual comparison between the AIDE and Scenario-Emo datasets. The Scenario-Emo offers a more favorable viewpoint for capturing driver facial expressions.

B. Data Pre-processing

Data preprocessing consists of two main components: facial expression extraction and data augmentation. For facial expression extraction, we crop facial regions based solely on the landmarks detected by the Multitask Cascade Convolutional Network (MTCNN) [58], without applying additional alignment or geometric transformations. This simple strategy ensures consistent facial feature extraction across the Scenario-Emo and AIDE datasets, enabling fair performance comparison. For data augmentation, we apply three basic transformations to both facial and traffic inputs in both datasets, which include RandomCrop, ColorJitter, and RandomHorizontalFlip. These augmentations help reduce overfitting and improve the robustness of the model.

C. Implementation Details

The model is implemented using the PyTorch framework (version 2.8.0) and trained on an NVIDIA GeForce RTX 5090 GPU with 32 GB of memory. The AdamW optimizer is adopted for network optimization with an initial learning rate of $1e-3$ and a weight decay of $5e-4$. Training is conducted over 100 epochs, with the first 5 epochs used for warm-up. The batch size is set to 24, and the learning rate is scheduled using cosine annealing. For each modality, 16 frames are uniformly sampled from each video and resized to $224 \times 224 \times 3$. To ensure fair comparisons, both the Scenario-Emo and AIDE datasets are evaluated under identical experimental settings.

D. Evaluation Metrics

Given the class imbalance in the datasets, we adopt two widely used metrics in the wild emotion recognition tasks [59], [60]: Unweighted Average Recall (UAR), which is the average of recall across all classes, and Weighted Average Recall (WAR), which corresponds to overall accuracy like Acc. Their definitions are as follows:

$$UAR = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FN_i} \quad (31)$$

$$WAR = \sum_{i=1}^C \frac{N_i}{N} \cdot \frac{TP_i}{TP_i + FN_i} \quad (32)$$

Where C denotes the total number of classes, N represents the total number of samples, and N_i is the number of samples in class i . TP_i and FN_i represent the number of true positives and false negatives for class i , respectively.

VI. RESULTS AND DISCUSSION

A. Comparison with the State-of-the-Art Results

To evaluate the effectiveness of TFI²F-Net, we compare it with several state-of-the-art models on both the Scenario-Emo and AIDE datasets. The comparison results are presented in Table III and Table IV. Following the taxonomy defined in [54], the comparison models are categorized into two types: 2D-pattern and 3D-pattern. The 2D-pattern models extract features using frame encoders and apply 1D convolution layers for temporal modeling across frames. In contrast, the 3D-pattern models incorporate spatiotemporal modeling within their network architectures. For fair comparison, the traffic and facial features produced by each method are fused using the cross attention fusion module proposed in [54] to obtain multimodal representations. Fig. 10 shows the confusion matrices of TFI²F-Net on both datasets.

Table III reports the detailed performance comparison on the Scenario-Emo dataset, including individual accuracy across the four emotion categories and overall metrics. TFI²F-Net achieves a significant performance gain. Specifically, it improves the global metrics of UAR and WAR by 5.87% and

6.16%, respectively, compared to the current best model, while maintaining the lowest parameter count among all 3D models. Furthermore, TFI²F-Net yields the highest accuracy for happiness, anxiety, and anger, surpassing the second-best results by 4.00%, 5.00%, and 7.67%, respectively. These indicate that the intra-inter frame interaction modeling of traffic and facial modalities effectively improves driver emotion recognition, especially for subtle emotions like anxiety and anger, where facial cues are often less distinct.

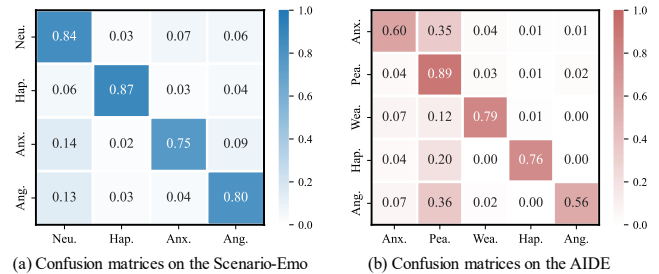


Fig. 10. Confusion Matrices of TFI²F-Net on the (a) Scenario-Emo and (b) AIDE datasets.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SCENARIO-EMO DATASET. BOLD DENOTES THE BEST.
UNDERLINE DENOTES THE SECOND BEST.

Pattern	Method	Neutral	Happiness	Anxiety	Anger	UAR	WAR	Params
2D	Resnet-34 [55]	79.67	82.67	53.89	63.67	69.97	73.48	46.52 M
	Inception-V4 [61]	85.83	72.67	55.00	62.00	68.88	73.77	90.88 M
	MobileNet-V2 [62]	77.50	80.00	67.78	63.33	72.15	73.70	9.71 M
	ShuffleNet-V2 [63]	80.67	83.33	60.00	68.33	73.08	75.87	16.74 M
	CVT-13 [64]	79.67	67.33	55.56	53.00	63.89	68.04	43.58 M
	CMT-XS [65]	86.33	78.00	57.78	51.00	68.28	73.12	33.55 M
	Axia-deeplab [66]	84.67	76.67	63.89	56.67	70.47	74.13	14.84 M
	PartialNet [67]	82.50	78.67	59.44	59.33	69.99	73.62	25.36 M
3D	I3D [68]	85.50	65.33	57.78	35.00	60.90	66.52	57.66 M
	Resnet2p1d [69]	89.50	60.67	50.56	46.33	61.76	68.77	70.79 M
	SlowFast [70]	82.83	67.33	56.67	51.00	64.46	69.13	73.42 M
	TimeSFormer [71]	<u>86.83</u>	58.33	41.11	50.00	59.07	66.67	247.27 M
	Former-DFER [59]	77.00	<u>83.33</u>	<u>70.00</u>	<u>72.33</u>	<u>75.67</u>	<u>76.45</u>	42.19 M
	M3DFEL [60]	75.33	79.00	46.11	60.00	65.11	68.99	98.64 M
	IAL [72]	81.33	69.00	66.11	71.33	71.94	74.49	46.18 M
	LSGTNet [73]	84.00	71.33	49.44	71.67	69.11	74.06	37.38 M
Ours	TFI ² F-Net	83.33	87.33	75.00	80.00	81.54	82.61	36.06 M

Additionally, Table IV presents the performance of different models on the AIDE public dataset in terms of UAR and WAR. Similar to the results observed on the Scenario-Emo dataset, TFI²F-Net achieves the best overall performance, with improvements of 9.46% in UAR and 5.25% in WAR, effectively mitigating the issue of class imbalance in driver emotion samples. Notably, across both datasets, models that combine convolutional and attention-based architectures consistently deliver strong results, such as Former-DFER and Axia-deeplab. This finding highlights the effectiveness of joint local-global modeling for driver emotion recognition and further motivates the encoder design adopted in TFI²F-Net.

Notably, TFI²F-Net achieves higher accuracy than MMTL-UniAD (76.67%) using only two view data, whereas input of MMTL-UniAD relies on four views, four modalities, and four supervision task labels [27].

To further evaluate the performance stability of the proposed TFI²F-Net, we conducted repeated experiments under five different dataset split seeds. Considering the computational cost, ShuffleNet-V2 and Former-DFER were selected as representative 2D- and 3D-pattern comparison methods, respectively. The comparison results are reported as mean $\pm$ standard deviation in Fig. 11.

The results on both the Scenario-Emo and AIDE datasets

demonstrate that TFI²F-Net consistently outperforms the two comparison methods. Specifically, it achieves UAR/WAR scores of 81.03% $\pm$ 1.10% / 82.33% $\pm$ 0.82% on the Scenario-Emo dataset and 71.96% $\pm$ 1.91% / 79.87% $\pm$ 1.38% on the AIDE dataset. Compared with the second-best method, Former-DFER, the proposed method improves UAR by 6.45% and 8.69%, and WAR by 5.97% and 5.00% on the two datasets, respectively. The Wilcoxon signed-rank test further confirms the statistical significance of these improvements.

TABLE IV

OVERALL COMPARISON RESULTS ON THE SCENARIO-EMO AND AIDE DATASET. BOLD DENOTES THE BEST.

UNDERLINE DENOTES THE SECOND BEST

Method	Scenario-Emo		AIDE	
	UAR	WAR	UAR	WAR
Resnet-34	69.97	73.48	55.24	69.62
Inception-V4	68.88	73.77	50.07	66.83
MobileNet-V2	72.15	73.70	58.01	72.91
ShuffleNet-V2	73.08	75.87	62.27	74.06
CVT-13	63.89	68.04	41.99	65.19
CMT-XS	68.28	73.12	58.18	68.47
Axia-deeplab	70.47	74.13	60.00	72.25
PartialNet	69.99	73.62	48.65	69.62
I3D	60.90	66.52	46.37	67.32
Resnet2p1d	61.76	68.77	45.55	67.49
SlowFast	64.46	69.13	48.11	70.11
TimeSFormer	59.07	66.67	50.88	68.47
Former-DFER	75.67	76.45	<u>62.43</u>	75.21
M3DFEL	65.11	68.99	48.40	69.62
IAL	71.94	74.49	56.23	70.94
LSGTNet	69.11	74.06	57.79	71.76
MMTL-UniAD [27]	—	—	—	76.67
TFI ² F-Net (Ours)	81.54	82.61	71.89	80.46

Note: The MMTL-UniAD [27] relies on four views, four modalities, and four supervision task labels, whereas other methods only use front-view and inside-view data.

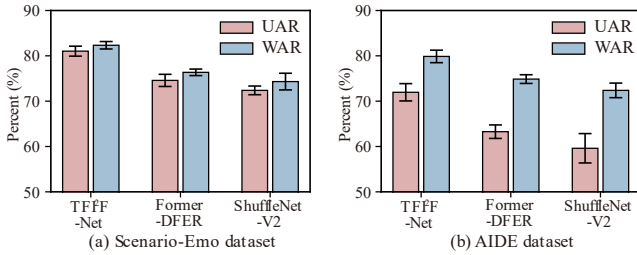


Fig. 11. Comparison of UAR and WAR among different models on the (a) Scenario-Emo and (b) AIDE datasets. Results are reported as mean $\pm$ standard deviation.

B. t-SNE Visual Analysis

To further investigate the discriminative capability of the proposed TFI²F-Net across different emotion categories, we employed t-SNE to visualize the two-dimensional projections of the multimodal fusion representations, with Former-DFER serving as the comparison model. The visualization results of testing set on the Scenario-Emo and AIDE datasets are shown in Fig. 12 and Fig. 13, respectively.

On the Scenario-Emo dataset, both models demonstrate relatively clear separation for the happiness category, which is mainly attributed to its typically obvious changes in facial expressions. However, for neutral, anxiety, and anger categories, Former-DFER exhibits noticeable feature overlap, particularly between anger and the other emotions. In contrast, TFI²F-Net benefits from its stagewise intra- and inter-frame interaction design, which not only reduces the overlap area of anger with other emotions but also enhances the intra-class compactness across all four categories.

On the AIDE dataset, TFI²F-Net consistently demonstrates clearer decision boundaries among different emotions, whereas Former-DFER fails to form consistent clustering patterns. In addition, cross-dataset observations indicate that models trained on the Scenario-Emo dataset generally achieve better separation, which can be attributed to its more optimal camera setup and the three-stage collaborative labeling strategy. In summary, the proposed TFI²F-Net significantly distinguishes between different emotion categories.

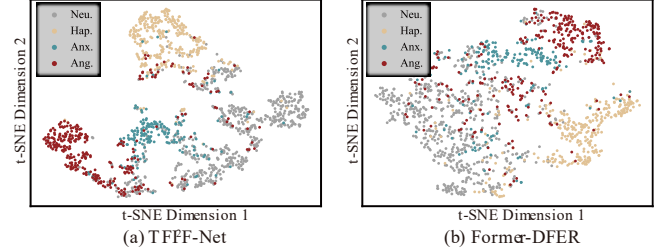


Fig. 12. t-SNE visualization comparing (a) TFI²F-Net with (b) Former-DFER on the Scenario-Emo dataset.

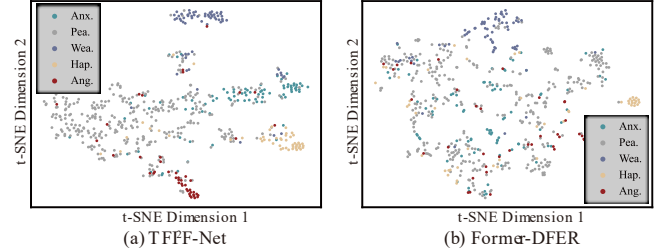


Fig. 13. t-SNE visualization comparing (a) TFI²F-Net with (b) Former-DFER on the AIDE dataset.

C. Ablation Experiment

To evaluate the necessity of incorporating both scenario and facial modalities for driver emotion recognition, we conducted modality ablation experiments on the Scenario-Emo and AIDE datasets. The results are summarized in Table V. Specifically, we retained the original architecture of TFI²F-Net and replaced one modality with random noise to isolate the individual contribution of the other. The results show that removing either modality leads to a significant performance drop: UAR decreased by 9.78% and 8.74%, and WAR dropped by 9.35% and 8.37% on Scenario-Emo and AIDE, respectively. Moreover, the similar performance of each modality in isolation further highlights the critical role of traffic context in recognizing driver emotions task.

However, we also observed modality-specific performance differences across the two datasets. In Scenario-Emo, the

single facial modality achieved higher accuracy, whereas in AIDE, the traffic modality performed better independently. We attribute this difference to the annotation strategy adopted in each dataset. Unlike Scenario-Emo, where labels were provided by the participants themselves, AIDE relies on external annotators, who may have placed greater emphasis on contextual traffic cues during the labeling process.

TABLE V
RESULTS OF MODALITY ABLATION EXPERIMENT

Modality	Scenario-Emo		AIDE	
	UAR	WAR	UAR	WAR
Traffic data	67.81	70.29	63.15	72.09
Facial data	71.76	73.26	57.43	71.10
Traffic-Facial data	81.54	82.61	71.89	80.46

To evaluate the effectiveness of each component in TFI²F-Net, we conducted ablation studies on the Scenario-Emo and AIDE datasets, covering three main stages and six constituent modules. The baseline model uses only convolutional layers as modality encoders, followed by a concatenation strategy to form the multimodal representation. For clarity, the module analysis was performed by incrementally adding each module. For all architectures, the checkpoint with the lowest validation loss was selected for performance evaluation. The results are summarized in Tables VI and VII.

TABLE VI
RESULTS OF THREE STAGES ABLATION EXPERIMENT

Architecture			Scenario-Emo		AIDE	
IFCME	IFRSF	MOL	UAR	WAR	UAR	WAR
×	×	×	69.75	74.06	61.81	73.56
✓	×	×	76.39	76.23	67.64	78.14
×	✓	×	75.64	77.61	62.45	75.21
×	×	✓	73.22	75.43	66.28	75.86
✓	✓	×	79.40	80.22	70.27	78.98
✓	×	✓	76.64	77.39	68.39	77.01
×	✓	✓	77.94	80.80	66.73	76.35
✓	✓	✓	81.54	82.61	71.89	80.46

Note: ✓ and × denote **with** and **without** the corresponding stage, respectively.

TABLE VII
RESULTS OF SIX MODULES ABLATION EXPERIMENT

Architecture	Scenario-Emo		AIDE	
	UAR	WAR	UAR	WAR
Baseline	69.75	74.06	61.81	73.56
+ DPLA module	75.72	75.87	65.64	76.35
+ DPAF module	76.39	76.23	67.64	78.14
+ IMMFR module	78.82	79.57	69.05	79.15
+ MSGF module	79.40	80.22	70.27	78.98
+ Separation loss	81.11	80.58	68.67	78.33
+ Unimodal loss	81.54	82.61	71.89	80.46

The results in Table VI show that removing any stage consistently degrades recognition performance compared with the complete architecture. Specifically, the UAR/WAR drops are 2.14%–11.79%/1.81%–8.55% and 1.62%–10.08%/1.48%–6.90% on the Scenario-Emo and AIDE datasets, respectively.

This degradation is mainly caused by insufficient modality interaction in the alternative components, which weakens the extraction of aligned emotion feature from facial expressions and traffic context. Notably, Table VII shows that the separation loss yields performance gains on the Scenario-Emo but leads to a slight performance drop on the AIDE dataset. This may be attributed to the lower reliability of facial cues and the greater discriminability of the traffic cues in AIDE. Under such conditions, stronger semantic disentanglement may introduce unstable guidance from the weaker facial modality and interfere with traffic-dominant representations. Overall, these results demonstrate the importance of modality interaction in feature encoding and multi-frame aggregation for alleviating semantic heterogeneity and temporal inconsistency.

To investigate the impact of different weight configurations in the joint loss function, we tested seven sets of coefficient combinations for α , β and γ on the Scenario-Emo and AIDE datasets. The comparison results are presented in Table VIII. The results show that the proposed multi-objective learning framework is sensitive to the weighting of individual loss. Improper weight settings can lead to a performance drop of up to 2.41% in UAR and 3.28% in WAR.

TABLE VIII

EXPERIMENT RESULTS UNDER DIFFERENT LOSS WEIGHT. BOLD DENOTES THE BEST. UNDERLINE DENOTES THE SECOND BEST

Loss Weight			Scenario-Emo		AIDE	
α	β	γ	UAR	WAR	UAR	WAR
0.8	0.8	0.8	79.24	80.51	69.48	77.34
0.8	0.8	0.5	79.44	80.22	70.44	77.18
0.8	0.5	0.5	80.54	<u>82.03</u>	70.86	78.97
0.5	0.5	0.5	81.54	82.61	71.89	<u>80.46</u>
0.5	0.5	0.2	<u>81.15</u>	81.45	<u>71.69</u>	80.79
0.5	0.2	0.2	80.31	80.80	70.82	79.47
0.2	0.5	0.2	79.36	81.59	71.53	79.15

Moreover, when the coefficients α , β , and γ are all set to 0.5, the model achieves the highest UAR on both datasets. This result is intuitively reasonable: the auxiliary unimodal losses are designed to enhance the independent discriminative ability of each modality. Assigning them moderate weights helps maintain a balance between unimodal supervision and multimodal integration. Meanwhile, at this setting, the separation loss provides a controlled structural constraint that facilitates the disentanglement of modality-invariant and modality-specific semantics without excessively restricting representation learning.

D. Model variant comparison

To evaluate the contribution of each key module design, we develop two variants of the dual-path attention fusion module, two variants of the multi-frame temporal module, and two variants of the multimodal fusion module, and conduct comparative experiments on the Scenario-Emo and AIDE datasets. Specifically, the attention variants include Element-wise Addition and Recursive fusion (self-attention followed by cross-attention), the multi-frame variants include LSTM

and 1D-CNN, while the multimodal fusion variants adopt Cross-attention (CA) and MISA fusion, respectively.

Table IX presents the comparison results. Overall, all variants led to decreased performance in emotion recognition, highlighting the clear advantage of the proposed DPAF, IMMFR and IMSGF modules in the traffic-facial modality fusion model. Notably, replacing the IMMFR module with LSTM or 1D-CNN resulted in larger performance drops, with UAR decreases of up to 3.16% and 4.44% on the Scenario-Emo and AIDE datasets, respectively. These findings suggest that traffic and facial cues under naturalistic driving may have weak temporal continuity, where key frame reweighting is more effective than full-sequence modeling.

TABLE IX

EXPERIMENT RESULTS UNDER MODEL VARIANT. BOLD DENOTES THE BEST. UNDERLINE DENOTES THE SECOND BEST

Variant type	Scenario-Emo		AIDE	
	UAR	WAR	UAR	WAR
Attention variant - Addition	79.18	80.58	68.02	79.15
Attention variant - Recursive	<u>80.24</u>	81.09	69.18	78.65
Attention model - DPAF (ours)	81.54	82.61	71.89	80.46
Frame variant - LSTM	78.38	78.70	67.69	77.83
Frame variant - 1DCNN	78.69	79.20	67.45	78.16
Frame model - IMMFR (ours)	81.54	82.61	71.89	80.46
Fusion variant - CA	79.06	<u>81.45</u>	68.81	<u>79.97</u>
Fusion variant - MISA	79.68	81.01	<u>69.73</u>	77.18
Fusion model - IMSGF (ours)	81.54	82.61	71.89	80.46

E. Typical failure case analysis

To examine the applicability and limitations of TFI²F-Net, we visualize representative anger sample cases in Fig. 14, given the safety relevance of anger in driving. Specifically, Fig. 14(a) and Fig. 14(b) show correct predictions, whereas Fig. 14(c)-Fig. 14(f) present misclassified cases with representative facial and traffic frames.

As shown in Fig. 14(a), TFI²F-Net can recognize anger even when facial expressions are subtle, suggesting that the proposed multi-stage fusion strategy effectively leverages facial and traffic cues. Nevertheless, several failure cases reveal remaining challenges. In Fig. 14(c), anger is triggered by verbal recollection, while the open traffic scene and facial movements lead to misclassification. In Fig. 14(d), anger is triggered by a vehicle changing lanes on the left, but the absence of salient facial changes causes the model to predict a neutral state. In Fig. 14(e), the complex traffic scene biases the model toward anxiety, although the speech record indicates anger. In Fig. 14(f), a slow bicycle induces anger under moderate traffic density, while the low-resolution input limits accurate emotion prediction.

Overall, these failure cases highlight several challenging conditions for naturalistic driver emotion recognition. When emotion-related evidence is weak or ambiguous, the model may be affected by scene contextual associations, subtle facial responses, and insufficient perception of fine-grained traffic participants. These factors can reduce the discriminability of facial and traffic cues, leading to incorrect predictions in

ambiguous scenarios.

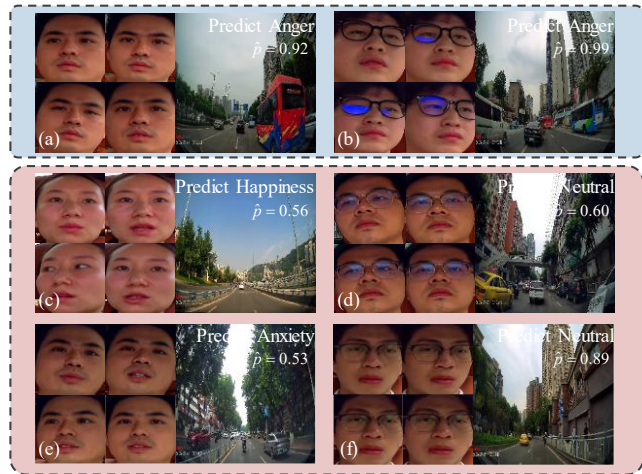


Fig. 14. Visualization of correct and incorrect cases of anger samples. $\hat{p}$ represents the confidence of prediction label.

VII. CONCLUSION

This paper introduces TFI²F-Net, a multimodal framework for driver emotion recognition under naturalistic driving conditions. Specifically, TFI²F-Net progressively models the interaction between facial expressions and traffic context through intra-frame cross-modal encoding and inter-frame reweighted semantic fusion, thereby addressing the challenges of modal semantic heterogeneity and temporal inconsistency. Auxiliary unimodal supervision and semantic separation constraints further improve modality discriminability and fusion stability. Additionally, we also introduce Scenario-Emo, a naturalistic scenario driver emotion dataset including synchronous outside-view and inside-view clips, providing an effective benchmark for this task. Extensive experiments on both the Scenario-Emo and AIDE datasets demonstrate that TFI²F-Net outperforms a range of baselines and state-of-the-art methods. Ablation studies and visual analysis further verify the effectiveness of the proposed modules and highlight the critical role of traffic context in improving recognition accuracy for naturalistic driver emotion. The proposed model and dataset may facilitate future affective interaction systems in intelligent cockpits, supporting improvements in driving safety and user experience.

Although this paper validates the effectiveness of the proposed TFI²F-Net for driver emotion recognition, there are still several limitations. First, TFI²F-Net relies on multi-stage interactive modeling of multimodal video sequences, which requires relatively high computational and storage resources, thus posing challenges for deployment in intelligent cockpits. Second, the failure case analysis reveals the complexity of driver emotions, suggesting that facial expressions and traffic context provide effective but not exhaustive cues for emotion recognition. In future research, we will investigate lightweight multimodal fusion algorithms to better balance recognition accuracy and model complexity. We will also incorporate additional emotion-related cues and explore causal reasoning strategies to enhance the robustness of emotion recognition.

REFERENCES

- [1] M. Jeon, B. N. Walker, and J.-B. Yim, "Effects of specific emotions on subjective judgment, driving performance, and perceived workload," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 24, pp. 197–209, May 2014.
- [2] Q. Zhang, W. Qu, and Y. Ge, "The effect of anger on pedestrian avoidance in a simulated driving task," *Accid. Anal. Prev.*, vol. 171, p. 106664, Jun. 2022.
- [3] W. H. Organization, *Global status report on road safety 2018*. World Health Organization, 2019.
- [4] T. A. Dingus et al., "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 10, pp. 2636–2641, Mar. 2016.
- [5] W. Li et al., "Intelligent cockpit for intelligent connected vehicles: Definition, taxonomy, technology and evaluation," *IEEE Trans. Intell. Veh.*, 2023.
- [6] M. Braun, F. Weber, and F. Alt, "Affective automotive user interfaces—Reviewing the state of driver affect research and emotion regulation in the car," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–26, Sep. 2022.
- [7] M. R. Manavand, M. H. Salarifar, M. Ghavami, and M. Taghipour-Gorjilolaie, "Driver's facial expression recognition by using deep local and global features," *Information Sciences*, vol. 692, p. 121658, Feb. 2025.
- [8] J. Li, L. Yang, C. Lv, Y. Chu, and Y. Liu, "GLF-STAF: A global-local-facial spatio-temporal attention fusion approach for driver emotion recognition," *IEEE Trans. Consumer Electron.*, pp. 1–1, 2025.
- [9] W.-Y. Hsu and T.-H. Chiang, "Cross-attribute feature-perceptive driver emotion recognition in real scenarios," *IEEE Trans. Intell. Transport. Syst.*, pp. 1–12, 2025.
- [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] S.-H. Kim, E. Jung, H. Shin, I.-B. Yang, and J. Woo, "Boosting weak learners with multi-agent reinforcement learning for enhanced stacking models: An application on driver emotion classification," *IEEE Trans. Intell. Transport. Syst.*, pp. 1–15, 2024.
- [12] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1541–1554, 2020.
- [13] Z. Dong, C. Hu, L. Zhu, X. Ji, and C. S. Lai, "A dual-pathway driver emotion classification network using multi-task learning strategy: A joint verification," *IEEE Internet Things J.*, pp. 1–1, 2025.
- [14] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *Journal of nonverbal behavior*, vol. 43, no. 2, pp. 133–160, 2019.
- [15] J. Mesken, M. P. Hagenzieker, T. Rothengatter, and D. De Waard, "Frequency, determinants, and consequences of different drivers' emotions: An on-the-road study using self-reports, (observed) behaviour, and physiology," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 10, no. 6, pp. 458–475, Nov. 2007.
- [16] E. Roidl, B. Frehse, and R. Höger, "Emotional states of drivers and the impact on speed, acceleration and traffic violations—A simulator study," *Accid. Anal. Prev.*, vol. 70, pp. 282–292, Sep. 2014.
- [17] M. Dittrich, "Why drivers feel the way they do: An on-the-road study using self-reports and geo-tagging," in *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Leeds United Kingdom: ACM, Sep. 2021, pp. 116–125.
- [18] A. Alsaid, J. D. Lee, S. I. Noeovich, and A. Chehade, "The effect of vehicle automation styles on drivers' emotional state," *IEEE Trans. Intell. Transport. Syst.*, vol. 24, no. 4, pp. 3963–3973, Apr. 2023.
- [19] G. Xiang et al., "Driver multi-task emotion recognition network based on multi-modal facial video analysis," *Pattern Recognition*, vol. 161, p. 111241, May 2025.
- [20] L. Yang, H. Yang, B.-B. Hu, Y. Wang, and C. Lv, "A robust driver emotion recognition method based on high-purity feature separation," *IEEE Trans. Intell. Transport. Syst.*, vol. 24, no. 12, pp. 15092–15104, Dec. 2023.
- [21] C. Hu et al., "MDEmoNet: A multimodal driver emotion recognition network for smart cockpit," in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA: IEEE, Jan. 2024, pp. 1–6.
- [22] L. Mou et al., "Driver emotion recognition with a hybrid attentional multimodal fusion framework," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2970–2981, Oct. 2023.
- [23] Y.-Z. Wu et al., "AI-enabled intelligent cockpit proactive affective interaction: Middle-level feature fusion dual-branch deep learning network for driver emotion recognition," *Adv. Manuf.*, Sep. 2024.
- [24] 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 J.*, vol. 11, no. 22, pp. 35812–35824, Nov. 2024.
- [25] S. Liu et al., "The empathetic car: Exploring emotion inference via driver behaviour and traffic context," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–34, Sep. 2021.
- [26] W. Liu et al., "UMD-net: A unified multi-task assistive driving network based on multimodal fusion," *IEEE Trans. Intell. Transport. Syst.*, pp. 1–14, 2025.
- [27] W. Liu et al., "MMTL-UniAD: A unified framework for multimodal and multi-task learning in assistive driving perception," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6864–6874.
- [28] D. Wang, S. Jia, X. Pei, C. Han, D. Yao, and D. Liu, "DERNet: Driver emotion recognition using onboard camera," *IEEE Intell. Transport. Syst. Mag.*, vol. 16, no. 2, pp. 117–132, Mar. 2024.
- [29] C. Bustos, A. Sole-Ribalta, N. Elhaoui, J. Borge-Holthoefer, A. Lapedriza, and R. Picard, "Analyzing the visual road scene for driver stress estimation," *IEEE Trans. Affective Comput.*, pp. 1–16, 2025.
- [30] W. Li et al., "Brain-inspired driver emotion detection for intelligent cockpits based on real driving data," *IEEE Intell. Transport. Syst. Mag.*, pp. 2–20, 2024.
- [31] H. Wu, Y. Sun, Y. Yang, and D. F. Wong, "Beyond simple fusion: Adaptive gated fusion for robust multimodal sentiment analysis," 2025, arXiv:2510.01677.
- [32] G. Tang, Y. Xie, K. Li, R. Liang, and L. Zhao, "Multimodal emotion recognition from facial expression and speech based on feature fusion," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16359–16373, 2023.
- [33] M. Yan, Z. Deng, B. He, C. Zou, J. Wu, and Z. Zhu, "Emotion classification with multichannel physiological signals using hybrid feature and adaptive decision fusion," *Biomedical Signal Processing and Control*, vol. 71, p. 103235, Jan. 2022.
- [34] B. T. Atmaja and M. Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 4482–4486.
- [35] Y. Zhao, X. Cao, J. Lin, D. Yu, and X. Cao, "Multimodal affective states recognition based on multiscale CNNs and biologically inspired decision fusion model," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1391–1403, Apr. 2023.
- [36] C. Cheng, W. Liu, X. Wang, L. Feng, and Z. Jia, "DISD-net: A dynamic interactive network with self-distillation for cross-subject multi-modal emotion recognition," *IEEE Trans. Multimedia*, vol. 27, pp. 4643–4655, 2025.
- [37] J. Zhang, Z. Zhang, and Z. Ju, "ATTSTF-net: Attention-based similarity fusion network for audio-visual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 4, pp. 3291–3303, Oct. 2025.
- [38] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9677–9696, Dec. 2024.
- [39] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
- [40] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1642–1651.
- [41] Y. Zhou, X. Liang, H. Chen, Y. Zhao, X. Chen, and L. Yu, "Triple disentangled representation learning for multimodal affective analysis," *Inf. Fusion*, vol. 114, p. 102663, Feb. 2025.
- [42] L. Zahara, P. Musa, E. Prasetyo Wibowo, I. Karim, and S. Bahri Musa, "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based raspberry pi," in *Proc. 5th Int. Conf. Inform. Comput.*, 2020, pp. 1–9.
- [43] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2017.

- [44]S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2584–2593.
- [45]P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Workshops*, 2010, pp. 94–101.
- [46]S. Koelstra et al., "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.
- [47]X. Jiang et al., "Dflew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2881–2889.
- [48]W. Li et al., "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 747–760, Jan. 2023.
- [49]W. Li et al., "A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks," *Sci. Data*, vol. 9, no. 1, p. 481, Aug. 2022.
- [50]K. D. Nobari and T. Bertram, "A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation," *Sci. Data*, vol. 11, no. 1, p. 327, Mar. 2024.
- [51]G. Xiang et al., "A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals," *Eng. Appl. Artif. Intell.*, vol. 130, p. 107772, Apr. 2024.
- [52]Z. Ma, M. Mahmoud, P. Robinson, E. Dias, and L. Skrypchuk, "Automatic detection of a driver's complex mental states," in *Proc. Int. Conf. Comput. Sci. Appl.*, Springer, 2017, pp. 678–691.
- [53]M. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors*, vol. 18, no. 12, p. 4270, Dec. 2018.
- [54]D. Yang et al., "AIDE: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 20402–20413.
- [55]K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
- [56]X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
- [57]A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.
- [58]K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- [59]Z. Zhao and Q. Liu, "Former-DFER: Dynamic facial expression recognition transformer," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1553–1561.
- [60]H. Wang et al., "Rethinking the learning paradigm for dynamic facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17958–17968.
- [61]C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1.
- [62]M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
- [63]N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.
- [64]H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.
- [65]J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12165–12175.
- [66]H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 108–126.
- [67]H. Huang, T. Xia, W. Zhao, and P. Ren, "Partial channel network: Compute fewer, perform better," 2025, arXiv:2502.01303.
- [68]J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.
- [69]D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.
- [70]C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.
- [71]G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," 2021, arXiv:2102.05095.
- [72]H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 1, pp. 67–75.
- [73]L. Wang, X. Kang, F. Ding, S. Nakagawa, and F. Ren, "A joint local spatial and global temporal CNN-transformer for dynamic facial expression recognition," *Appl. Soft. Comput.*, vol. 161, Aug. 2024, Art. no. 111680.



Chengmou Li received the B.Sc. degrees in vehicle engineering from Chongqing University, Chongqing, China, in 2020. He is currently working toward a Ph.D. degree in mechanical engineering with the college of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. His research interests include driver emotion recognition, multimodal learning, driver behavior modeling, and affective interaction in automotive intelligent cockpit.



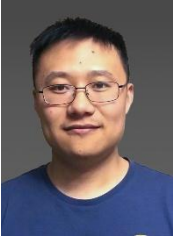
Zongxin Xiao received a B.Sc. degree in vehicle engineering from North University of China, Taiyuan, China, in 2020. She is currently working toward a Ph.D. degree in mechanical engineering with the college of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. Her research interests include new energy vehicles and explainable machine learning.



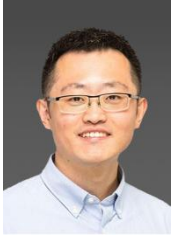
Yutian Zhang received the B.Sc. degree in robotics engineering from Chongqing University in Chongqing, China, in 2024. He is currently pursuing the M.Sc. degree in vehicle engineering with the Same University. His research interests include continual emotion learning and human-machine affective interaction.



Gang Guo received the Ph.D. degree in Mechanical Engineering from Chongqing University, Chongqing, China, in 1994. He is a Professor at the Department of Mechanical and Vehicle Engineering, Chongqing University, Deputy Director of Intelligent Cockpit Working Group of CSAE. He was the Dean of the College of Automotive Engineering at Chongqing University from 2016–2021. His research interests include intelligent cockpit human-machine interaction, user experience, driver behavior and states monitoring, and cockpit software development. He has authored and co-authored over 100 academic papers, published 2 academic books, obtained 8 invention patents, and obtained 12 software copyrights.



Zejian Deng received his Ph.D. degree in Mechanical and Mechatronics Engineering from the University of Waterloo in 2022. He is currently working as a Postdoctoral Fellow in Department of Data and Systems Engineering at The University of Hong Kong. His research interests include driver model, driving safety validations, decision-making in automated driving.



Yang Xing (Senior Member, IEEE) is currently a Senior Lecturer in Applied Artificial Intelligence for Engineering Cranfield University. Before joining Cranfield as Lecturer in 2021, he was a Research Associate with the University of Oxford and a Research Fellow with Nanyang Technological University. His research interests include intelligent multi-agent collaboration, and intelligent/autonomous vehicles. He is a senior member of IEEE and Associate Editor for IEEE Trans. on Intelligent Vehicles, IEEE Trans. on Neural Networks and Learning Systems, and IEEE Trans. on Automation Science and Engineering. He is the Secretary of IEEE VTS AdHoc Committee on Autonomous Vehicles. Currently he leads the HUMAX Lab at Cranfield University, which is focused on the design, testing, and validation of human-centred hybrid autonomous vehicles and transportation systems. His research interest focusses on human-autonomy collaboration and human-behaviours modelling.



Wenbo Li received Ph.D. degree in automotive engineering from Chongqing University, Chongqing, China, in 2021. From 2018 to 2020, he was a Visiting Ph.D. Student with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, Canada. From 2021 to 2023, he was a Postdoctoral Research Fellow at the School of Vehicle and Mobility at Tsinghua University, Beijing, China. He is currently an associate professor at the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. His research interests include intelligent vehicles, intelligent cockpits, human emotion and cognition, driver emotion detection and regulation, human-machine interaction, affective computing, and brain-computer interface.



DSPACE

<https://dspace.org/>

TFI2F-Net: Traffic-Facial Intra-Inter Frame Fusion Network for driver emotion recognition under naturalistic driving

Li, Chengmou; Xiao, Zongxin; Zhang, Yutian; Guo, Gang; Deng, Zejian ...

2026-12-31

Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/taffc.2026.3710335>

Li C, Xiao Z, Zhang Y, et al., (2026) TFI2F-Net: Traffic-Facial Intra-Inter Frame Fusion Network for driver emotion recognition under naturalistic driving. IEEE Transactions on Affective Computing, Available online 6 July 2026