Leveraging Eye Movement for Instructing Robust Video-based Facial Expression Recognition

Yuanyuan Liu, *Member, IEEE,* Lin Wei, Kejun Liu, Zijing Chen, *Member, IEEE,* Zhe Chen*, *Member, IEEE,* Chang Tang, *Senior Member, IEEE,* Jingying Chen, *Member, IEEE,* Shiguang Shan*, *Fellow, IEEE*

***Abstract*—Video-based facial expression recognition (VFER) is challenging due to variations caused by cultural background and expression camouflage. To tackle these problems, researchers introduced eye movement signals to complement visual information. However, existing methods either require expensive devices to capture high-quality eye movements or can only extract low-quality eye movements visually, making them ineffective in the real world. To address this, we propose an eye movement-instructed VFER (EM-VFER) that leverages high-quality eye movements to instruct the visual learning, obtaining robust performance without requiring costly devices during inference. Specifically, our EM-VFER operates in two stages: the high-quality eye movement pre-training stage and the eye movement-instructed video fine-tuning stage. In the pre-training, we compile an Eye-behavior-aided Multimodal Emotion Recognition (EMER) dataset and use it to train a multimodal Transformer. During the fine-tuning, we propose a novel progressive eye movement-instructed learning to take better advantage of the prior knowledge about high-quality eye movement signals from EMER. The instructed fine-tuning model could then make more robust predictions on downstream facial expression datasets. We evaluate our approach on three macro-expression datasets (DFEW, MAFW and Aff-wild2) and two micro-expression datasets (CASME III and CASME II). The results demonstrate that EM-VFER significantly outperforms existing methods. The code will be available.**

***Index Terms*—Video-based facial expression recognition, eye movement signals, pre-training, fine-tuning, instructed learning.**

I. Introduction

Video-based facial expression recognition (VFER) can effectively help understand and interpret emotional states during video-based communications. It involves the recognition and understanding of facial expressions from continuous frame images in a video. Currently, VFER is one of the important topics driving forces for the development of computer vision, emotion computing, and artificial intelligence technology [1]–[13]. An effective VFER model can facilitate many downstream human-computer interaction-based tasks, such as medical assistance and safe driving.

Most existing VFER research has focused on analyzing and recognizing facial expression-related dynamics from video sequences. Techniques such as optical flow analysis [14], keypoint tracking [15], and facial action unit analysis [16] are widely used to detect changes and dynamics in facial features across video frames to identify different facial expression states. Recently, robust sequence modeling models like Transformer networks [17], [18] have been developed for VFER. For instance, Ma *et al.* [17] proposed a unified spatio-temporal Transformer to jointly capture spatial and temporal dependencies, enhancing VFER. Despite the achievement, we found that they focus primarily on facial expression-related motion changes and still struggle to accurately capture weak, fake, or flickering expressions. For example, a fake smile may represent sadness but the smile-related facial movements would make VFER models tend to predict "happy" as the output.In addition to this, individuals from different cultural backgrounds and upbringings have differences in facial expressions, which may interfere with the recognition of real emotions [19]. Without sufficient training data, it would be very challenging to train a Transformer-based VFER to be aware of these nuanced details.

To overcome this problem, some more recent studies [20], [21] have demonstrated that objective physiological behavioral data like eye movements can provide complementary information for facial expression recognition. In the above example of the fake smile, the eyes of the person in the video would be likely to look down, thus incorporating this signal may help generate more accurate predictions about expressions. Some early methods for incorporating eye movement information typically include techniques such as optical flow fields of the eyes [22], which capture changes in the direction and speed of eye movements, or eye localization techniques based on Active Shape Models (ASM) [23], which monitor geometric changes in the eye region. Despite these attempts, these traditional methods may not adequately capture the dynamic features of eye movements in subtle temporal sequences, obtaining relative low-quality eye movement information. To address these shortcomings, recent literature [24], [25] employ specific and expensive sensing devices[1] to focus more on intrinsic high-quality eye movement signals, such as pupil diameter, eye gaze coordinates, and eye behaviors. These signals provide more

Yuanyuan Liu, Lin Wei, Kejun Liu, Chang Tang are with the School of Computer Science, China University of Geosciences (Wuhan), China. E-mail: liuyy, linw, liukejun, tangchang@cug.edu.cn.

Zijing Chen and Zhe Chen are with the School of Computing, Engineering and Mathematical Sciences, La Trobe University, Australia. They are also with the Cisco-La Trobe Centre for Artificial Intelligence and Internet of Things. E-mail: zijing.chen, zhe.chen@latrobe.edu.au.

Jingying Chen is with the National Engineering Laboratory for Educational Big Data and the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China, E-mail: chenjy@mail.ccnu.edu.cn.

Shiguang Shan is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, CAS, Beijing 100190, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: sgshan@ict.ac.cn.

*Corresponding author: Zhe Chen, Shiguang Shan

[1]https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3599859

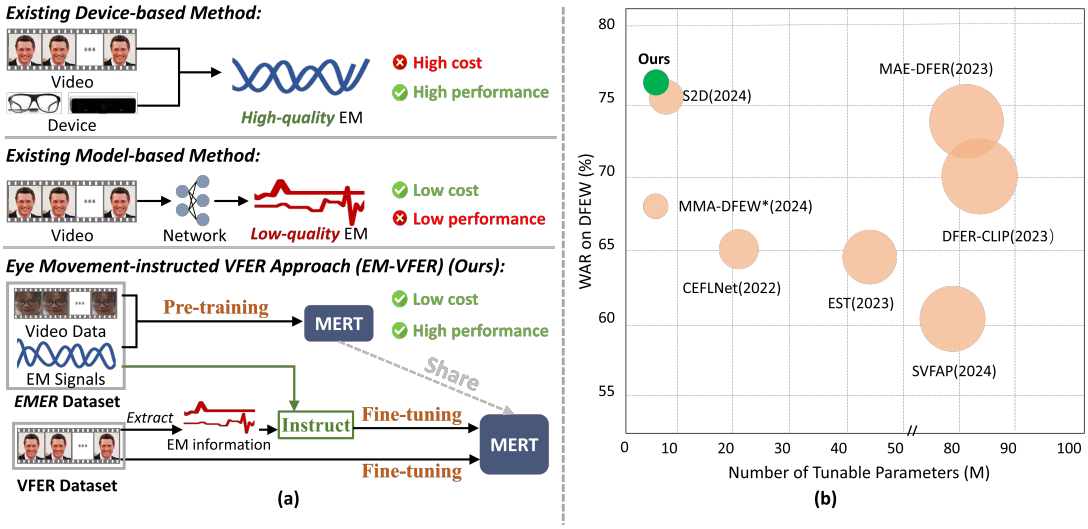

Fig. 1. The comparison of different methods with ours. (a) presents the pipeline of different methods. Previous methods rely on either expensive devices to extract high-quality EM signals or model-based methods to extract low-quality visual eye movement information. In contrast, our EM-VFER first leverages a high-quality EM dateset to instruct learning visual EM information on videos, and then fine-tunes on downstream VFER datasets purely relying on videos, therefore, reducing the reliance on expensive devices while maintaining high-quality eye movement information during inference. (b) shows the Weighted Average Recal (WAR) results and tunable parameters of different methods on the DFEW dataset, with the circle size indicating the quantity of tunable parameters and * indicates results obtained using only the Visual modality. Our method achieves high performance with fewer tunable parameters.

comprehensive and detailed information, thereby enhancing the accuracy and reliability of facial expression recognition. However, despite some improvements, existing eye movement-based methods generally either require expensive sensing devices to obtain high-quality eye movement signals or can only depend on model-based methods to extract low-quality visual eye movement information from videos with compromised benefits, which falls short in real-world practical applications.

To address the above problem, we attempt to maximize the benefits of eye movement signals by introducing a novel eye movement-instructed VFER (EM-VFER) approach. We aim to achieve a better trade-off between high-quality but costly eye-tracking devices and efficient but unreliable visual eye movement information. To achieve this, we apply a novel 2-stage training framework: *high-quality eye movement pre-training* and *eye movement-instructed video data fine-tuning*. To start with, we employ a multi-modal Transformer model, namely Multimodal Emotion Recognition Transformer (MERT), that accepts both eye movement signals and video frames to facilitate VFER. Using this model, we first introduce the high-quality eye movement pre-training stage. In this stage, we leverage the Eye Movement-assisted Multi-modal Emotion Recognition (EMER) dataset for pre-training, which consists of high-quality eye movement signals collected by eye-tracking devices. Using the EMER dataset, the MERT model could learn accurate eye movement patterns that can complement vision-based FER effectively. In the second stage, namely eye movement-instructed video data fine-tuning, we attempt to waive the necessity of high-quality eye movement devices and only use visual eye movement information on video data from common VFER datasets for fine-tuning. To deal with potentially significant noises in visual eye movement information, we propose to utilize the pre-trained prior knowledge about high-quality eye movement signals to help adapt the pre-trained MERT to common VFER data under a novel progressive eye movement-instructed learning (PEML) framework. In PEML, we attempt to estimate the alignment between the distributional representations of the high-quality eye movement signals and the visual eye movement information, and the alignment results are considered as instructing weights. By using the weights, we can instruct the learning of the model by revealing whether the current visual eye movement information is more reliable or more noisy, and we reduce the importance of noisy representations. Notably, the instructing weights, though primarily designed for fine-tuning, the knowledge gained in this process enhances the model's performance at inference without requiring high-quality eye movement signals. Built upon the instructing weights, we devise a progressive adaption process to gradually instruct the MERT model on video data, aiming to further lower the risk of forgetting the prior knowledge due to significant domain gaps. Lastly, after the 2-stage training, our model can deal with VFER data without requiring any specific eye-tracking devices, thus saving much computational costs and enhancing real-world applications.

In summary, our contributions are as follows:

- We propose a new effective and efficient framework for video facial expression recognition , namely EM-VFER, that utilizes high-quality eye movement signals to instruct learning on low-quality visual eye movement information during training while only uses video data for inference. To the best of our knowledge, our EM-VFER, for the first time, can effectively tackle both macro and micro VFER problems by leveraging eye movement signals.
- EM-VFER is a 2-stage training framework for VFER. In the first stage, we pretrain the model (MERT) on the EMER dataset, which provides high-quality, accurate eye movement patterns. In the second stage, we fine-tune

MERT on downstream VFER datasets, where only low-quality visual eye movement information is available. To address the distribution gap, we introduce PEML, which incorporates an eye movement-instructed weighting method for gradual fine-tuning. This approach mitigates the mismatch between pretraining and downstream data, ensuring high-quality eye movement feature distribution and significantly improving VFER performance without introducing additional eye tracking devises.

- Experimentally, we conducted extensive experiments on two VFER tasks, including macro-expression recognition (namely DFEW, MAFW and Aff-wild2 datasets) and micro-expression recognition (namely CASME III and CASME II datasets), showing that our approach achieves new state-of-the-art results. Specifically, we achieve average relative gains of 21.49% in UAR and 14.21% in WAR on DFEW and MAFW datasets, compared to the baseline method. Additionally, for micro-expression recognition, we observe average relative improvements of 16.36% in UAR and 11.23% in UF1, confirming the effectiveness and generality of our proposed approach. Code will be available at https://anonymous.4open.science/r/EM-VFER-7181.

II. Related Work

A. Video-based Facial Macro-Expression Recognition

Recent advances in video-based facial macro-expression recognition focus on dynamic models to enhance accuracy and robustness. Zhang *et al.* [26] proposed a joint network combining a PHRNN and an MSCNN to capture geometric-appearance, part-whole, and dynamic-static features, enhancing spatio-temporal expression recognition.Niu *et al.* [27] proposed the Four-player GroupGAN, which enhances facial expression recognition by improving the discriminative capability for weak expressions. Sun *et al.* [28] uses mask-based reconstruction pre-training, inspired by VideoMAE, for feature extraction from facial videos. Chen *et al.* [16] extend static models with temporal and landmark-guided modules. Zhang *et al.* [29] incorporates frame labels to improve discriminative feature learning and temporal fusion. Ma *et al.* [30] uses a local-global spatio-temporal Transformer to capture local interactions and long-range dependencies with compact loss regularization. Li *et al.* [31] combines global convolutional attention with intensity-aware loss to handle varying expression intensities. However, many methods still struggle to capture subtle, fake, or flickering expressions, limiting robustness in complex emotional recognition.

B. Video-based Facial Micro-expression Recognition

Micro-expression recognition involves detecting, recognizing, and classifying subtle facial expressions. Methods can be classified into manual machine learning and deep learning-based approaches. Manual methods focus on pixel-level changes, preserving detailed information for different facial expressions. For example, Liu *et al.* [32] proposed an optical flow-driven method considering both local motion and spatial position, while Huang *et al.* [33] used robust principal component analysis and localized binary patterns for spatio-temporal feature extraction. These methods are robust but computationally complex, often requiring feature selection to manage high-dimensionality.In contrast, deep learning-based methods have evolved from multi-stage training to end-to-end approaches. Sun *et al.* [34] transferred knowledge from a pre-trained neural network to a student network for micro-expression recognition, while Gupta *et al.* [35] used action units, landmarks, and appearance features with 2D CNNs to capture subtle deformations and analyze spatial and temporal behaviors. While deep learning methods capture deeper micro-expression features, they require large, realistic datasets for optimal performance.Most methods focus solely on facial expressions, neglecting that micro-expressions often coincide with other subtle cues, such as eye movements, in real-world scenarios. Relying only on facial expressions limits model performance and generalization.

C. Eye Movement-based Multimodal Emotion Recognition

Eye movement-based multimodal emotion recognition combines eye movement signals with other information sources, such as facial expressions and EEG, to enhance the accuracy and robustness of emotion recognition. Gong *et al.* [36] improved emotional information extraction by integrating emotion-related brain region signals with eye movement data. Wang *et al.* [37] proposed the ETF, based on a pure attention mechanism, which combines EEG and eye movement signals to better differentiate anger and surprise emotions. Wu *et al.* [38] incorporated head pose and eye movement signals to guide the use of facial features in continuous emotion recognition. Gong *et al.* [39] used EEG and eye movement signals to mitigate spurious correlations between different modalities. Despite these advances, practical applications are still limited by the high cost of equipment.

III. Method

A. Problem Definition and Overview

To improve the performance of VFER, we propose integrating eye movement information in an efficient yet effective way. In particular, we propose an eye movement-instructed VFER approach (EM-VFER) that utilizes high-quality eve movement signals to instruct the fine-tuning on the video-based VFER data, thus obtaining much more accurate expression recognition performance. As introduced previously, EM-VFER operates in two stages: *the high-quality eye movement pre-training stage* and *the eye movement-instructed video data fine-tuning stage*.

In the *high-quality eye movement pre-training stage*, we employ a multimodal network, namely Multimodal Emotion Recognition Transformer (MERT), to learn from high-fidelity eye movement signals. To facilitate pre-training, we compile a new pre-training *Eye-behavior-aided Multimodal Emotion Recognition (EMER) dataset*, denoted as $P = \left\{(V_i, E_i, y_i)\,|i = 1, ..., \ N^P\right\}$, where i indexes over samples, V_i represents the i-th facial video, E_i represents the corresponding i-th eye movement signals obtained from eye-tracking device, y_i denotes the true facial expression label, and N^P denotes the number of samples in P. Each

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3599859

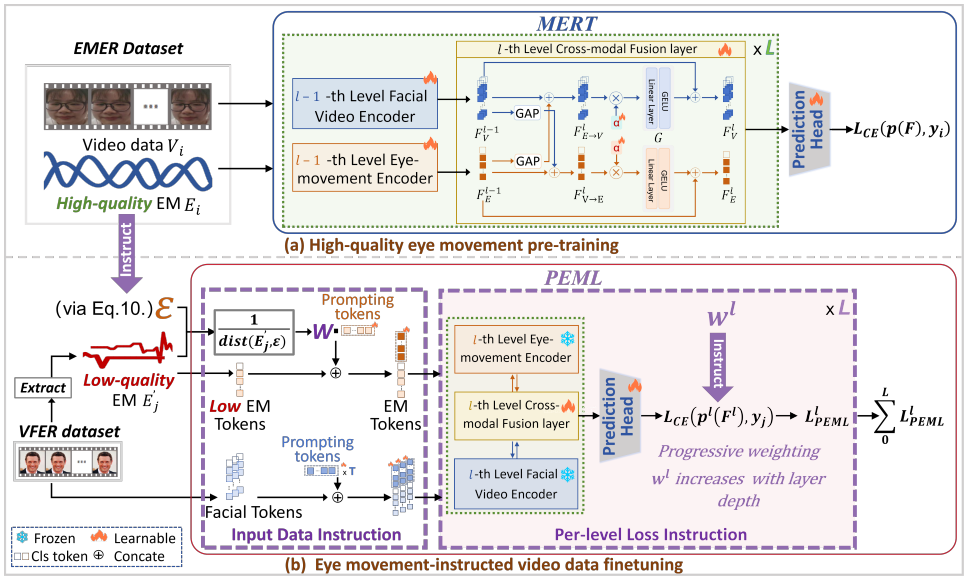

Fig. 2. The proposed EM-VFER framework consists of two stages: (a) high-quality eye movement pre-training and (b) eye movement-instructed video data fine-tuning. In the pre-training stage, MERT is trained on the EMER dataset, integrating facial video and high-quality eye movement (EM) signals via multi-level cross-modal fusion, enabling comprehensive feature alignment for VFER. In the fine-tuning stage, PEML addresses the gap between high- and low-quality EM through input data instruction, leveraging high-quality EM knowledge, and loss instruction, progressively adjusting feature weights. This allows the model to better utilize low-quality visual EM information, enhancing VFER performance.

eye movement signals $E_i = \{d_L, d_R, g_X, g_Y, A\}$ consists of the diameters (d_L, d_R) of the left and right pupils, eye gaze coordinates (g_X, g_Y), and eye behaviors $A = \{blinking, saccades, gazing\}$. During pre-training the MERT would learn to generate effective multimodal features for robust FER. After pre-training, we obtain optimized networks for encoding facial video input, eye movement input, and cross-modal fusion. In the meantime, we can extract distributional features of the high-quality eye-movement signals to help instruct the fine-tuning.

In the *eye movement-instructed video data fine-tuning stage*, the *downstream facial video dataset* is denoted as $D = \{(V_j, y_j) | j = 1, ..., N^D\}$, where j indexes over downstream facial video samples, V_j represents j-th facial video, y_j denotes the facial expression labels within D, and N^D is the number of samples in D. Here, the D does NOT have high-quality eye-movement signals. To take better advantage of pre-trained knowledge for video-based FER, we devise a novel progressive eye movement-instructed learning (PEML) to help fine-tune MERT. Due to the gap between high-quality eye movement signals from pretrained datasets and low-quality visual eye movement information on videos, it is challenging to directly incorporate prior knowledge about eye movements, thus our proposed progressive instructed learning aims to gradually adapt prior knowledge during the fine-tuning.

An overview of our 2-stage training framework is illustrated in Figure. 2. We will subsequently discuss both stages in detail.

B. High-quality Eye Movement Pre-training

1) Pre-training EMER dataset: To facilitate the high-quality eye movement pre-training, we first compile a new multimodal per-training dataset, EMER, which consists of facial videos V_i and high-quality eye movements E_i, collected by a high-definition camera and a Tobii Pro Fusion eye-tracking equipment[1]. The details for EMER can be seen the Sec.IV.A in the experimental part and the Appendix materials.

2) Multimodal Emotion Recognition Transformer (MERT): With the EMER, we devise a MERT to model both eye-movement signals and video data at the same time. As shown in Figure 2(a), MERT consists of three main components: modality-specific encoding, multi-level cross-modal fusion, and a prediction head. When pre-training, the MERT would learn to generate a multimodal feature F using high-quality eye-movement signals (denoted as E_i) and video data (denoted as V_i) in EMER:

$$F = f_{MERT}(\phi_V(V_i), \phi_E(E_i)), \tag{1}$$

where f_{MERT} is the multimodal fusion of MERT, ϕ_V is a facial video encoder and ϕ_E is an eye movement encoder. Then, a prediction head is added on top of F, and training is performed according to the label y_i, thus the loss function can be written as:

$$\mathcal{L}_{pre} = CE(p(F), y_i), \tag{2}$$

where $\mathcal{L}_{pre}$ is the pretraining loss, and CE is cross-entropy. To implement these, we first compile a high-quality eye movement dataset to provide E_i and V_i. We then devise the multimodal network MERT with a Transformer-based prediction head p. We will discuss them in this section.

Modality-specific Encoding: Given 2 types of input, we first employ two modality-specific encoders, *i.e.*, facial video encoder and eye-movement encoder. They extract facial expression features and eye-movement features, respectively. Formally, given the facial video $V_i \in \mathbb{R}^{T\times3\times224\times224}$ and eye-movement signals $E_i \in \mathbb{R}^{T\times5}$ as input, we employ two ViT-based encoders [40] as the facial video encoder ϕ_V and the eye-movement encoder ϕ_E, respectively, which are used

to extract face visual features $F_V \in \mathbb{R}^{T\times N_v\times 728}$ and eye-movement features $F_E \in \mathbb{R}^{N_e\times 728}$ as:

$$F_V = \phi_V(V_i), F_E = \phi_E(E_i), \tag{3}$$

where $N_v = 197$, $N_e = 257$ are the number of facial video patches and eye-movement tokens in ViT, respectively. T is the number of frames for the video. Notably, in this study, the ViT architecture in the two encoders consists of a 12-layer Transformer encoder.

Multi-level Cross-modal Fusion: Fusion is usually important for multimodal modeling [41], [42]. Here, we apply a deep cross-modal fusion strategy to enable better modeling performance. More specifically, we feed the information from each modality into the information of another modality, and we perform this cross-modality fusion for several levels in the encoding process. Figure 2(a) provides the detailed architecture of the l-th layer representation learning in this module.

Using the F_V^{l-1} and F_E^{l-1} at the $l-1$-th layer of ϕ_v and ϕ_E, we first employ a global average pooling (denoted as GAP), to obtain global facial expression and eye-movement features, represented as $GAP(F_V^{l-1})$ and $GAP(F_E^{l-1})$, respectively. Then, we integrate the global features with the original unimodal features to obtain the cross-modal emotion features:

$$F_V^l = F_V^{l-1} + G(\alpha \cdot F_V^{l-1} \oplus GAP(F_E^{l-1})), \tag{4}$$

$$F_E^l = F_E^{l-1} + G(\alpha \cdot F_E^{l-1} \oplus GAP(F_V^{l-1}), \tag{5}$$

where $\oplus$ is the element-wise addition operation, α denotes the learnable parameter that controls the cross-modal feature to fuse during training, and $G()$ contains a linear layer with an activation function. F_V^l and F_E^l are the l-th modality-enhanced emotion representations.

At the last layer, we generate the final multimodal feature F by adding the modality-specific features after aligning their sequence lengths:

$$F = f_{MERT}(F_V, F_E) = F_V^L + F_E^L, \tag{6}$$

where the L is the total number of layers. The layer number actually depends on the depth of the employed facial video encoder. If the facial video encoder has 12 layers, the F_E^{12} and F_V^{12} would be our final modality-specific encoding results.

Prediction Head: Based on the multimodal feature F, we can then perform robust FER by making predictions. We add an extra Transformer $Trans$ and a classifier C to achieve this:

$$p(F) = C(Trans(q/k/v = F)), \tag{7}$$

where $q/k/v$ are the query, key, and value tensors for the Transformer. Here, the $Trans$ mainly performs self-attention, thus the output of $p(F)$ has the same length as the input tokens. To perform classification, the element corresponding to the CLS token in $p(F)$ is passed through a classifier C, which generates the final prediction output.

C. Eye movement-instructed video data fine-tuning

In general, to better adapt the pretrained MERT to downstream datasets where only low-quality visual eye movement information can be extracted, we introduce the PEML in order to instruct the fine-tuning for MERT gradually. Through such progressive instruction, negative effects caused by significant domain gaps between pre-trained and fine-tuning datasets could be mitigated, thus improving the fine-tuning performance. This also helps take better advantage of prior high-quality eye movement signals, when learning on common FER datasets.

More specifically, our PEML instructs the learning on two components: (1) Instruction for Input Data Modeling; (2) Loss Instruction at Each Level of the MERT.

For Instruction for Input Data Modeling, we attempt to introduce extra learnable prompts [43] to divert the MERT from fitting too much on downstream eye movement information and also adjust the importance of the learnable prompts given an estimated quality of the current eye movement information on video data. In the literature on FER and related areas, adding learnable prompting tokens helps pretrained networks adapt to downstream datasets more easily. We refer readers to the papers [44] for more details. Here, we mainly describe how our approach is implemented based on these learnable prompting tokens. In particular, we assume that lower quality would make it difficult for the learnable prompt to effectively represent meaningful eye movement information for downstream tasks due to excessive noise.

For Loss Instruction at Each Level of the MERT, we add additional losses on top of the output of each level of the MERT and assigns these losses with increasing importance. That is, the lower levels of MERT would learn about downstream data more conservably while higher levels of MERT would learn about downstream data more aggressively. This helps the MERT adapt to downstream datasets more appropriately without introducing catastrophic forgetting problems. This pipeline is illustrated by Figure 2(b).

The following sections provide detailed introductions.

1) Instruction for Input Data Modeling: Formally, we introduce the notion l to represent the l-th level in MERT. Then, the multimodal feature we can obtain from MERT at l-th level can be written as: F^l. Similar to the final multimodal feature F as described in Eq.6, F^l fuses facial expression feature and eye-movement feature using related encoders ϕ_V and ϕ_E. Here, instead of directly using low-quality visual eye movement information E'_j on facial video data, we introduce instructions based on prior knowledge about high-quality eye movement signals, then we have:

$$F^l = f^l_{MERT}\Big(\phi_V(P_V \oplus V_j), \phi_E\Big(\Big(\mathcal{W}(E'_j, \mathcal{E}) \cdot P_E\Big) \oplus E'_j\Big)\Big), \tag{8}$$

where f^l_{MERT} denotes the implementation of Eq. 1 at the l-th network layer, and P_V and P_E are extra learnable prompting tokens for the facial video and eye movement modalities, respectively. Specifically, inspired by MMA-DFER [43], we introduce six randomly initialized prompting tokens per modality to facilitate feature learning at different depths. These tokens are prepended directly to the patch embedding sequence of each modality, positioned before the `CLS` token, i.e., at the very start of the Transformer input sequence. $\mathcal{E}$ represents the distributional features of high-quality eye-movement signals, $\mathcal{W}$ is an importance re-weighting function

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3599859

computed based on the alignment between the current eye movement feature E'_j and $\mathcal{E}$ at layer l, and $\oplus$ denotes concatenation. By concatenating P_V and P_E to their respective input sequences, the model encodes these prompting tokens as additional contextual cues, guiding subsequent layers to better attend to salient features within each modality. Moreover, $\mathcal{W}(E'_j, \mathcal{E})$ dynamically adjusts the learning weights: a larger discrepancy between E'_j and $\mathcal{E}$ indicates lower quality signals, resulting in smaller weights and reduced fine-tuning impact on such samples; conversely, higher quality signals receive larger weights to reinforce their influence. As trainable parameters, these prompting tokens are optimized during training, allowing the model to adapt dynamically to varying sample distributions and thereby enhancing multimodal fusion and generalization.

This mechanism enables instructed learning through high-quality eye movement-instructed weighting, designed to enhance model performance during fine-tuning. Importantly, it improves the model's inference capabilities without the need for high-quality eye movement signals at runtime.

High-quality eye movement-instructed weighting: As described in Eq.8, we instruct the modeling of input data by devising an importance weighting mechanism, implemented by $\mathcal{W}(E'_j, \mathcal{E})$, for the fine-tuning. Using extra learnable prompting tokens P_E for eye movement information on facial video data, our introduction of $\mathcal{W}(E'_j, \mathcal{E})$ aims to perform a general estimation about whether the current eye movement information E'_j has plenty of noises. If the current eye movement information on facial video is noisy, we suppose that it is meaningless to make the learnable prompts P_E represent these noisy points. On the contrary, if the current eye movement information has a roughly good quality, the P_E would be much more helpful for aiding the MERT to adapt to fine-tuning data. To estimate the level of noise in E'_j, it is straightforward to assume that E'_j can be cleaner if it is close to the distribution of high-quality eye movement signals (represented by $\mathcal{E}$) and *vice versa*. It is then reasonable to assign a smaller importance to model noisy eye movement information on videos. To fulfill this, we first formulate the $\mathcal{W}(E'_j, \mathcal{E})$ as:

$$\mathcal{W}(E'_j, \mathcal{E}) = \frac{1}{dist(E'_j, \mathcal{E})}, \tag{9}$$

where $dist$ is a distance estimation function. Therefore, the $\mathcal{W}(E'_j, \mathcal{E})$ would be higher if the distance is smaller and *vice versa*. We will present more details for the calculation of E'_j, $\mathcal{E}$ and $dist$ as below.

For current visual eye movement information E'_j, we employ the MTCNN [45] to detect 68 facial landmarks from the corresponding video V_j and calculate the eye landmarks by applying established techniques such as bilateral filtering [46], erosion [47], and binarization [48]. This process yields the 5-dimensional vector, capturing key aspects of the eye movements using the same format with high-quality eye movement signals. Then, to acquire a representation $\mathcal{E}$ that can describe the overall distributional features of high-quality eye movement signals, we found that the center of high-quality eye movement signals is already useful. That is, we calculate $\mathcal{E}$ as the average of all the high-quality eye movement signals:

$$\mathcal{E} = \frac{1}{N^P} \sum_{i=1}^{N^P} E_i, \tag{10}$$

where N^P is the number of samples in the EMER dataset. Regarding the distance estimation $dist$, we find that the simple Euclidean distance is already advantageous, thus

$$dist(E'_j, \mathcal{E}) = ||\mathcal{E} - E'_j||_2. \tag{11}$$

As a result, this formulation estimates the alignment between the current eye movement information E'_j and the high-quality eye movement signals from pre-trained datasets, and the worse alignment will have a lower importance during fine-tuning.

2) Loss Instruction at Each Level of the MERT: We further implement progressive learning by taking into account the network depth l , as illustrated in Figure 3. To better adapt pretrained MERT to downstream tasks, we gradually increase the importance of instructed learning loss through progressive weighting for fine-tuning MERT. As a result, our approach instructs less on lower-level features and more on higher-level features. This ensures that the pretrained network can adapt to low-quality visual eye movement information without being overwhelmed by a significant gap between fine-tuning data and pretrained data at a sudden. Then, for each level, we introduce an additional training loss:

$$\mathcal{L}^l_{PEML} = w^l \cdot \mathcal{L}_{CE}\Big(p^l(F^l), y_j\Big), \tag{12}$$

where $\mathcal{L}^l_{PEML}$ is the loss, w^l is the increasing loss weight, and p^l is an extra prediction head similar with Eq.7.

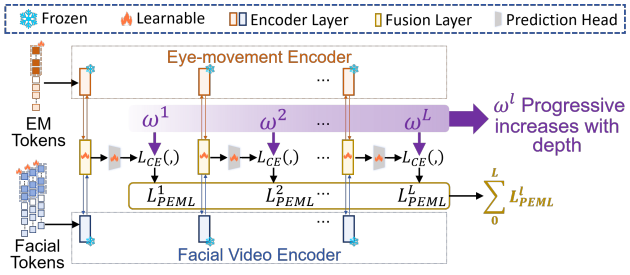

Fig. 3. Details of loss instruction at each level of the MERT in PEML.

Progressive weighting: The introduction of progressive weighting w^l aims to gradually bridge the gap between the pre-training dataset and downstream dataset. Eq. 12 describes the progressive weighting strategy employed in the MERT. More specifically, our progressive weighting happens during the cross-modality fusion network. At each layer, there is a weight w^l that controls the strength of learning. In the early stages, we attempt to assign extremely small weights to slightly guide the MERT, especially the fusion network, to learn patterns from downstream data for predicting facial expressions. At later stages, we assign a larger weight, and the final prediction has the largest weight. That is, we apply a series of w^l starting from a small value and growing into a larger value as the network goes deeper.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3599859

This progressive weighting in losses would support a gradual knowledge adaptation process. Early fusion stages would receive less adaptation pressure, preserving more prior knowledge, while deeper fusion stages gain higher weights, guiding the model toward downstream datasets. This is also more beneficial than only using a final loss. In fact, the final loss only trains the model as a whole, and the early fusion stages might not work well on addressing potentially significant domain gaps, especially when the pre-training dataset has high-quality eye movement signals while downstream datasets only have low-quality visual eye movement information on videos. Adding losses earlier may aid the MERT in dealing with such gaps more effectively.

3) More Details in Fine-tuning Stage and Inference: During the fine-tuning stage, to avoid catastrophic forgetting, we tend to fix the parameters in modality-specific encoders. Using notions from Eq. 8, we actually freeze all the parameters in facial encoder ϕ_V and the eye movement encoder ϕ_E. We achieve the fine-tuning mainly by training the extra learnable prompting tokens P_V and P_E for both modalities. However, we do adjust the weights of the network that implements cross-modality fusion.

In inference, given only a facial video as input, we first extract low-quality visual eye movement (EM) information and process it alongside facial video using fine-tuned extra learnable prompting tokens P_V and P_E. Then, frozen encoders ϕ_V and ϕ_E are used to capture facial expression and eye movement features, respectively. These features are fused through the multi-level cross-modal fusion module, which, along with the prediction head, is learned during the fine-tuning stage. Finally, classification is performed using the learned prediction head, enabling efficient and accurate VFER without the need for high-quality EM signals, reducing reliance on expensive devices.

IV. Experiments and Analysis

A. Datasets

To pre-train on high-quality eye-movement signals, the EMVFER framework utilized the EMER dataset. We then evaluated its performance on both macro- and micro-expression recognition tasks using five popular VFER datasets. For macro-expression recognition, we used MAFW [49], DFEW [50] and Aff-wild2 [51] datasets, and for micro-expression recognition, we selected CASME III [52] and CASME II [53] datasets.

1) Pre-training EMER Dataset: Figure 4 shows some examples and the collection pipeline of EMER, which is collected via a stimulus material-induced spontaneous emotion generation method [54], [55]. Specifically, 115 videos, each lasting 1-2 minutes, were retrieved from public databases and video platforms as emotional stimulus videos. Then, 121 participants from diverse backgrounds were required to view these videos to induce short-term and spontaneous emotional states in a laboratory setting. During this, Tobii Pro Fusion eye-tracking equipment[1] and high-definition camera were used to record eye movements, emotion-related gaze patterns, and facial expressions. After careful alignment, trimming, and filtering of the raw data, EMER compiles a total of 1,303 high-quality multimodal emotional data samples from 121 participants, predominantly covering 3 modalities: facial expression videos, eye movement sequences, and eye fixation maps.

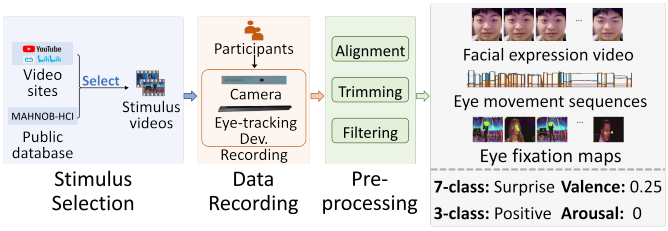

Fig. 4. Collection pipeline for the EMER dataset.

In addition, EMER provides comprehensive emotion-related annotations, as shown in Figure 4. The annotations for EMER include three types: (1) three coarse emotion labels, (2) seven fine-grained facial expression labels, (3) emotional arousal and valence scores. In this study, we mainly use facial expression annotations for deriving face expression changes and high-quality eye movement knowledge related to facial expressions. The details on EMER can be shown in the Appendix materials.

2) Down-stream Video-only Evaluation Dataset: Macro-Expression Recognition Datasets *MAFW* is the first large-scale, multimodal, and multi-label emotional database, comprising 10,045 videos. These videos are annotated with 11 single categories (anger, disgust, fear, happiness, sadness, surprise, contempt, anxiety, helplessness, disappointment, and neutral), 32 compound categories, and descriptive emotional texts. Following the official setup of MAFW, we adopted a 5-fold cross-validation as the evaluation scheme. *DFEW* comprises 11,697 videos, each video segment is individually annotated by 10 professional annotators under expert guidance and labeled as one of seven basic expressions (i.e., happiness, sadness, neutral, anger, surprise, disgust, and fear). Following the official setup of DFEW, we utilized a 5-fold cross-validation as the evaluation protocol. *Aff-wild2* contains 594 videos annotated with three affect models: dimensional, categorical, and action units. We focuses on the categorical annotations, which label each frame with one of eight emotions: anger, disgust, fear, happiness, sadness, surprise, neutral, and other.

Micro-Expression Recognition Datasets *CASME III* , officially known as $CAS(ME)^3$, is a third-generation spontaneous facial micro-expression database, with Part A containing 943 samples from 100 participants recorded using lab cameras at 30 fps and a resolution of 1280x720 pixels. While the original data is categorized into 7 emotions (happiness, anger, fear, disgust, surprise, other, and sadness), we followed prior works [56], [57] and grouped them into three broader categories for analysis. *CASME II* consists of 255 videos, elicited from 26 participants. The videos are recorded using Point Gray GRAS-03K2C camera which has a frame rate of 200fps. All the frames are cropped to 280×340 pixels. The videos are grouped into five categories: happiness, surprise, disgust, repression and others. Following exisiting methods [56], [57], we also grouped the videos into three categories.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3599859

B. Experimental Settings

Training Settings: We followed standard practice by extracting and resizing 16 uniformly sampled frames per video to 224x224 pixels [58]. During training, we utilized a learning rate of 1e-4 with cosine annealing for learning rate decay. The batch size was set to 2, and weight decay was configured at 1e-2. We employed the AdamW optimizer with default settings, and all experiments were conducted using PyTorch on a single NVIDIA GTX 4090 GPU.

Evaluation Metrics: *For macro-expression recognition*, consistent with previous approaches [16], [43], [59],we employ Weighted Average Recall (WAR) and Unweighted Average Recall (UAR) as the primary evaluation metrics for the DFEW and MAFW datasets, respectively, while using Accuracy (ACC) for the Aff-Wild2 dataset. WAR, which corresponds to accuracy, assesses the model's precision in predicting expressions. UAR computes the accuracy for each class and averages it across classes, normalizing for class imbalance. ACC is defined as the proportion of correctly predicted samples to the total number of samples, effectively measuring the model's overall classification performance. *For micro-expression recognition*, following the approach of previous studies [57], [60], we used UAR and Unweighted F1 score (UF1) to evaluate model performance. UF1 is employed to measure performance in multi-class tasks where class imbalances exist, complementing the accuracy assessment of individual classes. These evaluation metrics allow for a comprehensive analysis and assessment of the EM-VFER framework's performance across different tasks.

C. Macro-Expression Recognition Task

1) Performance on DFEW: In Table I, we present a comparative analysis of our method, EM-VFER, against several state-of-the-art approaches on the DFEW dataset for macro-expression recognition.

TABLE I
COMPARISON WITH THE SOTA METHODS FOR MACRO-EXPRESSION RECOGNITION ON DFEW. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED, AND * INDICATES RESULTS OBTAINED USING ONLY THE VISUAL MODALITY.

Method	Modality	WAR	UAR
T-MEP [61]	Audio & Visual	68.85	57.16
HiCMAE [62]	Audio & Visual	75.01	63.76
UMBEnet [63]	Audio & Visual	74.83	62.23
DFER-CLIP [64]	Visual & Text	71.25	59.61
VAEmo [65]	Audio &Visual	75.78	64.02
CEFLNet [66]	Visual	65.35	51.14
EST [18]	Visual	65.85	53.43
IAL [31]	Visual	69.24	55.71
SVFAP [28]	Visual	74.27	62.83
MAE-DFER [1]	Visual	74.43	63.41
MMA-DFER* [43]	Visual	67.15	54.34
OUS [16]	Visual	74.10	60.94
S2D [16]	Visual	75.98	62.57
EM-VFER(Ours)	Visual	**76.43**	**65.83**

Compared to other existing visual-only methods, EM-VFER outperforms most current visual approaches. Our method improves upon the visual modality of MMA-DFER [43] by 13.82% in WAR and 21.14% in UAR. In comparison to the current best visual-only method, S2D [16], our method shows a relative improvement of 5.21% in UAR. These results demonstrate that EM-VFER exhibits strong robustness when using only the visual modality. Additionally, while the multimodal VAEmo [65] achieves the WAR (75.78%) and UAR (64.02%), EM-VFER, using only the visual modality, achieves comparable results with a WAR of 76.43% and UAR of 65.83%. This shows that EM-VFER effectively captures macro-expression information in a unimodal setting, approaching the performance of multimodal methods.

To further valid the effectiveness of our EM-VFER on DFEW, Figure 5 compares the confusion matrices for the visual-only modality MMA-DFER method [43] and our approach. The matrix shows that EM-VFER consistently achieves high classification accuracy across most facial expression categories, with a particularly notable improvement in distinguishing the "disgust" category.

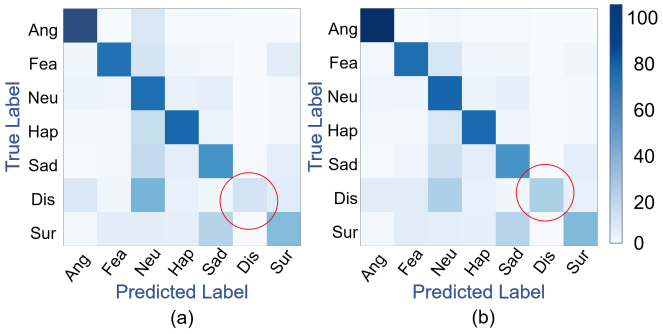

Fig. 5. Confusion matrices of the MMA-DFER method (a) and our EM-VFER approach (b) on the DFEW dataset. Hap, Sad, Ang, Sur, Fea, Dis, Neu are the abbreviations of the corresponding expression labels. The deeper colors indicate higher accuracy.

2) Performance on MAFW: We evaluated EM-VFER on the MAFW dataset for macro-expression recognition and compared it with state-of-the-art methods, as shown in Table II.

TABLE II
COMPARISON WITH THE SOTA METHODS FOR MACRO-EXPRESSION RECOGNITION ON MAFW. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED, AND * INDICATES RESULTS OBTAINED USING ONLY THE VISUAL MODALITY.

Method	Modality	WAR	UAR
T-ESFL [49]	Audio & Visual	48.70	33.35
T-MEP [61]	Audio & Visual	51.15	37.17
HiCMAE [62]	Audio & Visual	56.17	42.65
UMBEnet [63]	Audio & Visual	57.25	**46.92**
DFER-CLIP [64]	Visual & Text	52.55	38.89
VAEmo [65]	Audio &Visual	**58.91**	45.67
SVFAP [28]	Visual	54.28	41.19
MAE-DFER [1]	Visual	54.31	41.62
MMA-DFER* [43]	Visual	50.38	36.24
S2D [16]	Visual	56.20	39.87
FineCLIPER [67]	Visual	56.91	45.01
EM-VFER(Ours)	Visual	57.73	44.15

EM-VFER outperformed other visual-only methods, achieving a 1.44% relative improvement in WAR compared to the state-of-the-art FineCLIPER [67]. This demonstrates EM-VFER's ability to effectively extract complex facial expression features. Compared to MMA-DFER [43], EM-VFER shows

significant gains in the visual modality, with a 14.59% improvement in WAR and 21.83% in UAR [43]. Notably, EM-VFER also achieved competitive results against multimodal methods, with a WAR of 57.73% and UAR of 44.15%, relying solely on the visual modality. Although HiCMAE [62] has a higher UAR (42.65%), its WAR (56.17%) did not lead to a notable improvement in overall accuracy, indicating limitations in recognizing certain categories. While VAEmo achieves the best performance with a WAR of 58.91% and a UAR of 45.67%, it utilizes two modalities, whereas our method relies on a single modality, demonstrating the advantages and effectiveness of our unimodal approach.

In addition, as shown in Figure 6, our model's confusion matrix reveals superior performance compared to the visual-only MMA-DFER in recognizing categories with fewer samples. This highlights the model's strong generalization ability on imbalanced datasets, effectively identifying categories with limited training samples.

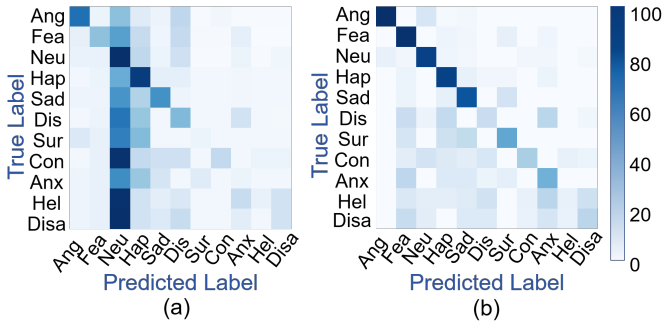

Fig. 6. Confusion matrices of the MMA-DFER method (a) and our approach (b) on the MAFW dataset. Ang, Fea, Neu, Hap, Sad, Dis, Sur, Con, Anx, Hel, Disa are the abbreviations of the corresponding expression labels. The deeper colors indicate higher accuracy.

3) Performance on Aff-Wild2: To thoroughly evaluate EM-VFER's robustness in complex real-world scenarios and its capability to capture subtle expression variations, we conduct experiments on the Aff-Wild2 dataset, which offers frame-level annotations and presents diverse in-the-wild challenges.

TABLE III
COMPARISON WITH THE SOTA METHODS FOR MACRO-EXPRESSION RECOGNITION ON AFF-WILD2. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED, AND * INDICATES RESULTS OBTAINED USING ONLY THE VISUAL MODALITY.

Method	ACC
DMUE [68]	63.64
Tr.FER [69]	68.92
RUL [70]	62.37
Eff.Face [71]	62.21
F2Exp [72]	66.34
POSTER [73]	67.74
EAC [74]	63.54
L.OFER [75]	66.02
LA-Net [76]	66.76
DAN [77]	65.82
POSTER++ [78]	69.18
GReFEL [59]	72.48
MMA-DFER* [43]	68.18
EM-VFER(Ours)	**74.24**

As shown in Table III, EM-VFER achieves state-of-the-art performance with an accuracy of 74.24%, surpassing the current best method, GReFEL (72.48%), by approximately 2.43% relative improvement. Compared to the visual-only baseline MMA-DFER (68.18%), EM-VFER demonstrates a significant 6.06% relative gain, highlighting the effectiveness of our proposed approach.

In alignment with previous datasets, we also present the corresponding confusion matrix, as shown in the Figure 7. From the matrix, it is evident that, compared to the visual-only MMA-DFER, our method achieves superior performance in recognizing several categories, further demonstrating the efficacy of the proposed prompting mechanism and modality-aware design.

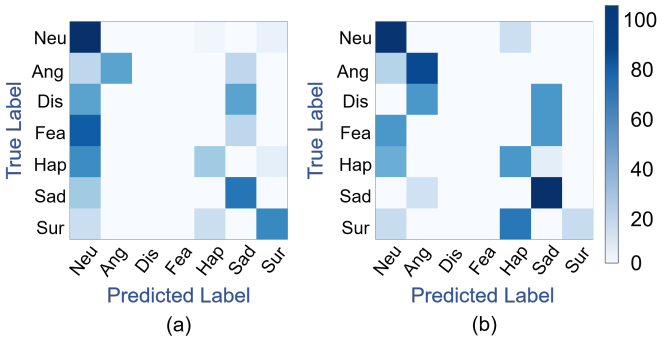

Fig. 7. Confusion matrices of the MMA-DFER method (a) and our EM-VFER approach (b) on the Aff-wild2 dataset.

D. Micro-expression Recognition Task

1) Performance on CASME III: The experimental results on the CASME III dataset (Table IV) demonstrate the effectiveness of EM-VFER, achieving a UF1 score of 65.21% and a UAR score of 66.90%.

TABLE IV
COMPARISON WITH THE SOTA METHODS FOR MICRO-EXPRESSION RECOGNITION TASK ON CASME III. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED, AND * INDICATES RESULTS OBTAINED USING ONLY THE VISUAL MODALITY.

Method	Modality	UF1	UAR
RCN [79]	Visual	38.93	39.28
STSTNet [60]	Visual	37.95	37.92
FeatRef [80]	Visual	34.93	34.13
μ_BERT [57]	Visual	56.04	<u>61.25</u>
HTNet [81]	Visual	<u>57.67</u>	54.15
MMA-DFER* [43]	Visual	54.02	52.61
EM-VFER	Visual	65.21	66.90

This marks significant performance improvements, with relative gains of 16.36% in UF1 and 9.22% in UAR compared to the state-of-the-art μ_BERT [57]. EM-VFER also outperforms the visual-only MMA-DFER [43], which achieved UF1 and UAR scores of 54.0% and 52.6%, respectively. These results highlight EM-VFER's ability to capture subtle micro-expressions and its superior capacity for extracting authentic facial expression features. The integration of high-quality eye movement signals and a progressive fine-tuning strategy further enhances its capability to discern complex emotional states.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3599859

Figure 8 shows EM-VFER's superior performance over MMA-DFER [43] on the CASME III dataset, with higher recognition rates in "Surprise," "Positive," and "Negative" categories. The darker regions highlight its accuracy in capturing subtle expression changes. Eye movement guidance and fine-tuning reduce misclassifications, showcasing EM-VFER's effectiveness in micro-expression recognition and emotional feature extraction.

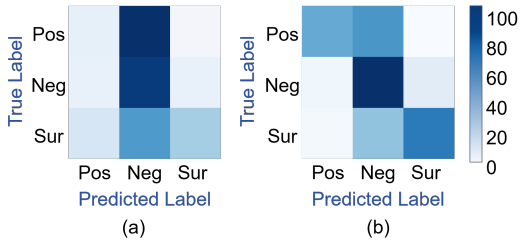

Fig. 8. Confusion matrices of the MMA-DFER method (a) and our approach (b) on the CASME III dataset. Sur, Pos, Neg are the abbreviations of the corresponding labels. The right-side legend displays the relationship between color and prediction accuracy; deeper colors indicate higher accuracy.

2) Performance on CASME II: Table V presents a comparative evaluation of EM-VFER on the CASME II dataset, showing superior performance with a UF1 of 90.53% and a UAR of 92.33%, outperforming all other methods. Notably, EM-VFER significantly improves upon the visual-only MMA-DFER [43] , which achieved UF1 of 89.01% and UAR of 87.5%. These results highlight EM-VFER's higher accuracy and consistency in capturing subtle facial expression changes.

TABLE V
COMPARISON WITH SOTA METHODS FOR MICRO-EXPRESSION RECOGNITION ON THE CASME II DATASET. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED, AND * INDICATES RESULTS OBTAINED USING ONLY THE VISUAL MODALITY.

Method	Modality	UF1	UAR
OFF-ApexNet [82]	Visual	87.64	86.80
Graph-TCN [83]	Visual	86.48	88.71
GACNN [84]	Visual	89.66	86.95
RCN [79]	Visual	81.23	85.12
STSTNet [60]	Visual	83.82	86.86
FeatRef [80]	Visual	89.15	88.73
μ_BERT [57]	Visual	<u>90.34</u>	<u>89.14</u>
MMA-DFER* [43]	Visual	89.01	87.50
EM-VFER (Ours)	Visual	**90.53**	**92.33**

In addition, as shown in the Figure 9, EM-VFER outperforms the MMA-DFER [43] in classification accuracy. Its ability to capture subtle micro-expression variations in CASME II is enhanced by leveraging high-quality eye movement for instructing, leading to more reliable recognition.

E. Ablation Studies and Analysis

1) Effects of Different Modules: We conducted ablation experiments on DFEW (macro-expression) and CASME III (micro-expression), respectivley, to evaluate the effectiveness of key modules in our EM-VFER, as shown in Table VI. From the table, introducing low-quality visual eye movement information resulted in a slight decrease in WAR and UAR, indicating that noise from such data hinders the model's ability

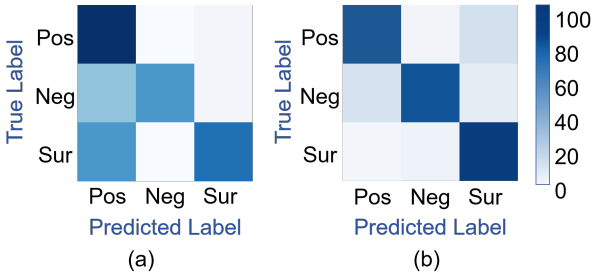

Fig. 9. Confusion matrices of the MMA-DFER method (a) and our EM-VFER approach (b) on the CASMEII dataset.

to capture effective emotional features. In contrast, pre-training with high-quality eye movement signals enhanced the model's ability to recognize complex facial expressions, improving WAR from 67.15% to 73.34% and UAR from 54.34% to 62.18% on DFEW. This approach better captures the subtle relationship between facial expressions and eye movement signals, providing a solid foundation for fine-tuning. In the fine-tuning phase, we employed a PEML mechanism to optimize the model, enabling it to better adapt to downstream datasets with low-quality visual eye movement information, ultimately improving recognition performance.

2) Different Pre-trained Models (baselines) with Our PEML: Table VII presents the performance enhancement of the PEML module in VFER by integrating it into various pre-trained models (baselines). As shown in Table VII, the PEML framework significantly enhances VFER performance across diverse pre-trained models for both macro- and micro-expression recognition tasks, confirming its generalizability and adaptability. Compared to MMA-DFER [43], PEML yields notable improvements: 16.20% in WAR and 29.10% in UF1 on MAFW and CASME III, respectively, along with 25.07% and 25.68% in UAR across both datasets. In the CLIP model [85], PEML achieves a 22.26% increase in WAR on MAFW. By progressively guiding the model to learn high-quality eye-movement features, PEML improves recognition accuracy and generalization, even under low-quality or noisy data, enhancing the model's ability to extract relevant features and improve discriminative performance.

3) Comparison of PEML with Other Domain Adaptation Methods: To comprehensively evaluate PEML's performance in adapting to low-quality eye movement data, particularly in comparison with existing mainstream domain adaptation methods, we conducted the following comparative experiments, with results summarized in Table VIII. It shows that PEML outperforms other domain adaptation methods on the MAFW and CASME III datasets, achieving the highest WAR and UAR scores, demonstrating its superior robustness in adapting to low-quality downstream eye movement data.

4) Different Computation Methods for Eye Movement Distributional Features $\mathcal{E}$ in PEML: Table IX shows the impact of various methods for computing eye movement distributional features $\mathcal{E}$ (see Eq. 10) in PEML, including multivariate_normal [89] and Mean methods [90] . The multivariate normal method fits high-quality eye movement signals to a Gaussian distribution and uses the probability density function

TABLE VI
THE IMPACT OF DIFFERENT MODULES. VIDEO-BASED FACIAL EXPRESSION: USING ONLY FACIAL VIDEO DATA FOR VFER; LOW-QUALITY EM: VISUAL EYE MOVEMENT INFORMATION ON FACIAL VIDEOS VIA CURRENT MODELS; EMER PRE-TRAINING: PRE-TRAINING ON THE EMER DATASET; PEML FINE-TUNING: FINE-TUNING ON DOWNSTREAM DATASETS VIA PROPOSED PEML METHOD.

Video-based Facial Expression	Low-quality Visual EM	EMER Pre-training	PEML Fine-tuning	DFEW (Macro)		CASME III (Micro)	
				WAR	UAR	UF1	UAR
✓				67.15	54.34	52.61	54.02
✓	✓			66.78	52.89	50.51	53.23
✓	✓	✓		73.34	62.18	63.50	63.91
✓	✓	✓	✓	76.43	65.83	65.21	66.90

TABLE VII
THE IMPACT OF DIFFERENT PRE-TRAINED BASELINES WITH OUR PEML.

Basline	MAFW (Macro)		CASME III (Micro)	
	WAR	UAR	UF1	UAR
CLIP [85]	19.36	14.33	9.99	17.48
CLIP [85]+PEML	23.67	17.39	12.32	19.67
NORM-TR [86]	48.19	44.1	35.04	39.00
NORM-TR [86]+PEML	53.2	48.80	38.14	43.60
MMA-DFER	49.68	35.30	50.51	53.23
MMA-DFER [43] +PEML	57.73	44.15	65.21	66.90

TABLE VIII
COMPARISON OF PEML WITH OTHER DOMAIN ADAPTATION METHODS

Baseline	MAFW (Macro)		CASME III (Micro)	
	WAR	UAR	UF1	UAR
Baseline (No DA)	73.34	62.18	63.50	63.91
Adversarial Alignment [87]	73.08	64.79	64.15	65.27
Contrastive Learning [88]	72.19	64.37	64.03	62.77
PEML (Ours)	76.43	65.83	65.21	66.90

(PDF) to calculate weights $\mathcal{W}(E'_j, \mathcal{E})$. While this method leverages distribution characteristics to improve model accuracy, noise and data complexity may reduce its effectiveness. In contrast, the mean method computes the average of high-quality features and uses the inverse of the Euclidean distance to assign weights. Experimental results show that the mean method outperforms the multivariate normal method on the DFEW and CASME III datasets.

TABLE IX
THE IMPACT OF DIFFERENT COMPUTATION METHODS FOR EYE MOVEMENT DISTRIBUTIONAL FEATURES $\mathcal{E}$.

Method	DFEW (Macro)		CASME III (Micro)	
	WAR	UAR	UF1	UAR
multivariate_normal	76.27	65.54	64.98	66.19
Mean	76.43	65.83	65.21	66.90

5) Different Distance Estimation Functions $dist$ in Eq. 11: In this section, we evaluated the impact of various distance computation functions $dist$ in Eq. 11 on model performance, as depicted in Table X. The Inverse Cosine Similarity [91] achieved WAR and UAR scores of 76.02% and 63.52% on the DFEW and CASME III datasets, respectively. While it captures angular relationships well, it struggles to differentiate samples in high-dimensional spaces. In contrast, the Euclidean Distance method improves performance, with WAR and UAR rising to 76.43% and 65.83%, respectively. This suggests that Euclidean Distance better preserves the geometric structure of the data, enabling more accurate differentiation in high-dimensional contexts, making it more suitable for distinguishing subtle variations in eye movement information.

TABLE X
THE IMPACT OF DIFFERENT DISTANCE COMPUTATION METHODS $dist$ IN EQ. 11.

Method	DFEW (Macro)		CASME III (Micro)	
	WAR	UAR	UF1	UAR
Inverse Cosine Similarity	76.02	63.52	63.70	65.70
Euclidean distance	76.43	65.83	65.21	66.90

6) Computational Cost and Model Performance Analysis: In this section, we evaluate the computational complexity and performance trade-offs of our proposed model compared to several baseline methods on the EMER dataset, as summarized in Table XI.

TABLE XI
COMPUTATIONAL COST AND MODEL PERFORMANCE ANALYSIS ON EMER DATASET

Method	Modality	Simple	Real-time	Params	FLOPs	WAR	UAR
MLP [92]	V	Yes	No	3.8M	13.69G	30.29	20.63
ResNet18 [93]	V	Yes	No	6.3M	15.68G	35.58	29.66
MMA-DFER [43]	V	No	No	7.3M	617.73G	50.01	37.66
CLIP [85]	V	No	No	84.75M	805.80G	21.08	16.53
C3D_LSTM [94], [95]	V+E	Yes	Yes	10.63M	50.21G	46.21	28.04
NORM-TR [86]	V+E	No	Yes	12.87M	63.38G	50.27	30.16
Ours	V	No	Yes	7.62M	308.86G	51.36	47.89

Our model achieves an optimal balance between accuracy and efficiency. While lightweight baselines (MLP [92], ResNet18 [93]) are efficient, they suffer from poor accuracy. In contrast, MMA-DFER [43] offer better performance but at the cost of high computational overhead and lack of real-time capability. With only 7.62M parameters, our model supports real-time inference and achieves competitive WAR and UAR. Notably, it outperforms real-time multimodal models like C3D_LSTM [94], [95] and NORM-TR [86] using only visual input, highlighting its scalability and deployment ease in resource-limited scenarios.

7) Individual sensitivity analysis: To assess whether individual differences in eye movement introduce prediction bias, we conducted grouped experiments by gender on the MAFW and CASME III datasets. The results are shown in Table XII.EM-VFER shows consistent performance across genders, demonstrating robustness to individual differences and common real-world variations like culture, facial features, and recording conditions.

We further examined age-related effects on model predictions. Due to the narrow age range in the micro-expression

TABLE XII
PERFORMANCE COMPARISON OF MICRO-EXPRESSION AND MACRO-EXPRESSION RECOGNITION BY GENDER

Gender	MAFW (Macro)			CASME III (Micro)		
	Sample	WAR	UAR	Sample	WAR	UAR
Male	5378	57.23	43.08	454	64.21	65.12
Female	3794	55.97	42.89	228	62.99	64.08
Male+Female	9172	56.69	43.68	682	64.79	66.34

dataset (mean 22, standard deviation 1.6), analysis was limited to the MAFW dataset. Results are shown in Table XIII.It shows minor performance differences across age groups, with the 26–35 group achieving the highest WAR (57.59%) and the 46–55 group the lowest WAR (56.62%). These results indicate that age-related eye movement differences have minimal impact on EM-VFER's predictions, demonstrating the model's robustness to individual variability.

TABLE XIII
PERFORMANCE COMPARISON OF MICRO-EXPRESSION BY AGE GROUP ON MAFW

Age Group	Sample	WAR	UAR
26-35	1500	57.59	43.68
36-45	1500	57.01	43.87
46-55	1500	56.62	43.01
26-55	4500	57.35	44.28

F. Visualization Analysis

1) Emotional Attention Visualization on Macro-expressions and Micro-expressions: In order to gain a deeper understanding of whether the model effectively utilizes key cues in VFER, we explore the regions of interest by visualizing the attention maps, as shown in Figure 10.

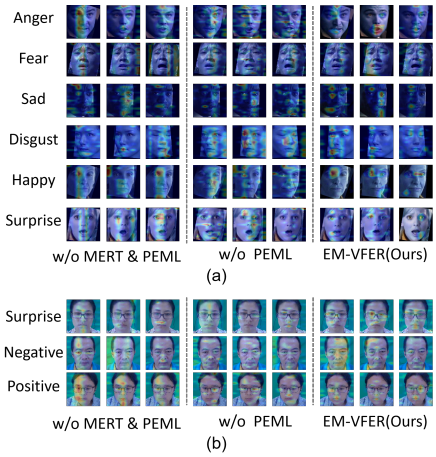

Fig. 10. Emotional attention visualization results on Macro-expression dataset (DFEW)(a) and Micro-expression dataset (CASME III)(b).

Preliminary analyses reveal that when the model relies exclusively on facial expressions for recognition, it predominantly attends to prominent changes in facial features, such as the movements of the mouth and eyebrows (refer to the first column of Figure 10(a) and Figure 10(b). While this emphasis facilitates the identification of overt emotional states, it may inadvertently lead to the oversight of subtle eye movements, which are equally vital for conveying emotional nuances. This observation underscores the potential for the model to miss critical visual cues in the emotional recognition process, consequently impacting its overall performance. After applying our framework, the attention maps show improved focus on both facial regions and subtle eye movements (second and third columns in both figures). This enhancement allows the model to capture a broader range of emotional cues, boosting its accuracy in recognizing complex emotional states. Comparing the attention maps before and after integrating the PEML module further confirms its effectiveness in improving the model's sensitivity to key emotional cues, significantly enhancing performance in both macro- and micro-expression tasks.

2) Visualization on Expression Feature Distribution: To better understand the feature distributions across different facial expression categories, we applied PCA dimensionality reduction to the emotional features fused with eye movement and facial expression data (F_V^L) from our multi-level cross-modal fusion modules. The results for the macro-expression dataset (DFEW) are shown in Figure 11(a), while those for the micro-expression dataset (CASME III) are presented in Figure 11(b).

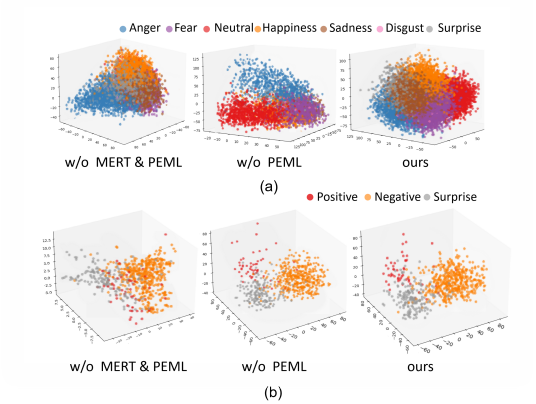

Fig. 11. Emotional feature visualization with and without key modules in our framework on the Macro-expression dataset (DFEW)(a) and Micro-expression dataset (CASME III)(b).

In the experiments, removing the PEML module resulted in a more scattered distribution of emotional features, blurring the boundaries between categories. This highlights the importance of PEML, which enhances eye movement patterns in visual expression features, enabling clearer inter-class separation. Further removal of the MERT module caused a significant loss in the structural integrity of the feature distribution, weakening the clustering effect of facial expression categories and increasing overlap. This suggests that the MERT module, by leveraging multimodal data, strengthens emotional feature extraction and provides high-quality eye movement signals that guide PEML. These findings underscore the crucial role of the PEML and MERT modules in improving recognition accuracy and enhancing category distinction.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3599859

3) Visualization on Eye-movement Feature Distribution: To investigate the distribution characteristics of eye-movement features across different facial expression categories, we applied our method with and without the pre-trained MERT and instructed PEML modules on the DFEW and CASME III datasets, using LDA [96] for visualization. Figure 12(a) displays the results of DFEW, while 12(b) shows the results of CASME III.

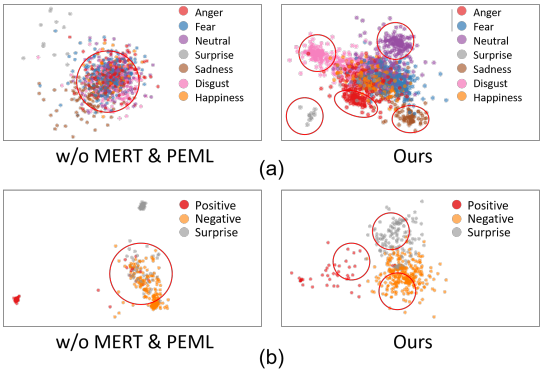

Fig. 12. Eye-movement feature distribution on Macro-expression dataset (DFEW)(a) and Micro-expression dataset (CASME III)(b).

The results show that adding the MERT and PEML modules improves feature distribution, leading to more distinct and well-separated features on both the DFEW and CASME III datasets. The model with these modules achieves better differentiation between facial expression categories, both for macro- and micro-expression recognition, while the model without them exhibits significant feature overlap and poor separation. These observations highlight the key role of the MERT and PEML modules in enhancing feature extraction, classification performance, and generalization.

4) Visualization on Distribution Differences between Various Eye-Movement Features: To verify that PEML learns a distribution that more closely aligns with high-quality, device-collected eye movement signals, we performed a comparative analysis using the DFEW and CASME III datasets.

On the DFEW dataset, as shown in Figure 13(a), the right side of the figure demonstrates that the L-th layer low-quality visual eye movement features $\phi_E\left(\left(\mathcal{W}(E'_j,\mathcal{E})\cdot P_E\right)\oplus E'_j\right)\big)$ (represented in yellow) processed by EM-VFER significantly overlap with the high-quality eye movement features $\phi_E(E_i)$ (represented in gray) extracted from the EMER dataset in multiple regions, with peak positions being very close. This indicates that PEML effectively aligns the distribution of low-quality visual eye movement features with that of the high-quality data. In contrast, the l-th layer low-quality visual eye movement features $\phi_E(E'_j)$ (represented in blue) that were processed by EM-VFER without PEML show less overlap with the high-quality features, as shown on the left side of Figure 13(a). This further confirms that PEML progressively adjusts the distribution of low-quality data to make it more closely resemble the high-quality eye movement signals collected by the device, thereby providing more reliable input features for the VFER task.

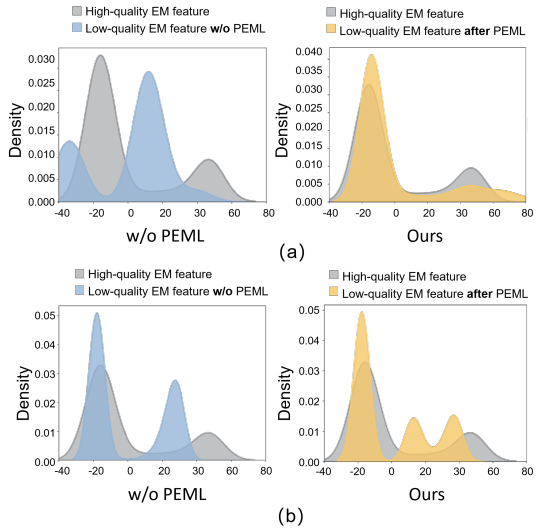

Fig. 13. Distributions of eye-movement features extracted by different methods on Macro-expression dataset (DFEW) (a) and Micro-expression dataset (CASME III) (b)

In addition, as shown in Figure 13(b), the distribution differences of eye movement information in the CASME III dataset further validate the effectiveness of PEML. Despite the low-quality visual eye movement features processed by EM-VFER in the CASME III dataset showing three peaks with reduced peak heights (shown on the right), this indicates that PEML enhances the diversity of the data. Although overall feature performance may decline, this diversity provides valuable information for the model, helping it adapt to different visual tasks. This suggests that PEML can improve the performance of low-quality data, making it better reflect complex situations and align more closely with high-quality device-collected data.

5) Visualization on Loss Performance Curves: Figure 14 illustrates the loss performance curves of the model on the DFEW and CASME III datasets as key modules are progressively integrated into the EM-VFER framework. These curves highlight the impact of each module on the model's convergence speed and stability.

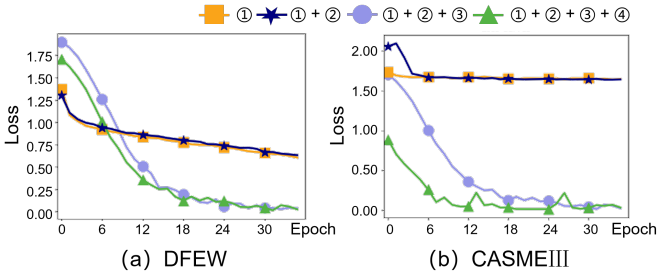

Fig. 14. Loss performance curves on DFEW and CASME III, respectively. ①: Video-based Facial Expression; ②: Low-quality Visual EM; ③: EMER Pre-training; ④: PEML Fine-tuning.

As shown in Figure 14(a) for the DFEW dataset, the models that rely solely on facial video data or incorporate low-quality visual eye movement information exhibit lower initial loss values and show a more slow decline during training. In contrast, our method, which is pre-trained with high-quality

eye movement signals from the EMER dataset, achieves faster convergence, as evidenced by a more rapid decrease in the loss curve. For the CASME III dataset (Figure 14(b)), the initial loss values are higher for all methods except ours. After pre-training with high-quality eye movement signals from the pre-trained EMER dataset, the other methods exhibit a faster loss reduction than when using low-quality visual eye movement information. However, our approach consistently outperforms the others, demonstrating both a faster and more stable loss reduction, which highlights its superior convergence speed and overall performance.

6) Case Visualization and Analysis: To comprehensively evaluate EM-VFER, we analyze both correct and incorrect predictions, focusing on rare or ambiguous expressions (see Figure 15). In Figure 15(e), the subject's eye movements exhibit prolonged fixations with few saccades, typically characteristic of sadness, but the concurrent furrowed brows and eye tension resemble anger, causing the model to misclassify "sad" as "angry." Similarly, in Figure 15(f), ambiguous cues such as pupil size variation, slight brow furrowing, and down-turned mouth corners increase the difficulty of classification. In contrast, Figure 15(c) presents a correctly classified ambiguous case where subtle cues, such as pupil size changes and slight brow furrowing, are accurately interpreted, reflecting the model's sensitivity to fine-grained features.

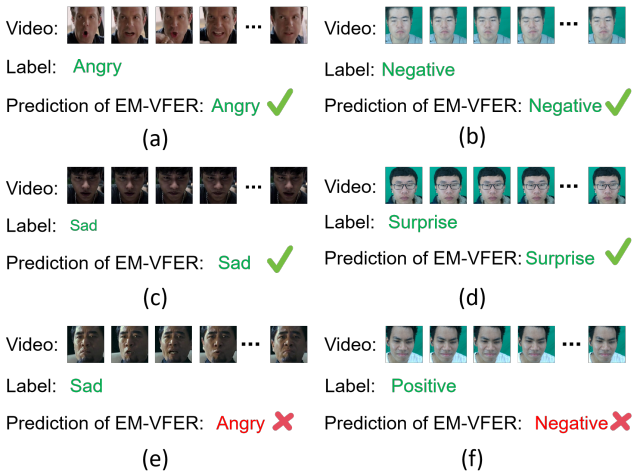

Fig. 15. Visualization of successful cases and failures. (a), (c), and (e) are from DFEW, while (b), (d), and (f) are from CASME III.

V. Conclusion

In this paper, we propose a novel Eye Movement-Instructed VFER (EM-VFER) approach that consists of two stages: the high-quality eye movement pre-training stage and the eye movement-instructed video data fine-tuning stage. In the pre-training stage, we introduce an Eye Movement-assisted Multi-modal Emotion Recognition (EMER) dataset, which is used to train the Multi-modal Emotion Recognition Transformer (MERT) model. This stage facilitates the extraction of meaningful features from both eye movement signals and facial video data. In the fine-tuning stage, we introduce a Progressive Eye Movement-Instructed Learning (PEML) strategy. PEML gradually incorporates eye movement signals into the learning process, guiding the model to refine its understanding of VFER cues. This fine-tuning process leverages the prior knowledge gained from the pre-training phase, where high-quality eye movement signals is used to inform the model's learning. The combination of these two stages significantly improves feature extraction and model performance on both macro-expression and micro-expression recognition tasks. Extensive evaluations demonstrate that EM-VFER outperforms existing methods, showing its promising potential for practical applications. However, we also identify challenges related to individual differences in eye movement patterns and emotional expressions, as well as unknown and diverse noise in open-world environments. In future work, we plan to integrate the multimodal foundation large model and causal inference models to enhance the robustness of eye movement signal learning in open environments, aiming to improve the generalizability and transferability of our models across open-world scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

References

[1] L. Sun, Z. Lian, B. Liu, and J. Tao, "Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition," in *ACM MM*. Association for Computing Machinery, 2023, p. 6110–6121.

[2] D. Kollias, P. Tzirakis, A. Baird, A. Cowen, and S. Zafeiriou, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges," in *CVPR*, 2023, pp. 5888–5897.

[3] D. Kollias, "Multi-label compound expression recognition: C-expr database & network," in *CVPR*, 2023, pp. 5589–5598.

[4] D. Kollias and S. Zafeiriou, "Analysing affective behavior in the second abaw2 competition," in *ICCV*, 2021, pp. 3652–3660.

[5] D. Kollias, "Abaw: Learning from synthetic data & multi-task learning challenges," in *ECCV*. Springer, 2023, pp. 157–172.

[6] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the first abaw 2020 competition," in *FG*, 2020, pp. 794–800.

[7] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution matching for heterogeneous multi-task learning: a large-scale face study," *arXiv preprint arXiv:2105.03790*, 2021.

[8] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.

[9] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," *arXiv preprint arXiv:1910.11111*, 2019.

[10] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *IJCV*, pp. 1–23, 2019.

[11] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild'challenge," in *CVPR*. IEEE, 2017, pp. 1980–1987.

[12] D. Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *CVPR*, 2022, pp. 2328–2336.

[13] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv preprint arXiv:1910.04855*, 2019.

[14] Y. He, Z. Xu, L. Ma, and H. Li, "Micro-expression spotting based on optical flow features," *Pattern Recognit. Lett*, vol. 163, pp. 57–64, 2022.

[15] K. Sasaki, K. Watanabe, M. Hashimoto, and N. Nagata, "Person-invariant facial expression recognition based on coded movement direction of keypoints of facial parts," *Ieej Transactions on Electronics, Information and Systems*, vol. 138, pp. 611–618, 2018.

[16] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *IEEE Trans. Affective Comput.*, 2024.

[17] F. Ma, B. Sun, and S. Li, "Spatio-temporal transformer for dynamic facial expression recognition in the wild," *arXiv preprint arXiv:2205.04749*, 2022.

[18] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, and Y. Zhan, "Expression snippet transformer for robust video-based facial expression recognition," *Pattern Recognition*, vol. 138, p. 109368, 2023.

[19] E. Wnuk and J. Wodowski, "Culture shapes how we describe facial expressions," *Scientific Reports*, vol. 14, no. 1, p. 21589, 2024.

[20] A. Sarkar, I. Chatterjee, A. Dhar, J. Das, P. Roy, P. Ghosh, and S. Das, "Emoeyes: A machine learning exploration of emotional states through eye movement tracking in visual content," *ICCECE*, pp. 1–8, 2024.

[21] A. Abdou, E. Sood, P. Muller, and A. Bulling, "Gaze-enhanced cross-modal embeddings for emotion recognition," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, pp. 1 – 18, 2022.

[22] N. T. T. Nguyen, D. T. T. Nguyen, and B. T. Pham, "Micro-expression recognition based on the fusion between optical flow and dynamic image," *Proceedings of the 2021 5th International Conference on Machine Learning and Soft Computing*, 2021.

[23] R. A. Asmara, P. Choirina, C. Rahmad, A. Setiawan, F. Rahutomo, R. D. R. Yusron, and U. D. Rosiani, "Study of drmf and asm facial landmark point for micro expression recognition using klt tracking point feature," *Journal of Physics: Conference Series*, vol. 1402, 2019.

[24] H. Cao and F. Elliott, "Analysis of eye fixations during emotion recognition in talking faces," in *ACII*, 2021, pp. 1–7.

[25] T. Van Huynh, H.-J. Yang, G.-S. Lee, S.-H. Kim, and I.-S. Na, "Emotion recognition by integrating eye movement analysis and facial expression model," in *ICMLSC*, ser. ICMLSC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 166–169.

[26] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, 2017.

[27] W. Niu, K. Zhang, D. Li, and W. Luo, "Four-player groupgan for weak expression recognition via latent expression magnification," *Knowl-based Syst*, vol. 251, p. 109304, 2022.

[28] L. Sun, Z. Lian, K. Wang, Y. He, M. Xu, H. Sun, B. Liu, and J. Tao, "Svfap: Self-supervised video facial affect perceiver," *IEEE Trans. Affective Comput.*, 2024.

[29] Z. Zhang, X. Tian, Y. Zhang, K. Guo, and X. Xu, "Label-guided dynamic spatial-temporal fusion for video-based facial expression recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 10 503–10 513, 2024.

[30] F. Ma, B. Sun, and S. Li, "Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition," *ICASSP*, pp. 1–5, 2023.

[31] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," in *AAAI*, 2023.

[32] Y.-J. Liu, J. Zhang, W.-J. Yan, S. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affective Comput.*, vol. 7, pp. 299–310, 2016.

[33] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 32–47, 2019.

[34] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Trans. Affective Comput.*, vol. 13, no. 2, pp. 1037–1043, 2022.

[35] P. Gupta, "Merastc: Micro-expression recognition using effective feature encodings and 2d convolutional neural network," *IEEE Trans. Affective Comput.*, vol. 14, no. 2, pp. 1431–1441, 2023.

[36] X. Gong, C. L. P. Chen, and T. Zhang, "Cross-cultural emotion recognition with eeg and eye movement signals based on multiple stacked broad learning system," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 2, pp. 2014–2025, 2024.

[37] Y. Wang, W.-B. Jiang, R. Li, and B.-L. Lu, "Emotion transformer fusion: Complementary representation properties of eeg and eye movements on recognizing anger and surprise," in *BIBM*, 2021, pp. 1575–1578.

[38] S. Wu, Z. Du, W. Li, D. Huang, and Y. Wang, "Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze," in *ICMI*. New York, NY, USA: Association for Computing Machinery, 2019, p. 40–48.

[39] X. Gong, C. P. Chen, B. Hu, and T. Zhang, "Ciabl: Completeness-induced adaptative broad learning for cross-subject emotion recognition with eeg and eye movement signals," *IEEE Trans. Affective Comput.*, 2024.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[41] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[42] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Comput. Surv*, vol. 56, no. 9, pp. 1–36, 2024.

[43] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild," in *CVPR*, 2024, pp. 4673–4682.

[44] Y. Liu, Y. Huang, S. Liu, Y. Zhan, Z. Chen, and Z. Chen, "Open-set video-based facial expression recognition with human expression-sensitive prompting," *arXiv preprint arXiv:2404.17100*, 2024.

[45] 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, 2016.

[46] S. Paris, P. Kornprobst, J. Tumblin, F. Durand *et al.*, "Bilateral filtering: Theory and applications," *Found. Trends Comput. Graph. Vis.*, vol. 4, no. 1, pp. 1–73, 2009.

[47] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[48] J. Yousefi, "Image binarization using otsu thresholding algorithm," *Ontario, Canada: University of Guelph*, vol. 10, 2011.

[49] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan, "Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild," in *ACM MM*, 2022, pp. 24–32.

[50] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *ACM MM*, 2020, pp. 2881–2889.

[51] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.

[52] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "Cas(me)3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2782–2800, 2023.

[53] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

[54] W. Liu, J. Qiu, W. Zheng, and B. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 2, pp. 715–729, 2022.

[55] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, 2019.

[56] L. Zhou, X. Shao, and Q. rong Mao, "A survey of micro-expression recognition," *Image Vis. Comput.*, vol. 105, p. 104043, 2020.

[57] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, and K. Luu, "Micron-bert: Bert-based facial micro-expression recognition," *CVPR*, pp. 1482–1492, 2023.

[58] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *ArXiv*, vol. abs/2312.05447, 2023.

[59] A. T. Wasi, T. H. Rafi, R. Islam, K. Šerbetar, and D.-K. Chae, "Grefel: Geometry-aware reliable facial expression learning under bias and imbalanced data distribution," in *ACCV*, 2024, pp. 4368–4384.

[60] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *FG*. IEEE, 2019, pp. 1–5.

[61] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3192–3203, 2024.

[62] L. Sun, Z. Lian, B. Liu, and J. Tao, "Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition," *Inform Fusion*, vol. 108, p. 102382, 2024.

[63] X. Mai, J. Lin, H. Wang, Z. Tao, Y. Wang, S. Yan, X. Tong, J. Yu, B. Wang, Z. Zhou *et al.*, "All rivers run into the sea: Unified modality brain-inspired emotional central mechanism," in *ACM MM*, 2024.

[64] Z. Zhao and I. Patras, "Prompting visual-language models for dynamic facial expression recognition," in *BMVC*, 2023.

[65] H. Cheng, Z. Zhao, Y. He, Z. Hu, J. Li, M. Wang, and R. Hong, "Vaemo: Efficient representation learning for visual-audio emotion with knowledge injection," *arXiv preprint arXiv:2505.02331*, 2025.

[66] Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, and Z. Luo, "Clip-aware expressive feature learning for video-based facial expression recognition," *Information Sciences*, vol. 598, pp. 182–195, 2022.

[67] H. Chen, H. Huang, J. Dong, M. Zheng, and D. Shao, "Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters," in *ACM MM*, 2024, pp. 2301–2310.

[68] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *CVPR*, 2021, pp. 6248–6257.

[69] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *ICCV*, 2021, pp. 3601–3610.

[70] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *NIPS*, vol. 34, pp. 17 616–17 627, 2021.

[71] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *AAAI*, vol. 35, no. 4, 2021, pp. 3510–3519.

[72] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2exp: Combating data biases for facial expression recognition," in *CVPR*, 2022, pp. 20 291–20 300.

[73] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *ICCV*, 2023, pp. 3146–3155.

[74] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *ECCV*. Springer, 2022, pp. 418–434.

[75] I. Lee, E. Lee, and S. B. Yoo, "Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition," in *ICCV*, 2023, pp. 1536–1546.

[76] Z. Wu and J. Cui, "La-net: Landmark-aware learning for reliable facial expression recognition under label noise," in *ICCV*, 2023, pp. 20 698–20 707.

[77] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, 2023.

[78] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang, "Poster++: A simpler and stronger facial expression recognition network," *Pattern Recognition*, p. 110951, 2024.

[79] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 8590–8605, 2020.

[80] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognition*, vol. 122, p. 108275, 2022.

[81] Z. Wang, K. Zhang, W. Luo, and R. Sankaranarayana, "Htnet for micro-expression recognition," *Neurocomputing*, vol. 602, p. 128196, 2024.

[82] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "Off-apexnet on micro-expression recognition system," *SIGNAL PROCESS-IMAGE*, vol. 74, pp. 129–139, 2019.

[83] L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-tcn with a graph structured representation for micro-expression recognition," in *ACM MM*, 2020, pp. 2237–2245.

[84] A. J. R. Kumar and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *CVPR*, 2021, pp. 1511–1520.

[85] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021.

[86] Y. Liu, H. Zhang, Y. Zhan, Z. Chen, G. Yin, L. Wei, and Z. Chen, "Noise-resistant multimodal transformer for emotion recognition," *IJCV*, pp. 1–21, 2024.

[87] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *NIPS*, vol. 31, 2018.

[88] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NIPS*, vol. 33, pp. 18 661–18 673, 2020.

[89] Y. L. Tong, *The multivariate normal distribution*. Springer Science & Business Media, 2012.

[90] D. L. Mohr, W. J. Wilson, and R. J. Freund, *Statistical methods*. Academic Press, 2021.

[91] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

[92] H. Taud and J.-F. Mas, "Multilayer perceptron (mlp)," in *Geomatic approaches for modeling land change scenarios*. Springer, 2017, pp. 451–455.

[93] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[94] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[95] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[96] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," *Robust data mining*, pp. 27–33, 2013.

Yuanyuan Liu received the PhD degree from Central China Normal University. She is currently an associate professor at the China University of Geosciences (Wuhan). She was also a visiting scholar in Nanyang Techonological University, Singapore. Her research interests include computer vision and multimodal analysis. She has published various top conferences and journals, such as CVPR, ACM MM, IEEE T-VCG, PR, IEEE TGRS, CIKM, INS, NC, IEEE FGR and so on.

Lin Wei received the B.S. and M.S. degrees in software engineering from the China University of Geosciences, Wuhan, China, in 2022 and 2025, respectively. Her research focus on affective computing.

Kejun Liu received the B.S. degree in Network Engineering from Henan University in 2022; she is currently working toward the Ph.D. degree at China University of Geosciences in Wuhan, China. Her research interests include affective computing.

Zijing Chen is a Lecturer in the Department of Computer Science and Information Technology at La Trobe University, Australia, affiliated with the Cisco-La Trobe Centre for Artificial Intelligence and IoT. She received her Ph.D. from the University of Technology Sydney, Australia, in 2019. Her research interests include video processing, machine learning, computer vision, and artificial intelligence, with publications in high-quality journals and conferences.

Zhe Chen is a Lecturer in the Department of Computer Science and Information Technology at La Trobe University, Australia, affiliated with the Cisco-La Trobe Centre for Artificial Intelligence and IoT and the Australian Centre for AI in Medical Innovation. He received his Ph.D. from the University of Sydney in 2019. His research focuses on computer vision and artificial intelligence, with applications in healthcare, robotics, and related domains. His work is regularly published in top-tier venues such as CVPR and IJCV and has accumulated over 5,000 citations to date. Dr. Chen serves as a reviewer for leading journals and conferences in the field.

Chang Tang (Senior Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2016. He joined the AMRL Laboratory, University of Wollongong, Wollongong, NSW, Australia, from September 2014 and September 2015. He has published various peer-reviewed articles, including those in highly regarded journals and conferences, such as TPAMI, TMM, TGRS, ICCV, CVPR, AAAI, ACM MM. His research interests include machine learning and computer vision.

Jingying Chen received the BEng degree in electronics & information, the MEng degree in computer science from the Huazhong University of Science and Technology, China, and the PhD degree in computer engineering from Nanyang Technological University, Singapore. She is currently a professor at the Central Normal University of China. Her research includes pattern recognition, artificial intelligence, and intelligent systems.

Shiguang Shan received the MS degree in computer science from the Harbin Institute of Technology, China, in 1999, and the PhD degree in computer science from the ICT, CAS, Beijing, in 2004. He has been with ICT, CAS since 2002 and has been a professor since 2010. His research interests include image analysis, pattern recognition, and computer vision. He is a fellow of the IEEE.