



High-level LoRA and hierarchical fusion for enhanced micro-expression recognition

Zhiwen Shao^{1,2,3,4} · Yifan Cheng^{1,4} · Yong Zhou^{1,4} · Xiang Xiang² · Jian Li⁵ · Bing Liu^{1,4} · Dit-Yan Yeung³

Accepted: 1 October 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Micro-expression recognition (MER) remains challenging due to its subtle and fleeting nature. Existing methods often suffer from insufficient training data or rely on handcrafted features. Inspired by recent advancements in large language model fine-tuning and visual foundation models (VFMs), we propose HLoRA-MER, a novel framework that combines high-level low-rank adaptation (HLoRA) and a hierarchical fusion module (HFM). HLoRA fine-tunes the high-level layers of a VFM to capture facial muscle movement information, while HFM aggregates inter-frame and spatio-temporal features. Experiments on benchmark datasets demonstrate that HLoRA-MER outperforms state-of-the-art methods, achieving an F1-score of 84.24% and 83.07% on CASME II and SAMM, respectively, with only 197k trainable parameters. Our approach offers a promising solution for MER in both constrained and unconstrained scenarios. The code is available at https://github.com/CYF-cuber/HLoRA_MER_dinov2.

Keywords Micro-expression recognition · High-level low-rank adaptation · Hierarchical fusion

1 Introduction

Micro expression recognition (MER) is an affective computing task that has attracted increasing attention in recent years. It involves detecting and analyzing facial micro-expressions (MEs), which are very brief and subtle facial muscle actions that occur within a short duration of no more

than 500 milliseconds [1]. These MEs are often involuntary and can reveal genuine emotions or hidden intentions [2]. Due to their potential applications in healthcare and public security, the study of MER has become significant. However, it is challenging due to the subtle nature of MEs and the limited availability of large-scale labeled datasets [3, 4].

Earlier methods usually adopt hand-crafted features associated with MEs. LBP-TOP [6] and LBP-SIP [7] use co-occurrence statistics of motions. HOG [8] computes gradients of image pixels while HIGO [9] maintains the invariance of geometric and optical transformation of images. Bi-WOOF [10] and HOOF [11] apply optical flow that describes the motion pattern of each pixel. However, these methods are hard to model the characteristics of subtle and diverse MEs.

Recent solutions focus on deep neural networks. AU-GCN [12] improves MER by jointly training with facial action unit recognition task. Zhang *et al.* [13] constructed a short and long range relation based spatio-temporal transformer. In these methods, pre-extracted features like optical flow and key frames such as onset and apex frames are commonly required. Some other methods directly input raw frame images to deep networks so as to remove the above limitations. Reddy *et al.* [14] employed a 3D convolutional neural network (CNN) to capture spatial and temporal infor-

✉ Yifan Cheng
yifan_cheng@cumt.edu.cn

✉ Xiang Xiang
xex@hust.edu.cn

✉ Dit-Yan Yeung
dyyeung@cse.ust.hk

¹ School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

² MoE Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

³ Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon 999077, Hong Kong

⁴ Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou 221116, China

⁵ YouTu Lab, Tencent Incorporated, Shanghai 200233, China

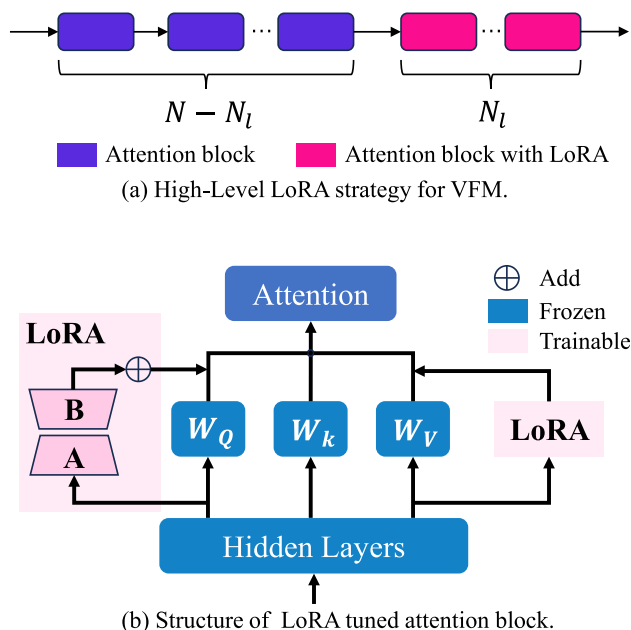


Fig. 1 Illustration of the proposed high-level LoRA strategy. **(a)** Most VFMs consist of a series of continuous attention blocks. For a VFM with N attention blocks, we fine-tune the last N_l ones. **(b)** W_Q and W_V in each fine-tuned attention block are implemented with LoRA [5], respectively. In the training stage, we keep the LoRA modules trainable while the others are frozen

mation. Xia *et al.* [15] used macro-expression recognition to facilitate MER. However, insufficient training data issue is not handled effectively.

With the development of large language models [16, 17], fine-tuning strategy for downstream tasks has been validated. In this paper, we propose a novel low-rank adaptation (LoRA) [5] based MER framework, which handles raw frame images directly and fine-tunes a visual foundation model (VFM) to suppress the data scarcity issue in MER. In particular, we first fine-tune the high-level attention blocks in a VFM, as illustrated in Fig. 1, which makes the VFM adapt to MER. Then, a hierarchical fusion module (HFM) is proposed to further fuse individual frame features from the VFM, in which an inter-frame feature aggregator and a spatio-temporal feature aggregator are utilized to model correlations in ME frames. This makes up for the shortcoming of VFMs which can only extract features from individual frames and are unable to model inter-frame and spatio-temporal information.

The main contributions of this work are threefold:

- We propose a simple yet powerful fine-tuning strategy for VFMs, which guides the attention based pre-trained models to adapt to visual downstream tasks. For the MER task, the proposed high-level LoRA leads VFM to focus more on facial regions related to MEs, which is beneficial for capturing ME motions.

- We propose an end-to-end MER framework with a hierarchical fusion module, which does not depend on pre-extracted hand-crafted features and key frames. To our knowledge, this is the first MER work based on fine-tuning VFMs.
- Extensive experiments on CASME II, SAMM, CAS(ME)³, and MEVIEW benchmarks demonstrate that our approach significantly outperforms the state-of-the-art MER methods in both constrained and unconstrained scenarios. Specifically, HLoRA-MER has 197k trainable parameters, and obtains 84.24% and 83.07% F1 results on CASME II and SAMM in terms of five categories, respectively.

2 Related work

2.1 Micro-expression recognition

Earlier MER methods focus on prior knowledge about local characteristics and motion patterns. LBP-TOP [6] considers co-occurrence statistics of motions in three directions, and LBP-SIP [7] is further proposed to avoid duplicated encoding in LBP-TOP. Another solution of hand-crafted features is based on histogram, like histogram of oriented (HOG) [8] and histogram of image gradient orientation (HIGO) [9]. Besides, optical flow describes the motion pattern of each pixel across frames, which has been widely used in MER. Bi-WOOF [10] uses it from onset frame and apex frame to represent a ME, and HOOF [11] is developed with using fuzzy membership function to collect motion directions.

However, these hand-crafted features only focus on partial characteristics associated with MEs, and often additionally rely on key frames of MEs. In recent years, considering the power of deep neural networks, Reddy *et al.* [14] introduced a 3D CNN to capture spatio-temporal information from raw image sequences for MER, and Wei *et al.* [18] proposed an attention-based magnification-adaptive network (AMAN) to magnify and focus on ME details. Since current deep networks suffer from small-scale and low-diversity ME datasets, some other approaches combine correlated auxiliary tasks or hand-crafted features with key frames. Xia *et al.* [15] jointly trained macro-expression recognition and MER with adversarial learning to align the feature distributions between macro-expressions and MEs. Zhang *et al.* [13] constructed a short and long range relation based spatio-temporal transformer with optical flow of the frames of ME videos. Additionally, recent researches on facial action unit detection [19, 20], human pose estimation [21, 22], gesture recognition [23, 24], and generative model [25, 26] have also influenced the development of MER.

All these methods suffer from insufficient training data, or dependence on hand-crafted features or key frames. In

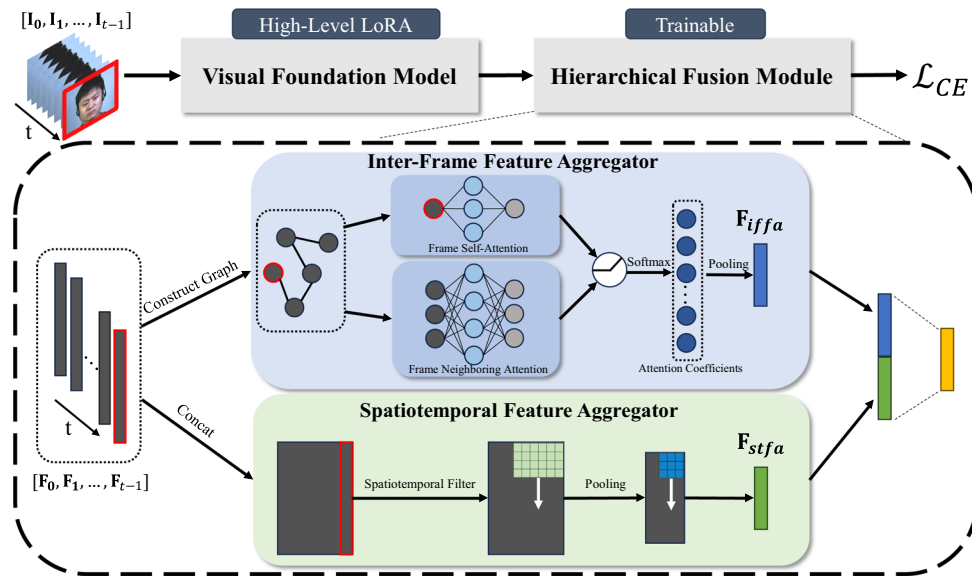


Fig. 2 The architecture of our proposed HLoRA-MER. Given a sequence of t frames, a VFM fine-tuned by HLoRA first extracts feature of each frame. Then, a hierarchical fusion module is applied to further aggregate the single-frame features, which consists of an inter-frame feature aggregator (IFFA) and a spatio-temporal feature aggregator (STFA). IFFA constructs a directed graph to model frame relation

information by calculating frame self-attention and frame neighboring attention for each node. STFA concatenates the single-frame features and simultaneously model local and temporal information by the spatio-temporal filter. Finally, the pooled inter-frame feature and spatio-temporal feature are fused to estimate the ME category

contrast, we apply VFM with excellent feature representation capability and fine-tune it to adapt to MER, which is beneficial for suppressing the problem of insufficient data. Moreover, our method handles raw image, which is more generalizable.

2.2 Parameter-efficient fine-tuning

Recently, as large models in natural language processing like BERT [16] and GPT [17] develops, VFMs trained on massive data have also demonstrated impressive performance on a wide range of tasks. CLIP [27] and BLIP [28] learn universal visual representations from multi-modal image-text data. DINOv2 [29, 30] models produce high-performance and robust visual features with self-supervised learning technology.

Parameter-efficient fine-tuning (PEFT) enables efficient model adaptation without extensively modifying the entire network. Adapter [31] compresses many visual domains in relatively small residual networks, with substantial parameter sharing between them. Visual prompt tuning [32] introduces task-specific learnable prompts in the input space, keeping the pre-trained backbone fixed, in which the tuned depth of VFM is also discussed. LoRA [5] is a flexible fine-tuning strategy, which allows training some dense layers in a neural network by optimizing rank decomposition matrices of the dense layers' change during adaptation instead, while keep-

ing the pre-trained weights frozen. Recently, an instruction tuning strategy [33] is proposed to enhance facial emotion understanding.

In our work, we fine-tune only high-level attention blocks of VFM to adapt to MER, in which each block is fine-tuned by LoRA. With further hierarchically fused features extracted by fine-tuned VFM, our method is effective at modeling transient and subtle ME motions.

3 Methodology

In this section, we elaborate the proposed method. As a classification task, MER involves distinguishing the emotion category of ME in a clip of video. Our HLoRA-MER consists of two main components, high-level LoRA (HLoRA) fine-tuned VFM and hierarchical fusion module (HFM), as shown in Fig. 2. We aim to predict ME category of an input video clip with t frames $\{I_0, I_1, \dots, I_{t-1}\}$. Firstly, the HLoRA fine-tuned VFM is applied to extract feature F_k of the k -th frame I_k in the input video, respectively. Then, the single-frame feature sequence $\{F_0, F_1, \dots, F_{t-1}\}$ is fed into the HFM to learn inter-frame feature and spatio-temporal feature, which are further fused for classification.

3.1 High-level LoRA for VFM

3.1.1 Preliminary: VFM and LoRA

VFM is always built upon Transformer [34], which is composed of continuous stacked blocks with each block including a multi-head self-attention mechanism. The self-attention is defined as

$$A = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q , K , and V denote queries and key-value pairs, respectively, and d_k is the dimension of the tokens. The basic VFM can be defined as

$$\mathcal{M}(x) = B_N(B_{N-1}(\cdots B_2(B_1(x)) \cdots)), \quad (2a)$$

$$B_i = B_i(A_i; O_i), \quad (2b)$$

where B_i is the i -th block in VFM, and A_i and O_i stand for the attention mechanism and the other components in B_i , respectively.

LoRA [5] hypothesizes the updates to the weights have a low “intrinsic rank” during adaptation. For a module with pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, it can be updated by representing the latter with a low rank decomposition $W_0 + \delta W = W_0 + BA$, where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and the rank $r \ll \min\{d, k\}$. For $h = W_0x$, the LoRA modified forward pass yields:

$$h = W_0x + \delta Wx = W_0x + BAx, \quad (3)$$

where x stands for the input of module, and W_0 is frozen while A and B are trainable during the training stage.

3.1.2 High-level LoRA

The goal of our proposed HLoRA strategy is to make VFMs adapt to downstream tasks efficiently.

In the attention mechanism, the query Q represents the current position or token that we want to attend, which can be regarded as tokens representing different patches or regions of the input image or feature map. To give more emphasis on certain regions that are more relevant for the specific task, we fine-tune the weight matrix W_Q with LoRA. The fine-tuned W_Q^L and Q^L can be expressed as

$$W_Q^L = W_Q + \delta W_Q = W_Q + B_Q A_Q, \quad (4a)$$

$$Q^L = W_Q^L x = W_Q x + B_Q A_Q x. \quad (4b)$$

Correspondingly, the value V carries the information or features associated with each token or region. V and W_V are

also fine-tuned to W_V^L and V^L :

$$W_V^L = W_V + \delta W_V = W_V + B_V A_V, \quad (5a)$$

$$V^L = W_V^L x = W_V x + B_V A_V x. \quad (5b)$$

By fine-tuning Q and V , as shown in Fig. 1(b), the low-rank adapted attention A^L is obtained:

$$A^L = \text{Attention}(Q^L, K, V^L) = \text{softmax}\left(\frac{Q^L K^T}{\sqrt{d_k}}\right)V^L. \quad (6)$$

VFMs demonstrate effective texture encoding capabilities, making them the preferred solution for several downstream tasks. As the deeper blocks capture more intricate representations of the data and provide the necessary information for downstream tasks, we believe these blocks with high-level features are more closely related to the specific task’s representations. Based on this assumption, as illustrated in Fig. 1a, HLoRA fine-tunes VFM to adapt to specific downstream tasks by fine-tuning only the last few blocks. The HLoRA fine-tuned VFM \mathcal{M}' is defined as

$$\mathcal{M}'(x) = B_N^L(\cdots B_{N-N_l+1}^L(B_{N-N_l}(\cdots B_1(x)) \cdots)), \quad (7a)$$

$$B_i^L = B_i(A_i^L; O_i), \quad (7b)$$

where N denotes the total number of attention blocks in VFM and N_l denotes the number of fine-tuned blocks.

3.2 Hierarchical fusion module

While HLoRA guides VFM to focus on relevant regions in the input image of specific task, such as facial muscle motion regions in ME frames, the single-frame feature sequence $\{\mathbf{F}_0, \mathbf{F}_1, \dots, \mathbf{F}_{t-1}\}$ with each $\mathbf{F}_i \in \mathbb{R}^d$ is passed into HFM to further aggregate inter-frame feature and spatio-temporal feature. Here, $\mathbf{F}_i = \mathcal{M}'(\mathbf{I}_i)$, and d denotes the dimension of the output of VFM.

3.2.1 Inter-frame feature aggregator

In the inter-frame feature aggregator (IFFA), we first construct a directed graph for the single-frame features. In particular, the graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertex set $\mathcal{V} = \{0, 1, \dots, t-1\}$ contains all the t single-frame features, and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Each edge feature is defined as $e_{ij} = \mathbf{F}_i - \mathbf{F}_j$, where $i \in \mathcal{V}$, and $j \in \mathcal{N}_i$. \mathcal{N}_i represents the set of indexes to neighboring frame features of \mathbf{F}_i , in which the neighbors are adaptively learned instead of being predefined as adjacent frames.

Inspired by [35], we model the inter-frame relationship by calculating self-attention and neighboring attention for

each frame, respectively. To transform the input features into higher-level features, a shared linear transformation parameterized by a weight matrix $W \in \mathbb{R}^{d' \times d}$ is applied to each node and edge feature. The self-coefficients and neighboring-coefficients are defined as

$$f'_i = W\mathbf{F}_i, \quad e'_{ij} = We_{ij}. \quad (8)$$

Each attention coefficient is obtained by fusing self-coefficients and neighboring-coefficients, and then following activation of ReLU [36] and normalization by softmax function:

$$a_{ij} = \frac{\exp(\text{ReLU}(f'_i + e'_{ij}))}{\sum_{u \in \mathcal{N}_i} \exp(\text{ReLU}(f'_i + e'_{iu}))}. \quad (9)$$

Afterward, each inter-frame feature \mathbf{F}'_i is calculated as

$$\mathbf{F}'_i = \text{ReLU} \left(\sum_{j \in \mathcal{N}_i} a_{ij} e'_{ij} \right). \quad (10)$$

Finally, the inter-frame features $\{\mathbf{F}'_0, \mathbf{F}'_1, \dots, \mathbf{F}'_{t-1}\}$ are further concatenated and aggregated to \mathbf{F}_{iffa} by pooling, as the output of IFFA. \mathbf{F}_{iffa} is defined as:

$$\mathbf{F}_{iffa} = \text{Pooling}(\text{Concat}(\mathbf{F}'_0, \mathbf{F}'_1, \dots, \mathbf{F}'_{t-1})). \quad (11)$$

3.2.2 Spatio-temporal feature aggregator

While IFFA models inter-frame relationship, spatio-temporal feature aggregator (STFA) extracts the spatial and temporal characteristics simultaneously. Single-frame features are first concatenated to $\mathbf{F}^c \in \mathbb{R}^{d \times t}$. Then, we apply a spatio-temporal filter $k \in \mathbb{R}^{t/2 \times t/2}$ using a filter operator \odot , to extract features from several frames with a rich spatio-temporal receptive field. Finally, a nonlinear function $h(\cdot)$ and pooling are exploited for activation and aggregation, respectively. Without loss of generality, the spatio-temporal feature \mathbf{F}_{stfa} is defined as

$$\mathbf{F}_{stfa} = \text{Pooling}(h(\mathbf{F}^c \odot k)). \quad (12)$$

In our experiments, spatio-temporal filter is concisely implemented by choosing \odot and $h(\cdot)$ as convolution and ReLU, respectively.

The output of hierarchical fusion module is defined as

$$Y_{HFM} = g(\mathbf{F}_{iffa} \oplus \mathbf{F}_{stfa}), \quad (13)$$

where $g(\cdot)$ is a linear layer and \oplus denotes a fusion operator. In our experiments, \oplus is set as concatenation.

Table 1 The number of videos for each ME category in CASME II [3], SAMM [4], and CAS(ME)³ [37]

Class	Dataset		
	CASME II	SAMM	CAS(ME) ³
Happiness	32	26	64
Anger	–	57	70
Contempt	–	12	–
Disgust	63	9	281
Fear	2	8	93
Repression	27	–	–
Surprise	28	15	201
Sadness	4	6	64
Others	99	26	170

The used five categories of CASME II and SAMM, as well as used seven categories of CAS(ME)³ are highlighted with its number in bold. “–” denotes this category is not included

Table 2 The number of videos for each ME category in CASME II [3], SAMM [4] and MEVIEW [38], in terms of three-classes evaluation

Class	Dataset		
	CASME II	SAMM	MEVIEW
Positive	32	26	6
Surprise	28	15	7
Negative	96	92	18

3.2.3 Loss function

The MER loss is defined as a cross-entropy loss:

$$\mathcal{L}_{CE} = - \sum_{j=0}^{m-1} p_j \log(\hat{p}_j), \quad (14)$$

where m denotes the number of ME categories, and \hat{p}_j denotes the predicted probability that the video sample is in the j -th category. p_j denotes the ground-truth probability, which is 1 if the video sample is in the j -th category and is 0 otherwise.

4 Experiments

4.1 Datasets and settings

4.1.1 Datasets

We evaluate our method on four popular spontaneous ME datasets, CASME II [3], SAMM [4], CAS(ME)³ [37], and MEVIEW [38].

Table 3 Comparison with state-of-the-art methods on CASME II [3] and SAMM [4] for five categories

Method	Paper	Type	CASME II		SAMM	
			ACC	F1	ACC	F1
SparseSampling [39]	TAFFC'17	NDL	49.00	51.00	–	–
Bi-WOOF [10]	SPIC'18	NDL+KF	58.85	61.00	–	–
HIGO+Mag [9]	TAFFC'18	NDL	67.21	–	–	–
FHOFO [11]	TAFFC'19	NDL	56.64	52.48	–	–
DSSN [40]	ICIP'19	DL+PF+KF	70.78	72.97	57.35	46.44
Graph-TCN [41]	MM'20	DL+RI+KF	73.98	72.46	75.00	69.85
MicroNet [42]	MM'20	DL+RI+KF	75.60	70.10	74.10	73.60
LGCcon [43]	TIP'21	DL+PF+KF	62.14	60.00	35.29	23.00
AU-GCN [12]	CVPRW'21	DL+PF+KF	74.27	70.47	74.26	70.45
KFC [44]	ICME'21	DL+PF+KF	72.76	73.75	63.24	57.09
GEME [45]	Neurocomputing'21	DL+PF	75.20	73.54	55.88	45.38
MERSiamC3D [46]	Neurocomputing'21	DL+PF+KF	81.89	83.00	68.75	64.00
MiNet&MaNet [15]	IJCAI'21	DL+RI	79.90	75.90	<u>76.70</u>	<u>76.40</u>
MER-Supcon [47]	PRL'22	DL+PF+KF	73.58	72.86	67.65	62.51
AMAN [18]	ICASSP'22	DL+RI	75.40	71.25	68.85	66.82
SLSTT [13]	TAFFC'22	DL+PF	75.81	75.30	72.39	64.00
Dynamic [48]	TAFFC'22	DL+RI+KF	72.61	67.00	–	–
Deep3DCANN [49]	INS'23	DL+RI	<u>86.00</u>	<u>84.00</u>	–	–
HLoRA-MER	Ours	DL+RI	86.34	84.24	83.94	83.07

DL, NDL, PF, RI, and KF denote deep learning-based methods, non-deep learning based methods, pre-extracted hand-crafted features, raw images, and key frames, respectively. “–” denotes the result is not reported in its paper. The best results are highlighted in bold, and the second best results are highlighted by an underline

- **CASME II** consists of 255 videos captured from 26 subjects in steady and high-intensity illumination. To elicit the MEs, subjects are induced to experience a high arousal with motivations to disguise. Each video is recorded with the frame rate of 200 frames per second (FPS) and the frame size of 280×340 . Consistent with most previous methods [15, 41], only samples of happiness, disgust, repression, surprise, and others categories are utilized.
- **SAMM** includes 159 videos at 200 FPS from 29 subjects, which are collected using gray-scale cameras in constrained lighting conditions without flickering. The MEs are elicited from stimuli tailored to each subject. Similar to the previous works [15, 41], five categories of happiness, anger, contempt, surprise, and others are selected for evaluation.
- **CAS(ME)³** [37] includes 1109 MEs and 3,490 macro-expressions from 100 subjects with the resolution of 1280×720 . Each subject is requested to watch 13 emotional stimuli and keep their faces expressionless. We conduct experiments on 943 ME videos from Part A, using seven categories (happiness, disgust, surprise, anger, fear, sadness, and others).
- **MEVIEW** is collected in unconstrained scenarios, mostly from poker games and TV interviews, in which 31 ME

videos from 16 subjects are captured. We select ME categories of positive, negative, and surprise for three-classes evaluation.

All video clips in four datasets are labeled with MEs, and the number of samples for each ME category in these datasets is summarized in Tables 1 and 2.

4.1.2 Evaluation metrics

Leave-one-subject-out (LOSO) cross-validation is applied in the single database evaluation, in which each subject is taken as the test set in turn while the remaining subjects are taken as the training set. Two popular metrics, accuracy (ACC) and unweighted F1-score (F1) are reported, in which the latter is defined as

$$F1 = \frac{1}{m} \sum_{j=0}^{m-1} \frac{2TP_j}{2TP_j + FP_j + FN_j}. \quad (15)$$

Here m is the total number of ME categories, N_j denotes the number of samples presenting the j -th ME category, N denotes the total number of samples, and TP_j , FP_j , and FN_j denote the number of true positives, false positives, and false negatives for the j -th category, respectively.

Table 4 Comparison with state-of-the-art methods on CASME II [3] and SAMM [4] for three categories

Method	Paper	Type	CASME II		SAMM	
			ACC	F1	ACC	F1
OFF-ApexNet [50]	SPIC'19	DL+PF+KF	88.28	86.97	68.18	54.23
STRCN-G [51]	TMM'19	DL+RI	80.30	74.70	78.60	<u>74.10</u>
AU-GACN [52]	MM'20	DL+RI	71.20	35.50	70.20	43.30
MER-Supcon [47]	PRL'22	DL+PF+KF	<u>89.65</u>	<u>88.06</u>	<u>81.20</u>	71.25
HLoRA-MER	Ours	DF+RI	92.31	90.94	90.83	86.71

The best results are highlighted in bold, and the second best results are highlighted by an underline

Table 5 Comparison with state-of-the-art methods on CAS(ME)³ [37]

Method	Paper	Type	F1	UAR
AlexNet [53]	NeurIPS'12	DL+KF	25.70	26.34
STSTNet [54]	FG'19	DL+PF+KF	37.95	37.92
RCN [55]	TIP'20	DL+PF+KF	39.28	38.93
FR [56]	PR'22	DL+PF+KF	34.93	34.13
HTNet [57]	Neurocomputing'24	DL+PF+KF	57.67	54.15
HLoRA-MER	Ours	DL+RI	58.96	55.23

The best results are highlighted in bold

The results of previous methods except for HTNet [57] are reported by [37]

To investigate the generalization ability of our method, we also perform holdout-database evaluation (HDE) [58] with CASME II and SAMM datasets, in which one dataset is used for training while the other dataset is used for testing. Following the settings in previous approaches [6, 59, 60], we report two metrics of weighted average recall (WAR) and unweighted average recall (UAR):

$$WAR = \sum_{j=0}^{m-1} \frac{TP_j}{N}, \quad UAR = \frac{1}{m} \sum_{j=0}^{m-1} \frac{TP_j}{N_j}. \quad (16)$$

In the following sections, ACC, F1, WAR, and UAR results are all reported in percentages, in which % is omitted for simplicity.

4.1.3 Implementation details

In our experiments, we extract a video clip with t frames as the input of our HLoRA-MER by uniformly space sampling from the raw video. Each frame image is aligned to $3 \times 224 \times 224$ via similarity transformation, in which facial shape is preserved without changing the ME. During training, each image is further horizontally flipped to improve the diversity of training data.

We apply the giant version of DINOv2 [29, 30] with 1100M parameters as our basic VFM. Our fine-tuned VFM is parameter-efficient with only 197k trainable parameters, as 0.18% of the basic VFM. The low-rank r is set to 16 in all fine-tuned layers, and N_l is set as 2. HLoRA-MER is implemented based on PyTorch [61], with a solver of Adam

[62], a fixed learning rate of 1×10^{-3} , and a mini-batch size of 36. The number of frames in the input video clip is set as $t = 8$. All the experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.2 Comparison with state-of-the-art methods

We compare our method HLoRA-MER against state-of-the-art MER methods under the same evaluation setting. These methods can be categorized into non-deep learning (NDL)-based methods and deep learning (DL)-based methods. The latter can be further divided into pre-extracted feature (PF)-based methods and raw image (RI)-based methods, based on the type of model input. Also, some methods depend on key frames (KF) including onset, apex, and offset frames of MEs.

4.2.1 Single database evaluation

Tables 3 and 4 show the comparison results on single datasets of CAMSE II and SAMM for five categories and three categories, respectively. It can be observed that DL-based methods often outperform NDL-based methods, attributed to the power of deep networks. In five emotion classes, HLoRA-MER achieves 86.34 ACC and 84.24 F1-score on CASME II, compared to Deep3DCANN [49]'s 86.00 and 84.00. Attaining 83.94 ACC and 83.07 F1-score on SAMM, HLoRA-MER realizes almost 7 gains over previous best MiNet&MaNet [15] on both metrics. Similar improvement is seen in three categories. HLoRA-MER achieves 2.66 and 2.88 increases over MER-Supcon [47] on CASME II, and outperforms

Table 6 Comparison with state-of-the-art methods on MEVIEW [38]

Method	Paper	Type	ACC	F1
MicroExpSTCNN [14]	IJCNN'19	DL+RI	67.64	58.61
AU-GCN [12]	CVPRW'21	DL+PF+KF	70.96	67.69
SLSTT [13]	TAFFC'22	DL+PF	74.19	70.00
HLoRA-MER	Ours	DL+RI	83.87	80.51

The best results are highlighted in bold

Other methods are implemented using its released code

Table 7 Comparison with state-of-the-art methods for three ME categories (happiness, surprise, and others) of HDE. Avg. denotes the average results of two cross-dataset evaluations

Method	Paper	Type	C→S		S→C		Avg.	
			WAR	UAR	WAR	UAR	WAR	UAR
LBP-TOP [6]	TPAMI'07	NDL	33.8	32.7	23.2	31.6	28.5	32.2
3DHOG [63]	ICDP'09	NDL	35.3	26.9	37.3	18.7	36.3	22.8
MDMO [59]	TAFFC'16	NDL	44.1	34.9	26.5	<u>34.6</u>	35.3	34.8
Peng <i>et al.</i> [64]	FG'18	DL+RI+KF	48.5	38.2	38.4	32.2	43.5	35.2
Khor <i>et al.</i> [60]	FG'18	DL+PF+KF	54.4	<u>44.0</u>	<u>57.8</u>	33.7	<u>56.1</u>	<u>38.9</u>
HLoRA-MER	Ours	DL+RI	<u>52.0</u>	47.6	62.3	38.5	57.2	43.1

The best results are highlighted in bold, and the second best results are highlighted by an underline

The results of previous methods are reported by [65]. C→S denotes training on CASME II [3] and testing on SAMM [4], and vice versa

Table 8 Results of different fine-tuning strategies on SAMM [4]

Tuning strategy	Params	ACC	F1
Full Fine-Tuning	1100M	81.75	79.34
HLoRA	197k	83.94	83.07

The best results are highlighted in bold

the prior leading methods by a large margin on SAMM. Unlike some recent state-of-the-art methods like MERSi-amC3D [46] rely on hand-crafted features or key frames, our HLoRA-MER directly processes raw images and performs better on both benchmarks, because of the sensitive ME feature extraction capability of HLoRA fine-tuned VFM and the strong fusion capability of HFM.

Table 5 presents the comparison results on CAS(ME)³. It can be observed that our HLoRA-MER outperforms the prior leading methods, especially achieving 1.29 and 1.08 increases over HTNet [57] in terms of F1 and UAR, respectively. Note that HLoRA-MER is the only method directly processing raw images, independent of key frames or any other pre-extracted features. Without relying on additional information, HLoRA-MER still obtains the best performance.

Moreover, Table 6 illustrates the comparison results on the in-the-wild dataset MEVIEW. Due to the powerful robustness brought by the VFM fine-tuning strategy, our approach achieves the best performance even in complex and variable unconstrained scenarios. The results on various datasets demonstrate the effectiveness and scalability of our approach.

Table 9 ACC and F1 results for different variants of HLoRA-MER on SAMM [4]

Method	ACC	F1
HLoRA-MER	83.94	83.07
HLoRA-MER w/o STFA	80.29	78.74
HLoRA-MER w/o IFFA	78.10	76.82
HLoRA-MER w/o HFM	75.91	74.44

The best results are highlighted in bold

4.2.2 Holdout-database evaluation

Table 7 presents the HDE results of different works, in which the common three ME categories of happiness, surprise, and others for two datasets are used. It can be observed that our approach achieves the best average performance especially for the UAR metric, which demonstrates the strong generalization ability of our HLoRA-MER. This mainly attributes to excellent robustness of HLoRA fine-tuned VFM and strong aggregating capacity of HFM.

4.3 Ablation study

Our ablation studies aim to answer the follow questions: **Q1**. Is the HLoRA fine-tuning strategy better than full tuning? **Q2**. Is the number of fine-tuned attention blocks the more the better? **Q3**. How does each component of HFM performs on MER? These experiments are all evaluated on SAMM dataset in terms of five categories.

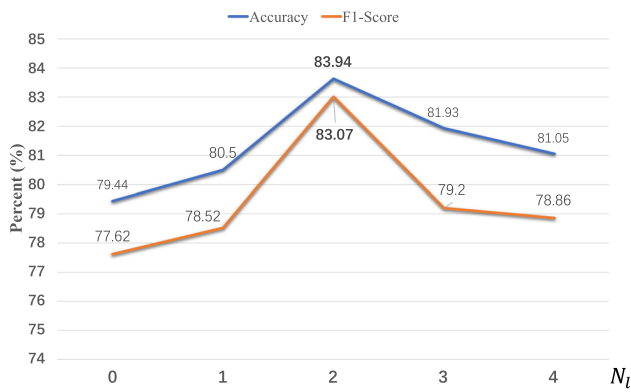


Fig. 3 Impact of the number of HLoRA fine-tuned blocks

4.3.1 Effectiveness of fine-tuning

We conduct experiments to evaluate different fine-tuning strategies for VFM, as shown in Table 8. The results demonstrated the superiority of HLoRA fine-tuning strategy over directly tuning full VFM layers. This is because full tuning

strategy may cause adverse effects to the pattern recognition abilities already embedded of VFM. Moreover, the fine-tuning strategy preserves the generalization ability of VFM and helps to prevent overfitting by modifying only a subset of parameters, especially in ME scenarios with limited training data.

4.3.2 Effectiveness of HLoRA

We explore the effect of different setting of hype-parameter N_l for MER. As illustrated in Fig. 3, HLoRA fine-tuned VFM with $N_l \geq 1$ performs better than VFM without fine-tuning ($N_l = 0$). This is because HLoRA improves VFM's compatibility to MER, in which the fine-tuned model can better capture relevant patterns and features those are crucial for performance. Moreover, with N_l increases, the result is not always better. Accuracy and F1-score peak at $N_l = 2$, which demonstrates fine-tuning only the last few blocks of VFM exactly benefits specific tasks.

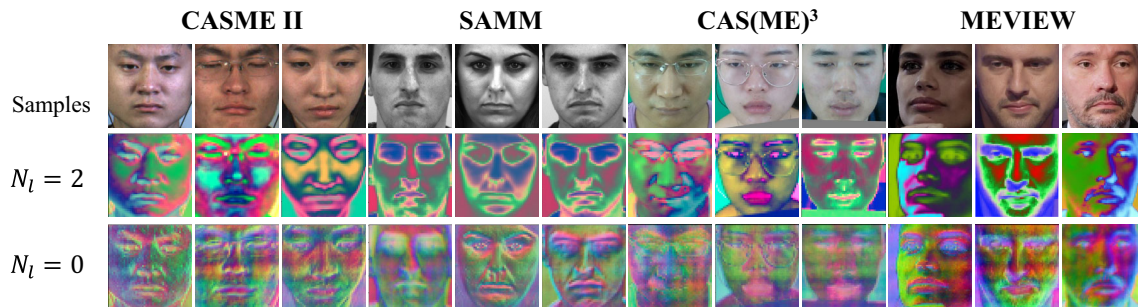


Fig. 4 Visualization of the first PCA component for methods with and without HLoRA fine-tuning on samples from CASME II [3], SAMM [4], CAS(ME)³ [37], and MEVIEW [38]

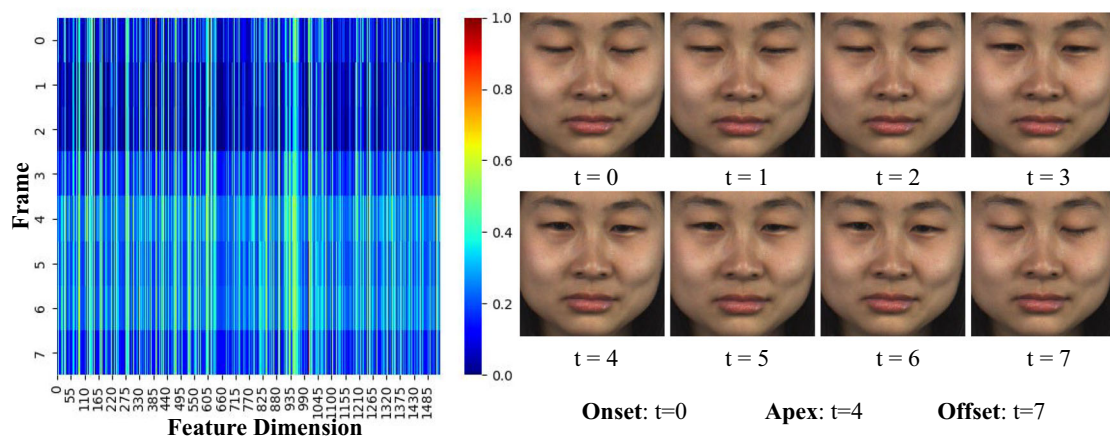


Fig. 5 Visualization of the graph attention maps of IFFA. The left part shows the heatmap of one video from CASME II [3], and the right part is the raw image of each frame




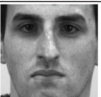
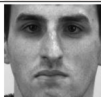

	Onset	Apex	Offset	GT	Prediction
CASME II 03_EP18_06				Disgust	Repression
SAMM 007_6_1				Happiness	Anger

Fig. 6 Failure cases of our HLoRA-MER on CASME II [3] and SAMM [4]. “GT” denotes ground truth

4.3.3 Effectiveness of HFM

Here, we evaluate the main components of HFM, including IFFA and STFA, in which N_l is set to 2. The results of different variants of HLoRA-MER are presented in Table 9. When removing the whole HFM and utilizing a linear layer instead, the ACC and F1 results of HLoRA-MER w/o HFM are remarkably decreased to 75.91 and 74.44, respectively. If we remain either IFFA or STFA, the results improve a lot. However, the performance is still worse than HLoRA-MER. This demonstrates the effectiveness of HFM with inter-frame and spatio-temporal feature learning.

4.4 Visualization

4.4.1 HLoRA

We show the results of principal component analysis (PCA) performed on the patch features of individual frames extracted by VFM. Figure 4 visualizes the first PCA component of HLoRA fine-tuned VFM with $N_l = 2$ and VFM without fine-tuning ($N_l = 0$) on several video frames from CASME II, SAMM, CAS(ME)³, and MEVIEW. We can observe that VFM without fine-tuning can show facial texture features, but it does not highlight the key facial regions related to MEs such as eyes and mouth corners. In contrast, our method enhances the representation of these facial regions in both constrained CASME II, SAMM, CAS(ME)³ and unconstrained MEVIEW, which demonstrates the effectiveness of our proposed HLoRA.

4.4.2 IFFA

Figure 5 illustrates the insights of our proposed IFFA, in which the heatmap visualizes the graph attentions in IFFA for eight equally spaced frames. We can observe that frames 3, 4, 5, and 6 obtain higher attention weights than other frames. Since frame 4 is the apex frame, and frames 3, 5 and 6 are the closest, these four frames have higher ME intensities than others. This demonstrates that IFFA can analyze the relationships between frames and focus on frames with higher expression intensity, which possess strong features for micro-expression recognition.

4.5 Limitations

According to the above experiments, our method significantly outperforms the previous works. However, there are a few failure cases, as shown in Fig. 6. We notice that mistakenly recognized videos are very challenging because of their extremely subtle motions. These samples often have very low ME intensities even in apex frames, which are difficult to distinguish their emotion categories. The recognition of extremely low-intensity MEs remains a major challenge in the MER task.

5 Conclusion

In this paper, we have proposed a LoRA-based fine-tuning strategy for VFMs and have applied it to MER to extract feature from each single frame. Besides, we have developed a hierarchical fusion module to aggregate inter-frame and spatio-temporal features. Our framework does not rely on pre-extracted hand-crafted features and key frames, which is a promising solution to MER with good applicability in both constrained and unconstrained scenarios.

We have compared our method with state-of-the-art works on the challenging CASME II, SAMM, CAS(ME)³, and MEVIEW benchmarks. It is shown that our method achieves competitive results to previous works for both single dataset evaluation and cross-dataset evaluation. Besides, we have conducted an ablation study which indicates that main components in our framework are all beneficial for MER. Moreover, the visualization results demonstrate that our method can guide VFM to capture facial subtle muscle actions related to MEs.

Author Contributions Material preparation, data collection and analysis were mostly performed by Zhiwen Shao. The HLoRA-MER framework was originally proposed by Zhiwen Shao and was improved by Yifan Cheng. Xiang Xiang and Dit-Yan Yeung, leaders of this project, delved into specific discussions of the feasibility. Yong Zhou, Jian Li, and Bing Liu were involved in partial experimental designs and paper revision. The manuscript was written by Zhiwen Shao. All authors read and approved the manuscript.

Funding This work was supported in part by the National Natural Science Foundation of China under Grants 62472424 and 62106268, in part by the Opening Fund of Key Laboratory of Image Processing and Intelligent Control (Huazhong University of Science and Technology), Ministry of Education, China, in part by the Natural Science Foundation of Hubei Province under Grant 2022CFB823, in part by the China Postdoctoral Science Foundation under Grant 2023M732223, and in part by the Hong Kong Scholars Program under Grant XJ2023037. It was also supported in part by the National Natural Science Foundation of China under Grants 62272461 and 62276266, and in part by the HUST Independent Innovation Research Fund under Grant 2021XXJS096.

Data Availability This study uses four publicly available ME datasets, including CASME II, SAMM, CAS(ME)³, and MEVIEW. They can be downloaded at <http://casme.psych.ac.cn/casme/c2>, <https://helward>.

mmu.ac.uk/STAFF/M.Yap/dataset.php, <http://casme.psych.ac.cn/casme/c4>, and <https://cmp.felk.cvut.cz/~cechj/ME>, respectively.

Declarations

Conflict of interest Zhiwen Shao has been serving as an editorial board member in TVC journal. Except for this conflict, the authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yan, W.J., Wu, Q., Liang, J., Chen, Y.H., Fu, X.: How fast are the leaked facial expressions: The duration of micro-expressions. *J. Nonverbal Behav.* **37**(4), 217–230 (2013)
2. Ekman, P.: *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage* (Revised Edition). WW Norton & Company, New York (2009)
3. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **9**(1), e86,041 (2014)
4. Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H.: Samm: A spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput.* **9**(1), 116–129 (2016)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: *International Conference on Learning Representations* (2022)
6. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 915–928 (2007)
7. Wang, Y., See, J., Phan, R.C.W., Oh, Y.H.: Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PLoS ONE* **10**(5), e0124,674 (2015)
8. Davison, A.K., Yap, M.H., Lansley, C.: Micro-facial movement detection using individualised baselines and histogram-based descriptors. In: *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1864–1869. IEEE (2015)
9. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikainen, M.: Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Trans. Affect. Comput.* **9**(4), 563–577 (2018)
10. Liong, S.T., See, J., Wong, K., Phan, R.C.W.: Less is more: Micro-expression recognition from video using apex frame. *Signal Process. Image Commun.* **62**, 82–92 (2018)
11. Happy, S., Routray, A.: Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Trans. Affect. Comput.* **10**(3), 394–406 (2019)
12. Lei, L., Chen, T., Li, S., Li, J.: Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1571–1580. IEEE (2021)
13. Zhang, L., Hong, X., Arandjelović, O., Zhao, G.: Short and long range relation based Spatio-temporal transformer for micro-expression recognition. *IEEE Trans. Affect. Comput.* **13**(4), 1973–1985 (2022)
14. Reddy, S.P.T., Karri, S.T., Dubey, S.R., Mukherjee, S.: Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. In: *International Joint Conference on Neural Networks*, pp. 1–8. IEEE (2019)
15. Xia, B., Wang, S.: Micro-expression recognition enhanced by macro-expression from spatial-temporal domain. In: *International Joint Conference on Artificial Intelligence*, pp. 1186–1193 (2021)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
17. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in Neural Information Processing Systems*, pp. 1877–1901. Curran Associates, Inc (2020)
18. Wei, M., Zheng, W., Zong, Y., Jiang, X., Lu, C., Liu, J.: A novel micro-expression recognition approach using attention-based magnification-adaptive networks. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2420–2424. IEEE (2022)
19. Liu, X., Yuan, K., Niu, X., Shi, J., Yu, Z., Yue, H., Yang, J.: Multi-scale promoted self-adjusting correlation learning for facial action unit detection. *IEEE Trans. Affect. Comput.* (2024)
20. Yuan, K., Yu, Z., Liu, X., Xie, W., Yue, H., Yang, J.: Auformer: Vision transformers are parameter-efficient facial action unit detectors. In: *European Conference on Computer Vision*. Springer (2024)
21. Kamel, A., Liu, B., Li, P., Sheng, B.: An investigation of 3d human pose estimation for learning tai chi: A human factor perspective. *Int. J. Human-Comput. Interact.* **35**(4–5), 427–439 (2019)
22. Aouaidjia, K., Sheng, B., Li, P., Kim, J., Feng, D.D.: Efficient body motion quantification and similarity evaluation using 3-d joints skeleton coordinates. *IEEE Trans. Syst. Man Cybern. Syst.* **51**(5), 2774–2788 (2019)
23. Zeghoud, S., Ali, S.G., Ertugrul, E., Kamel, A., Sheng, B., Li, P., Chi, X., Kim, J., Mao, L.: Real-time spatial normalization for dynamic gesture classification. *The Visual Computer* pp. 1–13 (2022)
24. Karambakhsh, A., Kamel, A., Sheng, B., Li, P., Yang, P., Feng, D.D.: Deep gesture interaction for augmented anatomy learning. *Int. J. Inf. Manage.* **45**, 328–336 (2019)
25. Li, P., Sheng, B., Chen, C.P.: Face sketch synthesis using regularized broad learning system. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(10), 5346–5360 (2021)
26. Yao, J., Chen, J., Niu, L., Sheng, B.: Scene-aware human pose generation using transformer. In: *ACM International Conference on Multimedia*, pp. 2847–2855 (2023)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*, pp. 8748–8763. PMLR (2021)
28. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International Conference on Machine Learning*, pp. 12,888–12,900. PMLR (2022)
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., et al.: Dinov2: Learning robust visual features without supervision (2023). [arXiv:2304.07193](https://arxiv.org/abs/2304.07193)
30. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers (2023). [arXiv:2309.16588](https://arxiv.org/abs/2309.16588)
31. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. *Advances in Neural Information Processing Systems*, pp. 506–516. Curran Associates, Inc (2017)
32. Jia, M., Tang, L., Chen, B.C., Cardie, C., Longie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: *European Conference on Computer Vision*, pp. 709–727. Springer (2022)
33. Xing, B., Yu, Z., Liu, X., Yuan, K., Ye, Q., Xie, W., Yue, H., Yang, J., Kälviäinen, H.: Emo-llama: Enhancing facial emo-

- tion understanding with instruction tuning (2024). arXiv preprint [arXiv:2408.11424](https://arxiv.org/abs/2408.11424)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008. Curran Associates, Inc. (2017)
 35. Chen, C., Fragonara, L.Z., Tsourdos, A.: Gapointnet: Graph attention based point neural network for exploiting local feature of point cloud. *Neurocomputing* **438**, 122–132 (2021)
 36. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: *International Conference on Machine Learning*, pp. 807–814. PMLR (2010)
 37. Li, J., Dong, Z., Lu, S., Wang, S.J., Yan, W.J., Ma, Y., Liu, Y., Huang, C., Fu, X.: Cas(me)³: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(3), 2782–2800 (2022)
 38. Husak, P., Cech, J., Matas, J.: Spotting facial micro-expressions “in the wild”. *Computer Vision Winter Workshop*, pp. 1–9 (2017)
 39. Le Ngo, A.C., See, J., Phan, R.C.W.: Sparsity in dynamics of spontaneous subtle emotions: analysis and application. *IEEE Trans. Affect. Comput.* **8**(3), 396–411 (2017)
 40. Khor, H.Q., See, J., Liong, S.T., Phan, R.C., Lin, W.: Dual-stream shallow networks for facial micro-expression recognition. In: *IEEE International Conference on Image Processing*, pp. 36–40. IEEE (2019)
 41. Lei, L., Li, J., Chen, T., Li, S.: A novel graph-tcn with a graph structured representation for micro-expression recognition. In: *ACM International Conference on Multimedia*, pp. 2237–2245 (2020)
 42. Xia, B., Wang, W., Wang, S., Chen, E.: Learning from macro-expression: a micro-expression recognition framework. In: *ACM International Conference on Multimedia*, pp. 2936–2944 (2020)
 43. Li, Y., Huang, X., Zhao, G.: Joint local and global information learning with single apex frame detection for micro-expression recognition. *IEEE Trans. Image Process.* **30**, 249–263 (2021)
 44. Su, Y., Zhang, J., Liu, J., Zhai, G.: Key facial components guided micro-expression recognition based on first & second-order motion. In: *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE (2021)
 45. Nie, X., Takalkar, M.A., Duan, M., Zhang, H., Xu, M.: Geme: Dual-stream multi-task gender-based micro-expression recognition. *Neurocomputing* **427**, 13–28 (2021)
 46. Zhao, S., Tao, H., Zhang, Y., Xu, T., Zhang, K., Hao, Z., Chen, E.: A two-stage 3d cnn based learning method for spontaneous micro-expression recognition. *Neurocomputing* **448**, 276–289 (2021)
 47. Zhi, R., Hu, J., Wan, F.: Micro-expression recognition with supervised contrastive learning. *Pattern Recogn. Lett.* **163**, 25–31 (2022)
 48. Sun, B., Cao, S., Li, D., He, J., Yu, L.: Dynamic micro-expression recognition using knowledge distillation. *IEEE Trans. Affect. Comput.* **13**(2), 1037–1043 (2022)
 49. Thuseethan, S., Rajasegarar, S., Yearwood, J.: Deep3dcann: A deep 3dcnn-ann framework for spontaneous micro-expression recognition. *Inf. Sci.* **630**, 341–355 (2023)
 50. Gan, Y.S., Liong, S.T., Yau, W.C., Huang, Y.C., Tan, L.K.: Off-apexnet on micro-expression recognition system. *Signal Process. Image Commun.* **74**, 129–139 (2019)
 51. Xia, Z., Hong, X., Gao, X., Feng, X., Zhao, G.: Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Trans. Multimed.* **22**(3), 626–640 (2019)
 52. Xie, H.X., Lo, L., Shuai, H.H., Cheng, W.H.: Au-assisted graph attention convolutional network for micro-expression recognition. In: *ACM International Conference on Multimedia*, pp. 2871–2880. ACM (2020)
 53. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105. Curran Associates, Inc. (2012)
 54. Liong, S.T., Gan, Y.S., See, J., Khor, H.Q., Huang, Y.C.: Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In: *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–5. IEEE (2019)
 55. Xia, Z., Peng, W., Khor, H.Q., Feng, X., Zhao, G.: Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Trans. Image Process.* **29**, 8590–8605 (2020)
 56. Zhou, L., Mao, Q., Huang, X., Zhang, F., Zhang, Z.: Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recogn.* **122**, 108,275 (2022)
 57. Wang, Z., Zhang, K., Luo, W., Sankaranarayanan, R.: Htnet for micro-expression recognition. *Neurocomputing* **602**, 128196 (2024)
 58. Yap, M.H., See, J., Hong, X., Wang, S.J.: Facial micro-expressions grand challenge 2018 summary. In: *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 675–678. IEEE (2018)
 59. Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X.: A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affect. Comput.* **7**(4), 299–310 (2016)
 60. Khor, H.Q., See, J., Phan, R.C.W., Lin, W.: Enriched long-term recurrent convolutional network for facial micro-expression recognition. In: *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 667–674. IEEE (2018)
 61. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., et al.: Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*, pp. 8024–8035. Curran Associates, Inc. (2019)
 62. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: *International Conference on Learning Representations* (2015)
 63. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In: *International Conference on Imaging for Crime Detection and Prevention*, pp. 1–6 (2009)
 64. Peng, M., Wu, Z., Zhang, Z., Chen, T.: From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In: *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 657–661. IEEE (2018)
 65. Yap, M.H., See, J., Hong, X., Wang, S.J.: Facial micro-expressions grand challenge 2018 summary. In: *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 675–678. IEEE (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Zhiwen Shao is currently an Associate Professor with the China University of Mining and Technology, as well as a Postdoctoral Fellow with the Shanghai Jiao Tong University and the Hong Kong University of Science and Technology. He received the B.Eng. degree and the Ph.D. degree in Computer Science and Technology from the Northwestern Polytechnical University and the Shanghai Jiao Tong University in 2015 and 2020, respectively. His research interests lie in computer

vision and affective computing. He has served as an Area Chair for ACM MM 2024, an Associate Editor for TVC, and a Publication Chair for CGI 2023.



Xiang Xiang received the B.S. degree from Wuhan University, China, in 2009, the M.S. degree from Institute of Computing Technology, Chinese Academy of Sciences, China, in 2012, the M.S.E. and Ph.D. degrees from Johns Hopkins University, USA, in 2014 and 2018, respectively, all in computer science. He is currently an Associate Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China. His research interests are computer vision

and machine learning with a focus on representation learning for video understanding, facial analysis, visual sensing and biomedical imaging.



Yifan Cheng received the B.Eng. degree from the School of Computer Science and Technology, China University of Mining and Technology, China in 2022, where he is currently pursuing the M.S. degree under the supervision of Dr. Zhiwen Shao. His current research interest lies in facial expression recognition.



Jian Li received his M.S. degree in 2019 from the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He is currently a Senior Researcher at Tencent YouTu Lab. His research interests include computer vision and artificial intelligence.



Yong Zhou received the M.S. and Ph.D. degrees in Control Theory and Control Engineering from the China University of Mining and Technology, China, in 2003 and 2006, respectively. He is currently a Professor with the School of Computer Science and Technology, China University of Mining and Technology, China. His research interests include machine learning, intelligence optimization, and data mining. He has been serving as an Associate Editor for ACM TOMM.



Bing Liu received the B.S., M.S., and Ph.D. degrees in 2002, 2005, and 2013, respectively, from the China University of Mining and Technology, Xuzhou, China. He is currently an Associate Professor at the School of Computer Science and Technology, China University of Mining and Technology, China. His current research interests include natural language processing, image understanding, and deep learning.



Dit-Yan Yeung received his B.Eng. degree in Electrical Engineering and MPhil degree in Computer Science from the University of Hong Kong and Ph.D. degree in Computer Science from the University of Southern California. He started his academic career as an Assistant Professor at the Illinois Institute of Technology in Chicago. He then joined the Hong Kong University of Science and Technology where he is now a Chair Professor at the Department of Computer science and Engineer-

ing. His research interests are primarily in computational and statistical approaches to machine learning and artificial intelligence. He is also interested in developing novel machine learning models for various applications particularly in computer vision, education, and recommender systems.