HKAFER: Achieve Visual Parameter-Efficient Fine-Tuning via Heterogeneous Kronecker Adaptation for Facial Expression Recognition

Yu Gao[1], Haoyu Ji[2], Zhiyong Wang[1, 2], Wenze Huang[2], Qian Dong[2], Zhihao Yang[2], Xueting Liu[3], Weihong Ren[1, 2], Honghai Liu[1, 2*]

[1]School of Biomedical Engineering, Harbin Institute of Technology, Shenzhen
[2]State Key Laboratory of Robotics and System, School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen
[3]The Department of Electronic and Electrical Engineering, Southern University of Science and Technology
{25B363008, 190310321, 190320418, 24B953002}@stu.hit.edu.cn, jihaoyu1224@gmail.com,
{wangzhiyong, renweihong, honghai.liu}@hit.edu.cn, liuxt2023@mail.sustech.edu.cn,

Abstract

Facial Expression Recognition (FER) seeks to classify affective states from facial images, which remains a challenging problem due to variations in real-world conditions. FER task becomes particularly complex when handling unconstrained environments characterized by partial occlusions, different head poses, and so on. To address the above problems, current approaches rely on extensive learnable parameters and complex model architectures, which inevitably lead to overfitting and cause the FER model to focus on non-discriminative facial regions. In this work, we propose an HKAFER model that can adaptively enhance visual expression representations through efficiently fine-tuning the image encoder in large Visual Foundation Models (VFMs) and Vision-Language Models (VLMs). Specifically, we establish Heterogeneous Kronecker Adaptation (HeKA), which consists of multi-scale adapters based on Kronecker product in a parallel manner, offering significantly diverse subspaces to learn the incremental matrices. Besides, we also propose Dual-Branch Interactive Router (DBIR) to dynamically assign the weights of adapters, which promotes collaboration and information flow among them. In this way, our HKAFER can effectively capture robust spatial features and the regional associations. Experimental results demonstrate that our proposed model not only outperforms state-of-the-art methods on several FER benchmarks but also uses significantly fewer trainable parameters.

Introduction

Facial expression serves as a pivotal means of emotional communication and has wide-ranging real-world applications, including sentiment analysis for social media monitoring (Savchenko, Savchenko, and Makarov 2022), human-robot interaction in service robotics (Hu et al. 2024), and psychological assessment in medical diagnosis (Ye et al. 2025), reflecting its importance in advancing both social and technological domains.

Although Facial Expression Recognition (FER) has achieved significant progress recently, it is still a challenging task due to inter-class similarity and intra-class diversity. To address the above problems, recent methods (Zhang et al. 2024; Wang et al. 2025) generally design complex model architectures and introduce extensive learnable parameters, promoting FER models to learn advanced semantic representations. Furthermore, some approaches explore regularization strategies (Xue et al. 2022; Lee et al. 2025; Gao et al. 2025) or introduce prior knowledge (Wu and Cui 2023; Zhou et al. 2024) to learn more robust and discriminative facial features. However, they still retain redundant parameters and suffer from the risk of overfitting, especially under the unconstrained conditions of occlusions and pose variations. Therefore, the following is a critical challenge in the field of FER:

How to extract more discriminative visual expression representations while utilizing less trainable parameters?

Recently, large pre-trained models have excelled in Natural Language Processing (NLP) and Computer Vision (CV) tasks. To mitigate the high computational cost associated with fine-tuning large pre-trained models, many efforts have explored Parameter-Efficient Fine-Tuning (PEFT) strategies (Hu et al. 2022; Jia et al. 2022; Fu et al. 2023; Ding et al. 2023; Jie and Deng 2023). Among them, Low-Rank Adaptation (LoRA) (Hu et al. 2022) stands out for its strong interpretability and ease of deployment. As shown in Fig.1 (a), LoRA works by freezing the pre-trained weights and representing the updated parameters as the product of two low-

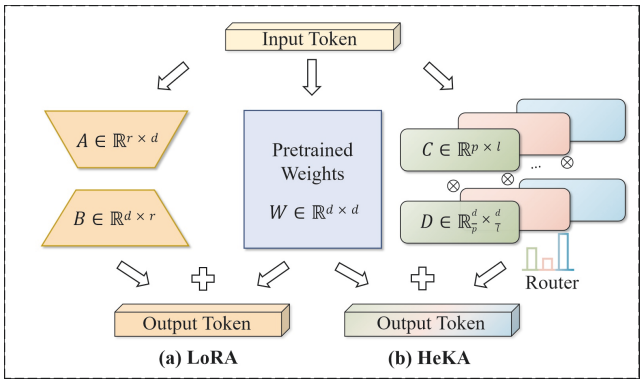


Figure 1: Comparison of different PEFT approaches.

*Corresponding author

rank matrices. By decomposing the incremental matrices, LoRA achieves significant parameter efficiency while maintaining performance comparable to full fine-tuning. However, on the one hand, for the fine-grained classification task FER, simple linear transformations cannot effectively represent the detailed facial information in the low-dimension space. On the other hand, LoRA assigns the fixed low-rank matrix to every module of the pre-trained model, ignoring significant space for compressing the trainable parameters. In the field of FER, some methods also attempt PEFT strategies. For example, CLIPER (Li et al. 2024) and CEPrompt (Zhou et al. 2024) introduce learnable prompts to achieve the alignment of textual and visual inputs. However, prompt-based methods struggle to deeply improve the generalization ability of FER models. When facing complex scenarios with occlusions, relying solely on learnable input prompts may not effectively guide FER models for accurate predictions. Considering that different from the textual description, which can accurately and clearly express the salient features of facial expressions, visual appearance often includes much redundant information. Hence, our purpose is to achieve visual PEFT for FER, not limited to Visual Foundation Models (VFMs) or Vision-Language Models (VLMs). To solve the above problems, we propose an HKAFER model that can adaptively enhance visual expression representations through efficiently fine-tuning Self-Attention and FFN modules in the image encdoer of large VFMs and VLMs (e.g., DINOv2 (Oquab et al. 2023) and CLIP (Radford et al. 2021)). Specifically, as shown in Fig.1 (b), we establish Heterogeneous Kronecker Adaptation (HeKA), which consists of multi-scale adapters based on Kronecker product in a parallel manner. To enhance the collaboration of adapters, we further propose Dual-Branch Interactive Router (DBIR) to dynamically allocate the weights and avoid concentrating redundant areas, e.g., occluded facial regions and the background. In this way, HKAFER decomposes the incremental matrix into multiple subspaces, effectively capturing discriminative regions and the associations between them.

To summarize, our contributions are three-fold:

- The existing strategies are inefficient for the FER task, which not only struggle to handle unconstrained conditions like occlusions and pose variations, but also retain numerous redundant parameters. Using a parallel architecture, we decompose the incremental matrices using multi-scale adapters based on Kronecker product, effectively capturing discriminative facial regions and the associations between them.
- To enhance the collaboration of adapters, we further propose Dual-Branch Interactive Router (DBIR) to dynamically allocate their weights. DBIR prevents adapters from concentrating on redundant facial regions in the embedding space, promoting the FER model to extract robust representations.
- Experimental results show that our proposed HKAFER model achieves superior performance on several FER benchmarks, especially under unconstrained conditions, not only outperforming the state-of-the-art methods, but also with much fewer learnable parameters.

Related Works

Facial Expression Recognition

With the development of deep learning techniques, FER has made remarkable progress. These methods can primarily be categorized as either attention-based or prior-based approaches. Attention-based methods (Xue, Wang, and Guo 2021; Xue et al. 2022; Li et al. 2023; Gao et al. 2025), dependent on complex module designs, enable the FER model to focus on critical facial regions. E.g., FG-AGR (Li et al. 2023) learns the associations between global and local attention features via graph convolutional networks to boost performance in FER. To further improve the generalization ability, some methods attempt to incorporate regularization into the attention mechanism. E.g., CAFE (Zhang et al. 2024) learns sigmoid masks based on the fixed facial features and separate the channels to reduce overfitting. However, the above methods all ignore the potential semantic associations between different expression categories. Thus, some methods attempt to introduce prior knowledge to refine FER features. For example, LDLVA (Le et al. 2023) and LA-Net (Wu and Cui 2023) aim to reconstruct label distributions by leveraging neighboring samples in the valence-arousal space and landmark space, respectively. However, to acquire prior knowledge, these methods not only incorporate substantial learnable parameters, but also become entangled with inevitable noise in the process. In contrast to conventional FER techniques, our HKAFER model innovatively addresses the aforementioned challenges by freezing the backbone network of large VFMs and VLMs as a static repository of world knowledge. This allows for an efficient fine-tuning approach focusing solely on the critical parameters, thereby achieving superior FER performance while significantly reducing the number of trainable parameters.

Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) approaches are designed to selectively adjust a subset of parameters or incorporate extra trainable parameters (Rebuffi, Bilen, and Vedaldi 2017; Houlsby et al. 2019; Zhou et al. 2022a,b; Hu et al. 2022; Jie and Deng 2023) in the large pre-trained models. Among these methods, LoRA (Hu et al. 2022) has been widely used due to its strong interpretability and ease of deployment. LoRA (Hu et al. 2022) represents weight updates as the product of two low-rank matrices, significantly decreasing the number of trainable parameters and memory needs during training. DoRA (Liu et al. 2024) decomposes the pre-trained weights into magnitude and direction, thereby exerting more refined control over the learning process. To further improve the performance, some LoRA-based methods attempt multi-branch architecture. For example, MeLoRA (Ren et al. 2024) maintains a higher rank by concatenating mini LoRAs along the diagonal to construct an equivalent block diagonal LoRA matrix. However, it only employ multiple identical LoRA modules in parallel and fail to effectively explore diverse subspaces. Recently, Kronecker product has been employed to enhance the PEFT methods. KronA (Edalati et al. 2022) simply replaces low-rank projections with Kronecker product. However, the uti-

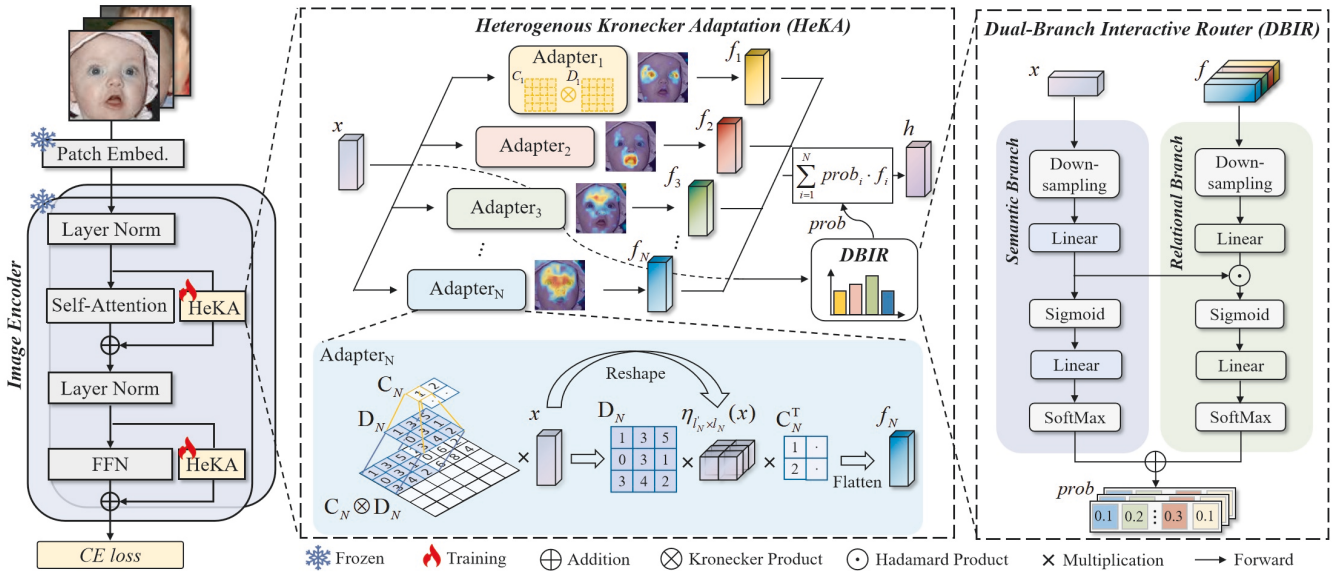


Figure 2: The overall pipeline of HAKFER model. Input facial images are fed into Image Encoder to extract features for the recognition. During training, the pre-trained backbone is frozen, with only Heterogeneous Kronecker Adaptation (HeKA) is optimized to learn the updated matrices of Self-Attention and FFN modules in Image Encoder. Then, Dual-Branch Interactive Router (DBIR) is employed to aggregate the features from different adapters dynamically.

lization of a single Kronecker mode may potentially limit the capacity to dynamically focus on crucial regions and investigate the interactions among different regions.

In the field of FER, some methods also attempt PEFT strategies. For example, following CoOp (Zhou et al. 2022b), DFER-CLIP (Zhao and Patras 2023) and CLIPER (Li et al. 2024) introduce learnable contexts for each expression descriptor, while neglecting the alignment of visual semantic information. CEPrompt (Zhou et al. 2024) further adopts a two-stage approach to align textual and visual prompts. However, these prompt-based methods still retain redundant learning parameters and restrict the generalization ability of FER models, especially under unconstrained conditions. To achieve visual PEFT in FER, our HKAFER establishes Heterogeneous Kronecker Adaptation (HeKA), which consists of multi-scale adapters based on Kronecker product in a parallel manner, offering significantly diverse subspaces to learn the incremental matrices.

Method

Preliminary

LoRA (Hu et al. 2022) has been widely used as a PEFT strategy in large pre-trained models. LoRA decomposes the updated matrix into the product of two low-rank matrices, reducing the number of trainable parameters. Formally, it is defined as:

$$h = \mathbf{W}x + \Delta\mathbf{W}x = \mathbf{W}x + \frac{\alpha}{r}\mathbf{B}\mathbf{A}x, \tag{1}$$

where $x \in \mathbb{R}^{d_1}$ and $h \in \mathbb{R}^{d_2}$ are the input and output token. $\mathbf{W} \in \mathbb{R}^{d_2 \times d_1}$ and $\Delta\mathbf{W} \in \mathbb{R}^{d_2 \times d_1}$ are the pre-trained and updated parameter matrices. $\mathbf{A} \in \mathbb{R}^{r \times d_1}$ and $\mathbf{B} \in \mathbb{R}^{d_2 \times r}$ are the low-rank matrices. α is the scaling coefficient. Compared with full fine-tuning, LoRA reduces the number of parameters to $r(d_1 + d_2) \ll d_1 \times d_2$, since $r \ll \min(d_1, d_2)$.

Heterogeneous Kronecker Adaptation

Although LoRA significantly reduces computational cost, its performance is constrained by the hyperparameter rank. With a too little rank, LoRA fails to express the incremental matrix well. However, with an overly large rank, it causes model overfitting due to redundant parameters. Meanwhile, simple linear transformations struggle to effectively capture detailed visual FER representations. Hence, as shown in Fig. 2, we propose Heterogeneous Kronecker Adaptation (HeKA) to decompose the incremental matrices using multi-scale adapters based on Kronecker product. We further analyze our motivation with three lemmas:

Lemma 1. *Let $\Delta\mathbf{W}_r$ be low-rank reconstruction matrices of $\Delta\mathbf{W}$, where the rank is r. Then, we have:*

$$\|\Delta\mathbf{W} - \Delta\mathbf{W}_r\|_F \geq \sqrt{\sum_{i=r+1}^{d} \sigma_i^2} = \epsilon_{\text{low}} > \epsilon_{\text{full}} = 0, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm, σ_i are the singular values arranged in descending order, ϵ_{low} and ϵ_{full} are the lower bound reconstruction errors from low-rank and full-rank reconstruction matrices.

Lemma 2. *Let $\mathbf{BA}$ with the rank r is the reconstruction matrix of $\Delta\mathbf{W}$. Then, we have:*

$$\exists \mathbf{A} \neq \mathbf{A}', \mathbf{B} \neq \mathbf{B}' \quad \text{s.t.} \quad \mathbf{BA} = \mathbf{B}'\mathbf{A}', \tag{3}$$

where $\mathbf{B}'\mathbf{A}'$ is another reconstruction matrix with the same rank as $\mathbf{BA}$.

Eq. (2) indicates that the performance of LoRA is limited to its rank. However, as the rank increases, the number of parameters grows exponentially. Eq. (3) indicates that there are multiple decomposition methods with the same rank for a matrix. While using multi-LoRA may potentially lead to better performance, it will also further increase the parameter count. Considering the limit of LoRA, we introduce Kronecker product (Henderson, Pukelsheim, and Searle 1983) to achieve full-rank decomposition with fewer parameters. Given two matrices $\mathbf{C}$ of size $q \times l$ and $\mathbf{D}$ of size $q' \times l'$, their Kronecker product $\mathbf{C} \otimes \mathbf{D}$ is a matrix of size $qq' \times ll'$. The formal mathematical form is:

$$\mathbf{C} \otimes \mathbf{D} = \begin{pmatrix} c_{1,1}\mathbf{D} & \cdots & c_{1,l}\mathbf{D} \\ \vdots & \ddots & \vdots \\ c_{q,1}\mathbf{D} & \cdots & c_{q,l}\mathbf{D} \end{pmatrix}, \tag{4}$$

where $\otimes$ is Kronecker product, $c_{q,l}$ is in the q-th row and l-th column of matrix $\mathbf{C}$.

Lemma 3. *The rank of $\mathbf{C} \otimes \mathbf{D}$ is the product of the ranks of $\mathbf{C}$ and $\mathbf{D}$. Formally,*

$$rank(\mathbf{C} \otimes \mathbf{D}) = rank(\mathbf{C}) \times \ rank(\mathbf{D}), \tag{5}$$

where $\otimes$ is Kronecker product.

Eq. (5) indicates that compared with LoRA, Kronecker product can produce a higher-rank matrix while maintaining two lower-rank decomposed matrices. Assuming that $d_1 \approx d_2$, when performing a full-rank decomposition on the matrix $\Delta\mathbf{W}$ with rank d, Kronecker product requires at least the number of parameters to $2\sqrt{d_1 d_2}$, which is less than the number of $d(d_1+d_2)$ from LoRA. In detail, Kronecker product can save parameters, since it splits an increment matrix into identical blocks and assigns each block its own weight. This regularization method can effectively make the FER model learn the specific facial representations and prevent overfitting. However, a single Kronecker mode may restrict the ability to adaptively concentrate on critical regions and explore regional interaction. This may lead to poor performance, especially under unconstrained conditions.

As shown in Eq. (3), there are also multiple full-rank decomposition methods for a matrix. Hence, we attempt to decompose the incremental matrices to multi-scale adapters based on Kronecker product in parallel. In this way, HeKA can provide multiple subspaces to learn robust facial representations adaptively. Formally, we first establish each adapter based on Kronecker product:

$$\begin{aligned} f_i &= \text{Adapter}_i(x) \\ &= (\mathbf{C}_i \otimes \mathbf{D}_i)x \\ &= \text{Flatten}(\mathbf{D}_i \eta_{l_i' \times l_i}(x)\mathbf{C}_i^\top), \end{aligned} \tag{6}$$

where $f_i \in \mathbb{R}^{d_2}$ is the output of the i-th adapter. $\mathbf{C}_i \in \mathbb{R}^{q_i \times l_i}$ and $\mathbf{D}_i \in \mathbb{R}^{q_i' \times l_i'}$ are two decomposed matrices that make up $\text{Adapter}_i(\cdot) \in \mathbb{R}^{d_2 \times d_1}$ and initialized using Kaiming initialization and zero initialization, respectively. Based on Eq. (4), $d_1 = l_i \times l_i'$ and $d_2 = q_i \times q_i'$. $\eta_{l_i' \times l_i}(\cdot)$ is an operation that reshapes a vector into a matrix of size $l_i' \times l_i$ and $\text{Flatten}(\cdot)$ is another operation that reshapes a matrix into a vector by stacking its columns. Through each adapter, the FER model can capture specific facial regional features. We further adopt multiple different adapters and assign dynamic weights to each one, which is beneficial for learning complex FER representations. Formally,

$$\begin{aligned} h &= \mathbf{W}x + \Delta\mathbf{W}x \\ &= \mathbf{W}x + \beta \sum_{i=1}^{N} prob_i \cdot f_i, \end{aligned} \tag{7}$$

where, $prob_i$ is the learnable weight for the i-th adapter from our router mechanism, N is the number of adapters, and β is the scaling coefficient. Following LoRA, β is set to 2.

Dual-Branch Interactive Router

The existing gating mechanisms simply generate weights based on the input token, without effective collaboration among the adapters. This can easily lead to redundant learning. However, complex interaction modules can also introduce a large number of trainable parameters, which contraries to the intention of PEFT. To solve the above challenges, we propose Dual-Branch Interactive Router (DBIR), which consists of Semantic Branch (SB) and Relational Branch (RB), mapping the input token and adapter tokens to the low-dimension space for dynamic interaction. Specifically, in SB, we first do the down-sampling operation to reduce computational complexity:

$$x^{\downarrow} = \textbf{AvgPool}(\eta_{\gamma d_1 \times \frac{1}{\gamma}}(x), -1), \tag{8}$$

where $x^{\downarrow} \in \mathbb{R}^{\gamma d_1}$ is the input token after the down-sampling operation. γ is the down-sampling rate and $\textbf{AvgPool}(\cdot, -1)$ is the average pooling operation of the last dimension. Then, the routing weights are calculated based on the low-dimension token $x^{\downarrow}$ and we apply a softmax normalization:

$$logit^s = \mathbf{W}_2^s \cdot \textbf{Sigmoid}(\mathbf{W}_1^s x^{\downarrow}), \tag{9}$$

$$prob^s = \frac{\exp(logit_i^s)}{\sum_{i=1}^{N} \exp(logit_i^s)}, \tag{10}$$

where $\mathbf{W}_1^s \in \mathbb{R}^{2N \times \gamma d_1}$ and $\mathbf{W}_2^s \in \mathbb{R}^{N \times 2N}$ are the trainable projection matrices. $logit^s \in \mathbb{R}^N$ and $prob^s \in \mathbb{R}^N$ are the weights from SB before and after softmax normalization. In RB, we utilize $x^{\downarrow}$ to interact with each adapter token, implicitly promoting their collaboration with few parameters. Formally, we also first do the down-sampling operation on each adapter token:

$$f_i^{\downarrow} = \textbf{AvgPool}(\eta_{\gamma d_2 \times \frac{1}{\gamma}}(f_i), -1), \tag{11}$$

where $f_i^{\downarrow} \in \mathbb{R}^{\gamma d_2}$ is the output token of i-th adapter after the down-sampling operation. Furthermore, $f_i^{\downarrow}$ is mapped to the same low-dimension space as $x^{\downarrow}$ for interaction and we also apply a softmax normalization::

$$logit_i^r = \mathbf{W}_2^r \cdot \textbf{Sigmoid}(\mathbf{W}_1^r f_i^{\downarrow} \odot \mathbf{W}_1^s x^{\downarrow}), \tag{12}$$

$$prob^r = \frac{\exp(logit_i^r)}{\sum_{i=1}^{N} \exp(logit_i^r)}, \tag{13}$$

where $\mathbf{W}_1^r \in \mathbb{R}^{2N \times \gamma d_2}$ and $\mathbf{W}_2^r \in \mathbb{R}^{N \times 2N}$ are the trainable projection matrices. $logit^r \in \mathbb{R}^N$ and $prob^r \in \mathbb{R}^N$ are the weights from RB before and after softmax normalization. $\odot$ is Hadamard product. The final weights for each adapter is calculated by:

$$prob = \lambda \cdot prob^s + (1 - \lambda) \cdot prob^r, \tag{14}$$

where $prob \in \mathbb{R}^N$ and λ is the learnable parameter to balance two types of weights dynamically.

Objective Function

For our proposed HKAFER model, the backbone network is frozen. We utilize HeKA to fine-tune Self-Attention and FFN modules of the image encoder in the backbone. All HeKA modules are jointly trained in an end-to-end manner. Given a facial image, for VFMs, we calculate the probability of the k-th class using a softmax normalization defined as:

$$p^k = \frac{\exp(z^k)}{\sum_{k=1}^{K} \exp(z^k)}, \tag{15}$$

where z_i^k represents the the probability of the k-th class for a facial image before normalization. For VLMs, we follow CLIP (Radford et al. 2021) and apply the contrastive learning to align the visual and textual tokens in a unified embedding space. We calculate the similarity between the visual class token v_i and textual class token t_k:

$$p^k = \frac{\exp(\cos(v, t_k)/\tau)}{\sum_{k=1}^{K} \exp(\cos(v, t_k)/\tau)}, \tag{16}$$

where $\cos(\cdot, \cdot)$ is the cosine similarity and τ (set to 0.01) is a temperature parameter. In general, the Cross-Entropy loss is calculated as follow:

$$\mathcal{L} = -\sum_{k=1}^{K} y \cdot \log p^k, \tag{17}$$

where y is the corresponding label for a facial image.

Experiment

Datasets

RAF-DB (Li, Deng, and Du 2017) contains 29,672 facial images with seven basic expression classes (Neutral, Happiness, Surprise, Sadness, Anger, Disgust, and Fear). There are 12,271 images used for training and 3,068 images used for testing.

AffectNet (Mollahosseini, Hasani, and Mahoor 2017) is the largest FER dataset so far. We utilize the version with the same seven expressions as RAF-DB, containing 283,901 for training and 3,500 for testing.

ExpW (Zhanpeng Zhang and Tang 2015) contains 91,793 facial images with the same seven basic expressions as RAF-DB. We randomly split the dataset into 75% used for training, 10% used for validation, and 15% used for testing.

SFEW 2.0 (Dhall et al. 2014) contains 958 facial images for training, 436 images for validation and 372 images for test, with the same seven basic expressions as RAF-DB.

FERPlus (Barsoum et al. 2016) is extended from FER2013 (Goodfellow et al. 2013), 28,709 for training, 3,589 for validation, and 3,589 for test, with the eight emotion categories (plus Contempt).

Implementation Details

We select DINOv2-ViT-L/14 (Oquab et al. 2023) and CLIP-ViT-L/14 (Radford et al. 2021) as the VFM-based and VLM-based backbones, respectively. For training, all facial images are resized to 224 × 224 pixels. We perform some commonly used augmentation operations, including random horizontal flipping, random rotation, random cropping, random erasing and so on. All the models are trained using Pytorch framework with a batch size 4 for 60 epochs on one NVIDIA RTX A6000 GPU. The Adam optimizer is adopted with an initial learning rate of $1e^{-4}$ and a weight decay of $1e^{-4}$, and the learning rate is decayed by a factor of 0.9 after each epoch. The hyperparameter γ in Eq. (8) and Eq. (11) is set to 6.25%. The number of adapters is set to 4 and the sizes of $\mathbf{C}_i$ are set to (32,32), (32,64), (64,32) and (64,64). For the text input, we use the same textual prompts as DFER-CLIP (Zhao and Patras 2023). We choose accuracy as the evaluation metric.

Comparison with State-of-the-Art Methods

As shown in Tab. 1, we compare our proposed HKAFER model with the state-of-the-art methods on five common FER benchmarks. It is obvious that our HKAFER performs well in both VFM-based and VLM-based methods. Among VFM-based methods, our HKAFER achieves the best accuracy, except for AffectNet and FERPlus. The reason may be that the backbones of other methods have been pre-trained on large facial datasets. Among VLM-based methods, our HKAFER outperforms all existing methods. E.g., HKAFER surpasses CEPrompt (Zhou et al. 2024) by 0.53% and 0.33% on RAF-DB and AffectNet, with only 1.32M trainable parameters. Moreover, our HKAFER achieves the best accuracy of 65.83%, 76.13% and 91.91% on SFEW 2.0, ExpW and FERPlus datasets, respectively. The reason is that our HKAFER can achieve visual parameter-efficient fine-tuning by decomposing the incremental matrix into diverse subspaces via HeKA, capturing critical facial regions and the associations between them. The experimental results demonstrate that our HKAFER can achieve excellent performance on various FER datasets while utilizing only a small number of trainable parameters.

Evaluation on Occlusion and Pose Variation

We also conduct experimental evaluations on occlusion and variant-pose datasets, as shown in Tab. 2. Following CEPrompt (Zhou et al. 2024), we select CLIP as the backbone. The original training images are unchanged, while those in the test set that involve occlusions and various poses are chosen. Obviously, our HKAFER outperforms all existing methods by a large margin. For example, HKAFER exceeds CEPrompt by 2.12%, 1.96% and 1.83% on RAF-DB, under Occlusion, Pose ($\geq 30^\circ$) and Pose ($\geq 45^\circ$) circumstances, respectively. The reason is that our HKAFER utilizes multi-scale adapters based on Kronecker product to efficiently fine-tuning large pre-trained models, offering significantly diverse subspaces to learn discriminative representations. Besides, DBIR dynamically allocates the weights to avoid concentrating redundant areas.

Method	Backbones	#Params	RAF-DB	AffectNet	SFEW 2.0	ExpW	FERPlus
VFM-based methods							
RAN (Wang et al. 2020)	ResNet18+VGG16	-	86.90	59.50	-	-	89.16
TransFER (Xue, Wang, and Guo 2021)	IR50+ViT	-	90.91	66.23	-	-	90.83
PAT-Res101 (Cai et al. 2022)	ResNet101	-	88.43	-	57.57	72.93	-
EAC (Zhang et al. 2022)	ResNet18	23.52M	89.99	65.32	-	-	89.64
APViT (Xue et al. 2022)	IR50+ViT	-	91.98	66.91	61.92	73.48	90.86
DAN (Wen et al. 2023)	ResNet18	19.72M	89.70	65.69	57.88	-	-
FG-AGR (Li et al. 2023)	CovNeXt	-	90.81	64.91	-	-	91.09
LA-Net (Wu and Cui 2023)	ResNet50	55.44M	91.56	67.60	-	-	91.78
JADFER (Gao et al. 2025)	IR50+ViT	77.89M	92.31	68.06	63.53	74.24	90.66
QCS (Wang et al. 2025)	IR50+ViT	66.58M	92.50	67.94	-	-	91.41
HKAFER (Ours)	DINOv2-ViT-L/14	1.76M	92.70	67.66	64.22	74.81	91.12
VLM-based methods							
CAFE (Zhang et al. 2024)	CLIP-ViT-B/32	11.18M	88.72	64.87	53.79	-	89.51
CLIPER (Li et al. 2024)	CLIP-ViT-B/16	-	91.61	66.29	-	-	-
CEPrompt (Zhou et al. 2024)	CLIP-ViT-B/16	32.29M	92.43	67.29	-	74.37	-
CEPrompt (Zhou et al. 2024)	CLIP-ViT-L/14	56.50M	93.96	68.21	-	-	-
HKAFER (Ours)	CLIP-ViT-L/14	1.32M	**94.49**	**68.54**	**65.83**	**76.13**	**91.91**

Table 1: Performance comparison (%) with the state-of-the-art methods on RAF-DB, AffectNet, SFEW 2.0, ExpW and FERPlus. #Params indicates the trainable parameters. The best result is marked with bold and the second best result is underlined.

Method	Occlusion	Pose ($\geq 30°$)	Pose ($\geq 45°$)
RAF-DB			
RAN (2020)	82.72	86.74	85.20
FG-AGR (2023)	88.15	91.02	90.50
CEPrompt (2024)	90.13	91.70	91.35
JADFER (2025)	88.98	91.34	90.68
HKAFER (Ours)	**92.25**	**93.66**	**93.18**
AffectNet			
RAN (2020)	58.50	53.90	53.19
FG-AGR (2023)	64.24	61.26	61.15
CEPrompt (2024)	66.83	64.28	63.44
JADFER (2025)	65.15	61.47	62.26
HKAFER (Ours)	**67.40**	**65.16**	**64.64**
FERPlus			
RAN (2020)	83.63	82.23	80.40
FG-AGR (2023)	85.79	88.38	87.52
JADFER (2025)	87.77	89.74	89.10
HKAFER (Ours)	**90.25**	**91.37**	**91.00**

Table 2: Performance comparison (%) with the state-of-the-art methods on RAF-DB, AffectNet and FERPlus datasets with realistic occlusions and variant poses.

Evaluation on Cross Datasets

To further demonstrate the generalization ability of our HKAFER, we also perform a cross-dataset evaluation with the backbone of CLIP. 'R→A' indicates the FER model is trained on RAF-DB and evaluated on AffectNet and 'A→R' represents the opposite dataset setting. As shown in Tab. 3, our HKAFER surpasses JADFER by an average accuracy of 5.17%. The reason is that HKAFER can focus on different critical regions by efficiently fine-tuning the image encoder via multiple heterogeneous matrices based on Kronecker product.

Method	R→A	A→R	Mean
STSN (2021)	48.49	76.99	62.74
KTN (2021)	49.60	76.53	63.07
FG-AGR (2023)	48.86	74.79	61.83
JADFER (2025)	51.74	78.39	65.07
HKAFER (Ours)	**55.83**	**84.65**	**70.24**

Table 3: Cross-dataset evaluation comparisons (%) on RAF-DB and AffectNet.

Ablation Studies

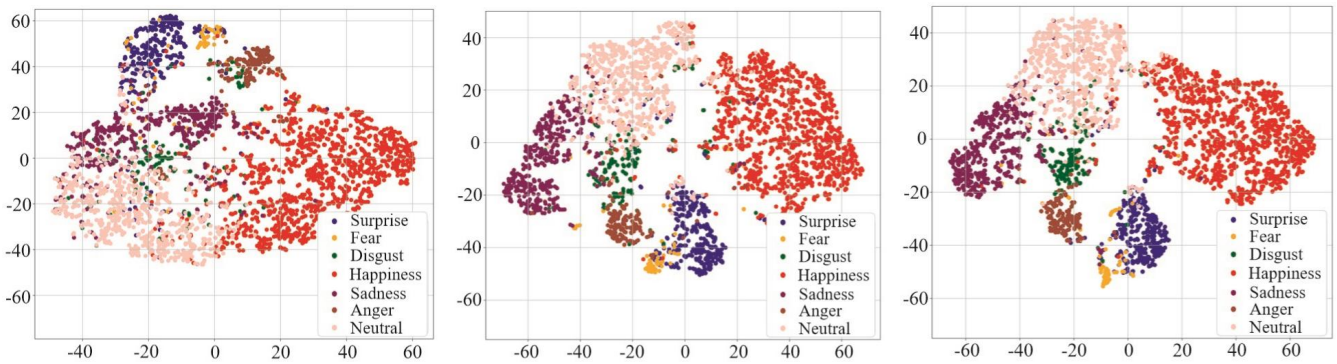
For simplicity, all the ablation studies are conducted on the RAF-DB and ExpW datasets, using DINOv2-ViT-L/14 as the backbone.

Effect of Dual-Branch Interactive Router. We also perform ablation studies to demonstrate the effectiveness of Dual-Branch Interactive Router (DBIR). As shown in Tab. 4, our DBIR achieves a 0.42% performance improvement on RAF-DB at the cost of increasing trainable parameters by only 0.21M, compared with the method of directly adding tokens from different adapters (denoted as 'Sum.').

Method	#Params	RAF-DB	ExpW
Sum.	1.55M	92.28	74.35
SB	1.64M	92.44	74.61
RB	1.75M	92.54	74.69
DBIR	1.76M	**92.70**	**74.81**

Table 4: Effect (%) of Dual-Branch Interactive Router on RAF-DB and ExpW datasets.

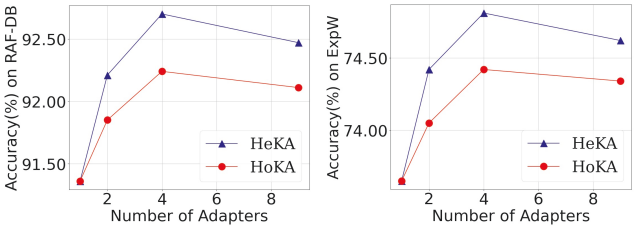
Effect of Heterogeneous Kronecker Adaptation. We conduct a series of ablation experiments to compare our



(a) LoRA (Hu et al. 2022) (b) HKAFER with HoKA (c) HKAFER with HeKA

Figure 3: Visualization of expression representations on the test set of RAF-DB using t-SNE.

HeKA with Low-Rank Adaptation (LoRA) and Homogeneous Kronecker Adaptation (HoKA). As shown in Fig. 4, we set the number of adapters N to 1, 2, 4, and 9, and report their best performance, respectively. When N is set to 4, which corresponds critical facial regions, the best results are achieved. As shown in Tab. 5, our HeKA achieves an accuracy of 0.72% higher than LoRA$_{r=32}$ on RAF-DB.



(a) Performance on RAF-DB (b) Performance on ExpW

Figure 4: Effect of the number of adapters.

Method	#Params	RAF-DB	ExpW
Full Fine-tune	304.37M	89.99	73.22
LoRA$_{r=32}$	12.58M	91.98	74.18
HoKA$_{N=4}$	1.78M	92.24	74.42
HeKA$_{N=4}$	1.76M	**92.70**	**74.81**

Table 5: Effect (%) of Heterogeneous Kronecker Adaptation on RAF-DB and ExpW.

Visualization

To intuitively show the effectiveness of our HKAFER model, we employ t-SNE (Van der Maaten and Hinton 2008) to visualize the expression representations. As shown in Fig. 3, we can observe that our HKAFER with HeKA can achieve relatively better intra-class compactness and inter-class separation. In addition, we use Grad-Cam++ (Chattopadhay et al. 2018) to visualize the attention maps extracted from the different adapters of HeKA in the last *atten.qkv* module. As shown in Fig. 5, the first row shows the original images for each category. The second to fifth rows represent the attention maps extracted from each adapter. It can be seen that the four adapters focus on different critical facial regions.

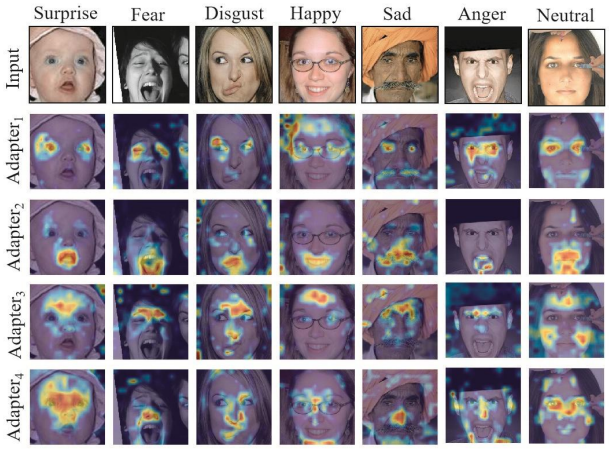


Figure 5: Attention maps of some examples on the test set of RAF-DB.

Conclusion

In this work, we propose an HKAFER model, which achieves visual parameter-efficient fine-tuning for Facial Expression Recognition (FER) via Heterogeneous Kronecker Adaptation (HeKA). HeKA consists of multi-scale adapters based on Kronecker product in a parallel manner, offering significantly diverse subspaces to learn the incremental matrices. To adaptively aggregate features from each adapter, we further propose Dual-Branch Interactive Router (DBIR) to assign the weights and prevent the FER model from focusing on redundant regions. Experimental results demonstrate that our proposed HKAFER model outperforms the state-of-the-art methods on several FER benchmarks.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4703200; in part by the National Natural Science Foundation of China under Grant 62261160652, Grant 52275013, Grant 62206075, Grant 61733011, Grant 62503139, and Grant 62573163; in part by Shenzhen Science and Technology Program under Grant JCYJ20240813105137049; in part by the Science and Technology Development Fund (FDCT), Macau, SAR, under Grant 0095/2022/AFJ; in part by the GuangDong Basic and Applied Basic Research Foundation under Grant 2024A1515012028; and in part by the Shenzhen Medical Research Fund under Grant A2502034.

References

Barsoum, E.; Zhang, C.; Ferrer, C. C.; and Zhang, Z. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction*, 279–283.

Cai, J.; Meng, Z.; Khan, A. S.; Li, Z.; O'Reilly, J.; and Tong, Y. 2022. Probabilistic Attribute Tree Structured Convolutional Neural Networks for Facial Expression Recognition in the Wild. *IEEE Transactions on Affective Computing*, 1–1.

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *IEEE Winter Conference on Applications of Computer Vision*, 839–847.

Dhall, A.; Goecke, R.; Joshi, J.; Sikka, K.; and Gedeon, T. 2014. Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol. In *International Conference on Multimodal Interaction,*, 461–466.

Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3): 220–235.

Edalati, A.; Tahaei, M.; Kobyzev, I.; Nia, V. P.; Clark, J. J.; and Rezagholizadeh, M. 2022. Krona: Parameter efficient tuning with kronecker adapter. *arXiv preprint arXiv:2212.10650*.

Fu, Z.; Yang, H.; So, A. M.-C.; Lam, W.; Bing, L.; and Collier, N. 2023. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12799–12807.

Gao, Y.; Ren, W.; Jiang, W.; Dong, Q.; Nie, W.; Wu, W.; and Liu, H. 2025. JADFER: Exploring Spatial-Contextual Interaction With Joint Attention Dropping for Facial Expression Recognition. *IEEE Transactions on Affective Computing*, 16(2): 655–668.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, 117–124.

Henderson, H. V.; Pukelsheim, F.; and Searle, S. R. 1983. On the history of the Kronecker product. *Linear and Multilinear Algebra*, 14(2): 113–120.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2): 3.

Hu, Y.; Chen, B.; Lin, J.; Wang, Y.; Wang, Y.; Mehlman, C.; and Lipson, H. 2024. Human-robot facial coexpression. *Science Robotics*, 9(88): eadi4724.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Jie, S.; and Deng, Z.-H. 2023. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *AAAI Conference on Artificial Intelligence*, volume 37, 1060–1068.

Le, N.; Nguyen, K.; Tran, Q.; Tjiputra, E.; Le, B.; and Nguyen, A. 2023. Uncertainty-aware label distribution learning for facial expression recognition. In *the IEEE Winter Conference on Applications of Computer Vision*, 6088–6097.

Lee, J.; Choi, Y.; Kim, H.; Kim, I.-J.; and Nam, G. P. 2025. Navigating label ambiguity for facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4517–4525.

Li, C.; Li, X.; Wang, X.; Huang, D.; Liu, Z.; and Liao, L. 2023. FG-AGR: Fine-Grained Associative Graph Representation for Facial Expression Recognition in the Wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Li, H.; Niu, H.; Zhu, Z.; and Zhao, F. 2024. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *IEEE International Conference on Multimedia and Expo*, 1–6.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2852–2861.

Liu, S.-Y.; Wang, C.-Y.; Yin, H.; Molchanov, P.; Wang, Y.-C. F.; Cheng, K.-T.; and Chen, M.-H. 2024. Dora: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Rebuffi, S.-A.; Bilen, H.; and Vedaldi, A. 2017. Learning multiple visual domains with residual adapters. *Advances in Neural Information Processing Systems*, 30.

Ren, P.; Shi, C.; Wu, S.; Zhang, M.; Ren, Z.; Rijke, M.; Chen, Z.; and Pei, J. 2024. MELoRA: Mini-Ensemble Low-Rank Adapters for Parameter-Efficient Fine-Tuning. In *Annual Meeting of the Association for Computational Linguistics*, 3052–3064.

Savchenko, A. V.; Savchenko, L. V.; and Makarov, I. 2022. Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network. *IEEE Transactions on Affective Computing*, 13(4): 2132–2143.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Wang, C.; Chen, L.; Wang, L.; Li, Z.; and Lv, X. 2025. QCS: Feature refining from quadruplet cross similarity for facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7563–7572.

Wang, K.; Peng, X.; Yang, J.; Meng, D.; and Qiao, Y. 2020. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29: 4057–4069.

Wen, Z.; Lin, W.; Wang, T.; and Xu, G. 2023. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2): 199.

Wu, Z.; and Cui, J. 2023. LA-Net: Landmark-aware learning for reliable facial expression recognition under label noise. In *the IEEE International Conference on Computer Vision*, 20698–20707.

Xue, F.; Wang, Q.; and Guo, G. 2021. Transfer: Learning relation-aware facial expression representations with transformers. In *IEEE International Conference on Computer Vision*, 3601–3610.

Xue, F.; Wang, Q.; Tan, Z.; Ma, Z.; and Guo, G. 2022. Vision Transformer with Attentive Pooling for Robust Facial Expression Recognition. *IEEE Transactions on Affective Computing*.

Ye, J.; Yu, Y.; Wang, Q.; Liu, G.; Li, W.; Zeng, A.; Zhang, Y.; Liu, Y.; and Zheng, Y. 2025. CmdVIT: A Voluntary Facial Expression Recognition Model for Complex Mental Disorders. *IEEE Transactions on Image Processing*, 1–1.

Zhang, Y.; Wang, C.; Ling, X.; and Deng, W. 2022. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision*, 418–434.

Zhang, Y.; Zheng, X.; Liang, C.; Hu, J.; and Deng, W. 2024. Generalizable Facial Expression Recognition. In *European Conference on Computer Vision*, 231–248.

Zhanpeng Zhang, C. C. L., Ping Luo; and Tang, X. 2015. Learning Social Relation Traits from Face Images. In *International Conference on Computer Vision*.

Zhao, Z.; and Patras, I. 2023. Prompting Visual-Language Models for Dynamic Facial Expression Recognition. In *British Machine Vision Conference*, 1–14.

Zhou, H.; Huang, S.; Zhang, F.; and Xu, C. 2024. Ceprompt: Cross-modal emotion-aware prompting for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.