2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

MLPER: Multi-Level Prompts for Adaptively Enhancing Vision-Language Emotion Recognition

Yu Gao, Weihong Ren, Xinglong Xu, Yan Wang, Zhiyong Wang and Honghai Liu, *Fellow, IEEE*

***Abstract*— In the field of robotics, vision-based Emotion Recognition (ER) has achieved significant progress, but it still faces the challenge of poor generalization ability under unconstrained conditions (e.g., occlusions and pose variations). In this work, we propose MLPER model, which introduces Vision-Language Model for Emotion Recognition to learn discriminative representations adaptively. Specifically, different from typically leveraging a hand-crafted prompt (e.g., "a photo of a [class] person"), we first establish Multi-Level Prompts from three aspects: *facial expression*, *human posture* and *situational condition* using large language models, like ChatGPT. Correspondingly, we extract the visual tokens from three levels: the face, body, and context. Further, to achieve fine-grained alignment at each level, we adopt textual tokens from the positive and the hard negative to query visual tokens, predicting whether a pair of image and text is matched. Experimental results demonstrate that our MLPER model outperforms the state-of-the-art methods on several ER benchmarks, especially under the conditions of occlusions and pose variations.**

I. INTRODUCTION

Emotion Recognition (ER), an interdisciplinary domain bridging psychology, computer science, and artificial intelligence, has emerged as a crucial area of research and application, especially in the realm of robotics [1], [2]. E.g., in healthcare, robots equipped with emotion recognition capability can provide better support and care to patients by understanding their emotional states and responding appropriately. In education, robots can adjust their teaching methods based on the emotional responses of students, thus improving learning outcomes. Furthermore, in customer service, robots can identify and adapt to the emotional needs of customers, offering personalized experiences that can lead to higher satisfaction.

Although vision-based ER methods have achieved significant progress recently, it is still facing the challenge of poor generalization ability under unconstrained conditions of occlusions and pose variations. To solve the challenge, many methods are proposed which can be divided into two categories: face-based methods and context-aware methods.

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 Grants 62206075, 61733011, and 62261160652, in part by the GuangDong Basic and Applied Basic Research Foundation under Grant 2024A1515012028, and in part by the Shenzhen Science and Technology Program under Grants RCBS20221008093220004 and GXWD20231129125006001.

Yu Gao, Weihong Ren, Xinglong Xu, Yan Wang, Zhiyong Wang and Honghai Liu are with the School of Mechanical Engineering and Automation, State Key Lab of Robotics and Systems, Harbin Institute of Technology, Shenzhen 518055, China. (e-mail: 22S153190@stu.hit.edu.cn, renweihong@hit.edu.cn, 22S053029@stu.hit.edu.cn, 22S153227@stu.hit.edu.cn, wangzhiyong@hit.edu.cn, honghai.liu@hit.edu.cn)

(Corresponding author: Weihong Ren)

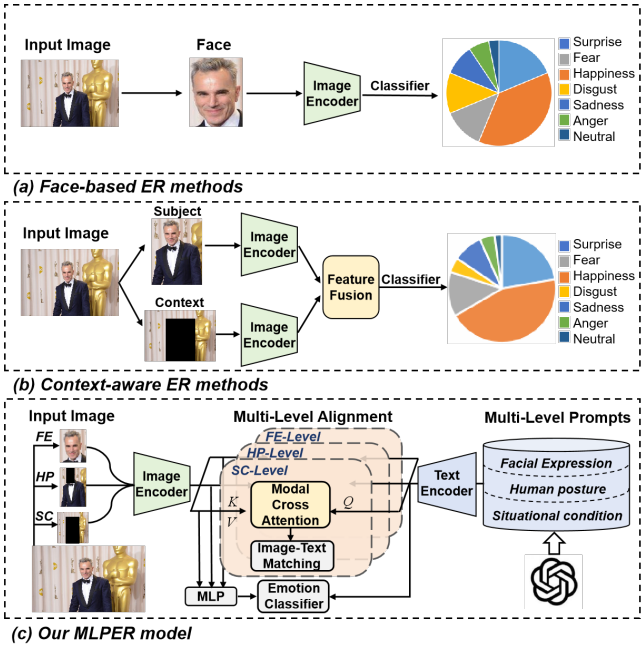


Fig. 1: Comparison of different ER approaches. The face-based ER methods (a) adopt the facial expression to directly predict the emotion, ignoring the information of the posture and context. The context-ware ER methods (b) fuse the features of the subject and the context, failing to explore fine-grained representations. Hence, In contrast, our proposed MLPER establishes Multi-Level Prompts from three aspects: *facial expression*, *human posture* and *situational condition*, which can provide informative and clear clues to distinguish the final emotion.

As shown in Fig. 1 (a), the face-based methods extract the facial region from each input image for Facial Expression Recognition (FER). E.g., APViT [3] guides the FER model to focus on crucial facial regions by retaining only a limited set of tokens, which also inevitably discards potentially valuable information. In order to utilize more informative prior, LA-Net [4] performs label refurbishment for each sample according to their neighbors in the landmark space. However, the face-based methods only take advantage of the facial information to recognize the emotion, ignoring the important roles of the posture and context. Hence, some works [5], [6], [7] have attempted to incorporate the situational context to enhance the understanding of emotional states for Context-Aware Emotion Recognition (CAER). As shown in Fig. 1(b), the context-aware methods usually separate the input image into the subject and the context, and further fuse the features from them to identify the emotion. E.g., EMOT-Net [6] crops out the body as the subject and takes the entire image as input for providing the necessary contextual support. Further,

TABLE I: Multi-Level Prompts for each emotion. **FE**, **HP** and **SC** represent *facial expression*, *human posture* and *situational condition*, respectively.

Emotion	FE	HP	SC
Surprise	Widened eyes, raised eyebrows, an open mouth	Reflexive body gestures	Unexpected events or reveals
Fear	Parted lips, furrowed brows	Tense body posture, protective gestures	Threatening or alarming situations
Disgust	Wrinkled noses, narrowed eyes	A withdrawn body posture	Repulsive or distasteful situations
Happiness	Smiles, squinted eyes, relaxed eyebrows	Open body language, clapping or hugging	Joyful moments, celebrations
Sadness	Downturned mouths, drooping eyelids	Covering the face or the head bowed down	Sorrowful events, disappointments
Anger	Furrowed brows, a clenched jaw	Pointing or fists clenched	Frustrating situations, confrontations
Neutral	Relaxed, non-expressive faces	Minimal dramatic body language	Calm and ordinary settings

EmotiCon [7] uses depth maps to model the proximity-based socio-dynamic interactions of the subject and the context. However, under unconstrained conditions (e.g., occlusions and pose variations), they incorporate extra priors, which also introduces unpredictable noise, negatively impacting the learning of robust representations.

Recently, Vision-Language Models (VLM) [8], [9] have made significant progress in boosting the performance of many downstream computer vision tasks [10], [11], [12] through the application of contrastive learning techniques on pairs of images and text. E.g., CLIP [8] leverages 400 million image-text pairs to map the visual and textual tokens to a unified embedding space, which also brings inspiration to ER. As shown in Fig. 1(c), different from the traditional ER methods, our ER method introduces VLM as the prior clues to mine the distinguished representations, adaptively.

Our contribution is therefore to explore a new vision-language pipeline to improve ER. First, different from the hand-crafted template (e.g., a photo of a [**class**] person), we establish Multi-Level Prompts from three aspects: *facial expression*, *human posture* and *situational condition* using large language models. Correspondingly, we extract the visual tokens from three levels: the face, body, and context. Further, to achieve fine-grained alignment at each level, we adopt textual tokens from the positive and the hard negative to query visual tokens, predicting whether a pair of image and text is matched. Overall, our main contributions can be summarized as follows:

1) The existing ER methods typically fail to handle unconstrained conditions (e.g., occlusions and pose variations) without introducing extra prior knowledge. Hence, we introduce Vision-Language Model (VLM) to learn discriminative representations adaptively.
2) Different from typically leveraging a hand-crafted prompt, we establish Multi-Level Prompts from three aspects: *facial expression*, *human posture* and *situational condition* using large language models, like ChatGPT.
3) To finely align text prompts and images at each level, we adopt textual tokens from the positive and the hard negative to query visual tokens, predicting whether a pair of image and text is matched to eliminate modal heterogeneity.
4) Experimental results show that our proposed MLPER model achieves superior performance on several ER benchmarks, outperforming the state-of-the-art methods, especially under the conditions of occlusions and pose variations.

II. Related work

In this section, we briefly review the related works on *Emotion Recognition* and *Vision-Language Model*.

A. Emotion Recognition

Emotion Recognition (ER), which focuses on interpreting human expressions, has continued to be an important research area for many years. Early attempts [13], [14] rely on hand-crafted features to describe and recognize different emotions. With the advancement of deep learning technologies, recent vision-based ER methods are mainly divided into two categories: face-based and context-aware methods. The face-based methods extract the facial region from each input image for Facial Expression Recognition (FER). E.g. WSCFER [2] attempts to simultaneously learn robust FER representations by pulling augmented samples of the same image together while pushing apart instance samples from different classes. Similarly, EAC [15] utilizes flip semantic consistency to learn robust representations adaptively. However, the face-based methods ignore the valuable information of the posture and context for ER. Hence, the context-aware methods separate the input image into the subject and the context to well learn the whole emotion representations. E.g., CCIM [16] integrates the subject and contextual information for ER through causal reasoning, while also inevitably introduces noise. Additionally, these methods struggle with limited generalization capabilities in unconstrained scenarios. In order to improve the generalization ability of ER, we introduce Vision-Language Model for ER, and establish Multi-Level Prompts to provides more distinguished details for ER.

B. Vision-Language Model

Vision-Language Models (VLMs) [8], [9] leverage a large number of web-sourced image-text pairs to integrate visual and textual elements within a unified embedding space. This approach has been effectively utilized in a variety of downstream tasks in computer vision, such as video understanding [12], person re-identification [11] and more. To enhance ER through VLM, CLIPER [17] leverages a simple hand-crafted prompt ("a photo of [class]") to enhance global facial expression representations. Despite its success, the existing VLM-based techniques primarily concentrate on holistic semantic alignment, thereby neglecting the learning

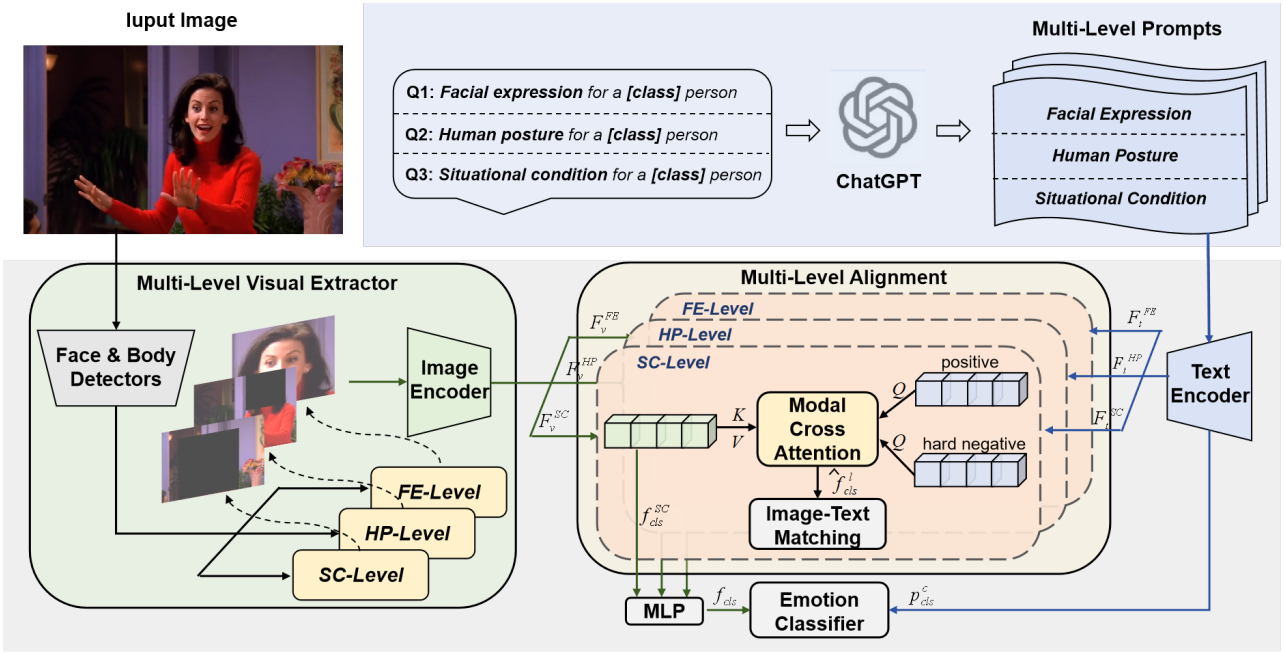


Fig. 2: The pipeline of MLPER model. First, to provide detailed descriptions for each emotion category, the Multi-Level Prompts from three aspects: *facial expression*, *human posture* and *situational condition* are generated by large language models, like ChatGPT. Correspondingly, one input image is cropped with the face, the body and the context part. Then, the cropped images and their corresponding prompts are fed into image and text encoders, respectively, to obtain the embedded tokens. To well align vision and text, the positive (matched) and the hard negative (not matched) textual embedded tokens query the visual embedded tokens at each level with Modal Cross Attention, to predict whether a pair of image and text is matched, respectively. In addition, the visual class tokens from three levels are concatenated and fed into Multi-Layer Perceptron (MLP). Then, the similarity between the total visual class and textual class tokens is maximized.

ambiguity introduced by modal heterogeneity in unconstrained environments. Thus, in this work, we aim to address this gap by focusing on the fine alignment of vision and text, offering an innovative perspective for refining ER features.

III. METHOD

The overall architecture of our proposed MLPER is illustrated in Fig. 2. First, to provide detailed descriptions for each emotion category, the Multi-Level Prompts from three aspects: *facial expression*, *human posture* and *situational condition* are generated by large language models, like ChatGPT. Correspondingly, one input image is cropped with the face, the body and the context part. These segmented images, along with their respective prompts, are processed by image and text encoders to produce embedded tokens. To achieve a seamless integration of visual and textual information, Modal Cross Attention is employed, where the embedded tokens from positive (matched) and hard negative (mismatched) text prompts are used to query the visual embedded tokens at each level, determining the congruence of image-text pairs. In addition, the visual class tokens from three levels are concatenated and fed into Multi-Layer Perceptron (MLP). Then, the similarity between the total visual class and textual class tokens is maximized, which used to predict the final emotion.

A. Multi-Level Prompts

The existing Vision-Language Models (VLM) typically adopt a hand-crafted template (e.g., a photo of a [**class**]) to obtain textual knowledge, which lacks fine-grained prompts for ER, and can not solve the problem of modal heterogeneity, thus leading to poor generalization under unconstrained conditions. In this work, we find that *facial expression*, *human posture* and *situational condition* are three main factors for ER. Thus, we aim to propose Multi-Level Prompts to guide ER model to explore the distinguish representations.

Upon examining a large number of ER images, we think that the emotion of a person is affected by three key elements: *facial expression*, *human posture* and *situational condition*. Besides, we ask ChatGPT[1] "How to judge the emotion of the person in an image from visual perspective?", and the response can also be summarized as the above three factors. Therefore, we create three distinct question templates for each emotion category:

Q1: When the emotion of a person is [**class**], what are the *facial expression* representations?

Q2: How to judge whether the emotion of a person is [**class**] from *human posture* perspective?

Q3: What is the *situational condition* for the emotion of a person is [**class**]?

Through ChatGPT's answers, we can achieve fine-grained prompts from *facial expression*, *human posture* and *situational condition*, respectively. In TAB. I, we list all emotion categories. Taking the emotion "Surprise" for example, the prompts can be summarized as "Alongside **unexpected events or reveals**, the **surprised** person has **widened eyes,**

[1]https://chat.openai.com

raised eyebrows, an open mouth and **reflexive body gestures**." that considering all the three factors, which provides more details than the hand-crafted prompt "a photo of a surprised person".

B. Multi-Level Visual Extractor

To better align the vision and the text, we utilize extra face and body detectors to crop out the images from three levels: the face, body and context. Then, we use the Vision-Language Model CLIP (ViT-L/14) [8] as the backbone, and its image encoder is adopted to extract visual tokens from the cropped images of different levels, which can be expressed as following:

$$F_v^l = \left[f_{cls}^l; f_1^l; f_2^l; \ldots; f_j^l; \ldots; f_N^l\right] = \mathcal{F}_v^l(I^l), \quad (1)$$

where l is from one of the three levels: *facial expression* (FE), *human posture* (HP) and *situational condition* (SC). At l level, I^l is the cropped image, $\mathcal{F}_v^l$ represents the image encoder, N is the number of visual tokens, f_{cls}^l represents the visual class token and F_v^l is the concatenation of all the visual tokens.

C. Multi-Level Alignment

To achieve the fine-grained alignment of vision and text, we propose Multi-Level Alignment. Following the human intuition, one emotion may be similar to different categories on different levels. E.g., "Fear" has the same open mouth as "Surprise" on FE, while also shows the curled body similar to "Sadness" on HP. To promote ER model concentrate on distinguished representations, we also select the most similar textual tokens among the mismatched ones as the hard negative at each level. Hence, we adopt the positive (matched) and the hard negative (not matched) textual embedded tokens to query the visual embedded tokens at each level with Modal Cross Attention, to predict whether a pair of image and text is matched, respectively. Specifically, to better explore the interaction between vision and text, we select the most similar textual tokens among the mismatched ones as the hard negative at each level. For the text encoder, the textual tokens extracted by the text encoder are expressed as following:

$$F_t^l = \left[p_{cls}^l; p_1^l; p_2^l; \ldots; p_i^l; \ldots; p_M^l\right] = \mathcal{F}_t(T^l), \quad (2)$$

where T^l are the corresponding text prompts from the positive and the hard negative for the image I^l, $\mathcal{F}_t$ represents the text encoder, M is the number of textual tokens, p_{cls}^l means the textual class token and F_t^l is the concatenation of all the textual tokens.

Then, we employ Modal Cross Attention to explore the interaction between visual and textual tokens. At l level, the queries arise from the textual tokens of the positive and negative F_t^l, and the keys and values are from the visual tokens F_v^l:

$$Q_i = p_i^l \mathbf{W}^Q, K_j = f_j^l \mathbf{W}^K, V_j = f_j^l \mathbf{W}^V, \quad (3)$$

where $i \in \{cls, 1, 2, 3, \cdots, M\}, j \in \{cls, 1, 2, 3, \cdots, N\}$. The matrices $\mathbf{W}^Q \in \mathbb{R}^{d \times d_q}$, and $\mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_v}$ are linear projections. Each textual token p_i^l is refined by the visual token f_j^l. Formally,

$$\begin{aligned} m_{i,j} =& \frac{\exp\left(\gamma_{i,j}\right)}{\sum_{j=cls}^{N} \exp\left(\gamma_{i,j}\right)}, \\ \gamma_{i,j} =& \frac{Q_i K_j^\top}{\sqrt{d_k}}, \end{aligned} \quad (4)$$

where the attention weight $m_{i,j}$ represents the cross similarity between the textual token p_i^l and the visual token f_j^l. The i-th token can be refined by:

$$\begin{aligned} \hat{f}_i^l =& \mathbf{FFN}\left(\mathbf{Att}\left(Q_i, K, V\right)\right) \\ =& \mathbf{FFN}\left(\sum_{j=cls}^{N} m_{i,j} V_j\right), \end{aligned} \quad (5)$$

where $\mathbf{FFN}(\cdot)$ is a feed-forward network. The refined tokens can be obtained by aggregating pixel context information of all positions:

$$\hat{F}^l = \left[\hat{f}_{cls}^l; \hat{f}_1^l; \hat{f}_2^l; \ldots; \hat{f}_i^l; \ldots; \hat{f}_M^l\right]. \quad (6)$$

Further, we use $\hat{f}_{cls}^l$ to get the probability z^l ("1" for matched, "0" for mismatched) with a linear layer, predicting whether the textual prompts T^l and the image I^l are a matched pair. Image-Text Matching is optimized with a Binary Cross-Entropy loss based on z^l. Formally, it is defined as

$$\mathcal{L}_{itm} = \sum_{l \in \{FE, HP, SC\}} \lambda^l \cdot [-y^l \cdot \log\left(z^l\right) - \left(1 - y^l\right) \log\left(1 - z^l\right)], \quad (7)$$

where λ^l is the learnable adaptive coefficient and y^l indicates the Ground Truth (GT) modal source on l level ("1" for matched, "0" for mismatched).

Besides, we also apply the contrastive learning to align the visual and textual tokens in a unified embedding space. We concatenate the visual class tokens from three levels as the total visual class token with a linear layer:

$$f_{cls} = \mathbf{MLP}(\mathbf{Concat}(f_{cls}^{FE}, f_{cls}^{HP}, f_{cls}^{SC})), \quad (8)$$

where f_{cls} is the total visual class token, $\mathbf{MLP}$ is Multi-layer Perceptron, $\mathbf{Contact}$ denotes the concatenation operation and $f_{cls}^{FE}, f_{cls}^{HP}$ and f_{cls}^{SC} represents the visual class tokens from FE, HP and SC, respectively. Correspondingly, as shown in Sec. III-A, we combine text prompts from three levels to construct a complete sentence T^c describing each emotion, and extract the total textual class token:

$$F_t^c = [p_{cls}^c; p_1^c; p_2^c; \ldots; p_M^c] = \mathcal{F}_t(T^c), \quad (9)$$

where p_{cls}^c is the total textual class token. Then, We calculate the similarity between the total visual class token f_{cls} and the total textual class token p_{cls}^c:

$$\mathbf{Sim}(f_{cls}, p_{cls}^c) = \frac{\langle f_{cls}, p_{cls}^c \rangle}{\|f_{cls}\|_2 \|p_{cls}^c\|_2}. \quad (10)$$

Then, the similarity is used to predict the emotion, by a Classification loss $\mathcal{L}_{cls}$,

$$\mathcal{L}_{cls} = \sum_{c=1}^{C} [-y \cdot \log \left(\mathbf{Sim} \left(f_{cls}, p_{cls}^{c} \right) / \tau \right)], \tag{11}$$

where C is the number of the total emoiton classes in a given dataset, and y is the corresponding one-hot GT label for one input image.

D. Overall Objective Function

For the proposed MLPER model, all the parameters are jointly trained in an end-to-end manner. The overall objective function is:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{itm}, \tag{12}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{itm}$ represent the Classification loss and the Image-Text Matching loss, respectively, and λ is the hyper-parameter to balance the two objective functions.

IV. EXPERIMENTS

A. Datasets

CAER-S [5] includes $70,000$ static images extracted from video clips of 79 TV shows. These images are annotated with 7 emotion categories: Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral. Following the previous works [5], [6], [7], we randomly split the dataset into 70% used for training, 10% used for validation, and 20% used for test.

RAF-DB [18], a real-world expression dataset, consists of $29,672$ facial images annotated by approximately 40 trained workers. For our studies, we utilize the single-label subset of RAF-DB, with the same emotion categories in CAER-S. This subset includes $12,271$ images for training and $3,068$ images for test.

AffectNet [19] is the largest FER dataset so far, where about $420,000$ images are manually annotated. Previous works [20], [21], [3], [4] vary in two different versions of AffectNet-7 and AffectNet-8 (with Contempt). AffectNet-7 contains $283,901$ images used for training and $3,500$ images used for test, while AffectNet-8 contains $287,651$ images used for training and $4,000$ images used for test.

FED-RO [22] is a FER dataset with real occlusions in the wild. There is no repeated image included in RAF-DB or AffectNet. There are 400 images in total, with the same seven expressions as RAF-DB.

Following the previous works [16], [3], the classification accuracy is used for evaluation.

B. Implementation Details

We use YOLOX [23] to detect the face and body part. For training, all the images are resized to 224×224 pixels. We perform three commonly used augmentation operations containing random horizontal flip, random erasing and random color changes. The pre-trained CLIP (ViT-L/14) is used as the backbone. All the models are trained using Pytorch framework with a batch size 16 for 60 epochs. The SGD optimizer is adopted with an initial learning rate of $2e^{-5}$, a momentum of 0.9 and a weight decay of $1e^{-4}$, and the learning rate is decayed by a factor of 0.9 after each epoch.

C. Comparison with State-of-the-Art Methods

As shown in TAB. II-III, we compare our proposed MLPER model with the state-of-the-art methods on CAER-S, RAF-DB and AffectNet-7, respectively. We also illustrate the confusion matrices in Fig. 3.

Results on CAER-S. As shown in TAB. II, our proposed MLPER model achieves the accuracy of 91.95%, outperforming the current state-of-the-art CCIM [16] by 0.78%. As illustrated in Fig. 3 (a), our proposed MLPER model demonstrates similar discrimination capabilities across each emotion category, with the accuracy ranging from 87% to 97%. The reason is that our MLPER can concentrate on distinguished representations of each category.

TABLE II: Performance comparison (%) with the state-of-the-art methods on CAER-S.

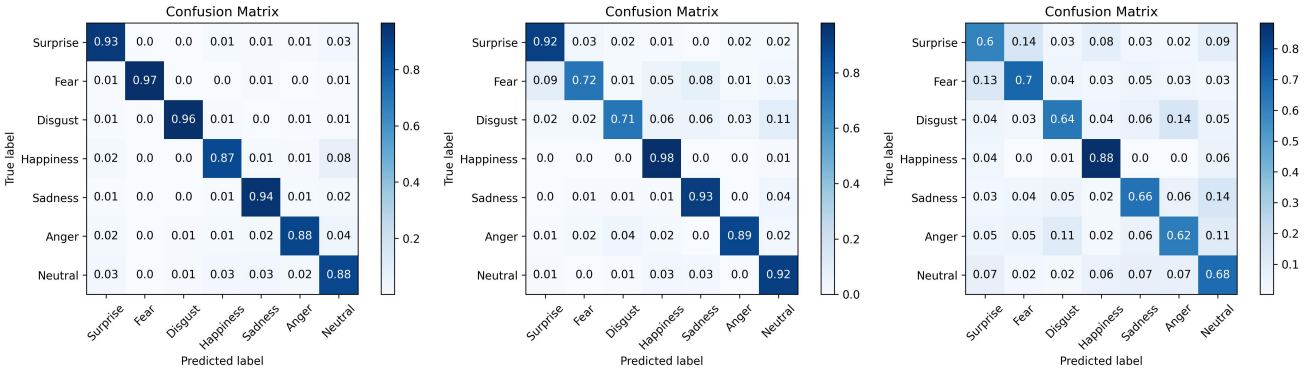
Method	CAER-S
CAER-Net [5]	73.51
EMOT-Net [6]	74.51
SIB-Net [24]	74.56
GNN-CNN [25]	77.21
RRLA [26]	84.82
EmotiCon [7]	88.65
VRD [27]	90.49
CCIM [16]	91.17
MLPER (Ours)	**91.95**

Results on RAF-DB. As shown in TAB. III, our proposed MLPER model performs the best, and achieves the accuracy of 92.70%, outperforming the current state-of-the-art APViT [3] by 0.72%. From Fig. 3 (b), "Happiness" has the highest accuracy with 98%, while "Disgust" has the lowest accuracy with 71%. This discrepancy of the accuracy may be attributed to the relative scarcity of negative expressions (e.g., Anger, Disgust, Fear, and Sadness) in RAF-DB dataset.

Results on AffectNet-7. Given the uneven distribution between the training and test sets, we adopt the same sampling strategy utilized by DAN [20] to ensure consistency and fairness in our evaluations. For AffectNet-7, our proposed MLPER achieves the accuracy of 68.23%, which is better than WSCFER [2] by 0.53%. As shown in Fig. 3 (c), our proposed MLPER demonstrates comparable discrimination capabilities (ranging from 60% to 70%) across each category, with the exception of "Happiness" (the highest is 88%).

TABLE III: Performance comparison (%) with the state-of-the-art methods on RAF-DB and AffectNet-7.

Method	RAF-DB	AffectNet-7
DAN [20]	89.70	65.69
EAC [15]	89.99	65.32
LDLVA [28]	90.51	66.23
TransFER [21]	90.91	66.23
GAAVE [29]	91.53	66.11
LA-Net [4]	91.56	67.60
WSCFER [2]	91.72	67.71
APViT [3]	91.98	66.91
MLPER (Ours)	**92.70**	**68.23**



(a) Confusion Matrix on CAER-S (b) Confusion Matrix on RAF-DB (c) Confusion Matrix on AffectNet-7

Fig. 3: Confusion matrices of our MLPER model on several datasets.

D. Evaluation on Realistic Occlusions and Pose Variations

In order to demonstrate the generalization ability of our proposed MLPER, we also perform experiments on the realistic occlusion and variant pose datasets. For FED-RO, we combine RAF-DB and AffectNet as the training set, following the previous works [30], [22], [31]. As shown in TAB. IV-VI, our MLPER all achieves the best performance. The reason is that our MLPER can concentrate on the discriminative visual tokens guided by Multi-Level Prompts.

TABLE IV: Performance comparison (%) with the state-of-the-art methods on occlusion and variant pose RAF-DB.

Method	Occlusion	Pose ($\geq 30°$)	Pose ($\geq 45°$)
RAN [30]	82.72	86.74	85.20
MA-Net [31]	83.65	87.89	87.99
VTFF [32]	83.95	87.97	88.35
FG-AGR [33]	88.15	91.02	90.50
MLPER (Ours)	**89.39**	**91.50**	**91.04**

TABLE V: Performance comparison (%) with the state-of-the-art methods on occlusion and variant pose AffectNet-8.

Method	Occlusion	Pose ($\geq 30°$)	Pose ($\geq 45°$)
RAN [30]	58.50	53.90	53.19
MA-Net [31]	59.59	57.51	57.78
VTFF [32]	62.98	60.61	61.00
FG-AGR [33]	64.24	61.26	61.15
MLPER (Ours)	**64.86**	**61.63**	**62.87**

TABLE VI: Performance comparison (%) with the state-of-the-art methods on FED-RO.

Method	FED-RO
gACNN [22]	66.50
RAN [30]	67.98
MA-Net [31]	70.00
MLPER (Ours)	**76.50**

E. Evaluation on Cross-Dataset

To further demonstrate the generalization ability of our MLPER, we also perform a cross-dataset evaluation. We conduct training of the ER model using the train set of RAF-DB and AffectNet-7 separately, and subsequently assess its performance on the test set of the other. As shown in TAB. VII, our MLPER also achieves the best. The reason is that our MLPER adopts Multi-Level Prompts to guide the ER model to focue on distinguished regions. Besides, MLPER achieves fine-grained alignment between vision and text, and thus has strong generalization ability.

TABLE VII: Cross-dataset evaluation comparisons (%) on RAF-DB and AffectNet-7.

Method	Train RAF-DB	Test AffectNet-7	Train AffectNet-7	Test RAF-DB
DLP-CNN[34]	38.37		72.43	
IL-CNN [35]	39.31		73.99	
LDAM [36]	46.51		74.09	
STSN [37]	48.49		76.99	
FG-AGR [33]	48.86		74.79	
KTN [37]	49.60		76.53	
MLPER (Ours)	**54.06**		**82.29**	

F. Ablation Studies

For simplicity, all the ablation studies are conducted on the CAER-S datasets.

Effect of Multi-Level Prompts. To validate the MLPER model, an ablation study is designed to investigate the effects of Multi-Level Prompts: *facial expression*, *human posture* and *situational condition*. The results are summarized in TAB. VIII. To analyze the effect of our fine-grained prompts, we also compare MLPER with the hand-crafted prompt from CLIP (the top row in TAB. VIII). As observed, on CAER-S dataset, each level can significantly improve the baseline model CLIP [8]. E.g., FE improves the baseline model from 89.50% to 90.17%.

TABLE VIII: Performance comparison (%) of Multi-Level Prompts in MLPER on CAER-S dataset. FE, HP and SC represent *facial expression*, *human posture* and *situational condition*, respectively.

FE	HP	SC	CAER-S
-	-	-	89.50
✓	-	-	90.17
-	✓	-	90.07
-	-	✓	89.96
✓	✓	-	90.58
✓	-	✓	90.40
-	✓	✓	90.32
✓	✓	✓	**90.88**

Effect of Multi-Level Alignment. To achieve finely vision-text alignment, we propose to fuse the visual tokens and textual tokens from the positve and the hard negative at each level, then predicting whether the pair of image and text is matched. Hence, an ablation study is designed to investigate the effects of the alignment at each level. As shown in TAB. IX, our proposed Multi-Level Alignment actually promote modal interaction, thereby eliminating modal differences. E.g., the alignment of FE improves the accuracy from 90.88% to 91.20%.

TABLE IX: Performance comparison (%) of Multi-Level Alignment in MLPER on CAER-S dataset. FE, HP and SC represent *facial expression*, *human posture* and *situational condition*, respectively.

FE	HP	SC	CAER-S
-	-	-	90.88
✓	-	-	91.20
-	✓	-	91.17
-	-	✓	91.00
✓	✓	-	91.63
✓	-	✓	91.50
-	✓	✓	91.40
✓	✓	✓	**91.95**

Effect of the hyperparameter λ. In order to balance the Classification loss and the Image-Text Matching loss, we set a hyper-parameter λ in (12). To evaluate the effect of λ, we set it to $0, 0.5, 1, 1.5, 2$, respectively. The results are summarized in TAB. X. It can be found that MLPER achieves the best results when $\lambda = 1$. Thus, for comparison with other state-of-the-art methods, we set λ to 1 on all the datasets.

TABLE X: Effect of the parameter λ on CAER-S dataset.

λ	0	0.5	1	1.5	2
Acc (%)	90.88	91.68	**91.95**	91.80	91.57

Effect of different backbones. To evaluate the effect of different backbones, we replace ViT-L/14 with ViT-B/32 and ViT-L/16 to extract the basic tokens. As shown in Tab XI, with different backbones, MLPER all significantly enhances the accuracy, compared to the baselines. E.g., MLPER improves ViT-B/32 from 86.85% to 90.09%. With ViT-B/16 as the backbone, MLPER achieves the accuracy of 91.30%, outperforming the current state-of-the-art CCIM [16], which adopts ResNet152 as the backbone with the similar parameter quantities, thereby demonstrating the effectiveness of MLPER.

TABLE XI: Performance comparison (%) of the different backbones in MLPER on CAER-S dataset.

Method	CAER-S
ViT-B/32 Baseline	86.85
ViT-B/16 Baseline	88.95
ViT-L/14 Baseline	89.50
ViT-B/32 + MLPER	90.09
ViT-B/16 + MLPER	91.30
ViT-L/14 + MLPER	**91.95**

G. Visualization

We use Grad-Cam to visualize the attention maps extracted from the last layer of the image encoder, as shown in Fig. 4. The first row shows the original images with the GT labels. The second and the third rows show the attention maps from the baseline CLIP [8] and our proposed MLPER, respectively. The baseline model often concentrates on non-discriminative regions, resulting in incorrect predictions. Conversely, our MLPER effectively identifies essential areas under the guidance of Multi-Level Prompts, thereby acquiring discriminative representations.

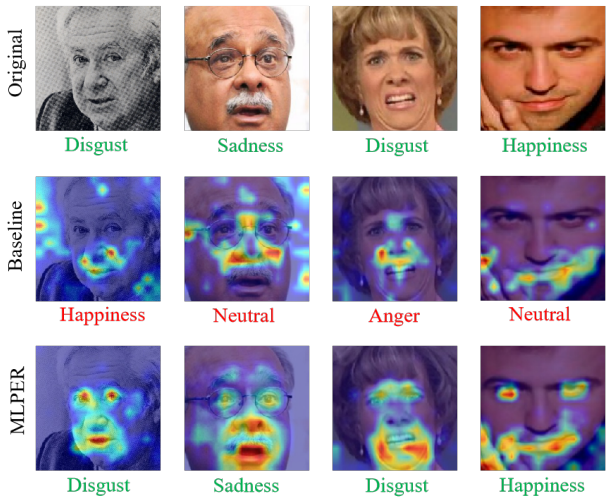
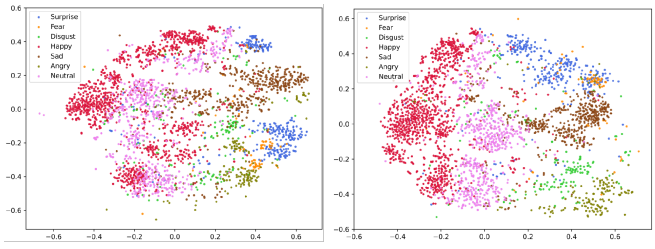


Fig. 4: Attention maps of some examples on the test set of RAF-DB and AffectNet.

Besides, we employ t-SNE to visualize the representations extracted by the baseline CLIP [8] and our proposed MLPER on the test set of RAF-DB, as shown in Fig. 5. As observed, the representations extracted from MLPER can achieve relatively better intra-class compactness and inter-class separation. In contrast, the expression representations extracted from the baseline [8] are relatively diffused.



(a) Baseline [8] (b) MLPER (Ours)

Fig. 5: Visualization of expression representations on the test set of RAF-DB using t-SNE.

V. CONCLUSIONS

Emotion Recognition plays a pivotal role in the robotics field, enhancing human-robot interaction by enabling robots to interpret and respond to human emotions, thereby significantly improving their effectiveness in assistance, healthcare, education, and customer service applications. In this

work, we propose MLPER model, which introduces Vision-Language Model for Emotion Recognition to learn discriminative representations adaptively. Specifically, we first establish Multi-Level Prompts from three aspects: *facial expression*, *human posture* and *situational condition* using large language models, like ChatGPT. Correspondingly, we extract the visual tokens from three levels: the face, body, and context. Further, to achieve fine-grained alignment at each level, we adopt textual tokens from the positive and the hard negative to query visual tokens, predicting whether a pair of image and text is matched. Experimental results demonstrate that our MLPER model outperforms the state-of-the-art methods on several ER benchmarks.

REFERENCES

[1] S. Akhyani, M. Abbasi, M. Chen, and A. Lim, "Towards inclusive hri: Using sim2real to address underrepresentation in emotion expression recognition," in *IEEE International Conference on Intelligent Robots and Systems*, 2022, pp. 9132–9139.

[2] W. Nie, B. Chen, W. Wu, X. Xu, W. Ren, and H. Liu, "Wscfer: Improving facial expression representations by weak supervised contrastive learning," in *IEEE International Conference on Intelligent Robots and Systems*, 2023, pp. 9816–9823.

[3] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Transactions on Affective Computing*, 2022.

[4] Z. Wu and J. Cui, "La-net: Landmark-aware learning for reliable facial expression recognition under label noise," in *IEEE International Conference on Computer Vision*, 2023, pp. 20 698–20 707.

[5] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *IEEE International Conference on Computer Vision*, 2019, pp. 10 143–10 152.

[6] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1667–1675.

[7] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 234–14 243.

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

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022, pp. 12 888–12 900.

[10] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Zero-shot temporal action detection via vision-language prompting," in *European Conference on Computer Vision*. Springer, 2022, pp. 681–697.

[11] S. Li, L. Sun, and Q. Li, "Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels," in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1405–1413.

[12] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[13] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[14] P. C. Ng and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

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

[16] D. Yang, Z. Chen, Y. Wang, S. Wang, M. Li, S. Liu, X. Zhao, S. Huang, Z. Dong, P. Zhai, *et al.*, "Context de-confounded emotion recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 005–19 015.

[17] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Cliper: A unified vision-language framework for in-the-wild facial expression recognition," *arXiv preprint arXiv:2303.00193*, 2023.

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

[19] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

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

[21] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *IEEE International Conference on Computer Vision*, 2021, pp. 3601–3610.

[22] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[24] X. Li, X. Peng, and C. Ding, "Sequential interactive biased network for context-aware emotion recognition," in *IEEE International Joint Conference on Biometrics*, 2021, pp. 1–6.

[25] M. Zhang, Y. Liang, and H. Ma, "Context-aware affective graph reasoning for emotion recognition," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2019, pp. 151–156.

[26] W. Li, X. Dong, and Y. Wang, "Human emotion recognition with relational region-level analysis," *IEEE Transactions on Affective Computing*, 2021.

[27] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "Context-aware emotion recognition based on visual relationship detection," *IEEE Access*, vol. 9, pp. 90 465–90 474, 2021.

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

[29] J. Zheng, B. Li, S. Zhang, S. Wu, L. Cao, and S. Ding, "Attack can benefit: An adversarial approach to recognizing facial expressions under noisy annotations," in *the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3660–3668.

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

[31] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021.

[32] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, 2023.

[33] C. Li, X. Li, X. Wang, D. Huang, Z. Liu, and L. Liao, "Fg-agr: Fine-grained associative graph representation for facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[34] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018.

[35] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *IEEE International Conference on Automatic Face Gesture Recognition*, 2018, pp. 302–309.

[36] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[37] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via cf labels and distillation," *IEEE Transactions on Image Processing*, vol. 30, pp. 2016–2028, 2021.