Prompting Visual-Language Models for Dynamic Facial Expression Recognition

Zengqun Zhao
zengqun.zhao@qmul.ac.uk

Ioannis Patras
i.patras@qmul.ac.uk

School of Electronic Engineering and
Computer Science,
Queen Mary University of London,
London, E1 4NS, UK

Abstract

This paper presents a novel visual-language model called DFER-CLIP, which is based on the CLIP model and designed for in-the-wild Dynamic Facial Expression Recognition (DFER). Specifically, the proposed DFER-CLIP consists of a visual part and a textual part. For the visual part, based on the CLIP image encoder, a temporal model consisting of several Transformer encoders is introduced for extracting temporal facial expression features, and the final feature embedding is obtained as a learnable "class" token. For the textual part, we use as inputs textual descriptions of the facial behaviour that is related to the classes (facial expressions) that we are interested in recognising – those descriptions are generated using large language models, like ChatGPT. This, in contrast to works that use only the class names and more accurately captures the relationship between them. Alongside the textual description, we introduce a learnable token which helps the model learn relevant context information for each expression during training. Extensive experiments demonstrate the effectiveness of the proposed method and show that our DFER-CLIP also achieves state-of-the-art results compared with the current supervised DFER methods on the DFEW, FERV39k, and MAFW benchmarks. Code is publicly available at https://github.com/zengqunzhao/DFER-CLIP.

1 Introduction

Facial expression is an important aspect of daily conversation and communication [6]. Because of its application in various fields, such as human-computer interaction (HCI) [8], driving assistance [46], and mental health [2, 11], facial expression recognition (FER) attracts increasing attention from researchers in psychology, computer science, linguistics, neuroscience, and related disciplines [5]. When using a discrete emotion model [9], FER aims to classify an image or video sequence into one of several basic emotions, i.e., neutral, happiness, sadness, surprise, fear, disgust, and anger. Traditional facial expression recognition approaches have mainly focused on static images or video frames, which do not capture the temporal dynamics of facial expressions [24, 48]. However, the recognition of dynamic facial expressions involves capturing the changes in facial movements over time, which can provide more information for accurate emotion recognition. Therefore, the recognition of dynamic facial expressions has become an increasingly important research area within the field of computer vision and affective computing [24, 43].
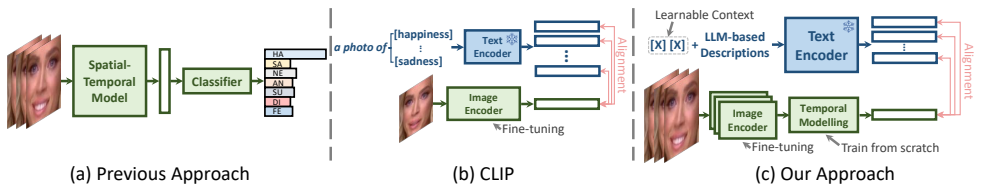
Figure 1: Illustration of the (a) previous approach for DFER, which relies on a classifier to predict the facial expression label. (b) Standard vision-language model CLIP. (c) Based on the CLIP, we propose a DFER-CLIP, which further models temporal facial features and incorporates fine-grained descriptors.

Early studies for dynamic facial expression recognition (DFER) are mainly focused on lab-controlled conditions [24], in which the faces are all frontal and have no occlusions. However, in the real world, facial expressions always occur accompanied by variations such as illumination, non-frontal pose, and occlusion. Therefore, works on in-the-wild DFER are getting popular in recent years [16, 27, 28, 44, 47]. The main focus is trying to learn discriminative and robust spatial-temporal feature representations for the DFER task. These methods can be roughly categorized into 3DCNN-based [16, 19], CNN-RNN-based [1, 26, 45], and Transformer-based [10, 22, 28, 47], in which Transformer-based methods have achieved state-of-the-art performance. In addition, many in-the-wild datasets were proposed in recent years for DFER [16, 27, 44]. However, the high cost of labelling facial expression data limits the amount of training data available for DFER models. This can have an impact on the model's performance, especially if the training data does not fully capture the diversity of facial expressions that may occur in the real world.

Recently, vision-language pre-training (V-LP) models such as CLIP [33] have emerged as promising alternatives for visual representation learning. The main idea of the V-LP models is to align large connections of images and raw texts using two separate encoders, so as to learn semantic information between visual and textual data. The success of V-LP models has led to a growing interest in using them for various downstream computer vision tasks, such as video understanding [17, 39], image synthesis [36], and semantic segmentation [29]. Given the powerful representation learning capabilities of V-LP models, a relevant question is how to best exploit their potential for the DFER task. One proposed approach is to fine-tune the image encoder of a V-LP model on DFER datasets. However, this approach faces two challenges that must be addressed. Firstly, while the original CLIP model can recognize objects and scenes in images, it may not be as effective in recognizing subtle facial expressions, which require more fine-grained descriptors and modelling of the similarities between expressions. Secondly, learning robust temporal facial features to understand emotions is crucial for DFER. Unfortunately, the standard CLIP visual encoder encodes static images and therefore lacks ways of capturing temporal information.

To address these challenges, based on the CLIP model, we propose a novel architecture, namely the DFER-CLIP model. An overview of the differences between DFER-CLIP with previous approaches and CLIP is shown in Fig. 1 and the detailed architecture is shown in Fig. 2. The proposed method mainly consists of a visual part and a textual part. Regarding the visual part, based on the CLIP image encoder, we introduce a temporal model, which consists of several Transformer encoders [7], for modelling the temporal facial features. The final video-level facial features are obtained by a learnable class token. Regarding the textual
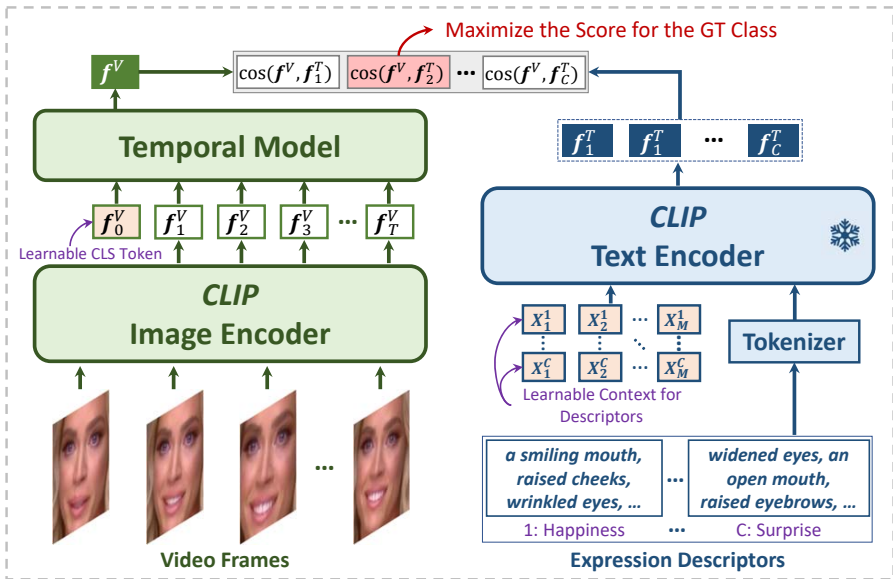
Figure 2: The structure of the proposed DFER-CLIP model. cos() denotes cosine similarity. M denotes the learnable context number. C denotes expression classes.

part, considering that different facial expressions have both common properties and unique or special properties at the level of local behaviour [49, 51], we utilize descriptions related to facial behaviour instead of class names for the text encoder. As a result, the text embedding can provide more detailed and precise information about the specific movements or positions of muscles involved in each expression. Furthermore, inspired by CoOp [53], we adopt the learnable prompt as a context for descriptors of each class – this does not require experts to design context words and allows the model to learn relevant context information for each expression during training. To evaluate our DFER-CLIP model, we conducted experiments on three datasets. The results show that the temporal model can clearly enhance performance, and adopting the fine-grained expression descriptions with learnable context is superior to class-level prompts. Furthermore, compared with the current supervised DFER methods, the proposed DFER-CLIP achieves state-of-the-art results on DFEW, FERV39k, and MAFW benchmarks.

2 Related Work

2.1 DFER in the Wild

For in-the-wild DFER, deep learning-based methods have achieved superior performance to methods using hand-crafted features in recent years [24]. Those methods are mainly focusing on learning robust spatial-temporal facial expression features, and can be roughly categorized into 3DCNN-based [16, 19], CNN-RNN-based [1, 26, 45], and Transformer-based [11, 22, 28, 47] methods. To incorporate more modalities for emotion understanding, multi-modal learning is also prevalent in DFER in the wild [4, 52]. Zhao et al. [47] first

introduced the Transformer into DFER and achieved state-of-the-art results. After that, many transformer-based methods [20, 22, 40] were proposed to further enhance the performance. In the latest work, Li *et al.* [21] proposed a model called CLIPER that uses visual language models for the FER task, achieving state-of-the-art performance. However, the model is limited in its ability to capture temporal information for dynamic FER and uses a two-stage training paradigm instead of end-to-end training. By contrast, our DFER-CLIP model is capable of learning spatial-temporal features and is trained in an end-to-end manner.

2.2 Visual-Language Models

Vision-language models have made significant progress in zero-shot performance through contrastive learning on pairs of images and text [17, 31, 35, 38, 39, 52, 53]. This approach leverages a vast and virtually limitless supply of image-text pairs collected from the internet, making the process both highly efficient and cost-effective. The most recent vision-language learning frameworks CLIP [38] and ALIGN [15] benefit from 400 million and 1.8 billion image-text pairs collected from the Internet. These large-scale vision-language models are then applied to various downstream computer vision tasks and achieve stage-of-the-art performance [17, 33, 34, 36, 39]. For the visual-language modes, the researchers usually utilize a prompt (e.g., "a photo of a") along with the name of the class to generate a text descriptor for each class. Zhou *et al.* [53] introduced a learnable continuous prompt method named context optimization (CoOp), which can adaptively adjust according to different datasets. Furthermore, Zhou *et el.* [52] proposed conditional context optimization (CoCoOp) based on CoOp, which can generate a prompt conditioned on each image and make better inferences on unseen classes. Li *et al.* [23] proposed a new V-LP model named BLIP which achieves state-of-the-art performance on both understanding-based and generation-based tasks. BLIP trains a multimodal mixture of encoder-decoder models on large-scale noisy image-text pairs. In the latest work, Li *et al.* [25] proposed a simple and more efficient method for training CLIP, named Fast Language-Image Pre-training (FLIP). They randomly mask out and remove a large portion of image patches during training, allowing more image-text pairs can be used for training under the same overheads. Our method is designed based on the CLIP model.

3 Methods

3.1 Overview

The structure of the DFER-CLIP is shown in Fig. 2. As we can see, the proposed DFER-CLIP consists of a visual part and a textual part. Regarding the visual part, based on the CLIP image encoder, we introduce a temporal model, which consists of several Transformer Encoders, for modelling the temporal facial features. The final video-level facial features are obtained by a learnable class token. Regarding the textual part, we utilize descriptions related to facial behaviour instead of class names for the text encoder. Furthermore, we adopt the learnable prompt as a context for descriptors of each class, which does not require experts to design context words and allows the model to learn relevant context information for each expression during training. We will first introduce the detail of the DFER-CLIP model and then show how we build textural descriptions for facial expressions.

3.2 DFER-CLIP

Regarding the visual part, on top of the CLIP image encoder, the temporal model consisting of several ViT encoder layers is adopted for modelling the temporal relationship. Each encoder also consists of a multi-headed self-attention and feed-forward network, which are all trained from scratch. The frame-level features are first learnt by the shared CLIP visual encoder. Then all of the frame-level features along with an additional learnable class token will feed into the temporal model, in which the learnable position embedding is added to encode the temporal position.

Concretely, given a facial video, we sample T frames of the size $H \times W$ so as to form an input $x \in \mathbb{R}^{T \times 3 \times H \times W}$. For each frame x_i, we first utilize a shared CLIP image encoder $f(\cdot)$ to extract feature vectors $\boldsymbol{f}_i^v \in \mathbb{R}^L$, where $i \in \{1, 2, \cdots, T\}$, L is the length of the feature vectors. Then T feature vectors will feed into a temporal model $g(\cdot)$ for learning temporal features, and the final visual representation $\boldsymbol{f}^V \in \mathbb{R}^L$ can be obtained:

$$\boldsymbol{f}_i^v = f(x_i) \tag{1}$$

$$\boldsymbol{f}^V = g(\boldsymbol{f}_0^v + \boldsymbol{e}_0, \boldsymbol{f}_1^v + \boldsymbol{e}_1, ..., \boldsymbol{f}_T^v + \boldsymbol{e}_T) \tag{2}$$

where $\boldsymbol{f}_0^v$ is a special learnable vector for the class token, $\boldsymbol{e}$ represents a learnable position embedding added to encode the temporal position.

Regarding the textual part, we utilize descriptions related to facial behaviour instead of class names for the text encoder. Furthermore, we adopt the learnable prompt as a context for descriptors of each class. The form of the prompt is as follows:

$$\boldsymbol{P}_k = [p]_k^1 [p]_k^2 \cdots [p]_k^M [Tokenizer(description)]_k \tag{3}$$

where M is a hyperparameter specifying the number of context tokens, $k \in \{1, 2, \cdots, C\}$, C is the number of the facial expression classes, and each $[p]_k^m$, $m \in \{1, 2, \cdots, M\}$, is a vector with the same dimension as word embeddings. Here, we adopt class-specific context where context vectors are independent of each description. By forwarding a prompt $\boldsymbol{P}_k$ to the text encoder $h(\cdot)$, we can obtain C classification weight vector $\boldsymbol{f}_k^T \in \mathbb{R}^L$ representing a visual concept:

$$\boldsymbol{f}_k^T = h(\boldsymbol{P}_k) \tag{4}$$

Then the prediction probability can be computed as:

$$p(y = k|x) = \frac{exp(cos(\boldsymbol{f}^V, \boldsymbol{f}_k^T)/\tau)}{\sum_{k'=1}^{C} exp(cos(\boldsymbol{f}^V, \boldsymbol{f}_{k'}^T)/\tau)} \tag{5}$$

where τ is a temperature parameter learned by CLIP and $cos(\cdot, \cdot)$ denotes cosine similarity.

During the training phase, the CLIP text encoder is fixed and we fine-tune the CLIP image encoder. The temporal model, learnable class token, and learnable context are all learned from scratch. DFER-CLIP is trained end to end and the cross-entropy loss is adopted for measuring the distance between the prediction and the ground-truth labels.

3.3 Building Descriptions

Previous research has indicated that human facial expressions share some common properties while also possessing unique or distinctive features [5,1]. This suggests that there are

similarities in the movements of local facial muscles across different expressions. For example, the expressions of happiness and surprise both involve the raising of the eyebrows, and the expressions of sadness and anger both involve the lowering of the eyebrows and the furrowing of the forehead. Considering the fact that the CLIP text encoder learned semantic information from natural language text [38], we propose taking the facial expression action description as the input for the text encoder. Instead of manually designing the facial expression description, we prompt a large language model such as ChatGPT to automatically generate descriptions based on contextual information. We prompt the language model with the input:

Q: What are useful visual features for the facial expression of {class name}?

A: Some useful visual features for facial expressions of {class name} include: ...

All the generated descriptors of each facial expression class will be combined to form a comprehensive description. The descriptions for the classification of the seven basic emotions and of the classes in the datasets we used can be found in the *supplementary material*.

4 Experiments

4.1 Datasets

To evaluate our method, we conduct experiments on three popular DFER benchmarks.

DFEW [16] dataset contains 11,697 in-the-wild video clips, and all the samples have been split into five same-size parts without overlap. Each video is individually annotated by ten annotators under professional guidance and assigned to one of the seven basic expressions (i.e., happiness, sadness, neutral, anger, surprise, disgust, and fear). The video clips are collected from over 1,500 movies worldwide, covering various challenging interferences, such as extreme illuminations, occlusions, and variant head pose. The 5-fold cross-validation is adopted as an evaluation protocol.

FERV39k [44] dataset contains 38,935 in-the-wild video clips, which is currently the largest in-the-wild DFER dataset. All the video clips were collected from 4 scenarios, which can be further divided into 22 fine-grained scenarios, such as crime, daily life, speech, and war. Each clip is annotated by 30 individual annotators and assigned to one of the seven basic expressions as DFEW. Video clips of all scenes are randomly shuffled and split into training (80%), and testing (20%) without overlapping.

MAFW [27] dataset contains 10,045 in-the-wild video clips, which is the first large-scale, multi-modal, multi-label affective database with 11 single-expression categories (anger, disgust, fear, happiness, sadness, surprise, contempt, anxiety, helplessness, disappointment, and neutral), 32 multiple-expression categories, and emotional descriptive texts. The 5-fold cross-validation is adopted as an evaluation protocol.

4.2 Implementation Details

Training Setting: For the visual input, all the fixed 16 frames of sequence in our experiments follow the sampling strategy in [16, 20, 21, 22, 30, 47], and then resized to 224×224. The random resized crop, horizontal flipping, random rotation, and color jittering are used to avoid overfitting. Considering the high computational cost of video data, ViT-B/32-based CLIP [38] is employed in our DFER-CLIP. The maximum number of textual tokens is 77 (following the official CLIP design), and the temperature hyper-parameter τ is set to 0.01.

Number of the Context Prompts	Depth of the Temporal Model	DFEW		FERV39k		MAFW	
		UAR	WAR	UAR	WAR	UAR	WAR
4	✗	56.91	**69.01**	40.26	**50.96**	38.03	50.62
8	✗	**57.39**	69.00	**40.64**	50.92	**38.51**	**50.91**
16	✗	57.32	68.96	40.22	50.64	37.98	50.40
8	1	**59.61**	**71.25**	**41.27**	**51.65**	**39.89**	**52.55**
8	2	58.87	70.92	40.41	51.08	39.13	52.10
8	3	58.64	70.80	40.35	50.98	38.90	51.86

Table 1: Evaluation of the learnable context prompt numbers & the temporal model depths.

Our models are trained on the Tesla A100 GPU based on the open-source PyTorch platform. The parameters are optimized via the SGD optimizer with a mini-batch size of 48. The initial learning rate for the CLIP image encoder, temporal model, and learnable prompt is set as 1×10^{-5}, 1×10^{-2}, and 1×10^{-3}, respectively. The MultiStepLR with milestones of $\{30, 40\}$ and gamma of 0.1 is employed as a scheduler to adjust the learning rate. The DFER-CLIP model is trained for 50 epochs in an end-to-end manner. To obtain more stable and reliable results, we train models three times with different random seeds and then use the average as the final result.

Evaluation Metrics: Consistent with most previous methods [16, 20, 21, 22, 30, 47], the weighted average recall (WAR, i.e., accuracy) and the unweighted average recall (UAR, i.e., the accuracy per class divided by the number of classes without considering the number of instances per class) are utilized for evaluating the performance of methods.

4.3 Ablation Analysis

To evaluate the effectiveness of each component in our DFER-CLIP, we conduct exhaustive ablation analysis on all three benchmarks. Specifically, we will first demonstrate the effectiveness of the temporal model, the effect of the different number of context prompts, and the depth of the temporal model. We will then compare different training strategies, including classifier-based and text-based (classifier-free). Finally, we will compare different prompts for the DFER task.

Evaluation of Temporal Model & Context Prompts: As aforementioned, learning temporal facial features is crucial for video-based FER tasks. From Tab. 1, we can see that by adopting the temporal model, the UAR performance can be improved by 2.22%, 0.63%, and 1.38%, and WAR performance can be improved by 2.25%, 0.73%, and 1.64% on DFER, FERV39k, and MAFW datasets, respectively. In general, deeper models can get better performance, but in our DFER-CLIP, the best performance is obtained under the one-layer temporal model. This is because the temporal model is trained from scratch and may overfit if it is complex. This is also a consideration in the learnable context, in which more learnable vectors do not improve the results. We believe that increasing the learnable context number or temporal model depth will cause overfitting on the training data, resulting in a worse generalization performance on test data.

Evaluation of Different Training Strategies: The final recognition result in our DFER-CLIP is obtained by computing the similarities between visual embedding and all the textual class-level embeddings, which is different from the conventional classifier-based training

Strategies		DFEW		FERV39k		MAFW	
		UAR	WAR	UAR	WAR	UAR	WAR
Classifier-based	Linear Probe	45.46	57.40	32.47	43.72	30.74	42.95
	Fully Fine-Tuning (w/o TM)	55.70	68.41	39.64	50.77	37.53	50.48
	Fully Fine-Tuning (w/ TM)	58.28	70.27	40.55	51.22	38.39	50.92
Text-based (Classifier-free)	Zero-shot CLIP [58]	23.34	20.07	20.99	17.09	18.42	19.16
	Zero-shot FaRL [60]	23.14	31.54	21.67	25.65	14.18	11.78
	CoOp [53]	44.98	56.68	31.72	42.55	30.79	42.77
	Co-CoOp [52]	46.80	57.52	32.91	44.25	30.81	43.23
	DFER-CLIP (w/o TM) (Ours)	57.39	69.00	40.64	50.92	38.51	50.91
	DFER-CLIP (w TM) (Ours)	**59.61**	**71.25**	**41.27**	**51.65**	**39.89**	**52.55**

Table 2: Evaluation of different training strategies. Linear Probe: Adding a linear classifier to the CLIP image encoder and optimizing the parameters of the linear layer with the image encoder fixed. Fully fine-tuning: fine-tuning both the linear layer and the CLIP image encoder. Zero-shot CLIP/FaRL: Using the prompt "an expression of [class]". CoOp & Co-CoOp: Using the prompt "[learnable prompt] [class]". The image encoder in FaRL is ViT-B/16-based and the rest are all ViT-B/32-based. TM stands for the temporal model.

Prompts		DFEW		FERV39k		MAFW	
		UAR	WAR	UAR	WAR	UAR	WAR
w/o TM	a photo of [Class]	56.21	68.44	39.44	49.94	37.91	50.87
	an expression of [Class]	56.16	68.73	39.28	50.41	37.71	51.08
	[Learnable Prompt] [Class]	57.37	68.86	40.42	50.50	38.01	50.81
	[Learnable Prompt] [Descriptors]	57.39	69.00	40.64	50.92	38.51	50.91
w/ TM	[Learnable Prompt] [Class]	58.28	70.29	40.60	51.18	39.64	51.21
	[Learnable Prompt] [Descriptors]	**59.61**	**71.25**	**41.27**	**51.65**	**39.89**	**52.55**

Table 3: Evaluation of different prompts. We set the number of the learnable context prompts as 8 and with or without the temporal model for comparison.

strategies. Therefore, we compare DFER-CLIP with Linear Probe and Fully Fine-tuning methods. The results in Tab. 2 show that our method outperforms Fully Fine-tuning in UAR by 3.91%, 1.63%, and 2.36%, and in WAR by 2.84%, 0.88%, and 2.07% on DFER, FERV39k, and MAFW datasets, respectively. Even without the temporal model, our method is better than all the classifier-based methods. We also add a temporal model for the Fully Fine-tuning strategy, the results demonstrate our method is still superior to it. We also compare our method with zero-shot CLIP [58] and zero-shot FaRL [60], in which FaRL is pre-trained on the large-scale visual-language face data. The results in Tab. 2 show that fine-tuning the image encoder can improve the performance remarkably. CoOp [53] and Co-CoOp [52] are all using the learnable context, in which the Co-CoOp also add projected image features into the context prompt. The comparison between our method without a temporal model and these two methods in Tab. 2 show the effectiveness of our strategy.

Evaluation of Different Prompts: In a supervised learning setting, V-LP models differ from conventional classification models in that they can be prompted to design classifier-free models for prediction. This makes prompt engineering a critical aspect of adapting V-LP models for downstream tasks. We compare our method with two kinds of manually

Description Position	Class Specific	DFEW		FERV39k		MAFW	
		UAR	WAR	UAR	WAR	UAR	WAR
Middle	✗	58.91	71.19	40.36	51.22	39.55	52.54
Middle	✓	59.24	**71.33**	40.71	51.32	39.26	52.19
End	✗	58.97	71.12	40.50	51.21	39.46	52.45
End	✓	**59.61**	71.25	**41.27**	**51.65**	**39.89**	**52.55**

Table 4: Evaluation of different settings for learnable context. All the methods adopt 8 learnable context tokens, with the temporal model.

Methods	DFEW		FERV39k		MAFW	
	UAR	WAR	UAR	WAR	UAR	WAR
C3D [40]	42.74	53.54	22.68	31.69	31.17	42.25
P3D [57]	43.97	54.47	23.20	33.39	-	-
I3D-RGB [5]	43.40	54.27	30.17	38.78	-	-
3D ResNet18 [12]	46.52	58.27	26.67	37.57	-	-
R(2+1)D18 [40]	42.79	53.22	31.55	41.28	-	-
ResNet18-LSTM [13, 14]	51.32	63.85	30.92	42.95	28.08	39.38
ResNet18-ViT [7, 13]	55.76	67.56	38.35	48.43	35.80	47.72
EC-STFL [16] [MM'20]	45.35	56.51	-	-	-	-
Former-DFER [47] [MM'21]	53.69	65.70	37.20	46.85	31.16	43.27
NR-DFERNet [20] [arXiv'22]	54.21	68.19	33.99	45.97	-	
DPCNet [45] [MM'22]	57.11	66.32	-	-	-	-
T-ESFL [27] [MM'22]	-	-	-	-	33.28	48.18
EST [28] [PR'23]	53.94	65.85	-	-	-	-
LOGO-Former [30] [ICASSP'23]	54.21	66.98	38.22	48.13	-	-
IAL [22] [AAAI'23]	55.71	69.24	35.82	48.54	-	-
CLIPER [21] [arXiv'23]	57.56	70.84	41.23	51.34	-	-
M3DFEL [42] [CVPR'23]	56.10	69.25	35.94	47.67	-	-
AEN [18] [CVPRW'23]	56.66	69.37	38.18	47.88	-	-
DFER-CLIP (Ours)	**59.61**	**71.25**	**41.27**	**51.65**	**39.89**	**52.55**

Table 5: Comparison with the state-of-the-art methods.

designed prompts: "a photo of [class]" and "an expression of [class]". Results in Tab. 3 show that our method outperforms manually designed prompts on both DFEW and FERV39k datasets, but has a slightly lower WAR than the prompt of "an expression of [class]" on the MAFW dataset. We note that the MAFW dataset comprises eleven facial expression classes and that the video samples are imbalanced, with the proportions of the "contempt", "help-lessness", and "disappointment" expressions accounting for only 2.57%, 2.86% and 1.98%, respectively. The results in Tab. 3 demonstrate that the learning-based context consistently achieves the best WAR results. Furthermore, our method outperforms using the prompt of the class name with the learnable context approach, which indicates the effectiveness of using descriptions.

In our DFER-CLIP, the description is at the end of the prompt and the learnable context is class-specific, namely, each description will have its own learned context. We also conduct

experiments with different settings: a) put the description in the middle of the learnable context, and b) all the descriptions share the same learned context. The results on three datasets are shown in Tab. 4. The results indicate that positioning the description tokens to the end gains more improvement. Moreover, for the end position, adopting the class-specific is always better than a shared context prompt, achieving the best results.

4.4 Comparison with State-of-the-Art Methods

In this section, we compare our results with several state-of-the-art methods on the DFEW, FERV39k, and MAFW benchmarks. Consistent with previous methods [16, 21, 22, 27, 30, 47], the experiments on DFEW and MAFW are conducted under 5-fold cross-validation and use training and test set on FERV39k. Furthermore, we train models three times with different random seeds and then use the average for more stable and reliable results. The comparative performance in Tab. 5 demonstrates that the proposed DFER-CLIP outperforms the compared methods both in UAR and WAR. Specifically, compared with the previous best results, our method shows a UAR improvement of 2.05%, 0.04%, and 4.09% and a WAR improvement of 0.41%, 0.31%, and 4.37% on DFEW, FERV39k, and MAFW, respectively. It should be pointed out that FERV39k is the current largest DFER benchmark with 38,935 videos. Given this substantial scale, making significant enhancements becomes a formidable task.

5 Conclusion

This paper presents a novel visual-language model called DFER-CLIP for in-the-wild dynamic facial expression recognition. In the visual part, based on the CLIP image encoder, a temporal model consisting of several Transformer Encoders was introduced for modelling the temporal facial expression features. In the textual part, the expression descriptors related to facial behaviour were adopted as the textual input to capture the relationship between facial expressions and their underlying facial behaviour, in which the expression descriptors were generated by large language models like ChatGPT. The learnable contexts for these descriptors were also designed to help the model learn relevant context information for each expression during training. Extensive experiments demonstrate the effectiveness of each component in DFER-CLIP. Moreover, the proposed method achieves state-of-the-art results on three benchmarks.

Acknowledgements

References

[1] Wissam J Baddar and Yong Man Ro. Mode variational lstm robust to unseen modes of variation: Application to facial expression recognition. In *AAAI*, pages 3215–3223, 2019.

[2] Mina Bishay, Petar Palasek, Stefan Priebe, and Ioannis Patras. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE Transactions on Affective Computing*, 12(4):949–961, 2019.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[4] Junkai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Facial expression recognition in video with multiple feature fusion. *IEEE Transactions on Affective Computing*, 9(1): 38–50, 2016.

[5] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, 2016.

[6] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, 1998.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[8] Zoran Duric, Wayne D Gray, Ric Heishman, Fayin Li, Azriel Rosenfeld, Michael J Schoelles, Christian Schunn, and Harry Wechsler. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceeding of the IEEE*, 90(7):1272–1289, 2002.

[9] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[10] Niki Maria Foteinopoulou and Ioannis Patras. Learning from label relationships in human affect. In *MM*, pages 80–89, 2022.

[11] Niki Maria Foteinopoulou, Christos Tzelepis, and Ioannis Patras. Estimating continuous affect with label uncertainty. In *ACII*, pages 1–8, 2021.

[12] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021.

[16] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *MM*, pages 2881–2889, 2020.

[17] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124, 2022.

[18] Bokyeung Lee, Hyunuk Shin, Bonhwa Ku, and Hanseok Ko. Frame level emotion guided dynamic facial expression recognition with emotion grouping. In *CVPRW*, pages 5680–5690, 2023.

[19] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10143–10152, 2019.

[20] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, and Feng Zhao. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022.

[21] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. *arXiv preprint arXiv:2303.00193*, 2023.

[22] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *AAAI*, pages 67–75, 2023.

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[24] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2020.

[25] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023.

[26] Daizong Liu, Hongting Zhang, and Pan Zhou. Video-based facial expression recognition using graph convolutional networks. In *ICPR*, pages 607–614, 2021.

[27] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *MM*, 2022.

[28] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023.

[29] Chaofan Ma, Yuhuan Yang, Yanfeng Wang, Ya Zhang, and Weidi Xie. Open-vocabulary semantic segmentation with frozen vision-language models. In *BMVC*, pages 1–15, 2022.

[30] Fuyan Ma, Bin Sun, and Shutao Li. Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In *ICASSP*, pages 1–5, 2023.

[31] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.

[32] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI*, pages 1359–1367, 2020.

[33] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *ECCV*, pages 681–697, 2022.

[34] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, pages 1–18, 2022.

[35] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis A Nicolaou, and Ioannis Patras. Parts of speech-grounded subspaces in vision-language models. *arXiv preprint arXiv:2305.14053*, 2023.

[36] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *CVPR*, pages 2085–2094, 2021.

[37] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5533–5541, 2017.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[39] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, pages 6545–6554, 2023.

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *CVPR*, pages 4489–4497, 2015.

[41] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.

[42] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *CVPR*, pages 17958–17968, 2023.

[43] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 2022.

[44] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos. In *CVPR*, pages 20922–20931, 2022.

[45] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *MM*, pages 101–110, 2022.

[46] Torsten Wilhelm. Towards facial expression analysis in a driver assistance system. In *FG*, pages 1–4, 2019.

[47] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *MM*, pages 1553–1561, 2021.

[48] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.

[49] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *AAAI*, pages 3510–3519, 2021.

[50] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, pages 18697–18709, 2022.

[51] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In *CVPR*, pages 2562–2569, 2012.

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.