2024 IEEE International Joint Conference on Biometrics (IJCB) | 979-8-3503-6413-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCB62174.2024.10744485

CLIP-Guided Bidirectional Prompt and Semantic Supervision for Dynamic Facial Expression Recognition

Junliang Zhang[1], Xu Liu[1], Yu Liang[1], Xiaole Xian[1], Weicheng Xie[1]*, Linlin Shen[1], Siyang Song[2]
[1]School of Computer Science & Software Engineering, Shenzhen University, China
[2]School of Computing and Mathematical Sciences, University of Leicester, U.K.
{zhangjunliang2022@email., 2310275003@email., realmelo621@email., 2310275030@email., wcxie@, llshen@}szu.edu.cn, ss1535@leicester.ac.uk

Abstract

Due to the insufficient semantic information supervision in existing works for dynamic facial expression recognition (DFER), videos with similar facial changes but different expressions may be easily confused. Thanks to the potential textual information for semantic supervision, contrastive language-image pretraining (CLIP) model provides a new direction for DFER. However, pre-trained CLIP based on image-text pairs has difficulty in capturing temporal features in the video domain. Therefore, we propose a novel visual language model that captures and aggregates dynamic features of expressions in semantic supervision via Inter-Frame Interaction Transformer (InterFIT) and Multi-Scale Temporal Aggregation (MSTA). Furthermore, though prompt learning is often used in CLIP to enhance semantic supervision, previous studies have only focused on the role of textual prompts, ignoring the importance of visual prompts in facilitating the relationality between the two. Therefore, we designed a Bidirectional Enhanced Prompt (BiEhPro) to facilitate the learning of this relationality between text and visual cues in enhancing semantic supervision. Extensive experiments and ablation studies on three benchmark datasets, i.e., DFEW, FERV39K, and MAFW, validate the effectiveness of our modules and algorithm. Code is publicly available at https://github.com/JunLiangZ/CLIP-Guided-DFER.

1. Introduction

Automatic facial expression recognition (FER) has become a hotspot for researchers due to its applications in various fields. Since video can provide richer spatio-temporal facial pattern reflecting facial expressions, increased attentions have been drawn to dynamic facial expression recognition (DFER).

*Corresponding author

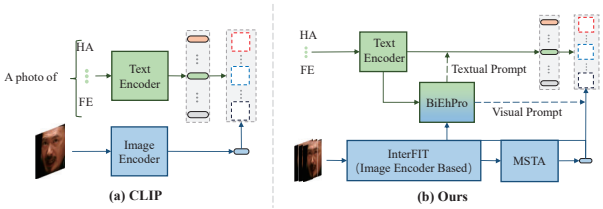

Figure 1. Comparison of traditional CLIP-guidance methods and ours for DFER. (a) Standard CLIP [35] designs a textual descriptor for each class and uses the cosine similarity between image and text embeddings for inference. (b) Our approach provides enhanced semantic supervision of CLIP for DFER, via the introduced inter-frame interaction, aggregation and bidirectional prompt learning. InterFIT, MSTA and BiEhPro denote Inter-Frame Interaction Transformer, Multi-Scale Temporal Aggregation and Bidirectional Enhanced Prompt.

With the popularity of deep learning and the availability of large-scale datasets (e.g., DFEW [17], FERV39K [45]), researchers have developed a variety of deep neural networks (DNNs) to address the challenges in DFER, e.g. 2D/3D convolutional neural networks (CNNs) [8, 19], recurrent neural networks (RNNs) [7, 38], and more advanced transformer-based architectures [22, 28, 44, 49]. However, these methods train the model based on manually-labeled tags, which may lack sufficient semantic supervision, thus limiting the model's capacity in understanding and differentiating facial expressions.

Recently, visual-language pre-training models such as CLIP [35] (shown in Fig. 1(a)) provide an effective solution, which use two independent encoders to achieve image-text alignment in millions of image-text pairs, and offer more semantic cues for supervising the visual representations. Though visual-language pre-training models are helpful in various visual tasks, such as image classification [51], detection [10, 2], and synthesis [33], it still faces certain challenges in DFER. **Challenge 1**: How to effec-

tively capture the temporal information of dynamic changes in facial expressions in terms of CLIP guidance. Since the duration of each expression is different, there will also be the problem of inconsistency between short-term and long-term temporal information [43]; and **Challenge 2**: Since the expression label names are abstract, using only the textual features of labels to provide semantic information may not correlate well with the complex visual features of expressions in the video. This motivates us to introduce prompts to enhance visual and textual correlation.

To address these challenges, Li et al. [21] designed a unified framework based on CLIP for both static and dynamic FER, i.e., CLIPER. However, the capacity of the model is limited in capturing temporal information. Zhao et al. [50] designed DFER-CLIP, incorporating a temporal encoder and a prompt descriptor generated by a large language model. However, this model only generates textual prompts but ignores visual prompts, which may result in a model that does not align visual and textual features well. As shown in Fig. 2, under only textual prompts (Only TP), the expression visual features result in lower similarity with the corresponding textual features, producing wrong classification (i.e., green bars). In addition, it also overlooks the problem of inconsistent information in DFER for short-term and long-term temporal sequences, which may result in the failure to capture the details of expressions appearing in different temporal sequences.

To this end, we propose a new method for modeling temporal cues of videos based on CLIP for addressing **Challenge 1**. As shown in Fig. 1(b), the InterFIT utilizes pre-training CLIP image encoders to generate frame-level representations, to enable information exchange between frames through interaction token mechanism. In this way, each interaction token not only describes the semantics of the current frame but also communicates with other frames to model their spatio-temporal dependencies. Then, MSTA aggregates frame-level features from different temporal scales to better capture short-term and long-term dynamic expression changes. Meanwhile, in order to enhance the correlation between text and visual information, we have developed a learnable BiEhPro module for generating both text and visual prompts (i.e., addressing **Challenge 2**). By generating prompts for both visual and text cues, BiEhPro facilitates a tight correlation between both (as shown by the red bar in Fig. 2), providing valuable semantic supervision for learning expression information. In summary, the novelties of our algorithm compared with related CLIP-based ones are shown in the supplementary material, and our contributions are outlined as follows:

- Based on the CLIP, we propose a new visual modeling method that introduces an interaction token mechanism to capture dynamic features while aggregating features from multiple temporal scales to obtain dynamic visual expression features.
- We design a Bidirectional Enhanced Prompt (BiEhPro) to generate textual prompt and visual prompt based on the mutual interaction of visual and textual features, thus enhancing the correlation of text and vision in semantic supervision.
- Extensive experimental results on three datasets (i.e., DFEW, FERV39K and MAFW) show that our method outperforms existing state-of-the-art (SOTA) methods in terms of both WAR and UAR.

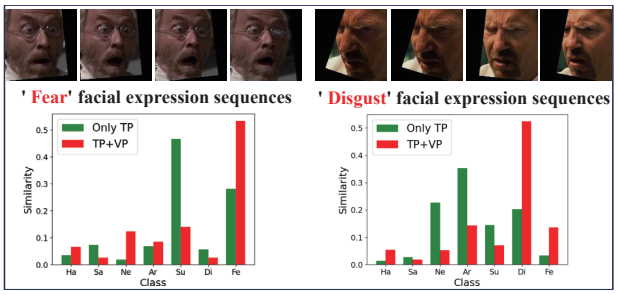

Figure 2. Visualization of the impact of different prompts. Different video expression sequences and corresponding similarity histograms in only text prompt (Only TP) and both textual and visual prompts (TP+VP). Compared to only textual prompts, combining textual and visual prompts makes the visual features closer to the corresponding textual features, resulting in maximum similarity. Ha: happy. Sa: sad. Ne: neutral. Ar: angry. Su:surprise. Di: disgust. Fe: fear.

2. Related Work

2.1. Dynamic Facial Expression Recognition

The DFER task is challenging because it requires dealing with the abstract nature of expressions and the dynamic features of videos. Although manual feature methods [4] and STLMBP [15] have shown reasonable effectiveness, they have limited applicability in specific scenarios. Thanks to the development of deep learning, there has been a trend towards extracting spatio-temporal features from sequences using 3D CNN [40, 41, 12], cascaded CNN-LSTM [8, 7, 38], and CNN-Transformer [49, 25] structures. Yu et al. [47] extracted local-global and spatio-temporal information based on a cascaded CNN-LSTM structure. Wang et al. [46] proposed a dual path multi-excitation collaborative network (DPCNet) to extract essential information about expressions from a limited number of key frames in a video. Li et al. [22] designed a plug-and-play module of global convolution-attention blocks and an intensity-aware loss for in-the-wild DFER. However, the aforementioned methods solely depend on visual information and utilize manually-labeled tags as supervision to train models, and the seman-

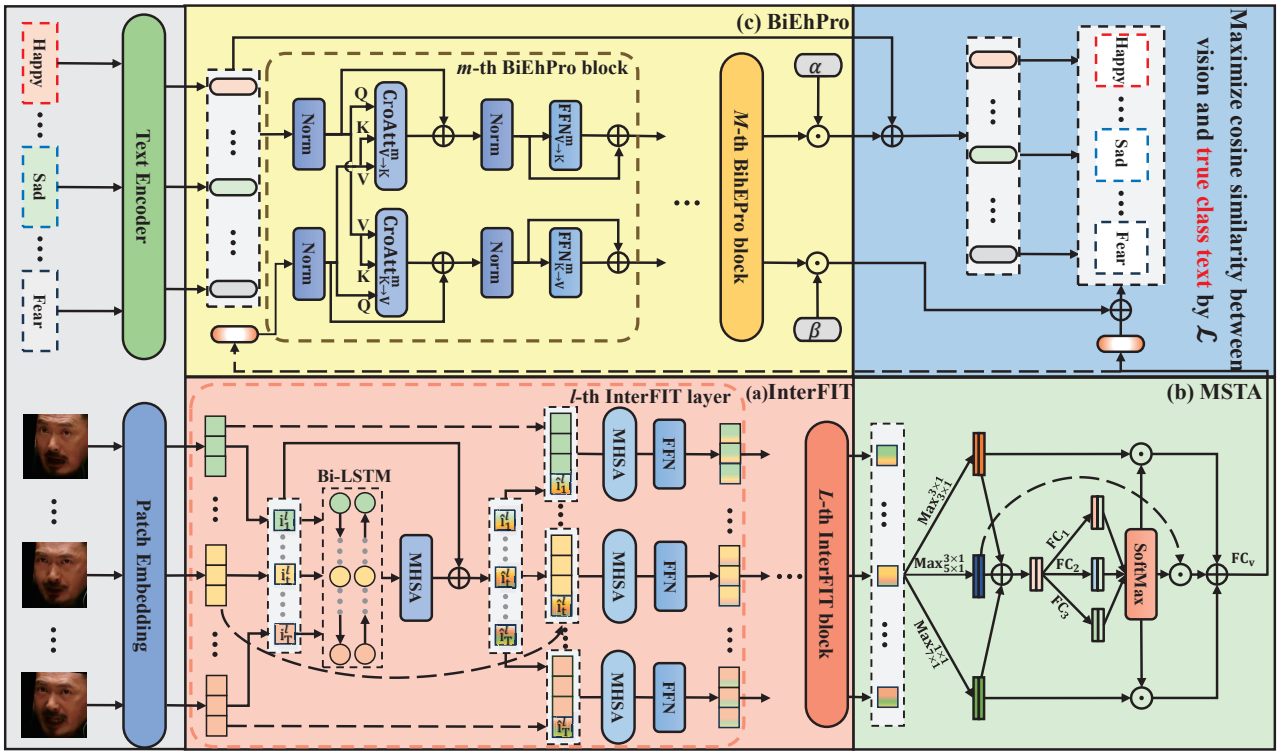

Figure 3. Overview of our method. (a) Inter-Frame Interaction Transformer (InterFIT) implements spatio-temporal modeling. (b) Multi-Scale Temporal Aggregation (MSTA) aggregates frame-level features from multiple temporal scales. (c) Bidirectional Enhanced Prompt (BiEhPro) enhances the correlation between textual and visual features. MHSA denotes Multi-Head Self-Attention. FFN denotes Feed-Forward Network. $\mathrm{Max}_{5\times1}^{3\times1}$ denotes the max pooling with kernel size of 5×1 and step size of 3×1. CroAtt denotes Cross-Attention.

tic cues were not fully explored in distinguishing between different expressions with similar facial movements.

Therefore, we introduced a visual-language model (e.g., CLIP [35]) to enhance the semantic information via textual supervision, to allow models to consider both visual features and semantic information during the learning process.

2.2. Visual-Language Models

Visual-language pretraining has received increasing attention in recent years, especially visual-language pretraining methods based on large-scale data contrastive learning [31, 16, 35]. These methods have shown appealing performance on various downstream tasks, including object detection [2], image segmentation [5, 23], and visual retrieval [36]. However, the absence of video-specific temporal cues in image-level pre-training makes it challenging to adapt visual-language pre-training models to videos. Therefore, recent studies [18, 32] apply CLIP to videos and show its prospect on DFER. Li et al. [21] proposed an unified framework for dynamic FER called CLIPER, while its capacity in capturing dynamic temporal information is limited. Zhao et al. [50] proposed DFER-CLIP to explore expression descriptors and transformer-based temporal modeling for FER. However, this model does not consider the imbalance between short-term and long-term temporal relationships in dynamic expressions, and may ignore the expression details that occur within a short-term temporal sequence [43].

To this end, we propose to enhance the pre-training CLIP for dynamic expressions, so as to allow models to encode frame-level temporal features and capture both long-term and short-term temporal features.

2.3. Prompt Learning

With the progress of large language models, prompt learning has become a new approach [11, 14, 27]. As opposed to traditional fixed prompts (e.g., "A photo of a $\{class\}$" [35]), prompt learning treats prompts as learnable parameters, aiming to find the most appropriate embedding [32]. This approach allows models to learn and adapt the prompt representation, and has been widely applied to various visual and visual-language models [18]. Zhou et al. [52] proposed context optimization (CoOp), which utilizes learnable vectors to model the context of prompts. They [51] further proposed conditional context optimization based on CoOp, which can be generalized to unseen classes and reduce sensitivity to category transformation. However, these methods only focus on prompt learning of

enhanced textual features, ignoring that for visual features.

Therefore, we propose to leverage text features to mutually reinforce visual features, so as to learn the correlation between text and vision via generating both textual and visual prompts. Based on these prompts, the corresponding textual and visual features can be both enhanced.

3. Proposed Method

Overview. As shown in Fig. 3, our approach aligns video and corresponding text representations by jointly training video and text encoding components. In the visual encoding part, we specially designed the Inter-Frame Interaction Transformer (InterFIT) and the Multi-Scale Temporal Aggregation (MSTA) to facilitate inter-frame information transfer, capture the dynamic features of expressions, and consequently extract their visual features. In the text encoding part, we utilize the pre-training text encoder in CLIP and propose a learnable Bidirectional Enhanced Prompt (BiEhPro) to extend it. This key idea is to utilize the mutual interaction between text and vision to generate corresponding textual and visual prompts, thus enhancing the model's understanding and recognition of expressions. Finally, based on the similarity between the visual and text features in the feature space, the classification category is determined based on the highest similarity.

Preliminary. Formally, given a video clip $X \in \mathbb{R}^{T\times C\times H\times W}$ consisting of T RGB images, where H and W denote the height and width of images, respectively, and C denotes the dimension. For each frame x_t, $t \in \{1,\cdots,T\}$ in the video clip X, we partition it into N non-overlapping patches of size $P\times P$, and then use a linear mapping to obtain an embedding representation for each patch $\left\{\mathbf{e}_{t,n}^{(0)}\right\}_{n=1}^{N} \in \mathbb{R}^D$, where $N = H\ W/P^2$, D denotes the dimension of each patch embedding. After that, we add a learnable embedding $\mathbf{e}_{t,0}^{(0)}$ to the sequence of embedded patches, which is referred to the class token. This token is represented as a frame-level representation in the final output of the encoder, whose input at frame t is denoted as:

$$\mathbf{z}_t^{(0)} = \left[\mathbf{e}_{t,0}^{(0)}, \mathbf{e}_{t,1}^{(0)}, \cdots, \mathbf{e}_{t,N}^{(0)}\right] + \mathbf{e}^s \tag{1}$$

where $\mathbf{e}^s$ represents the spatial position encoding [32].

3.1. Inter-Frame Interaction Transformer (InterFIT)

In order to capture dynamic features in a sequence of frames during the extraction of frame-level representations by the CLIP visual encoder, we introduced an interaction token mechanism in the CLIP visual encoder, which facilitates the interactive transmission of inter-frame temporal information. This InterFIT is shown in Fig. 3(a).

Specifically, the interaction token $\mathbf{i}_t^{(l)} \in \mathbb{R}^D$ for the t-th frame in the l-th layer is obtained by linearly transforming the class token $\mathbf{e}_{t,0}^{(l-1)}$ to make use of the global information in this frame. Then, InterFIT inputs all interaction tokens into a Bidirectional LSTM (Bi-LSTM) [37] and Multi-Head Self-Attention (MHSA), where MHSA uses a self-attention at the frame level to learn the temporal dependencies in the expression sequence, and its representation at the l-th layer is formulated as:

$$\widehat{\mathbf{I}}^{(l)} = \mathbf{I}^{(l)} + \mathrm{MHSA}\left(\mathrm{LN}\left(\text{Bi-LSTM}\left(\mathbf{I}^{(l)}\right)\right)\right) \tag{2}$$

where $\widehat{\mathbf{I}}^{(l)} = \left[\widehat{\mathbf{i}}_1^{(l)}, \cdots, \widehat{\mathbf{i}}_T^{(l)}\right]$, LN denotes the layer normalization [1], $l \in \{1,\cdots,L\}$ denotes the index of the InterFIT layer and L denotes the number of InterFIT layers.

Next, the interaction completion token $\widehat{\mathbf{i}}_t^{(l)}$ is connected to the corresponding frame token $\mathbf{z}_t^{(l-1)}$, which conveys global temporal information to each frame via the MHSA. This process aids in learning temporal and spatial interactions to capture global spatio-temporal dependencies, which can be represented as follows:

$$\left[\widehat{\mathbf{z}}_t^{(l)}, \bar{\mathbf{i}}_t^{(l)}\right] = \left[\mathbf{z}_t^{(l-1)}, \widehat{\mathbf{i}}_t^{(l)}\right] + \mathrm{MHSA}\left(\mathrm{LN}\left(\left[\mathbf{z}_t^{(l-1)}, \widehat{\mathbf{i}}_t^{(l)}\right]\right)\right) \tag{3}$$

where $[\cdot,\cdot]$ denotes the concatenation of frame tokens and interaction token features.

InterFIT implements the transfer of inter-frame information through L InterFIT layers to encode the global spatio-temporal information in a sequence. Finally, we utilize the class tokens, which contain global spatio-temporal information, as frame-level representations to acquire each frame-level feature of the video, i.e., $Z = [e_{1,0}^{(L)}, \cdots, e_{T,0}^{(L)}] \in \mathbb{R}^{T\times D}$.

3.2. Multi-Scale Temporal Aggregation (MSTA)

To aggregate frame-level feature of the video to obtain its final visual characterization, the duration of expression varies from video to video, thus simply aggregating the total duration would overlook some instantaneous subtle changes in expression. To this end, we introduce the MSTA as shown in Fig. 3(b), which aggregates frame-level features from different temporal scales to obtain richer visual representations.

Given all frame-level representations $Z \in \mathbb{R}^{T\times D}$ of a video clip, we first fuse the features of different temporal scales as:

$$\widehat{Z} = \textstyle\sum_{i=1}^{3} Z_i = \sum_{i=1}^{3} \mathrm{Max}_{ks_i}^{ss_i}\left(Z^{Tr}\right), \tag{4}$$

where $Z^{Tr} \in \mathbb{R}^{D\times T}$ is the transpose of Z, $\mathrm{Max}_{ks_i}^{ss_i}$ denotes a max pooling operation with kernel sizes of ks_i and the step size of ss_i. $ks_i \in \{5\times 1, 5\times 1, 7\times 1\}$; $ss_i \in \{3\times 1, 3\times 1, 1\times 1\}$. $\widehat{Z}$, Z_i have the same size of $(D, T/3)$. The

output of Eq. (4), $\widehat{Z}$, is an element-wise summation of the aggregation results for multiple temporal scales Z_i.

Then, the MSTA model produces frame-level weights as:

$$U_i = \mathrm{FC}_i\left(\widehat{Z}\right) \quad i \in \{1,2,3\} \tag{5}$$

where $\mathrm{FC}_i(\cdot)$ stands for a fully connected layers of Z_i generating the frame selection weighting tensor $U_i \in \mathbb{R}^{D\times T/3}$, respectively. The weights are further normalized as:

$$W_{s,c,t} = e^{U_{s,c,t}} / \textstyle\sum_i e^{U_{i,c,t}}, \quad s \in \{1,2,3\} \tag{6}$$

where $U_{s,c,t} \in \mathbb{R}^{1\times 1}$ denotes the value of the c-th channel of the t-th frame of U_s, $W_{s,c,t} \in \mathbb{R}^{1\times 1}$ is the weight of the c-th channel of the t-th frame of Z_s. Weighted features for different temporal scales are then obtained as:

$$Z_w = \textstyle\sum_{i=1}^{3} W_i \odot Z_i \tag{7}$$

where $W_i \in \mathbb{R}^{D\times \frac{T}{3}}$ are weight tensors computed according to Eq. (6) and $\odot$ denotes the element-wise multiplication.

Finally, we flatten $Z_w \in \mathbb{R}^{D\times \frac{T}{3}}$ and map it through a fully connected layer $\mathrm{FC_V}$ to obtain video-level features of dynamic expressions:

$$V = \mathrm{FC_V}\left(flatten\left(Z_w\right)\right) \tag{8}$$

where $V \in \mathbb{R}^D$, $\mathrm{FC_V}(\cdot)$ denotes a fully connected layer.

3.3. Bidirectional Enhanced Prompt (BiEhPro)

Previous works on DFER have focused on acquiring discriminative feature embedding supervised by manually-labeled tags [46, 28], which often neglect the underlying deeper semantics conveyed by the expressions. Text can provide rich semantic information, which is helpful to a model for learning the association between expressions and semantic meaning.

Thus, we use a pre-training text encoder from CLIP to obtain the text features:

$$K_j = TxtEnc(cls_j) \quad j \in \{0,1,\cdots,\#class-1\} \tag{9}$$

where $K_j \in \mathbb{R}^D$ denotes the textual features of the j-th class, $\#class$ denotes the number of classes, cls_j denotes the j-th expression class name, and $TxtEnc(\cdot)$ denotes the text encoder.

Based on this, as shown in Fig. 3(c), the learnable Bidirectional Enhanced Prompt (BiEhPro) module is proposed, leveraging textual features to mutually reinforce visual features so as to learn the correlation between text and vision. This process generates textual and visual prompts to enhance the understanding and recognition of expressions.

Specifically, BiEhPro contains M blocks. Each block consists of two Cross-Attention (CroAtt) and Feed-Forward Networks (FFN) for learning textual and visual prompts. The block at the m-th layer can be represented as:

$$\begin{aligned}
&\widetilde{K}_j^{(0)} = K_j \\
&\overline{K}_j^{(m)} = \widetilde{K}_j^{(m-1)} + \mathrm{CroAtt}_{V\to K_j}^{(m)}(\widetilde{K}_j^{(m-1)}, \widetilde{V}^{(m-1)}) \\
&\widetilde{K}_j^{(m)} = \overline{K}_j^{(m)} + \mathrm{FFN}_{V\to K_j}^{(m)}(\overline{K}_j^{(m)}) \\
&\widetilde{V}^{(0)} = V \\
&\overline{V}^{(m)} = \widetilde{V}^{(m-1)} + \mathrm{CroAtt}_{K_j\to V}^{(m)}(\widetilde{V}^{(m-1)}, \widetilde{K}_j^{(m-1)}) \\
&\widetilde{V}^{(m)} = \overline{V}^{(m)} + \mathrm{FFN}_{K_j\to V}^{(m)}(\overline{V}^{(m)})
\end{aligned} \tag{10}$$

where $m \in \{1,\cdots,M\}$, the textual representation K_j and visual representation V are obtained in Eq. (9) and Eq. (8), respectively, $\widetilde{K}_j^{(m)}$ is the text prompt generated at the m-th layer, and we obtain the text prompt at the m-th layer by using the text prompt $\widetilde{K}_j^{(m-1)}$ from the $(m-1)$-th layer as a query and the video prompt $\widetilde{V}^{(m-1)}$ as both the key and value, which allows the textual representation to extract relevant visual information from the video. Similarly, we obtain the visual prompt $\widetilde{V}^{(m)}$.

We then use the textual prompt $\widetilde{K}_j^{(M)}$ and visual prompt $\widetilde{V}^{(M)}$ generated after M BiEhPro blocks to extend the textual representation K_j and visual representation V as:

$$\begin{aligned}
\widehat{K}_j &= K_j + \alpha\widetilde{K}_j^{(M)} \\
\widehat{V} &= V + \beta\widetilde{V}^{(M)}
\end{aligned} \tag{11}$$

where α, β are also learnable parameters. Finally, we introduce a contrastive loss to optimize our goal (i.e., maximize $sim(\widehat{V},\widehat{K}_j)$ if $\widehat{V}$ and $\widehat{K}_j$ match, otherwise minimize it), based on the cosine similarity $sim(\widehat{V},\widehat{K}_j)$ between the visual and textual representations:

$$\mathcal{L} = -\textstyle\sum_{j=0}^{\#class-1} 1_{y=j} \log \frac{\exp\left(sim(\widehat{V},\widehat{K}_j)/\tau\right)}{\sum_{k=0}^{\#class-1}\exp\left(sim(\widehat{V},\widehat{K}_k)/\tau\right)} \tag{12}$$

where y denotes the true label, τ is the temperature parameter. For clarity, the pseudo-code and the time complexity of our algorithm are presented in the supplementary material.

4. Experiments

4.1. Datasets and Implementation Details

Datasets: We evaluated the performance of our method on three DFER datasets (i.e., DFEW [17], FERV39k [45], and MAFW [26]). We use DFEW and FERV39k for 7-class DFER tasks and MAFW for 11-class tasks.

Data Pre-processing: For the visual component of the input, we select $T = 8$ frames from a video clip. Due to the limited number of videos used to train the network, we employ data augmentation techniques, such as random cropping, color dithering, and image flipping in the training. We preprocess all the raw images based on facial landmarks and

resize them to 224x224. For the text part, we directly input the labeled text to learn the specific prompts.

Training Setting: The entire framework was implemented using PyTorch GPUs and all our models are trained on a single Tesla P100 GPU. We utilized CLIP's VIT-B/32 ($L = 12$, $P = 32$) as the backbone, keeping the text encoder fixed, and fine-tuning it by incorporating CLIP's visual encoder into the InterFIT. For all datasets, the model was trained for 100 epochs using the AdamW optimizer and cosine scheduler, with a 5-epoch warm-up. The learning rate, the minimum learning rate, the weight decay and the batch size were set to 2e-5, 2e-7, 5e-4 and 32, respectively. The number of bidirectionally enhanced prompt blocks, i.e., M, was set to 2.

In all experiments, we use weighted average recall (WAR) and unweighted average recall (UAR) as evaluation metrics. In the following analysis, our method is primarily evaluated on the DFEW dataset, and the performance is compared with fully supervised SOTA method.

4.2. Comparison with SOTA Methods

Results on large in-the-wild datasets. We compare our method with previous SOTA supervised methods on DFEW [17], FERV39k [45], and MAFW [26] in Tab. 1.

Tab. 1 shows that our method achieves the SOTA results in terms of UAR and WAR metrics on all the three datasets. Specifically, on the DFEW dataset, our method surpasses the baseline by 2.57% in terms of UAR and 2.62% in terms of WAR. Meanwhile, our method outperforms the method utilizing CLIP [35], i.e., DFER-CLIP [50] by a margin of 1.24% in terms of UAR and 1.33% in terms of WAR. These large improvements suggest that our method can effectively incorporate image and text-based pre-training of CLIP into DFER to enhance semantic supervision. For FERV39k, our method achieves SOTA performance, i.e. 41.43% (UAR) and 51.83% (WAR), outperforming the previous best method, DFER-CLIP, by a margin of 0.16% in terms of UAR and 0.18% in terms of WAR. For MAFW, our method outperforms current best-performance methods (i.e., DFER-CLIP [50] and T-MEP [48]) by 1.17% in terms of UAR and 1.53% in terms of WAR. These appealing performances of our algorithm reveal its powerful generalization capacity, even across diverse scenarios.

4.3. Ablation Studies

The baseline model uses the pre-trained CLIP as the backbone. Based on this, we conduct ablation studies to investigate the impact of the components in our approach, including Inter-Frame Interactive Transformer (InterFIT), Multi-Scale Temporal Aggregation (MSTA), and Bidirectional Enhanced Prompt (BiEhPro).

Evaluation of each component: Tab. 2 shows the performance evolution of our method for extending CLIP to the DFER task. We can observe that our proposed InterFIT can improve UAR and WAR by 0.77% and 1.58%, respectively, by modeling inter-frame spatio-temporal information. Then, attaching the MSTA for aggregating frame-level features of different temporal scales can further improve the UAR by 0.72% and WAR by 0.34%. This indicates that our module can effectively utilize the dynamic temporal information in a video while enabling the visual features to encode both long-term and short-term information. Finally, through the proposed BiEhPro, our approach can surpass the baseline UAR by 2.57% and WAR by 2.62%. For the reasons, the specific textual and visual prompts are generated to enhance the correlation between the two, and obtain better semantic supervision.

Impact of the text: To evaluate the impact of the text cues, we replace the text encoder with a randomly initialized fully connected layer as the classification head. Tab. 3 shows that the performance of the model degrades without the textual branch. The text information can improve the UAR by 0.44% and the WAR by 0.53%, respectively. This suggests that the semantic information included in the textual representation is helpful to complement expressions.

Evaluation of block number M in BiEhPro: In this section, we investigate the performance sensitivity against M and present the results in Tab. 4.

Tab. 4 shows that smaller shallow models are unable to effectively capture the correlation between text and visual priors due to their limited number of parameters. In addition, networks with larger M may lead to overfitting due to limited training data. Therefore, increasing the number of BiEhPro blocks may not always helpful to performance improvement.

Comparison with related prompt methods: We compare several existing prompting methods in Tab. 5, including CLIP (zero-shot) [35] and CLIP-based prompt learning methods such as CoOp [52], CoCoOp [51], and DFER-CLIP [50].

As shown in Tab. 5, our prompt method outperforms other related ones. In particular, our method outperforms DFER-CLIP, which is also applied to DFER, by 1.51% and 2.12% on UAR and WAR, respectively, with only textual prompts (i.e.,w/o InterFIT, MSTA, VP and w/o TM). In addition, the UAR and WAR are further improved by 0.41% and 0.52% under the concurrent use of textual and visual prompts (i.e.,w/o InterFIT, MSTA). With the entire method (i.e.,$w/$ InterFIT, MSTA and $w/$ TM), our modules achieve an improvement of 1.24% and 1.33% in terms of UAR and WAR, respectively. In addition, our approach enables the end-to-end online learning without the need of an additional large language models as in DFER-CLIP. Besides, our prompting approach not only considers textual prompts but also generates visual prompts, which enhances the correlation between textual and visual cues in semantic supervision.

Table 1. Comparison (%) with SOTA methods on DFEW. **Bold** denotes the best. Underline denotes the second best. The red text indicates the improvement of our method over the baseline.

Methods	Backbone	DFEW		FERV39k		MAFW	
		UAR	WAR	UAR	WAR	UAR	WAR
C3D$_{CVPR'15}$ [40]	C3D	42.74	53.54	22.68	31.69	31.17	42.25
P3D$_{ICCV'17}$ [34]	P3D	43.97	54.47	23.20	33.39	-	-
I3D-RGB$_{CVPR'17}$ [3]	Inflated 3D ConvNets	43.40	54.27	30.17	38.78	-	-
3D ResNet18$_{CVPR'18}$ [12]	ResNet18	46.52	58.27	26.67	37.57	-	-
R(2+1)D18$_{CVPR'18}$ [41]	R(2+1)D	42.79	53.22	31.55	41.28	-	-
ResNet18-LSTM [13, 9]	ResNet18-LSTM	51.32	63.85	30.92	42.95	28.08	39.38
ResNet18-ViT [6, 13]	ResNet18-ViT	55.76	66.56	38.35	48.43	35.80	47.72
EC-STFL$_{MM'20}$ [17]	C3D / P3D / et al.	45.35	56.51	-	-	-	-
Former-DFER$_{MM'21}$ [49]	Transformer	53.69	65.7	37.20	46.85	31.16	43.27
DPCNet$_{MM'22}$ [46]	ResNet50	57.11	66.32	-	-	-	-
T-ESFL$_{MM'22}$ [26]	ResNet-Transformer	57.11	66.32	-	-	33.28	48.18
EST$_{PR'22}$ [28]	ResNet18	53.94	65.85	-	-	-	-
Freq-HD$_{MM'23}$ [39]	VGG13-LSTM / et al.	46.85	55.68	33.07	45.26	-	-
LOGO-Former$_{ICASSP'23}$ [30]	ResNet18	54.21	66.98	38.22	48.13	-	-
IAL$_{AAAI'23}$ [22]	ResNet18	55.71	69.24	35.82	48.54	-	-
AEN$_{CVPRW'23}$ [20]	ResNet18	56.66	69.37	38.18	47.88	-	-
M3DFEL$_{CVPR'23}$ [43]	ResNet18-3D	56.10	69.25	35.94	47.67	-	-
MSCM$_{PR'23}$ [24]	ResNet18+TSM	58.49	70.16	-	-	-	-
T-MEP$_{TCSVT'23}$ [48]	CNN-Transformer	57.16	68.85	-	-	39.37	52.85
CLIPER$_{arXiv'23}$ [21]	CLIP	57.56	70.84	41.23	51.34	-	-
DFER-CLIP$_{BMCV'23}$ [50]	CLIP	59.61	71.25	41.27	51.65	39.89	52.55
LSTPNet$_{IVC'24}$ [29]	ResNet18-Transformer	60.18	71.16	40.63	50.07	-	-
Baseline	CLIP	58.28	69.96	39.52	49.80	39.13	52.31
Ours	CLIP	**60.85**(+2.57)	**72.58**(+2.62)	**41.43**(+1.91)	**51.83**(+2.03)	**41.06**(+1.93)	**54.38**(+2.07)

Figure 4. Visualization of text impact on DFEW. The first row is a heatmap generated without text imported (i.e. w/o text) and the prediction is labeled in blue. The second row is the heatmap generated with text imported (i.e., w/ text) and the prediction is labeled in red.

Table 2. Ablation study (%) of our modules on DFEW.

InterFIT	MSTA	BiEhPro	FLOPs(G)	UAR	WAR
×	×	×	23.58	58.28	69.96
✓	×	×	24.65	59.05	71.54
✓	✓	×	24.65	59.77	71.88
✓	✓	✓	24.71	**60.85**	**72.58**

Table 3. Ablation study (%) of text information on DFEW.

Method	FLOPs(G)	UAR	WAR
w/o text	24.65	60.41	72.05
$w/$ text	24.71	**60.85**	**72.58**

4.4. Visualization

Visualization of text impact. To shed light on the text cues for semantic supervision in DFER, we visualize the heatmaps of the model's attention on expression sequences with/without text information in Fig. 4.

Fig. 4 shows that the model without the supervision of text cues (i.e., w/o text) is unable to accurately focus on some easily confused expression sequences, e.g. predicting 'sad' as 'angry' or 'angry' as 'natural'. In contrast, after

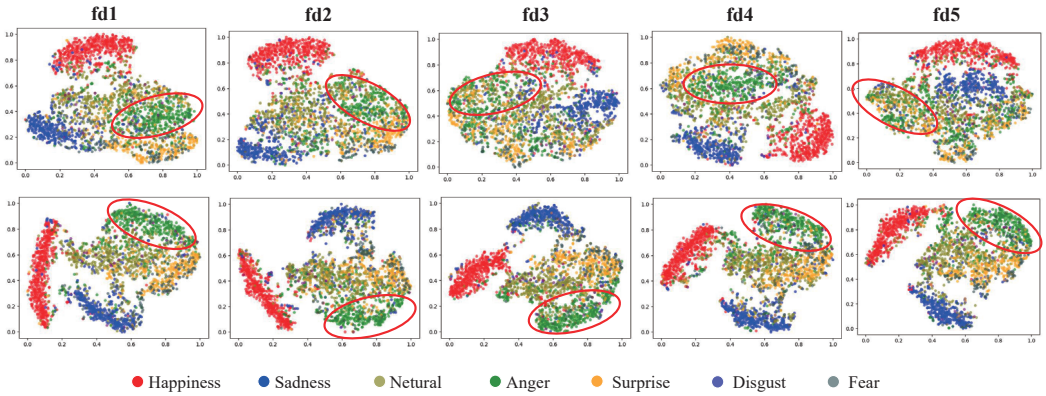

Figure 5. Feature distributions learned by the baseline (top) and our proposed method (bottom) on fd1∼fd5 of DFEW. The red boxes label the example regions specific to 'angry' class where the feature distributions of the baseline and ours differ much.

Table 4. Performance (%) sensitivity of the number of BiEhPro blocks (M) on DFEW. $M = 2$ is the default.

M	FLOPs(G)	UAR	WAR
1	24.68	60.22	72.45
2	24.71	**60.85**	**72.58**
3	24.74	59.44	72.38
4	24.77	59.92	72.28
5	24.80	59.59	72.21

Table 5. Comparison (%) of different prompting methods on DFEW. TP: text prompt. VP: visual prompt. TM: temporal model.

Method	Prompt	UAR	WAR
Zero-shot CLIP [35]	TP	23.34	20.07
CoOp [52]	TP	44.98	56.68
CoCoOp [51]	TP	46.80	57.52
DFER-CLIP(w/o TM) [50]	TP	57.39	69.00
DFER-CLIP($w/$ TM) [50]	TP	59.61	71.25
Ours(w/o InterFIT, MSTA, VP)	TP	58.90	71.12
Ours(w/o InterFIT, MSTA)	TP+VP	59.31	71.64
Ours($w/$ InterFIT, MSTA)	TP+VP	**60.85**	**72.58**

introducing text cues (i.e., w/ text), it can guide the model to focus on facial regions with higher expression characteristics, such as the mouth or cheeks, to enhance semantic supervision in expression understanding and recognition.

T-SNE Visualization. In this section, we use t-SNE [42] to visualize the distribution of dynamic expression features represented by the baseline and our method in Fig. 5. It shows that the features obtained by the baseline exhibit obvious overlaps between classes. In contrast, our method achieves clearer boundaries between classes (as shown in the red box in Fig. 5, our approach is clearer than the baseline in distinguishing 'angry' from other classes) and makes features distribute around cluster centers, i.e., learning more discriminative features for DFER.

5. Conclusion

Due to insufficient semantic supervision in current methods of dynamic facial expression recognition (DFER), it is difficult to distinguish confusing expression videos by relying solely on their labelled tags. Recently, the introduction of the visual-language pre-training model, i.e., CLIP provides a promising direction for effective semantic supervision. However, it remains a major challenge how to utilize CLIP to capture the dynamic changes in expressions. To this end, we design a CLIP-based Inter-Frame Interaction Transformer and Multi-Scale Temporal Aggregation, enabling CLIP to extract dynamic features of expressions while aggregating expression cues from multiple temporal scales. In addition, we also design Bidirectional Enhanced Prompt to enhance the complementarity of visual and text cues in semantic supervision. Extensive experiments validate the competitiveness of our method over the state of the arts on three widely-used benchmarks, i.e., DFEW, FERV39k and MAFW.

Although our method has shown effectiveness, it may still fail to effectively capture and recognize short or subtle facial expression changes, as shown in the supplementary material. In our future work, we will design a dedicated micro-expression detection model and integrate multiple modalities to improve our model.

Acknowledgment

The work was supported by the Natural Science Foundation of China under grants no. 62276170, 82261138629, the Science and Technology Project of Guangdong Province under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20220531101412030, Guangdong Provincial Key Laboratory under grant no. 2023B1212060076.

References

[1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516, 2013.

[5] J. Ding, N. Xue, G.-S. Xia, and D. Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474, 2015.

[8] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016.

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

[10] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[11] J. Guo, C. Wang, Y. Wu, E. Zhang, K. Wang, X. Xu, H. Shi, G. Huang, and S. Song. Zero-shot generative model adaptation via image-specific prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11494–11503, 2023.

[12] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] S. Huang, B. Gong, Y. Pan, J. Jiang, Y. Lv, Y. Li, and D. Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6565–6574, 2023.

[15] X. Huang, Q. He, X. Hong, G. Zhao, and M. Pietikainen. Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 514–520, 2014.

[16] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[17] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.

[18] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022.

[19] J. Kossaifi, A. Toisoul, A. Bulat, Y. Panagakis, T. M. Hospedales, and M. Pantic. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6060–6069, 2020.

[20] B. Lee, H. Shin, B. Ku, and H. Ko. Frame level emotion guided dynamic facial expression recognition with emotion grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5680–5690, 2023.

[21] 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.

[22] H. Li, H. Niu, Z. Zhu, and F. Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 67–75, 2023.

[23] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[24] T. Li, K.-L. Chan, and T. Tjahjadi. Multi-scale correlation module for video-based facial expression recognition in the wild. *Pattern Recognition*, 142:109691, 2023.

[25] Y. Li, Y. Gao, B. Chen, Z. Zhang, L. Zhu, and G. Lu. Jdman: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3312–3320, 2021.

[26] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 24–32, 2022.

[27] Y. Liu, Y. Lu, H. Liu, Y. An, Z. Xu, Z. Yao, B. Zhang, Z. Xiong, and C. Gui. Hierarchical prompt learning for

multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10888–10898, 2023.

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

[29] C. Lu, Y. Jiang, K. Fu, Q. Zhao, and H. Yang. Lstpnet: Long short-term perception network for dynamic facial expression recognition in the wild. *Image and Vision Computing*, 142:104915, 2024.

[30] F. Ma, B. Sun, and S. Li. Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[31] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[32] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022.

[33] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021.

[34] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[35] 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*, pages 8748–8763. PMLR, 2021.

[36] A. Sain, A. K. Bhunia, P. N. Chowdhury, S. Koley, T. Xiang, and Y.-Z. Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2023.

[37] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[38] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st international on multimodal sentiment analysis in real-life media challenge and workshop*, pages 27–34, 2020.

[39] Z. Tao, Y. Wang, Z. Chen, B. Wang, S. Yan, K. Jiang, S. Gao, and W. Zhang. Freq-hd: An interpretable frequency-based high-dynamics affective clip selection method for in-the-wild facial expression recognition in videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 843–852, 2023.

[40] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[41] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[42] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17958–17968, 2023.

[44] K. Wang, Z. Lian, L. Sun, B. Liu, J. Tao, and Y. Fan. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 75–80, 2022.

[45] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022.

[46] Y. Wang, Y. Sun, W. Song, S. Gao, Y. Huang, Z. Chen, W. Ge, and W. Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 101–110, 2022.

[47] M. Yu, H. Zheng, Z. Peng, J. Dong, and H. Du. Facial expression recognition based on a multi-task global-local network. *Pattern Recognition Letters*, 131:166–171, 2020.

[48] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[49] Z. Zhao and Q. Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021.

[50] Z. Zhao and I. Patras. Prompting visual-language models for dynamic facial expression recognition. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*, page 98. BMVA Press, 2023.

[51] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

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