Temporal Group Constrained Transformer with Deformable Landmark Attention for Video Dimensional Emotion Recognition

Weixin Li, Xiangjing Meng, Linmei Hu, and Xuan Dong

***Abstract*—Video dimensional emotion recognition aims to map human affect into the dimensional emotion space based on visual signals. Recent works notice that it is beneficial to locate key facial regions related to human emotion perception, as well as establish long-term temporal dependencies. While preliminary attempts have been made, there still exists much space for further improvements. In this paper, to better exploit key facial regions, we propose the Temporal cue guided Deformable Landmark Spatial (TDLS) transformer which attends to key facial regions in a data-dependent manner. We also propose the temporal cue guided frame representation learning to learn the spatial representation of each frame by considering features of other frames together. To better model temporal dependencies, we propose the Multi-layer Group Constrained Temporal (MGCT) transformer to summarize features of frames to multi-layer groups, perform group-to-group communications, and let group-level features guide the frame-level emotion recognition. We also introduce cross-clip representation learning to generate consistent results across different clips and videos. Extensive experiments are conducted on two benchmark datasets and superior results are achieved by our method compared to state-of-the-art approaches.**

***Index Terms*—Dimensional emotion recognition, group constrained temporal transformer, deformable landmark attention.**

I. Introduction

In recent years, the task of emotion recognition has gained significant attention from both academia and industry for its wide applications in human-computer interaction [1], healthcare [2], [3], driver fatigue monitoring [4], *etc*. The dimensional emotion model quantifies human affective behaviors in a continuous dimensional space, and allows for expressing and understanding complex emotions [5]–[9]. Given that videos are easily accessible in general and contain rich emotion-related cues, we focus on emotion recognition in videos in this paper, and aim at inferring dimensional emotion values, including valence and arousal, for single frames to describe fine-grained and variational human affect (as shown in Figure 1).

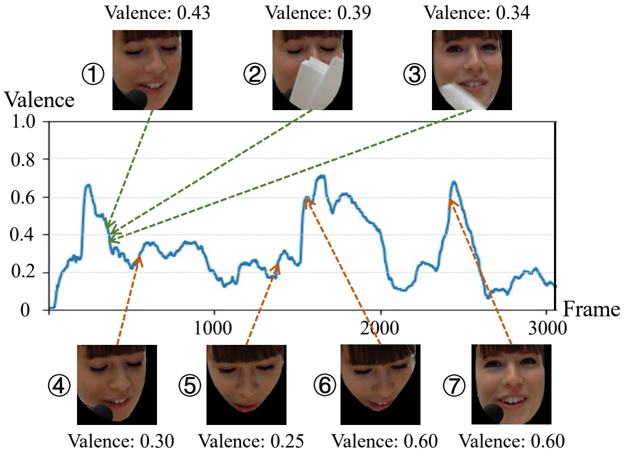


Fig. 1. Dimensional emotion variance of an example sample from the Recola dataset [21]. Frames ⑤ and ⑥ have very similar appearances, yet their valence values differ significantly. The valence value of the occluded frame ② is challenging to estimate on its own, but can be inferred from other frames ① and ③ nearby.

The keys to solving video dimensional emotion recognition are *spatial affective representation learning for each frame* and *inter-frame communication for exploiting their temporal correlations* [10]–[13]. For the former, existing works use various methods including hand-crafted features [10], [14], Convolutional Neural Network-based [12], [13], [15]–[17], or Transformer-based [18], [19] methods to capture emotion related cues from human faces. For the latter, recent efforts adopt Recurrent Neural Networks (RNNs) [15], [17] or Temporal Convolutional Neural Networks (TCNs) [11], [16], [20] to model frame dependency and achieve promising results.

It is also noticed that certain emotional prior knowledge is helpful to solve this emotion recognition problem. (1) Firstly, the perception of human emotions is closely related to appearances of key face components, *e.g.* eyes, mouth and nose, and different face components are complementary to each other [22]. To exploit this knowledge, recent methods [16], [19], [22], [23] generate attention maps to focus on key facial regions and obtain remarkable accuracy. However, the attention operations of these methods perform grid-based partition, which is data-agnostic and thus may not be optimal.

This work was supported in part by the National Key R&D Program of China under Grant 2022ZD0161901, the National Natural Science Foundation of China under Grant 62276018, and BUPT Kunpeng & Ascend Center of Cultivation.

Weixin Li and Xiangjing Meng are with the School of Computer Science and Engineering, Beihang University, Beijing, China (email: weixinli@buaa.edu.cn; 16061127@buaa.edu.cn). Linmei Hu is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China (email: hulinmei@bit.edu.cn). Xuan Dong is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China (email: dongxuan8811@bupt.edu.cn). (Corresponding author: Xuan Dong.)

Manuscript received 28 March 2024; revised 7 June 2025; accepted 8 September 2025.

(a) Grid-based attention (b) Deformable landmark attention

Fig. 2. Comparisons of grid-based attention and our deformable landmark attention.

As shown in Figure 2, only part of the points are concentrated in the emotion-related facial regions in grid-based attention, leaving space for further improvement. (2) Secondly, modeling long-term temporal frame dependency can help refine affective representations of single frames. Recent works [12], [24] use Transformer to perform frame-to-frame communications, achieving promising results. However, as shown in Figure 1, since some frames may have occlusions, *e.g.* frame ②, or appearances with ambiguous emotion states, *e.g.* frames ⑤ and ⑥, frame-to-frame correlation modeling may be wrong, as illustrated in Figure 3. In this case, the flickering problem may appear, *i.e.* the estimated results of some frames are inaccurate and shake sharply over their temporally neighboring frames.

In this paper, we provide some novel insights to address these challenges. (1) For spatial affective representation learning, we enable candidate points in the attention operation to be flexibly located on emotion-related positions of eyes, mouth and nose in a data-dependent manner, as depicted in Figure 2. (2) For temporal inter-frame correlation modeling, in addition to frame-level communication, we summarize the information of neighboring frames to groups, conduct group-to-group communications, and expand group-level information back to the frame level. In this way, the wrongly estimated correlation between individual frames due to occlusions or ambiguous appearances can be corrected based on the comprehensive group-level information as illustrated in Figure 3, thus correcting the estimated emotion recognition results of flickering frames and solving the flickering problem. Integrating these insights, we propose a fully transformer-based model for video dimensional emotion recognition. As illustrated in Figure 4, 1) we propose the Temporal cue guided Deformable Landmark Spatial (TDLS) transformer to extract affective representation of each frame. The deformable landmark attention module is proposed to locate emotion-related key face positions and establish their connections. We also propose the temporal cue guided frame representation learning to learn the spatial representation of each frame by considering representations of other frames together. 2) For temporal correlation modeling, we propose the Multi-layer Group Constrained Temporal (MGCT) transformer. The multi-layer group constrained transformer is built, where, besides letting different frames communicate directly, we extract features of frame groups in a multi-layer way and perform communications at multiple group-level layers as well. The group features have summarized temporal information in a larger field of view, thereby guiding the frame-level representation learning to avoid the flickering problem. To train the MGCT transformer, the cross-clip representation learning is proposed to generate consistent results over different clips and videos.

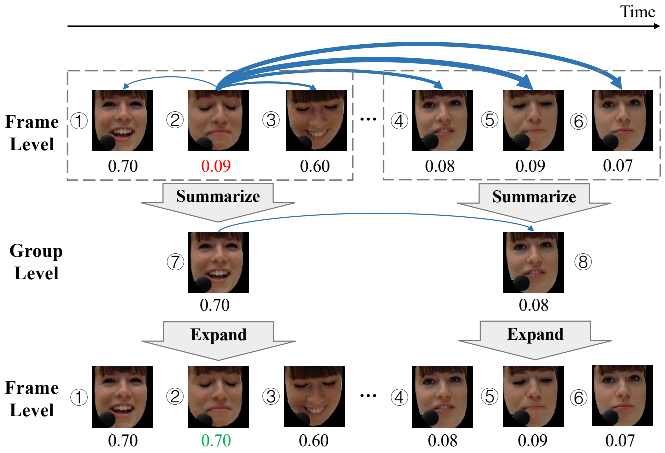


Fig. 3. In the initial frame-level processing, frame ② gets inaccurate estimation due to high visual similarity with frame ⑤. By leveraging expanded group-level information, frame ② can obtain a more accurate estimation in subsequent frame-level processing.

We demonstrate the effectiveness of the proposed method by extensive experiments on two widely used benchmark datasets, namely Recola [21] and SEWA [25]. The proposed method outperforms state-of-the-art approaches on both datasets.

In summary, the main contributions of this work include: (1) the TDLS transformer which contains the deformable landmark attention module to locate key face positions as the attention points in a data-dependent manner for spatial affective representation learning, (2) the MGCT transformer which models long-term temporal affective dependencies using contextual cues from multi-layer groups, (3) the temporal cue guided frame representation learning and cross-clip representation learning for avoiding overfitting problem during the training of the dimensional emotion regression model, and (4) a fully transformer-based spatio-temporal model for video dimensional emotion recognition, which achieves the state-of-the-art performance on Recola and SEWA datasets.

II. Related Works

A. Spatial Affective Representation Learning

Traditional spatial affective feature extraction methods are mostly handcrafted ones including Local Binary Patterns (LBP) [10], Local Gabor Binary Patterns from Three Orthogonal Planes (LBP-TOP) [14], Nonnegative Matrix Factorization [14], *etc*. Recently, deep learning based methods make breakthroughs in improving the performance of emotion recognition. Most of these methods take global facial images as input [11], [15], [20], neglecting that the perception of human emotions is also closely related to local key face regions. Some methods [26], [27] crop key face patches based on landmarks, and achieve the feature aggregation using learned weights for multiple local patches. Tellamekala *et al.* [28] utilize the 3D face shapes to recognize continuous emotions. However, these methods rely on the extraction of facial landmarks or 3D face shapes, which is easily effected by some factors *e.g.* illumination variations and occlusions. Xue et al. [22] and Hu et al. [16] propose to apply the self-attention mechanism to original feature maps. However, the discriminative regions useful for

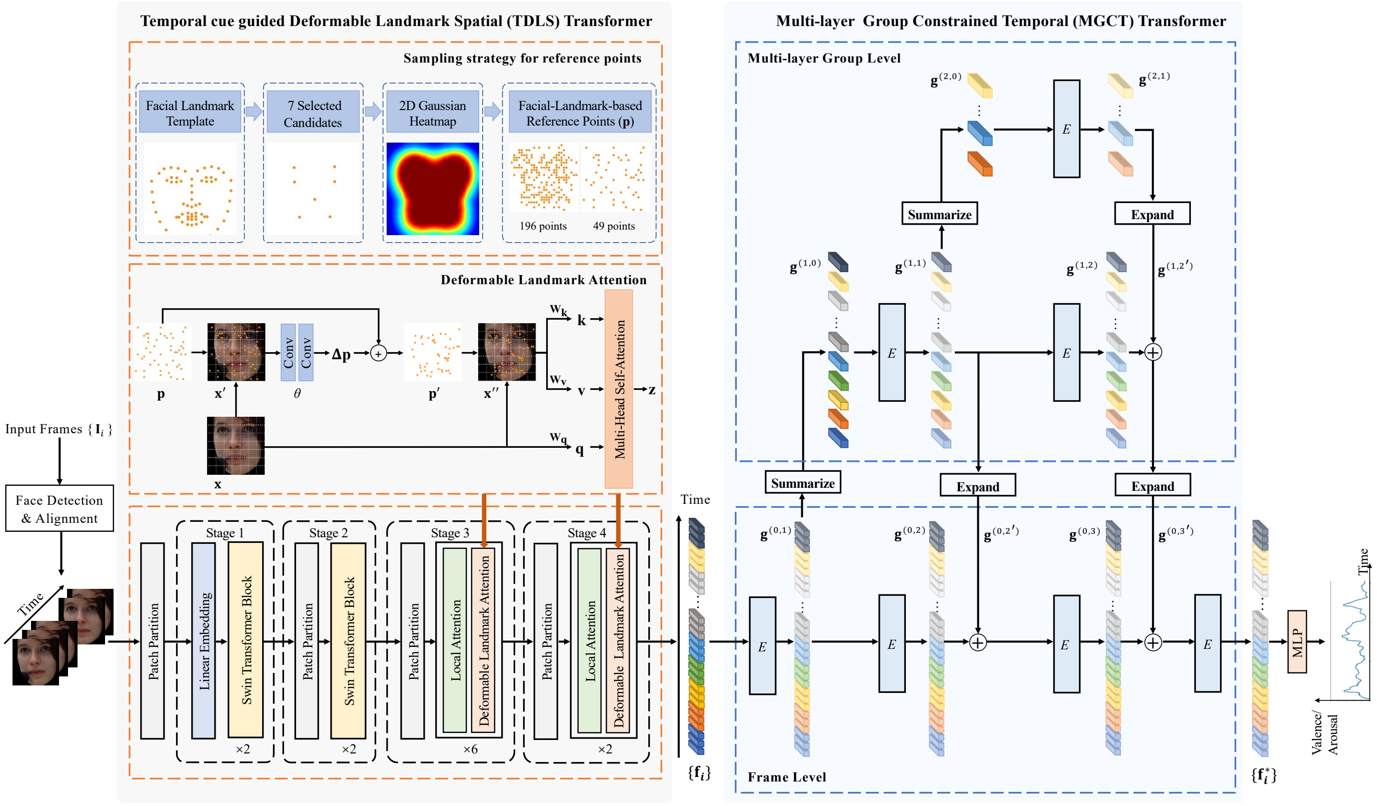


Fig. 4. An overview of our Temporal cue guided Deformable Landmark Spatial (TDLS) transformer and Multi-layer Group Constrained Temporal (MGCT) transformer. Given a frame sequence $\{\mathbf{I}_i\}$, facial images are first extracted and aligned, and then fed into TDLS transformer which can attend to emotion-related areas and generate an affective representation sequence $\{\mathbf{f}_i\}$. MGCT transformer models long-term temporal cues by exploiting group-level representations and generates refined representation sequence $\{\mathbf{f}_i^*\}$ for final prediction.

dimensional emotion recognition can hardly be well located in this straightforward way. Liu *et al.* design orientation tokens to explicitly encode basic orientation regions [29]. Xue *et al.* propose two attentive pooling modules to pool noisy features to avoid the influence of background [30]. To sum up, these methods explicitly emphasize the necessity of focusing on local key regions for emotion recognition, but leave much space for improvement in effective local region localization and spatial feature extraction.

B. Temporal Affective Dependency Modeling

Recently, RNNs and TCNs have demonstrated their ability to establish temporal dependencies for video dimensional emotion recognition. Chao et al. [31] utilize the Long Short-Term Memory (LSTM) network to model temporal correlations. Sun et al. [32] introduce self-attention in LSTM to improve its ability to capture long-term dependencies. Bose *et al.* [9] propose to predict the time-varying emotion state as a series of beta distributions based on LSTMs. However, RNNs generally have the gradient vanishing problem and probably cannot well capture the long-range semantics. Recent methods adopt TCNs [11], [16], [33] which expand receptive field of the convolution by dilated convolution or down-sampling, and learn affective fluctuation in a coarse-to-fine way. But the actual receptive field of TCNs is smaller than the theoretical receptive field [34], leading to their limitations in modeling long-term temporal dependencies.

Recently, transformers are applied to address various computer vision problems and achieve promising performance, including image understanding tasks *e.g.* image classification, object detection, and instance segmentation [35]–[39], video understanding tasks *e.g.* action recognition [40]–[43], *etc*. Compared with CNNs, RNNs and TCNs, transformers have a larger receptive field and are competitive in modeling long-range dependencies. Thus transformers are suitable to be employed to establish connections between local face regions, as well as capture temporal contextual cues from videos with long spans in our method. Transformers have been adopted for video understanding [40], video classification [44], the detection of Mild Cognitive Impairment [45], *etc*. Recent years also witness the utilization of large language and multimodal models for emotion recognition, *e.g.* Emotion-LLaMA [46], Video-LLaMA [47], but they are mostly dealing with the discrete emotion recognition task. One important reason is the lack of large-scale continuous video emotion recognition datasets.

Moreover, transformers are not suitable to be directly used for the video dimensional emotion recognition problem. First, existing vision transformers mostly adopt a grid-based attention pattern where potential attended regions are generated from uniform grids. Yet it is inevitable that emotion-irrelevant redundant regions exist in the grid-based attention, and some potential key regions are not fully considered, leading to unsatisfactory performance. Secondly, performing communications directly between frames in the video as is done by most existing transformers may cause the flickering problem for video dimensional emotion recognition. Thirdly, when training on video-based emotion recognition datasets which usually have small sizes, transformers with a large number of parameters usually suffer from the problem of overfitting. Our method fully considers and effectively solves these issues.

III. Methodology

We propose a fully transformer-based model for video dimensional emotion recognition. As shown in Figure 4, 1) to extract emotion-related features $\mathbf{f}_i$ for each frame $\mathbf{I}_i$, where $i = 1, ..., T$ is the frame index and T is the number of frames, we propose the Temporal cue guided Deformable Landmark Spatial (TDLS) transformer. In TDLS transformer, the deformable landmark attention module locates emotion-related key face positions and extracts their features in a data-dependent manner, as well as establishes connections between features of these positions in a global scope. This can help extract features in key areas of the face, comparing with grid-based attention from popular vision transformers, *e.g.* Swin Transformer [48], ViT [36], *etc*. We also propose the temporal cue guided frame representation learning to learn the spatial representation of each frame by considering features of other frames together. 2) To further make use of temporal information to boost the recognition accuracy of each frame, we propose the Multi-layer Group Constrained Temporal (MGCT) transformer to refine features $\{\mathbf{f}_i\}$ to $\{\mathbf{f}_i^*\}$. In MGCT transformer, we build the multi-layer group constrained transformer, which extracts features for groups of frames in a multi-layer way and performs communications between different groups. The group features have summarized temporal cues in a larger field of view. They can help correct the wrongly estimated correlation between individual frames obtained only using the frame-to-frame communications, thus alleviating the flicking problem. We also propose the cross-clip representation learning to generate consistent results over different clips for training the MGCT transformer.

A. TDLS Transformer

To better learn the spatio representations of each frame, we design the deformable landmark attention module to focus on emotion-related key face positions more effectively Specifically, the Swin Transformer is employed as the backbone of our TDLS Transformer, which consists of four stages with a structure of feature pyramid. We replace its shift-window attention module with the proposed deformable landmark attention in the third and fourth stages following [49].

We also learn the spatial representation of each frame by considering other frames in the video, to exploit the temporal cues in the spatial feature extraction stage. However, due to memory constraints, we have only a limited number of other frames available to be used for comparison with the current frame. To obtain sufficient samples for training, motivated by Momentum Contrast (MoCo) [50], we generate augmented data from these frames, and pull frames with similar valence/arousal values closer in a supervised way.

1) Deformable landmark attention: As shown in Figure 4, a set of reference points are firstly generated, which remain the same for all the input frames. Specifically, we firstly average the coordinates of facial landmarks of all images in the training set, and use the averaged landmark locations as the face template (with 68 landmarks in total). Secondly, from all landmarks in the face template, we select 7 representative ones with relatively scattered positions, which are located at the center of the two eyebrows, center of the under-eye contour, tip of the nose and two corners of the mouth, respectively. Thirdly, we use these 7 points as the mean value to generate an activation heatmap with the size of $H_S \times W_S$ that obeys the 2D Gaussian distribution, since their nearby areas are treated as more important than areas far away from them for emotion recognition [51]. Fourthly, from the heatmap, we sample N_r position coordinates $\mathbf{p}$, where N_r is the number of embeddings in the current stage. The sampling principle is that for each pixel in the heatmap, the probability of selecting its position is proportional to the pixel value. The sampled points $\mathbf{p}$ are used as the reference points. As Figure 4 shows, for the flattened input embeddings $\mathbf{x} \in \mathbb{R}^{N_r \times C}$ of a stage, where C denotes the embedding dimension, our deformable landmark attention module uses reference positions $\mathbf{p}$ as key positions to extract rough reshaped embeddings $\mathbf{x}'$, *i.e.*

$$\mathbf{x}' = \phi(\mathbf{x}; \mathbf{p}), \tag{1}$$

where $\phi(\cdot)$ is the bilinear interpolation function.

Since the fixed reference points may not be suitable for the current input image, we further feed $\mathbf{x}'$ into a sub-network θ, to learn the offset $\Delta\mathbf{p}$ of $\mathbf{p}$, *i.e.*

$$\Delta\mathbf{p} = \theta(\mathbf{x}'), \tag{2}$$

where θ is composed of two CNN layers and each layer consists of the 3×3 convolution and ReLU. After getting $\Delta\mathbf{p}$, we generate the deformed points $\mathbf{p}'$ by $\mathbf{p}' = \mathbf{p} + \Delta\mathbf{p}$, and the updated embeddings $\mathbf{x}''$ of $\mathbf{x}$ using the deformed points $\mathbf{p}'$ are computed by

$$\mathbf{x}'' = \phi(\mathbf{x}; \mathbf{p}'). \tag{3}$$

Thus the deformed points are obtained in a data-dependent way, and can be optimized within the TDLS transformer to improve the emotion recognition performance.

Next, we use $\mathbf{x}$ to generate query feature $\mathbf{q}$, and $\mathbf{x}''$ to generate key/value features $\mathbf{k}$ and $\mathbf{v}$, *i.e.*

$$\mathbf{q} = \mathbf{x}\mathbf{W_q}, \mathbf{k} = \mathbf{x}''\mathbf{W_k}, \mathbf{v} = \mathbf{x}''\mathbf{W_v}, \tag{4}$$

where $\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v} \in \mathbb{R}^{C \times C}$ denote the projection matrices. Multi-Head Self-Attention between $\mathbf{q}$, $\mathbf{k}$ and $\mathbf{v}$ is further performed to generate self attended feature $\mathbf{z}$, *i.e.*

$$\mathbf{z}^{(m)} = \sigma(\mathbf{q}^{(m)}\mathbf{k}^{(m)^\top}/\sqrt{d})\mathbf{v}^{(m)}, m = 1, ..., M, \tag{5}$$

$$\mathbf{z} = Concat(\mathbf{z}^{(1)}, ..., \mathbf{z}^{(m)})\mathbf{W_o}, \tag{6}$$

where $\mathbf{z}^{(m)}, \mathbf{q}^{(m)}, \mathbf{k}^{(m)}, \mathbf{v}^{(m)} \in \mathbb{R}^{N_r \times d}$ denote the embedding output, query, key, value from the m-th attention head respectively. $\mathbf{W_o} \in \mathbb{R}^{C \times C}$ denotes the projection matrix. $\sigma(\cdot)$ is the softmax function. $d = C/M$ is the dimension of each head and M denotes the number of attention heads.

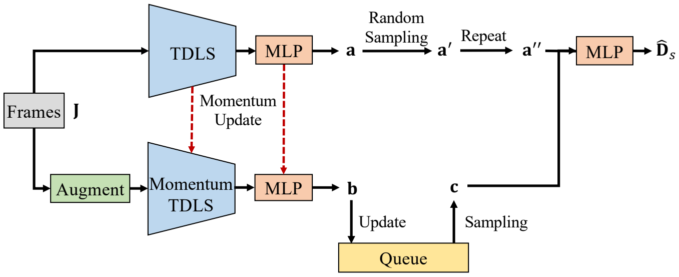


Fig. 5. Temporal cue guided frame representation learning. A regression task is proposed to predict distances between frames.

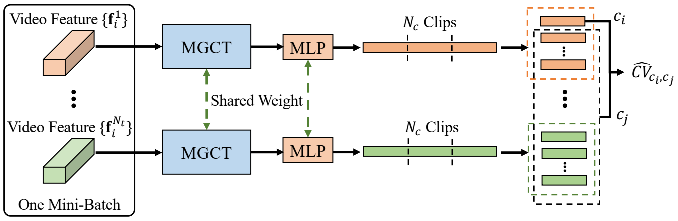

Video Feature $\{\mathbf{f}_i^1\}$
Video Feature $\{\mathbf{f}_i^{N_t}\}$
One Mini-Batch
MGCT
MLP
Shared Weight
MGCT
MLP
N_c Clips
N_c Clips
c_i
c_j
$\widetilde{CV}_{c_i,c_j}$

Fig. 6. Cross-clip representation learning. A regression task is proposed to predict distances between different clips.

2) Temporal cue guided frame representation learning: Given a frame, we obtain temporal cue by predicting the difference between it and other frames in the video to help learn its spatial feature. And we generate augmented data to obtain more samples for training.

As shown in Figure 5, for a set of input frames $\mathbf{J} = \{\mathbf{I}_i | i = 1, ..., bs\} \in \mathbb{R}^{bs \times 3 \times 224 \times 224}$ where bs is the number of input frames, we feed $\mathbf{J}$ into the TDLS Transformer and an MLP to get the result $\mathbf{a} \in \mathbb{R}^{bs \times D}$ where D is the dimension of each feature. The augmented frames of $\mathbf{J}$ are fed into the momentum version of the TDLS Transformer, named Momentum TDLS, as well as an MLP to get the result $\mathbf{b} \in \mathbb{R}^{bs \times D}$. The two encoders have the same network structure. Parameters in the TDLS Transformer are updated by gradient back propagation, while parameters in Momentum TDLS are updated in a moving average manner with a momentum parameter m_{MT} as $\xi \leftarrow m_{MT}\xi + (1 - m_{MT})\xi'$, where ξ' and ξ are parameters in the TDLS Transformer and Momentum TDLS, respectively.

We maintain a queue containing affective representations of all images in the training set, which is updated by $\mathbf{b}$. We perform random sampling in $\mathbf{a}$ to get $\mathbf{a}' \in \mathbb{R}^{1 \times D}$, and repeat $\mathbf{a}'$ to get $\mathbf{a}'' \in \mathbb{R}^{N_k \times D}$. We then sample N_k representations $\mathbf{c} \in \mathbb{R}^{N_k \times D}$ near the frame in $\mathbf{a}'$ from the queue. $\mathbf{a}''$ and $\mathbf{c}$ are concatenated and fed into another MLP for estimating the absolute difference of the valence or arousal values of $\mathbf{a}''$ and $\mathbf{c}$, named $\hat{\mathbf{D}}_s$. The loss for training is defined as:

$$L_{TG} = \sum_{i=1}^{N_k} |\hat{\mathbf{D}}_s^i - \mathbf{D}_s^i|^2, \tag{7}$$

where $\mathbf{D}_s$ is the absolute difference value between ground-truth labels of $\mathbf{a}'$ and each of the N_k elements in $\mathbf{c}$.

B. MGCT Transformer

As shown in Figure 4, we build the multi-layer group constrained transformer to exploit temporal information of $\{\mathbf{f}_i\}$ among different frames so as to obtain more accurate and refined features $\mathbf{f}_i^*$ for each frame i.

We construct multi-layer groups and perform self attention in each layer. Groups in different layers have different temporal resolutions and thus can represent temporal features from short-term to long-term. By performing group-to-group communications in different layers, detailed analysis of both short-term and long-term relationships can be conducted. By feeding group-level features in different layers back into the frame-level computation, the estimation of each frame is guided by the comprehensive multi-layer group information. In this way, the wrongly estimated frame-to-frame correlation can be corrected, and the flicking problem can thus be alleviated.

Moreover, we propose the cross-clip representation learning to let clips with similar affective fluctuations have similar representations, so as to help generate consistent prediction results and avoid overfitting in model training of MGCT transformer.

1) Multi-layer group constrained transformer: We use the transformer encoder layer [12], which consist of a Multi-Head Self-Attention layer and a Feed-Forward layer, to let features of different frames and groups communicate with each other, *i.e.* the processing of E in Figure 4, where E is a transformer encoder layer consisting of a multi-head attention and one add and normalization operation.

In the frame-level processing (*i.e.* level 0), the features $\mathbf{g}_i^{(0,0)}$ are fed into the transformer encoder layer to have direct frame-to-frame communications, where $\mathbf{g}_i^{(0,0)} = \mathbf{f}_i$.

In the group-level processing, at any level l ($l = 1, 2, ...$), we first *summarize* features of temporally neighboring frames at level $l - 1$, *i.e.* $\mathbf{g}^{(l-1,1)}$ by max-pooling to get the group feature:

$$\mathbf{g}^{(l,0)} = pooling(\mathbf{g}^{(l-1,1)}). \tag{8}$$

Secondly, $\mathbf{g}^{(l,0)}$ are fed into a transformer encoder layer E to have group-to-group communications and obtain the processed feature $\mathbf{g}^{(l,1)}$, which denotes the features processed by the 1-*st* E at level l, *i.e.* $\mathbf{g}^{(l,1)} = E(\mathbf{g}^{(l,0)})$.

The group-level processing is performed iteratively to obtain multi-layer group features. When *expanding* the group feature *e.g.* $\mathbf{g}^{(l,1)}$ at level l to level $l - 1$, the linear interpolation is used, *i.e.*

$$\mathbf{g}^{(l-1,2')} = \Phi(\mathbf{g}^{(l,1)}), \tag{9}$$

where $\Phi(\cdot)$ denotes the linear interpolation operation, and we use the notation $2'$ in the right position within the bracket to denote that the features $\mathbf{g}^{(l-1,2')}$ and $\mathbf{g}^{(l-1,2)}$ will be added before further processing.

During training, in order to achieve joint optimization for frame-level and different group-level layers, we perform supervision at all levels. In detail, for the output of the last transformer encoder layer in different levels, the prediction is obtained through an inference sub-network, and we down-

TABLE I
COMPARISON RESULTS OF OUR MODEL IN TERMS OF CCC WITH THE STATE-OF-THE-ART VIDEO-BASE DIMENSIONAL EMOTION RECOGNITION METHODS ON THE RECOLA DATASET. † INDICATES RE-EXPERIMENT WITH THE SAME EXPERIMENTAL SETUP AS OURS.

Methods	Features	Models	Arousal	Valence
Lee et al. [15]	Raw Face	3D-CNN + LSTM + STA	-	0.546
Wu et al. [20]	Raw Face + P&G	FER-P&G-Net	0.603	0.686
Du et al. [11]	Raw Face	SCE + TH-CNN	0.656	0.677
Du et al. † [11]	Raw Face	SCE + TH-CNN	0.662	0.684
Chen et al. [12]	Geometric + LGBP-TOP	CNN-TE	0.533	0.664
Chen et al. † [12]	Geometric + LGBP-TOP	CNN-TE	0.551	0.668
Chen et al. [12]	Geometric + LGBP-TOP + Audio	CNN-TE	**0.838**	0.681
Hu et al. [16]	Raw Face	TS-SATCN	0.659	0.690
Praveen et al. [13]	Raw Face	I3D	0.582	0.642
Jegorova et al. [52]	Raw Face + Audio	3Dconv+ResNet18+GRU	0.675	0.626
Tran et al. [53]	Raw Face + Audio	AW-HuBERT	0.701	0.653
Bose et al. [9]	Raw Face + Audio	LSTM+Beta Distribution Modeling	0.639	0.321
baseline	Raw Face	Swin + TE	0.597	0.674
Ours	Raw Face	TDLS + MGCT	**0.697**	**0.719**

sample the ground-truth labels to the corresponding resolution for supervision. The loss is thus defined as:

$$L_{MGCT} = \sum_{i=1}^{N_l} \sum_{t=1}^{T} \lambda_i \|\hat{\mathbf{y}}_t^i - \mathbf{y}_t^i\|^2, \tag{10}$$

where N_l is the number of levels, $\hat{\mathbf{y}}_t^i$ is the prediction result of the i-th level at time step t, and $\mathbf{y}_t^i$ is the corresponding ground-truth label. λ_i is a parameter used to balance the learning process of different levels.

2) Cross-clip representation learning: The problem of insufficient training samples also exist in the temporal learning stage. So we propose a regression task to predict distances between clips for data augmentation.

As Figure 6 shows, we segment each video into N_c clips. A mini-batch has N_t videos, and thus has $N_c \times N_t$ clips in total. For any pair of clips c_i and c_j in the mini-batch, we propose to compute the cross-clip prediction value, *i.e.* $\hat{\mathbf{CV}}_{c_i,c_j} = |\hat{\mathbf{y}}_{c_i} - \hat{\mathbf{y}}_{c_j}|$, and the corresponding ground-truth cross-clip prediction value, *i.e.* $\mathbf{CV}_{c_i,c_j} = |\mathbf{y}_{c_i} - \mathbf{y}_{c_j}|$.

In the training of each epoch, we randomly select one clip c_i in the mini-batch as the main clip, and compute the cross-clip prediction values between c_i and all the other clips c_j in the mini-batch to form the loss, *i.e.*

$$L_{CCL} = \sum_{c_j} |\hat{\mathbf{CV}}_{c_i,c_j} - \mathbf{CV}_{c_i,c_j}|^2. \tag{11}$$

The insight of L_{CCL} is to encourage the estimated values of different clips to be consistent in the affective fluctuation.

Above all, the total loss of our method is:

$$L = \alpha L_{MGCT} + \beta L_{TG} + \gamma L_{CCL}, \tag{12}$$

where α, β, γ are parameters for balancing different terms.

IV. EXPERIMENTS

A. Datasets

To demonstrate the effectiveness of our method, we conduct extensive experiments on two widely used benchmark datasets, *i.e.* Recola [21] and SEWA [25].

Recola is one multimodal corpus recording behaviors of subjects as they work in pairs remotely. It contains 27 recording sessions annotated with Arousal and Valence, equally split for training, validation, and testing. The duration of each recording is 5 minutes with 7,501 frames.

SEWA collects multimodal spontaneous and naturalistic human interactions. It contains 64 recording sessions annotated with Arousal, Valence, and Liking, with 34, 14 and 16 sessions for training, validation and testing respectively. The recording duration ranges from 40 seconds to 3 minutes.

Both datasets have been adopted for the Audio-Visual Emotion recognition Challenges (AVEC) [56]–[58]. For both datasets, we only use visual signals, and adopt their standard data splits.

B. Metrics

We follow the protocol in [56], [57], and evaluate the performance of different video dimensional emotion recognition methods using Concordance Correlation Coefficient (CCC) [59] in the Arousal and Valence dimensions. Since sample labels in the test set are not available in Recola and SEWA datasets, we train the methods on the training set, and report experimental results on the validation set.

C. Implementation Details

Network parameters. In our TDLS transformer, for the backbone Swin Transformer [48], the activated outputs from its penultimate linear layer are used as spatial features which have a dimension of 128. The size of input images is 224×224. OpenFace [60] is used for face detection, alignment and

TABLE II
COMPARISON RESULTS OF OUR MODEL IN TERMS OF CCC WITH STATE-OF-THE-ART METHODS ON THE SEWA DATASET. † INDICATES RE-EXPERIMENT WITH THE SAME EXPERIMENTAL SETUP AS OURS. ALL THESE METHODS USE THE RAW FACES AS INPUT.

Methods	Features	Models	Arousal	Valence
Lee et al. [15]	Raw Face	3D-CNN + LSTM + STA	-	0.612
Huang et al. [17]	Raw Face	3D-CNN + ConvLSTM	0.583	0.654
Du et al. [11]	Raw Face	SCE + TH-CNN	0.715	0.713
Du et al. † [11]	Raw Face	SCE + TH-CNN	0.719	0.720
Sanchez et al. [54]	Raw Face	AP	0.662	0.672
Toisoul et al. [23]	Raw Face	FAN + Attention	0.610	0.650
Tellamekala et al. [55]	Raw Face	HG-FAN + APs	0.650	0.710
Tellamekala et al. [28]	Raw Face	EmoFAN + GRU	0.568	0.715
Jegorova et al. [52]	Raw Face + Audio	3Dconv + ResNet18 + GRU	0.713	**0.771**
baseline	Raw Face	Swin + TE	0.663	0.704
Ours	Raw Face	TDLS + MGCT	**0.737**	**0.740**

TABLE III
COMPARISON RESULTS OF OUR MODEL IN TERMS OF CCC WITH VIDEO TRANSFORMERS ON RECOLA DATASET. † INDICATES THAT THE MODEL IS NOT PRE-TRAINED ON AFFECTNET DATASET AND HAS THE SAME EXPERIMENTAL SETTINGS AS VIDEO TRANSFORMERS.

Methods	Models	Arousal	Valence	Param ($\times 10^6$)
Bertasius et al. [40]	Timesformer	0.539	0.612	121.4
Arnab et al. [44]	ViViT	0.550	0.634	115.1
Ours†	TDLS + MGCT	0.594	0.663	60.7
Ours	TDLS + MGCT	**0.697**	**0.719**	60.7

TABLE IV
EXPLORATION OF ATTENTION PATTERNS IN TERMS OF CCC ON RECOLA DATASET.

Models	Arousal	Valence
Swin	0.374	0.563
FL-Swin	0.376	0.560
RP-Swin	0.377	0.565
TDLS	**0.381**	**0.567**

TABLE V
EXPLORATION OF ATTENTION PATTERNS IN TERMS OF CCC ON AFFECTNET DATASET.

Models	Arousal	Valence
Swin	0.531	0.586
RP-Swin	0.534	0.588
TDLS	**0.535**	**0.594**

landmark detection for the training set. The size of the 2D Gaussian heatmap is 28×28. We sample 196 and 49 points for the third and fourth stages of Swin Transformer respectively. In the temporal cue guided frame representation learning, the momentum parameter m_{MT} is set to 0.999, and the number for sampling, *i.e.* N_k, is set to 768. For the MGCT transformer, we have one frame level and two multi-layer group levels, whose output channels are 128, 256, and 512 respectively. The number of clips from one input is set to 3 in cross-clip representation learning. Taking the dataset size, video length/frequency, and computation efficiency into consideration, we use 768 and 400 time steps on Recola and SEWA datasets respectively.

TABLE VI
EXPLORATION OF TEMPORAL-CUE GUIDANCE IN TERMS OF CCC ON RECOLA DATASET.

Models	Arousal	Valence
Swin w/o temporal-cue guidance	0.374	0.563
Swin w/ temporal-cue guidance	0.409	0.576
TDLS w/o temporal-cue guidance	0.381	0.567
TDLS w/ temporal-cue guidance	**0.417**	**0.578**

TABLE VII
EXPLORATION OF MOMENTUM UPDATE STRATEGY IN TERMS OF CCC ON RECOLA DATASET.

Models	Arousal	Valence
Momentum	**0.417**	**0.578**
Non-momentum	0.402	0.573

Network training. We train our TDLS and MGCT transformers separately. During training, the mini-batch size is set to 256 for the TDLS transformer. The TDLS transformer is first pretrained on ImageNet-1K [61] and then fine-tuned on AffectNet dataset [62] with a learning rate of $1e^{-5}$. The obtained model is then fine-tuned on Recola or SEWA dataset with a learning rate of $1e^{-6}$. We perform zoom with a factor from 1.1 to 1.5, horizontal flip and color perturbation for data augmentation in temporal cue guided frame representation learning at the training phase. For the MGCT transformer, the mini-batch is set to 64, and the learning rate is $1e^{-4}$. The parameter λ_i is set to 1.2, 0.9 or 0.6 in Eq. 10 for outputs with different scales. Parameters α, β and γ in Eq. 12 are set to 1, 0.1 and 0.1 respectively. We adopt Adam [63] for optimization. Our training is conducted on three Nvidia GeForce RTX 3090 cards.

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

We compare the proposed model with several state-of-the-art video dimensional emotion recognition methods on Recola dataset in Table I and on SEWA dataset in Table II. In terms of spatial affective representation learning, Chen et al. [12] use the handcrafted LGBP-TOP feature, while the other comparison methods [9], [11], [13], [15]–[17], [23], [28], [52], [55] extract features based on CNNs. Tran et al. [53] learn the audio and visual representations using Transformers. Du et al. [11] pretrain their Spatial Convolutional Encoder (SCE) on the FERplus [64] dataset in advance, Praveen et al. [13] pretrain their I3D model on the Kinetics-400 dataset [65], Jegorova et al. [52] pretrain their model using the Lip Reading Sentences 3 dataset (LRS3) [66], Tellamekala et al. [55] use the VGG-Face database [67] for model pre-training, while the other works do not mention pre-training in their papers. In terms of temporal modeling, CNN-TE [12] use transformer encoder (TE) while earlier methods [11], [15]–[17], [52] are based on RNNs or TCNs. CNN-TE [12] only selects 8 of the 9 videos in the original training set as its training set. Affective Process (AP) [54], [55] models the affective fluctuation as a stochastic process. To make fair comparisons, we re-experiment the open-source state-of-the-art methods with the same experimental setups as ours. We pre-train the SCE and TH-CNN in [11] on ImageNet-1K and AffectNet datasets and then fine-tune on the target dataset. We retrain CNN-TE [12] on all the 9 videos of training set of Recola. To demonstrate the effectiveness of our method more intuitively, we design a fully transformer-based baseline model in which the spatial encoder is a Swin Transformer and the temporal encoder is a 4-layer transformer-encoder.

As we can see from the results reported on Recola dataset in Table I, our method achieves the best performance for Arousal and Valence respectively when using only visual signals. Using both visual and audio signals, the CCC score of CNN-TE [12] for Arousal reaches the highest value 0.838, since Arousal can be reflected more from the audio signals. But even with audio signals, its performance for Valence is still inferior to our method, and it achieves relatively poor performance when only using visual signals.

We can draw similar conclusions from the results on SEWA dataset in Table II. Our method achieves the state-of-the-art results on SEWA dataset with improvements for both Arousal and Valence using only visual signals. Using both visual and audio signals, the CCC score of the method by Jegorova et al. [52] for Valence is better than our method, while their method also pretrains on LRS3 dataset. We attribute our improved performance to transformer's powerful long-range modeling capabilities, which can be seen from results of the baseline model. In addition, the proposed deformable landmark attention and multi-layer group level guidance enable our model to significantly outperform the baseline model.

E. Comparison with Video Transformers

We propose a fully transformer-based model for video dimensional emotion recognition. Recently, some fully transformer-based methods are proposed for video classification tasks, *e.g.* TimeSformer [40] and ViViT [44]. We also compare with them to show the effectiveness of our method. ViViT extracts spatio-temporal tokens from input video, and encodes them by a series of transformer layers. To handle long token sequences, several efficient variants of ViViT are proposed, and we adopt the Factorised encoder model considering both the accuracy and computational costs. TimeSformer learns spatio-temporal video features from a sequence of frame-level patches based on self-attentions. Here we adopt the divided attention with the best performance in TimeSformer. We initialize ViViT and TimeSformer using a ViT image model [36] trained on ImageNet-1K. Due to memory constraints, we reduce the length of a single sample for video transformers to 80, and re-experiment our method under the same setting.

The comparison results on Recola dataset are shown in Table III. Our method reduces the parameters by about half compared with video transformers, but the performance improves by 8.0% and 4.6% in terms of CCC for Arousal and Valence respectively under the same setting. We argue that the reason for the poor performance of video transformers is that large-scale models commonly have overfitting problems on small datasets, which can seriously damages their performance. Our method locates key positions and establishes connections between them, and models long-term temporal dependencies with multi-layer groups, thus effectively solving the video dimensional emotion recognition problem.

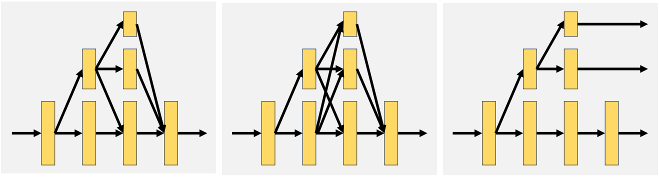

(a) Base (b) Summarize twice (c) No guidance

Fig. 7. Illustration of three different forms of MGCT transformer.

TABLE VIII
VALIDATION OF DIFFERENT FORMS OF MGCT TRANSFORMER ON RECOLA DATASET IN TERMS OF CCC.

Pattern	Arousal	Valence
Base	**0.697**	**0.719**
Summarize twice	0.681	0.711
No guidance	0.679	0.707

TABLE IX
EXPLORATION OF SUMMARIZATION PATTERN IN MGCT TRANSFORMER IN TERMS OF CCC ON RECOLA DATASET.

Pattern	Arousal	Valence
Max-pooling	**0.697**	**0.719**
Average-pooling	0.680	0.714
Concatenate	0.657	0.706

TABLE X
EXPLORATION OF GROUP SIZE IN TERMS OF CCC ON RECOLA DATASET.

Group Size	Arousal	Valence
3	0.662	0.702
4	**0.697**	**0.719**
5	0.691	0.705
6	0.672	0.698

TABLE XI
VALIDATION OF THE MGCT TRANSFORMER WITH DIFFERENT GROUP LEVELS IN TERMS OF CCC ON RECOLA DATASET.

Level			Arousal	Valence
Frame	1*st*-layer Group	2*nd*-layer Group		
✓			0.611	0.681
✓	✓		0.617	0.710
✓		✓	0.684	0.689
✓	✓	✓	**0.697**	**0.719**

TABLE XII
VALIDATION OF MULTI-LEVEL OUTPUTS IN OUR MGCT TRANSFORMER IN TERMS OF CCC ON RECOLA DATASET.

Level	Arousal	Valence
Frame	**0.697**	**0.719**
1*st*-layer Group	0.615	0.676
2*nd*-layer Group	0.576	0.645

TABLE XIII
EXPLORATION OF CROSS-CLIP REPRESENTATION LEARNING (CCL) IN TERMS OF CCC ON RECOLA DATASET.

Models	Arousal	Valence
TE w/o CCL	0.597	0.674
TE w/ CCL	0.611	0.681
MGCT w/o CCL	0.685	0.715
MGCT w/ CCL	**0.697**	**0.719**

F. Ablation Study

We also conduct ablation studies for our method to analyze the effectiveness of each component.

1) Ablation study for TDLS transformer: For the proposed TDLS transformer, we conduct experiments to evaluate contributions of the deformable landmark attention module, as well as the temporal cue guidance and momentum update strategies.

Effectiveness of deformable landmark attention. Table IV shows the results of Swin Transformer with different attention patterns on the Recola dataset, where the listed models in the first column are used as the spatial encoder. To avoid the influence of subsequent temporal modeling stage, features extracted by the spatial encoder are directly passed through an inference sub-network to obtain the final prediction results. RP-Swin is obtained by replacing points on the grids in Swin with the reference points obtained by our sampling strategy. RP-Swin improves by 0.8% for Arousal and 0.4% for Valence, which shows the rationality of our generated reference points. We argue that the improvement of RP-Swin compared with Swin comes from the avoidance of interference by irrelevant areas as well as the finer embeddings obtained through bilinear interpolation. With TDLS, the performance is improved by 1.9% for Arousal and 0.7% for Valence compared with Swin, which is attributed to the deformable landmark attention that intensively focuses on emotion-related regions in a data-dependent way. Adjusting the points based on the input facial image can contribute to more accurate attention results. We also compare with FL-Swin, which uses landmarks of test images obtained by external detectors as $\mathbf{p}'$. Its CCC scores are worse compared to TDLS, showing the effectiveness and necessity of our deformable landmark attention module integrated in the end-to-end Swin Transformer.

We also evaluate the deformable landmark attention module on AffectNet dataset. Table V shows the comparison results of transformers with different attention patterns on AffectNet dataset. The models are set in the same way as is done for Recola dataset. As shown in Table V, RP-Swin improves by 0.6% for Arousal and 0.3% for Valence on AffectNet dataset compared with Swin transformer. We argue that the improvement comes from the avoidance of the interference of irrelevant areas in RP-Swin. With our TDLS, the performance is improved by 0.8% for Arousal and 1.4% for Valence, which is attributed to the proposed deformable landmark attention.

Effectiveness of temporal cue guidance. The evaluation results of the temporal cue guidance strategy on Recola dataset are reported in Table VI. Here we still discard the temporal modeling stage. With temporal cue guidance, Swin transformer has an improvement of 9.4% for Arousal and 2.3% for Valence. The performance of TDLS with temporal cue guidance is improved by 9.4% for Arousal and 1.9% for Valence. From these results, it can be demonstrated that temporal cue can guide the model to learn affective fluctuation information so that more discriminative affective representations can be extracted. This can be proved from the experimental results that the improvement brought by temporal cue guidance for Arousal is much greater than that of Valence, since the recognition of Arousal requires more temporal information than Valence.

Effectiveness of momentum update. In TDLS transformer, we use the momentum update strategy to obtain more samples for comparison with the current frame under limited computing resources. To justify our choice, we also add the experiment with the non-momentum update strategy. The experimental results on the validation set of Recola dataset [21] are shown in Table VII. The CCC scores with non-momentum update strategy are lower than the scores achieved with momentum update strategy, showing its effectiveness.

2) Ablation study for MGCT transformer: For the proposed TDLS transformer, we conduct experiments to explore different forms of MGCT transformer, group-level feature fusion methods, group size and the number of levels in MGCT respectively. We also validate the multi-level outputs in MGCT and evaluate contributions of cross-clip representation learning.

Different forms of MGCT. We design three different forms of MGCT transformer as shown in Figure 7. In the first one as

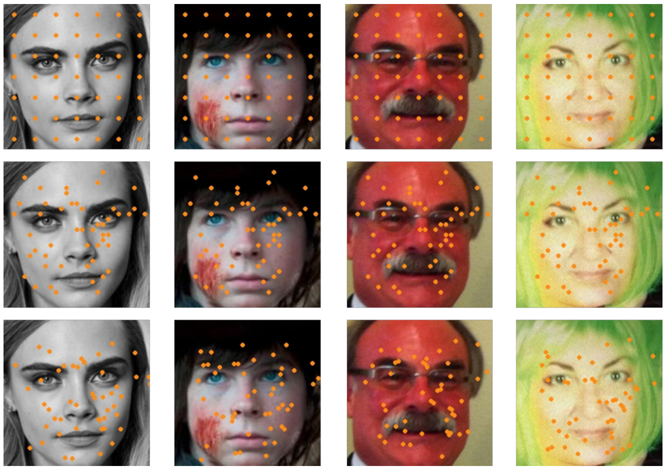


Fig. 8. Visualization results on AffectNet dataset [62] of sampling locations from different attention patterns in the spatial encoder at Stage 4. The points showed in the first two rows are generated from uniform grids and our sampling strategy respectively. The third row shows the learned sampling locations based on our deformable landmark attention.

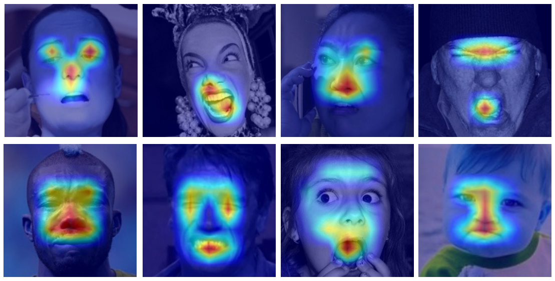


Fig. 9. Heatmap visualization of the TDLS transformer on AffectNet dataset.

shown in Figure 7(a), only when a new group level of feature sequence is generated for the first time, we summarize the feature sequence of its previous level, which is our default setting. In the second one as shown in Figure 7(b), at each layer of the group levels, feature sequences of all previous levels will be summarized. In the last one as shown in Figure 7 (c), we do not use any guidance from the multi-layer group layer.

The evaluation results of different forms of MGCT transformer on Recola dataset are reported in Table VIII. It can be seen that the best performance is achieved when the default setting is applied. More complex interactive mode does not necessarily lead to better performance, but may incur more parameters that are more difficult to optimize. The worst performance is achieved by the model without any guidance from the multi-layer group level, which demonstrates the necessity of group-level processing that summarizes effective group-level cues for alleviating the flickering problem.

Group-level feature fusion. In MGCT transformer, we use max-pooling for Summarize. We also conduct experiments with two other methods, *i.e.* average-pooling and concatenating for Summarize. The evaluation results on Recola dataset are reported in Table IX. It can be demonstrated that summarizing one group by max-pooling can help obtain better affective representation of the group.

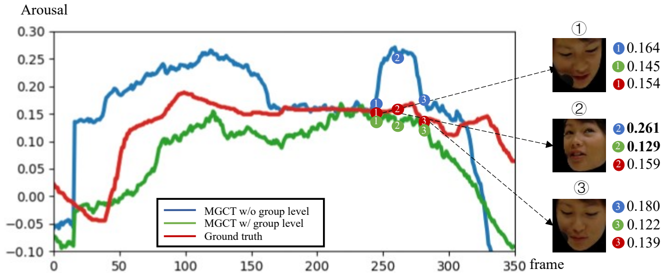
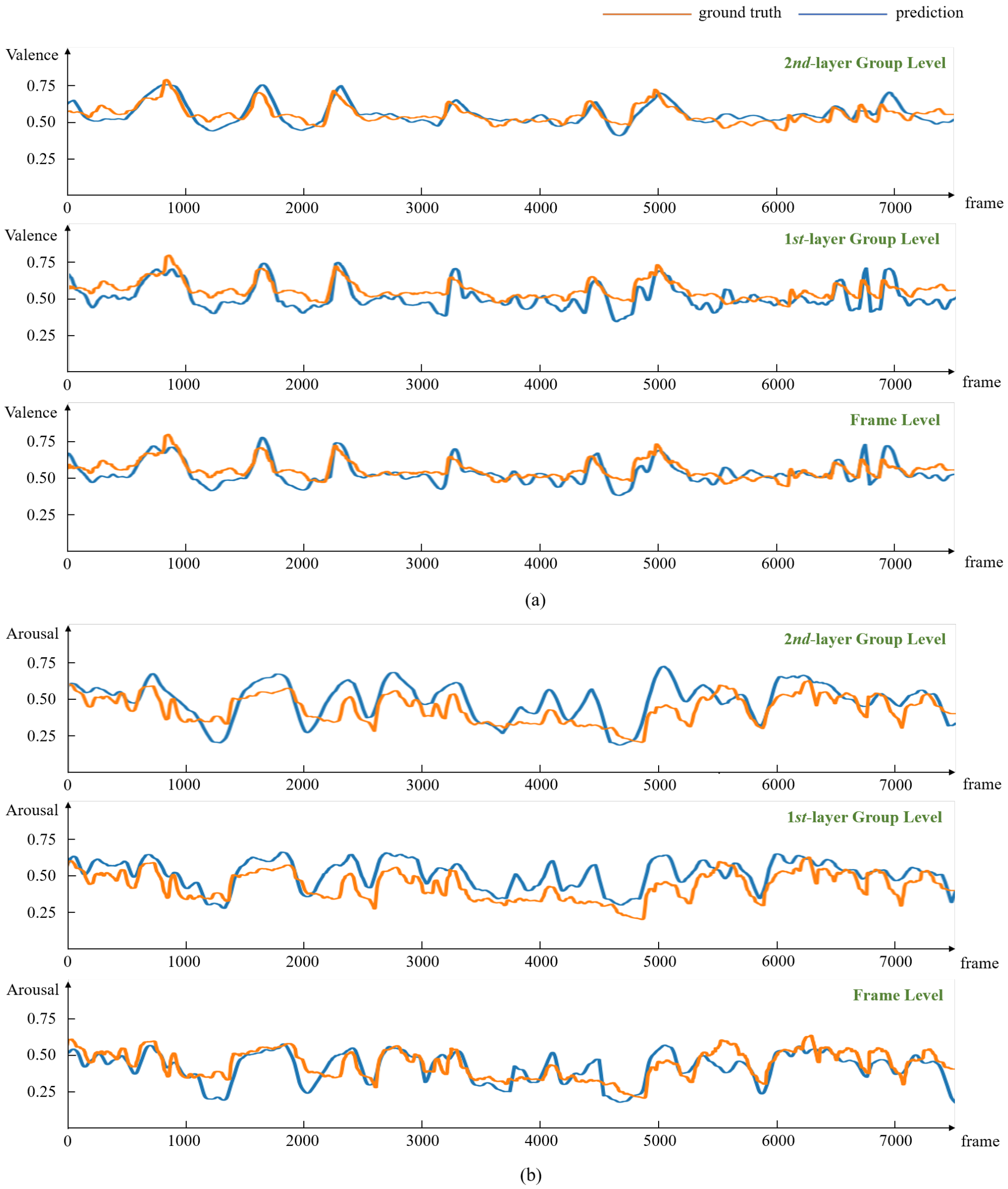


Fig. 10. Prediction curves of a test sample from the Recola dataset. Estimation of frame ② is inaccurate with only frame-level processing, but can be corrected when guided by group-level cues.

Group size. For the multi-layer group in MGCT transformer, in order to balance the accuracy it improves and the error it incurs, we perform ablation experiments on the number of representations used in each group. The evaluation results on Recola dataset are reported in Table X, which show that 4 is the most suitable size for one group.

Level number. To explore the impact of different level numbers in our MGCT transformer, we design four MGCTs respectively with different levels. When there is only the frame level, MGCT degenerates to the baseline temporal encoder. The validation results of different MGCTs on Recola dataset are shown in Table XI. We can find that when there is only the frame level, the performance is the worst. The performance of the model with the $2nd$-layer group level is better than that with the $1st$-layer group level in terms of CCC for Arousal. The corresponding results in terms of CCC for Valence are the opposite. We argue that the representations for Arousal have more noises than those for Valence, and the $2nd$-layer group level filters out more noises, resulting in better performance than the $1st$-layer group level. For Valence, the less affective detail loss is more important, so the performance of the $1st$-layer group level is better than that of the $2nd$-layer group level. The best performance is achieved when the model has all the levels, which demonstrates that features from three different levels are complementary to each other. We only test the performance of the proposed method for two levels due to computational constraints, since adding one level will increase the total computation costs, especially at higher levels.

Validation of multi-level outputs. Our MGCT transformer generates affective feature sequences in three different levels, *i.e.* the frame-level, the $1st$-layer group level, and the $2nd$-layer group level, which can be referred to as $\{\mathbf{f}_i^*\} \in \mathbb{R}^{T\times D}$, $\{\mathbf{g}^{(1,2)}\} \in \mathbb{R}^{(T/4)\times D}$, and $\{\mathbf{g}^{(2,1)}\} \in \mathbb{R}^{(T/16)\times(2\times D)}$ respectively, where T denotes the length of input image sequence and D denotes the feature dimension. All of them are fed into an inference sub-network respectively to generate the prediction results for arousal and valence, *i.e.* $\{\hat{\mathbf{y}}_t^1\} \in \mathbb{R}^{T\times 2}$, $\{\hat{\mathbf{y}}_t^2\} \in \mathbb{R}^{(T/4)\times 2}$, and $\{\hat{\mathbf{y}}_t^3\} \in \mathbb{R}^{(T/16)\times 2}$. Since the lengths of outputs in different levels are different, we down-sample the original ground-truth labels to the corresponding temporal resolution as the ground truth. We calculate the CCC scores for outputs from different levels of our MGCT transformer respectively on the validation set of Recola dataset as shown in Table XII. We can find that CCC becomes higher from the



(a)

(b)

Fig. 11. The prediction curves of valence and arousal of our method on Recola dataset. (a) illustrates the comparison between the predictions from different levels of our MGCT transformer and the ground truth in terms of Valence on the video dev_9.mp4. (b) illustrates the comparison between the predictions from different levels of our MGCT transformer and the ground truth in terms of Arousal on the video dev_7.mp4.

2*nd*-layer group level to the frame level. At the 1*st*-layer group level, the CCC score is comparable to the score of the final output, while CCC keeps increasing in the frame-level since more fine-grained emotion details are learned by aggregating the summarized group-level information.

Effectiveness of cross-clip representation learning. The results on Recola dataset reported in Table XIII demonstrate the effectiveness of our cross-clip representation learning strategy. We train different temporal transformers with features extracted from TDLS with temporal cue guidance. With our strategy, the performance of MGCT is improved by 1.8% for Arousal and 0.6% for Valence.

G. *Visualization results*

Deformable landmark attention module. To verify the effectiveness of our deformable landmark attention module, we visualize different sampling points on the original input image. We randomly choose four images from the AffectNet dataset [62], and show points generated from uniform grids, our proposed sampling strategy and our deformable landmark attention module in Figure 8. We can find that the grid-based attention pattern adopted in Swin Transformer attends varieties of regions that are useless for emotion recognition such as background and hair. Whereas our sampling strategy drastically reduces the number of points falling in the irrelevant areas. Moreover, in the images shown in the second and fourth columns, the subjects' hairs occlude parts of their faces, but the hand-designed attention pattern still attends to these occluded areas. On the contrary, our deformable landmark attention makes parts of the reference points that originally fall in the hair, background and other irrelevant regions shift to the key areas of the face, which demonstrates that the proposed deformable landmark attention can localize key facial regions related to emotion recognition adaptively in a data-dependent manner. We also visualize the outputs from the TDLS transformer to highlight the key facial regions that it focuses on in Figure 9, which shows that our method can attend on key regions that are closely related to human emotions.

Prediction curves. To visually compare the predictions generated by our method with the ground truth, we visualize the curves of predictions from different levels of our MGCT transformer and ground truth curves of random samples on Recola dataset [21] as shown in Figure 11. Figure 11 (a) shows a relatively simple scene, where the emotions are relatively stable for most of the time, and there are only a few sudden changes. Figure 11 (b) shows a scene with many emotional changes, which has higher requirements on the performance of the prediction model. We can find that our prediction curves fit the ground truth well, which shows the effectiveness of our method. From multi-layer group levels to frame level, the prediction results are improved, which is consistent with the results in Table XII.

Effectiveness of solving the flickering problem. To verify the effectiveness of our MGCT in relieving the flickering problem, we visualize the prediction curves of a test sample in Figure 10. Flickering problem exists in baseline Transformer with only frame-level communications. In MGCT, we summarize representations of frames into representations of multi-level groups. Refined group representations by group-level communication further guide learning of frame-level representations through expanding, helping correct the inaccurate frame correlations estimated in frame-to-frame communication and thus alleviating the flickering problem.

V. Discussion

We propose the Temporal cue guided Deformable Landmark Spatial (TDLS) transformer and the Multi-layer Group Constrained Temporal (MGCT) transformer to better exploit and model the spatio and temporal cues for dimensional emotion recognition. We use multi-layer frame groups to model long-term temporal affective dependencies. One deficiency of our model lies in the increase of model complexity as the number of layers increases. More layers have the potential to model contextual cues more effectively, so one future work is to design more computational efficient module for temporal cue learning. Another future work is to utilize data from more modalities besides the visual one, *i.e.* designing multimodal feature extraction and fusion modules for video dimensional emotion recognition. Moreover, foundation models can be exploited for this task.

VI. Conclusion

In this paper, we propose a temporal group constrained transformer with deformable landmark attention for video dimensional emotion recognition. We propose the TDLS transformer with deformable landmark attention which can flexibly attend to emotion-related positions when learning spatial representation. The MGCT transformer is proposed to further refine the extracted features by modeling temporal dependencies using group-level cues. In addition, the temporal cue guided frame representation learning and cross-clip representation learning are introduced for model training. The effectiveness of our approach is demonstrated by state-of-the-art results reported on two benchmark datasets.

References

[1] J. P. Sullins, "Robots, love, and sex: the ethics of building a love machine," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 398–409, 2012. 1

[2] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu, "3D human sensing, action and emotion recognition in robot assisted therapy of children with autism," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2158–2167. 1

[3] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 23–30. 1

[4] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2339–2352, 2018. 1

[5] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977. 1

[6] S. Hamann, "Mapping discrete and dimensional emotions onto the brain: controversies and consensus," *Trends in Cognitive Sciences*, vol. 16, no. 9, pp. 458–466, 2012. 1

[7] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell, "On the importance of both dimensional and discrete models of emotion," *Behavioral Sciences*, vol. 7, no. 4, p. 66, 2017. 1

[8] W. Zhou, J. Lu, C. Ling, W. Wang, and S. Liu, "Enhancing emotion recognition with pre-trained masked autoencoders and sequential learning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 4666–4672. 1

[9] D. Bose, V. Sethu, and E. Ambikairajah, "Continuous emotion ambiguity prediction: Modeling with beta distributions," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1684–1695, 2024. 1, 3, 6, 8

[10] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007. 1, 2

[11] Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang, "Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 565–578, 2019. 1, 2, 3, 6, 7, 8

[12] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2020. 1, 2, 5, 6, 8

[13] R. G. Praveen, P. Cardinal, and E. Granger, "Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 3, pp. 360–373, 2023. 1, 6, 8

[14] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2010. 1, 2

[15] J. Lee, S. Kim, S. Kiim, and K. Sohn, "Spatiotemporal attention based deep neural networks for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1513–1517. 1, 2, 6, 7, 8

[16] M. Hu, Q. Chu, X. Wang, L. He, and F. Ren, "A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video," *IEEE Signal Processing Letters*, vol. 28, pp. 698–702, 2021. 1, 2, 3, 6, 8

[17] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3D ConvLSTM networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 6837–6841. 1, 7, 8

[18] Z. Zhao and Q. Liu, "Former-DFER: Dynamic facial expression recognition transformer," in *ACM International Conference on Multimedia*, 2021, pp. 1553–1561. 1

[19] Q. Huang, C. Huang, X. Wang, and F. Jiang, "Facial expression recognition with grid-wise attention and visual transformer," *Information Sciences*, vol. 580, pp. 35–54, 2021. 1

[20] S. Wu, Z. Du, W. Li, D. Huang, and Y. Wang, "Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze," in *International Conference on Multimodal Interaction*, 2019, pp. 40–48. 1, 2, 6

[21] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8. 1, 2, 6, 9, 12

[22] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3601–3610. 1, 2

[23] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, 2021. 1, 7, 8

[24] C. Li, L. Xie, X. Wang, H. Pan, and Z. Wang, "A twin disentanglement transformer network with hierarchical-level feature reconstruction for robust multimodal emotion recognition," *Expert Systems with Applications*, vol. 264, p. 125822, 2025. 2

[25] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller *et al.*, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019. 2, 6

[26] 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, 2018. 2

[27] 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. 2

[28] M. K. Tellamekala, Ö. Sümer, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "Are 3D face shapes expressive enough for recognising continuous emotions and action unit intensities?" *IEEE Transactions on Affective Computing*, no. 2, pp. 535–548, 2023. 2, 7, 8

[29] H. Liu, C. Zhang, Y. Deng, T. Liu, Z. Zhang, and Y.-F. Li, "Orientation cues-aware facial relationship representation for head pose estimation via transformer," *IEEE Transactions on Image Processing*, vol. 32, pp. 6289–6302, 2023. 3

[30] 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*, vol. 14, no. 4, pp. 3244–3256, 2022. 3

[31] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 65–72. 3

[32] 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*, 2020, pp. 27–34. 3

[33] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," *arXiv preprint arXiv:1708.07050*, 2017. 3

[34] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 4898–4906. 3

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229. 3

[36] 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. 3, 4, 8

[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021, pp. 10 347–10 357. 3

[38] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1944–1953. 3

[39] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "MViTv2: Improved multiscale vision transformers for classification and detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814. 3

[40] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *International Conference on Machine Learning*, 2021, p. 4. 3, 7, 8

[41] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020. 3

[42] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3163–3172. 3

[43] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211. 3

[44] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846. 3, 7, 8

[45] J. Sun, H. H. Dodge, and M. H. Mahoor, "Mc-vivit: Multi-branch classifier-vivit to detect mild cognitive impairment in older adults using facial videos," *Expert systems with applications*, vol. 238, p. 121929, 2024. 3

[46] Z. Cheng, Z.-Q. Cheng, J.-Y. He, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 110 805–110 853, 2024. 3

[47] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 543–553. 3

[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022. 4, 6

[49] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4794–4803. 4

[50] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738. 4

[51] M. Hu, H. Wang, X. Wang, J. Yang, and R. Wang, "Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 176–185, 2019. 4

[52] M. Jegorova, S. Petridis, and M. Pantic, "SS-VAERR: Self-supervised apparent emotional reaction recognition from video," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–8. 6, 7, 8

[53] M. Tran, Y. Kim, C.-C. Su, C.-H. Kuo, and M. Soleymani, "SAAML: A framework for semi-supervised affective adaptation via metric learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6004–6015. 6, 8

[54] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective Processes: stochastic modelling of temporal context for emotion and facial expression recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9074–9084. 7, 8

[55] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, "Modelling stochastic context of audio-visual expressive behaviour with affective processes," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2290–2303, 2023. 7, 8

[56] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceeding of 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 3–8. 6

[57] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and chal-

lenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10. 6

[58] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3–9. 6

[59] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989. 6

[60] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 59–66. 6

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. 7

[62] 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. 7, 10, 11

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 7

[64] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283. 8

[65] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 8

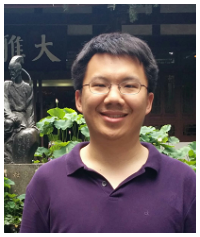
[66] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018. 8

[67] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, 2015. 8

Weixin Li received the PhD degree in computer science from the University of California, Los Angeles (UCLA), Los Angeles, California, in 2017. She is currently an associate professot with the School of Computer Science and Engineering (SCSE), Beihang University, Beijing, China. Her research interests include computer vision, pattern recognition, and multimodal learning.

Xiangjing Meng received the Bachelor and Master degrees in computer science and technology from the School of Computer Science and Engineering (SCSE), Beihang University, Beijing, China. His research interests include computer vision, and affective computing.

Linmei Hu works as an associate professor in the School of Computer Science and Technology, Beijing Institute of Technology. She received her Ph.D degree in Tsinghua University in 2018. Her research interests include Knowledge Graph, Natural Language Processing and Multimodal Content Analysis.



Xuan Dong received the BE degree in computer science from Beihang University, Beijing, China, in 2010, and the PhD degree in computer science from Tsinghua University, Beijing, China, in 2015. He is currently an associate professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China. His research interests include computer vision and computational photography.