Former-DFER: Dynamic Facial Expression Recognition Transformer

Zengqun Zhao[1,3], Qingshan Liu[1,2*]
{zqzhao,qsliu}@nuist.edu.cn
[1] Engineering Research Center of Digital Forensics, Ministry of Education
[2] School of Computer and Software, [3] School of Automation,
Nanjing University of Information Science & Technology, Nanjing, China

ABSTRACT

This paper proposes a dynamic facial expression recognition transformer (Former-DFER) for the in-the-wild scenario. Specifically, the proposed Former-DFER mainly consists of a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former). The CS-Former consists of five convolution blocks and N spatial encoders, which is designed to guide the network to learn occlusion- and pose-robust facial features from the spatial perspective. And the temporal transformer consists of M temporal encoders, which is designed to allow the network to learn contextual facial features from the temporal perspective. The heatmaps of the leaned facial features demonstrate that the proposed Former-DFER is capable of handling the issues such as occlusion, non-frontal pose, and head motion. And the visualization of the feature distribution shows that the proposed method can learn more discriminative facial features. Moreover, our Former-DFER also achieves state-of-the-art results on the DFEW and AFEW benchmarks. Code is available at https://github.com/zengqunzhao/Former-DFER.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Biometrics.*

KEYWORDS

Dynamic facial expression, in-the-wild facial expression recognition, spatio-temporal transformer, deep learning

ACM Reference Format:
Zengqun Zhao[1,3], Qingshan Liu[1,2]. 2021. Former-DFER: Dynamic Facial Expression Recognition Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475292

1 INTRODUCTION

Facial expression is one of the most crucial signals for human beings to convey their emotions, which plays a significant role in communications [13]. In recent years, automatic recognition of facial expression has become a hot topic for researchers because of its applications in various fields, such as the human-computer interaction (HCI) [8, 19], medical diagnosis [29, 39], driver assistance [69], and so on [2, 27]. Facial expression recognition (FER) aims to classify an image or video sequence into one of several basic emotions, i.e., neutral, happiness, sadness, surprise, fear, disgust, and anger. According to the data type, the FER can be divided into static FER (SFER) and dynamic FER (DFER), in which the SFER takes the still image as an input and the DFER takes the video sequence as an input. Previous work [44, 71, 72] has indicated that the natural facial event is dynamic and facial expression can be better described as the sequential variation in a dynamic process. Therefore, the DFER has received more and more attention in recent years.

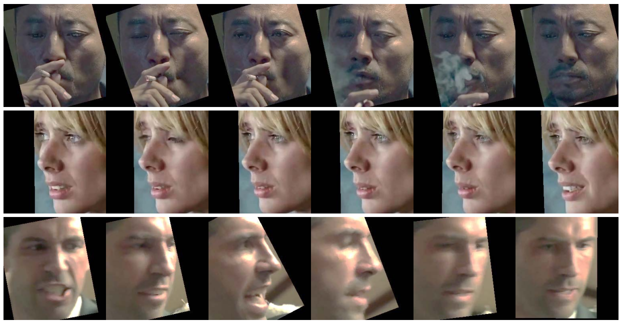

Figure 1: Three kinds of challenges, including occlusion (top), non-frontal pose (middle), and head motion (bottom), to DFER under the in-the-wild scenario. The sequences are sampled from the DFEW dataset.

According to the data scenario, the DFER can be also divided into lab-controlled and in-the-wild. Towards lab-controlled DFER, the datasets, such as CK+ [48], Oulu-CASIA [76], and MMI [53] are all recorded under laboratory conditions, in which all the facial images are frontal and without any occlusion. Over the past decade, there are many methods proposed for DFER under lab-controlled datasets, and these methods have achieved outstanding performance [65, 67, 73]. Towards in-the-wild DFER, as shown in Fig. 1, the video sequences come from the real-world scenario, which exists many challenges such as occlusion, non-frontal pose, and head motion. Due to the scenario gap between laboratory and real-world, the DFER models, designed based on the lab-collected datasets, cannot deal well with human expressions recognition under natural and uncontrolled conditions [42, 77, 78]. Moreover, with the collection of the large-scale facial expression datasets in the wild, such as

[*] Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475292

AFEW [17], Aff-Wild [74], CAER [36], and DFEW [30], the research focus of the DFER has been transferred from laboratory-controlled to challenging in-the-wild conditions [30].

For DFER in the wild, the early work is mainly proposed based on the hand-crafted features, such as the LBP-TOP [16], STLMBP [28], and HOG-TOP [9]. Instead of hand-crafted descriptors, a novel spatio-temporal manifold (STM) method [44] is proposed for modeling each video clip, and Liu *et al.* [45] also combine multiple kernel methods on the Riemannian manifold for DFER in the wild. With the prevalence of deep learning and collection of large-scale datasets, deep-learning-based methods have been becoming the dominant strategy. These methods are proposed based on the convolutional neural network (CNN) [1, 21, 49], recurrent neural network (RNN) [20, 38, 47, 52, 66], and 3D CNN [22, 30, 36, 64]. Moreover, ensemble learning and multi-modal learning are also prevalent in DFER in the wild [5, 10, 34, 37, 41, 50, 51, 75]. In the latest work, Baddar *et al.* [4] propose a mode variational LSTM, Liu *et al.* [43] introduce the Graph Convolutional Networks (GCN) into in-the-wild DFER, Kumar *et al.* [35] adopt the noisy student training method, and Jiang *et al.* [30] propose a novel EC-STFL Loss for DFER in the wild.

Although many methods have been proposed for in-the-wild DFER, the performance of these models are still weak, hindering the application of the in-the-wild DFER models for real world [40, 46]. Recently, the success of the transformer [63] models in natural language processing (NLP) has inspired researchers to employ the Transformer Encoder in computer vision tasks [32]. As a result, Transformer models have been successfully used for image recognition [18, 59], object detection [6, 81], image super-resolution [70], and video understanding [23, 58]. Inspired by it, we propose a dynamic facial expression recognition transformer (Former-DFER), the structure of the proposed method is shown in Fig. 2. From the spatial perspective, the facial patches split from a whole facial image can be viewed as a sequence of the visual words. From the temporal perspective, the facial clip is sequential, and each frame of the clip can be also viewed as a visual word. Furthermore, the self-attention mechanism in Transformer can learn correlation among regional facial features and correlation among temporal facial features, possessing a natural ability in addressing the occlusion, non-frontal pose, and head motion problems for DFER in the wild.

As shown in Fig. 2, the proposed Former-DFER mainly consists of two parts: a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former). The CS-Former consists of five convolution blocks and N spatial encoders, which is designed to guide the network to learn occlusion- and pose-robust facial features from the spatial perspective. And the T-Former consists of M temporal encoders, which is designed to allow the network to learn contextual facial features from the temporal perspective. The visualizations show that the proposed Former-DFER has the ability to handle the challenging issues and can encode more discriminative facial representation for recognition. Moreover, the quantitative results indicate that our Former-DFER has achieved state-of-the-art performance on two popular benchmarks. The contributions of our work can be summarized as follows:

- We propose a dynamic facial expression recognition transformer for the in-the-wild scenario. A convolutional spatial transformer and a temporal transformer are designed to guide the network extracting of robust facial features both spatial and temporal.
- To the best of our knowledge, we are the first to apply Transformers for dynamic facial expression recognition. The self-attention enables the network to be capable of learning contextual information.
- The heatmaps of the learned facial features indicate that the proposed method can handle the issues such as occlusion, non-frontal pose, and head motion. And the visualization of the feature distribution shows that the proposed method can learn more discriminative facial features. Moreover, the proposed method also achieves state-of-the-art results on two popular benchmarks, and the code is publicly available.

2 RELATED WORK

2.1 DFER in the Wild

In recent years, deep learning has been successfully adopted to recognize dynamic emotions in the wild, and these methods have achieved superior performance to methods of using hand-crafted features [12]. So we mainly review the deep-learning-based method.

For the CNN-RNN-based methods, the spatial facial features of each video frame are first learned by a CNN, and then all the frames' spatial features are processed by an RNN to learn temporal information among all frames. Under this paradigm, many methods [20, 38, 47, 52, 66] adopt the VGG [57] or ResNet [25] to extract spatial features and Long Short-Term Memory (LSTM) [26] or Gated Recurrent Unit (GRU) [11] to extract temporal features. In recent CNN-RNN-based work, Baddar *et al.* [4] propose a mode variational LSTM to encode spatio-temporal features robust to unseen modes of variation. Liu *et al.* [43] introduce a GCN layer into the common CNN-RNN-based model for video-based FER. Moreover, a multi-modal recurrent attention network (MRAN) [37] is also proposed to learn spatio-temporal attention maps for robust DFER in the wild.

For the 3D-CNN-based methods, the spatial and temporal feature representation of video sequences is extracted jointly through the 3D convolution. The early work [22, 34, 47, 64, 75] extracts spatial-temporal facial features by adopting a 3D-CNN directly, and such a spatial-temporal feature often combines with other kinds of facial features as a final representation. Recently, Lee *et al.* [38] propose a CAER-Net to exploit not only human facial expression but also context information in a joint and boosting manner. To reduce the overheads of the 3D networks in DFER, Kossaifi *et al.* [33] propose a CP-Higher-Order Convolution that allowing to train of a network on the images and using transduction to generalize to videos. The last work proposes a EC-STFL [30] for DFER, which can enforce the spatio-temporal deep neural networks to better learn discriminative features describing dynamic facial expressions in the wild.

2.2 Transformer

The Transformer is proposed by Vaswani *et al.* [63] for machine translation and has achieved the state of the art performance in many NLP tasks. The prevalence of the transformer networks in the NLP domain has sparked great interest in the computer vision community to adapt these models for vision learning tasks [32]. Regarding the image classification, Dosovitskiy *et al.* [18] propose a Vision Transformer (ViT) to treat an image as a sequence of patches

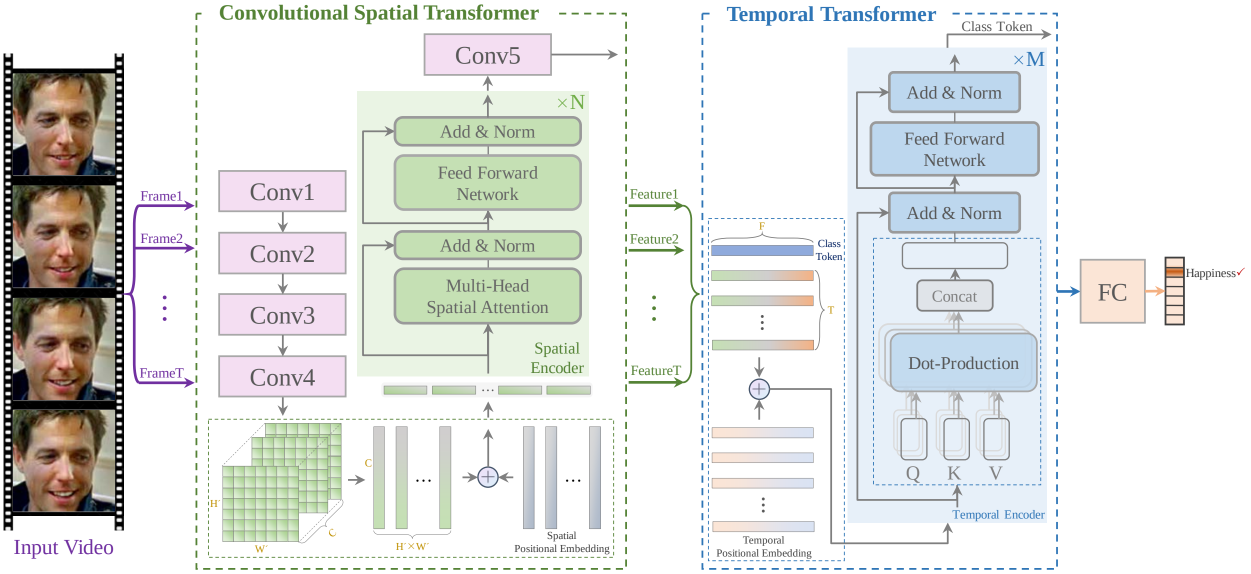

Figure 2: The structure of the proposed method.

and process it by a standard Transformer encoder as used in NLP, which indicates that Transformer has the ability in handling the vision classification task. Regarding object detection, the new models named DETR [6] and deformable DETR [81] are proposed to reason about the relations of the objects and the global image context to directly output the final set of predictions in parallel, which show a superior performance to other modern detectors. Regarding the image super-resolution, Yang *et al.* [70] propose a novel texture Transformer network for image super-Resolution, in which the low-resolution image and high-resolution images are formulated as queries and keys in a transformer. Moreover, the Transformer-based methods are also proposed for video understanding [23, 58].

3 THE FORMER-DFER MODEL

3.1 Overview

As shown in Fig. 2, the proposed Former-DFER mainly consists of a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former). For our Former-DFER, the fixed-length facial expression sequence is dynamically sampled from the raw video as an input. The CS-Former takes each frame as an input to extract spatial facial features, and then the T-Former takes all the frame's spatial features as an input to generate a discriminative feature representation. Finally, the recognition result is obtained by a full-connected (FC) network.

3.2 CS-Former

The convolutional spatial transformer (CS-Former) consists of five convolution blocks and N spatial encoders, and each spatial encoder consists of a multi-headed self-attention and feed forward network. The structure is shown in Fig. 2.

Input Clips: The CS-Former takes a clip $X \in \mathbb{R}^{T \times 3 \times H \times W}$ consisting of T RGB frames of the size of $H \times W$ as an input, and the clip is dynamically sampled from the original video. Specifically, we first split all frames into U segments regarding the training video samples and then select V frames in each segment randomly. Regarding the test video samples, we first split all frames into U segments and then selecting V frames in the mid of each segment. Hence, the length of the sampled clip is $T = U \times V$.

Convolutional Embedding: For each frame, we first utilize four convolution blocks to extract feature maps $M \in \mathbb{R}^{C \times H' \times W'}$. The acquired feature maps need to be flattened into a 1D sequence, and further can be fed for the S-Former (consists of N spatial encoders). Therefore, we then reshape $M \in \mathbb{R}^{C \times H' \times W'}$ into a flatted sequence $M^f \in \mathbb{R}^{Q \times C}$ ($Q = H'W'$). As a result, we can obtain Q visual word embeddings with the length of C, and each visual word embedding represents the features of the original facial patch with the size of $H/H' \times W/W'$. Then the input embedding to S-Former can be computed as:

$$\mathbf{z}_p^0 = \mathbf{m}_p^f + \mathbf{e}_p \tag{1}$$

where $\mathbf{e}_p \in \mathbb{R}^C$ represents a learnable position embedding added to encode the spatial position, $p \in \{1, 2, \cdots, Q\}$.

Query-Key-Value Computation: Our S-Former consists of N spatial encoders. At each encoder l, the query/key/value vector is computed for each visual world from the representation $\mathbf{z}_p^{l-1}$ encoded by the preceding block, which can be computed as:

$$\begin{aligned} \mathbf{q}_p^{(l,k)} &= W_Q^{(l,k)} LN(\mathbf{z}_p^{l-1}) \in \mathbb{R}^{C'} \\ \mathbf{k}_p^{(l,k)} &= W_K^{(l,k)} LN(\mathbf{z}_p^{l-1}) \in \mathbb{R}^{C'} \\ \mathbf{v}_p^{(l,k)} &= W_V^{(l,k)} LN(\mathbf{z}_p^{l-1}) \in \mathbb{R}^{C'} \end{aligned} \tag{2}$$

where $LN(\cdot)$ denotes layer normalization [3], $k \in \{1, \cdots, K\}$ represents the index of the multiple attention heads and K denotes the total number of attention heads, $C' = C/K$ denotes the latent dimensionality of each attention head.

Self-attention Computation: The self-attention weights $\boldsymbol{\lambda}_p^{(l,k)} \in \mathbb{R}^Q$ for each query p are computed by dot-production, which can be formulated as:

$$\boldsymbol{\lambda}_p^{(l,k)} = SM(\frac{\mathbf{q}_p^{(l,k)\top}}{\sqrt{C'}} \cdot \{\mathbf{k}_{p'}^{(l,k)}\}_{p'=1,\cdots,Q}) \tag{3}$$

where SM denotes the softmax activation function.

Encoding: To compute the encoding $\mathbf{z}_p^{(l,k)}$ at block l, we can first compute the weighted sum of value vectors by using self-attention coefficients from each attention head, which can be formulated as:

$$\mathbf{s}_p^{(l,k)} = \sum_{p'=1}^{Q} \boldsymbol{\lambda}_{p,p'}^{(l,k)} \mathbf{v}_{p'}^{(l,k)} \tag{4}$$

Then, the concatenation of the vectors from all attention heads is projected and passed through an MLP, in which the residual connections are employed. The operation can be formulated as:

$$\mathbf{z'}_p^l = W_O \begin{bmatrix} \mathbf{s}_p^{(l,1)} \\ \vdots \\ \mathbf{s}_p^{(l,K)} \end{bmatrix} + \mathbf{z}_p^{l-1} \tag{5}$$

$$\mathbf{z}_p^l = MLP(LN(\mathbf{z'}_p^l)) + \mathbf{z'}_p^l \tag{6}$$

Finally, the Q encodings $\mathbf{z}_p^N$ are concatenated at the spatial level to generate refined feature maps $M_r \in \mathbb{R}^{C \times H' \times W'}$. And each frame's feature embedding $\mathbf{x}'_t \in \mathbb{R}^F$ can be computed as:

$$\mathbf{x}'_t = GAP(g(M_r)) \tag{7}$$

where $g(\cdot)$ denotes convolution block, GAP denotes global average pooling, $t \in \{1, 2, \cdots, T\}$.

Because all the frames are sharing one CS-Former, given an input clip $X \in \mathbb{R}^{T \times 3 \times H \times W}$, the output $X' \in \mathbb{R}^{T \times F}$ can be obtained through a CS-Former. Regarding the proposed ST-Former, the first four convolution blocks are employed to learn facial features within a kernel which can be treated as a local operation, and the S-Former is employed to learn the correlation among all the kernels which can be treated as a global operation. The final convolution block is employed to refine the facial features.

3.3 T-Former

The temporal transformer (T-Former) consists of M temporal encoders, and each temporal encoder also consists of a multi-headed self-attention and feed forward network. The structure is shown in Fig. 2.

Embedding: Given an input $X' \in \mathbb{R}^{T \times F}$, the T spatial feature vectors can be obtained. Then, the input embeddings for T-Former can be computed as:

$$\mathbf{z}_{t'}^0 = \mathbf{x}'_{t'} + \mathbf{e}_{t'} \tag{8}$$

where $\mathbf{e}_{t'} \in \mathbb{R}^F$ represents a learnable position embedding added to encode the temporal position, $t' \in \{0, 1, \cdots, T\}$. Different from the S-Former, we add a special learnable vector $\mathbf{x}'_0 \in \mathbb{R}^F$ in the first position of the sequence to represent the embedding of the classification token (class token).

Query-Key-Value Computation: For T-Former, the query $\mathbf{q}_{t'}^{(l,k)}$, key $\mathbf{k}_{t'}^{(l,k)}$, and value $\mathbf{v}_{t'}^{(l,k)} \in \mathbb{R}^{F'}$ at each layer l can be computed using Equ. 2, where $F' = F/K$.

Self-attention Computation: The self-attention weight $\boldsymbol{\mu}_{t'}^{(l,k)} \in \mathbb{R}^T$ for each query t' can be computed as:

$$\boldsymbol{\mu}_{t'}^{(l,k)} = SM(\frac{\mathbf{q}_{t'}^{(l,k)\top}}{\sqrt{F'}} \cdot [\mathbf{k}_0^{(l,k)} \{\mathbf{k}_t^{(l,k)}\}_{t=1,\cdots,T}]) \tag{9}$$

Encoding: The encoding $\mathbf{z}_{t'}^{(l,k)}$ at block l can be computed as following equation:

$$\mathbf{s}_{t'}^{(l,k)} = \boldsymbol{\mu}_{t',0}^{(l,k)} \mathbf{v}_0^{(l,k)} + \sum_{t=1}^{T} \boldsymbol{\mu}_{t',t}^{(l,k)} \mathbf{v}_t^{(l,k)} \tag{10}$$

$$\mathbf{z'}_{t'}^l = W_I \begin{bmatrix} \mathbf{s}_{t'}^{(l,1)} \\ \vdots \\ \mathbf{s}_{t'}^{(l,K)} \end{bmatrix} + \mathbf{z}_{t'}^{l-1} \tag{11}$$

$$\mathbf{z}_{t'}^l = MLP(LN(\mathbf{z'}_{t'}^l)) + \mathbf{z'}_{t'}^l \tag{12}$$

Classification Embedding: The final clip embedding is obtained from the class token of the T-Former's final layer, and the final recognition results can be computed as:

$$\mathbf{y} = FC(\mathbf{z}_0^M) \in \mathbb{R}^J \tag{13}$$

where FC denotes full-connected network, J denotes the classes of the facial expression.

4 EXPERIMENTS

4.1 Datasets

DFEW: The DFEW [30] dataset is proposed recently, which is the current largest benchmark for DFER in the wild. The video clips are collected from over 1,500 movies worldwide, covering various challenging interferences, such as extreme illuminations, occlusions, and variant head pose. Moreover, each video in DFEW has been individually labeled ten times by the annotators under professional guidance and assigned to one of seven basic expressions, i.e., happiness, sadness, neutral, anger, surprise, disgust, and fear. The DFEW dataset includes 12,059 video clips, and all the samples have been split into five same-size parts without overlap. The 5-fold cross-validation is adopted as an evaluation protocol; in each fold (fd1~fd5), one part of the samples are used for testing, and the remaining for training. Finally, all the predicted labels are used to compute the evaluation metrics by comparing them with the ground truth.

AFEW: The AFEW [17] dataset served as an evaluation platform for the annual EmotiW from 2013 to 2019. AFEW contains video clips collected from different movies and TV serials with spontaneous expressions, illuminations, various head poses, and occlusions. Same as the DFEW, each video clip in AFEW is assigned to one of seven basic expressions. The AFEW dataset includes 1,809 video clips, and all the samples have been split into three splits: train (773 video clips), validation (383 video clips), and test (653 video clips). Since the test split is not publicly available, we train our model on train split and report results on validation split.

Method	Metrics (%)		Comlexity
	UAR	WAR	(GFLOPs)
ResNet18 + GRU (Baseline)	51.68	64.02	7.78
ResNet34 + GRU	52.05	63.96	15.47
ResNet18 + S-Former + GRU	52.66	64.97	9.01
ResNet18 + T-Former	52.86	64.86	7.88
ResNet18 + BAM + T-Former	53.48	64.92	7.89
ResNet18 + S-Former + T-Former	**53.69**	**65.70**	9.11

Table 1: Evaluation of each component in Former-DFER.

Method	Metrics (%)		Comlexity
	UAR	WAR	(GFLOPs)
ResNet18 + S-Former(w/o PE) + T-Former	53.30	65.04	9.11
ResNet18 + S-Former + T-Former(w/o PE)	52.07	64.47	9.11
ResNet18+ S-Former(conv5) + T-Former	53.05	65.41	9.59
ResNet18 + S-Former + T-Former(mean)	52.98	65.50	9.11
ResNet18 + S-Former + T-Former	**53.69**	**65.70**	9.11

Table 2: Evaluation of different settings for Former-DFER. PE denotes positional embedding.

Setting		Metrics		Complexity
N	M	UAR	WAR	(GLOPs)
1	1	53.27	65.33	8.22
3	1	53.28	65.60	9.04
1	3	53.62	65.44	8.29
3	3	**53.69**	**65.70**	9.11
6	3	53.26	65.06	10.35
3	6	53.23	65.22	9.22
6	6	53.60	65.44	10.45

Table 3: Evaluation of different volumes of layers.

4.2 Implementation Details

Data Pre-processing: For the AFEW dataset, the face region of the video frame is detected using RetinaFace [14], and then the face region is cropped and aligned according to the bounding box and landmarks. It should be noted that the RetinaFace is robust to occlusion and non-front pose. For the DFEW dataset, the video frame’s face region is publicly available, so we use the processed data directly. All the faces are resized to 112 × 112 pixels as an input. Due to the low-light issues that existed in AFEW, a pre-trained deep learning model, i.e., Enlighten-GAN [31], is used to enhance the light.

Training Setting: Our models are trained on one GeForce RTX 2080 Ti GPU based on the open-source PyTorch [55] platform, and parameters were optimized via the SGD optimizer. For the DFEW dataset, consistent with the EC-STFL [30], we train our model from scratch with a batch size of 32, initializing the learning rate as 0.01 and dividing it by ten every 40 epochs. The training operation is stopped in the 100th epoch. For the AFEW datasets, due to the quantity of the data samples is tiny, previous work pre-trained models on different datasets (both static and dynamic). To make a fair comparison, we first pre-train our model and other models on DFEW (fd1) and then fine-tuning on AFEW with the same setting (a batch size of 32 and a learning rate of 0.001, the training operation is stopped in the 20th epoch). In our method, the $U = 8$, $V = 2$. Hence, the length of the dynamically sampled sequence is 16. And the number of the self-attention heads $K = 8$.

Validation Metrics: Consistent with the EC-STFL [30], we choose the unweighted average recall (UAR, i.e., the accuracy per class divided by the number of classes without considerations of instances per class) and weighted average recall (WAR, i.e., accuracy) as the metrics. We hope to improve models’ performance both in UAR and WAR metrics.

4.3 Ablation Analysis

To validate the effectiveness of our Former-DFER, we conduct an ablation analysis on the DFEW [30] benchmark (5-fold cross-validation). In our experiments, each component’s effectiveness in Former-DFER, the setting of the S-Former and the T-former, and the number of the layers are studied, respectively. The ResNet18-GRU is employed as a baseline in our experiments.

Evaluation of Each Component: We first study the effectiveness of each component in our Former-DFER. The experimental results are shown in Tab. 1. The results indicate that the use of the S-Former can improve the UAR and WAR by 0.98% and 0.95%, respectively. And the use of the T-Former can improve the UAR and WAR by 1.18% and 0.84%, respectively. Moreover, if the S-Former and T-Former are both employed, the UAR and WAR will be enhanced by 2.01% and 1.68%, respectively. Due to our Former-DFER’s complexity is large than the baseline network, we adopt a deeper baseline work to compare with our model. The result shows that the performance of the deeper baseline network is still inferior to our model. Furthermore, we also replace our S-Former with a spatial-temporal attention module named BAM [54], and the results indicate that our S-Former is superior to BAM. The experimental results demonstrate that each part of the Former-DFER is practical, and guiding the network to learn contextual information is crucial for the DFER task. Such contextual information can effectively suppress the interference of the occlusion, non-frontal pose, and head motion issues.

Evaluation of Setting: We then study the performance of the different settings for our Former-DFER. The experimental results are shown in Tab. 2. In our method, the learnable position embeddings are adopted to encode spatial and temporal positions, and the results show that using the position embeddings is crucial to S-Former and T-Former. We also place the S-Former after the last convolutional block, namely extracting facial features by five convolutional blocks and obtaining final spatial facial features after S-Former, but the result is inferior. Moreover, for T-Former, we conduct experiments to obtain the final clip embedding by averaging all tokens’ outputs instead of the class token, however, it will entail a decrease both in UAR and WAR.

Evaluation of Depth: Finally, we study the effect of the different number of layers for our Former-DFER. The experimental results are shown in Tab. 3. The default number of layers for S-Former and T-Former both are 3 in our method. To study the effect of the layer’s number for our Former-DFER, we compare the shallower model

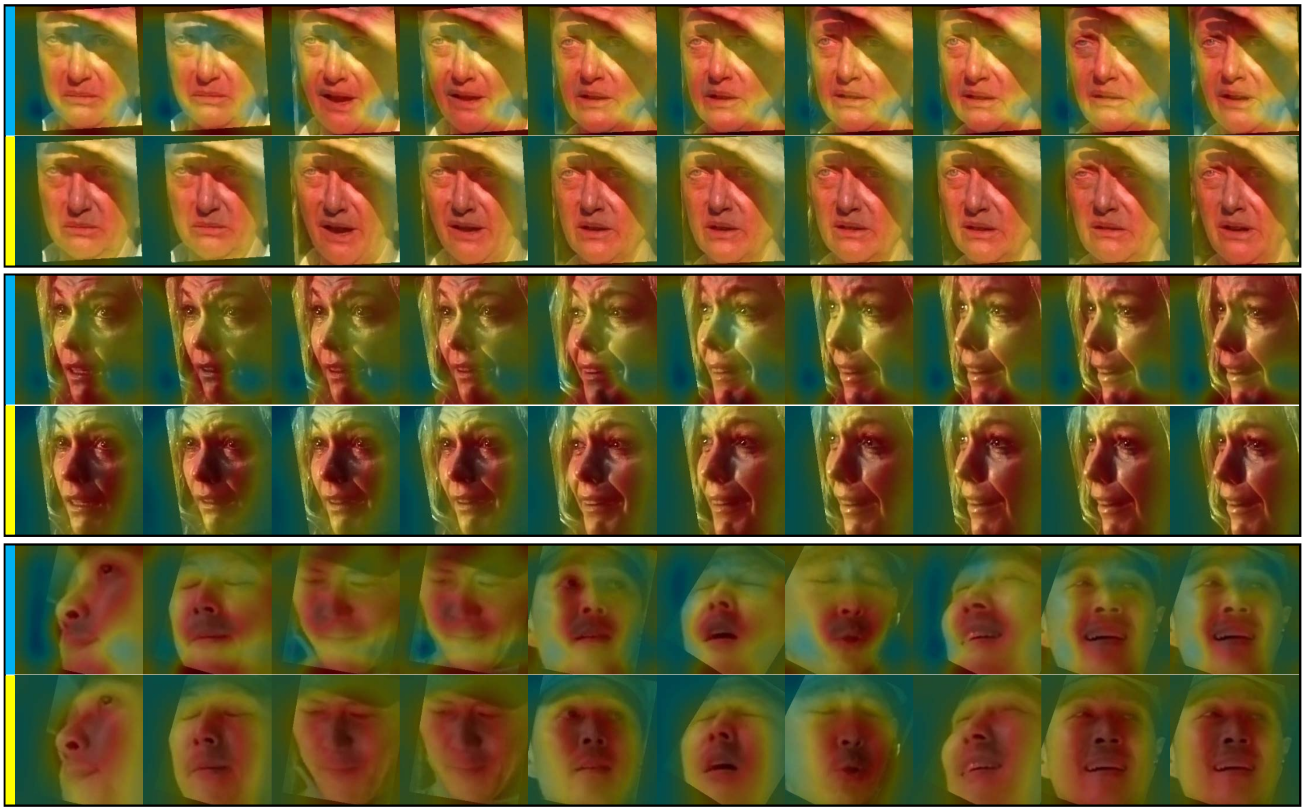

Figure 3: Visualization of the learned feature maps. There are three sequences are presented, which including the issues of the occlusion, non-frontal pose, and head motion, respectively. For each sequence, the images in the first row are heatmaps generated by the baseline, and the images in the second row are heatmaps generated by our Former-DFER.

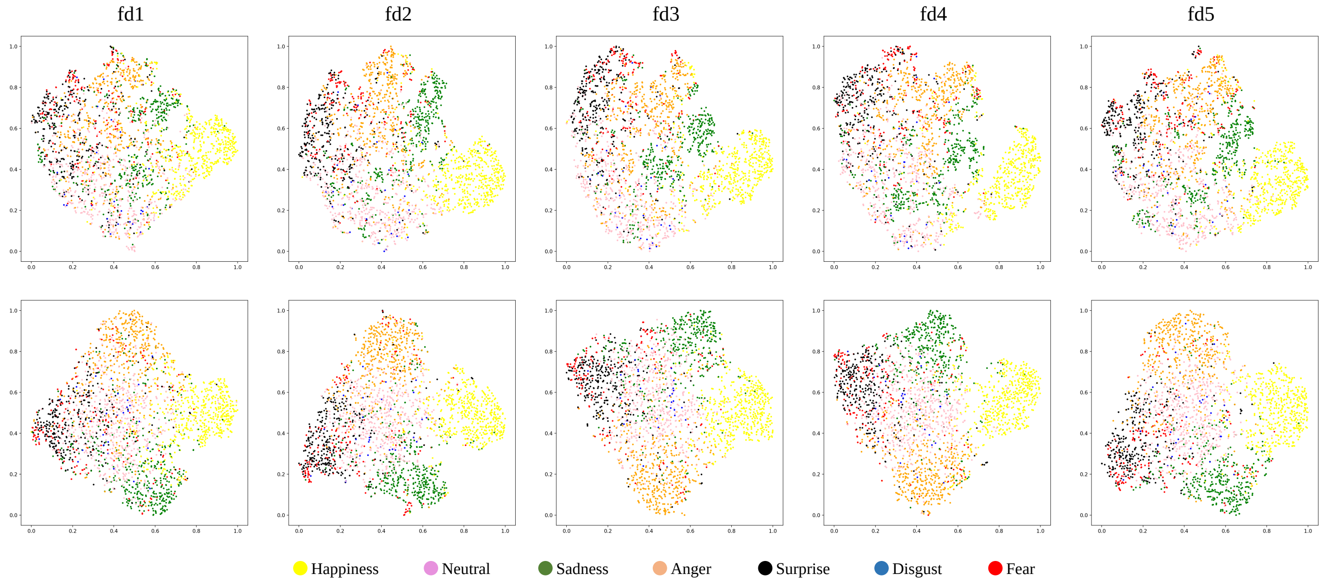

Figure 4: Illustration of feature distribution learned by the baseline (top) and our proposed Former-DFER (bottom) on fd1~fd5.

Method	Sample	Accuracy of Each Emotion (%)							Metrics (%)		Comlexity
	Strategies	Happiness	Sadness	Neutral	Anger	Surprise	Disgust	Fear	UAR	WAR	(GFLOPs)
C3D [60]	TI	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54	38.57
P3D [56]	TI	74.85	43.40	54.18	60.42	50.99	0.69	23.28	43.97	54.47	n/a
R(2+1)D18 [61]	TI	79.67	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22	42.36
3D Resnet18 [24]	TI	73.13	48.26	50.51	64.75	50.10	0.00	26.39	44.73	54.98	8.32
I3D-RGB [7]	TI	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.40	54.27	6.99
VGG11+LSTM [26, 57]	TI	76.89	37.65	58.04	60.70	43.70	0.00	19.73	42.39	53.70	31.65
ResNet18+LSTM [25, 26]	TI	78.00	40.65	53.77	56.83	45.00	**4.14**	21.62	42.86	53.08	7.78
3D R.18+Center Loss [24, 68]	TI	78.49	44.30	54.89	58.40	52.35	0.69	25.28	44.91	55.48	8.32
EC-STFL [30]	TI	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51	8.32
3D Resnet18 [24]	DS	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27	8.32
ResNet18+LSTM [25, 26]	DS	83.56	61.56	**68.27**	65.29	51.26	0.00	29.34	51.32	63.85	7.78
Resnet18+GRU [11, 25]	DS	82.87	**63.83**	65.06	68.51	52.00	0.86	30.14	51.68	64.02	7.78
Former-DFER (Ours)	DS	**84.05**	62.57	67.52	**70.03**	**56.43**	3.45	**31.78**	**53.69**	**65.70**	9.11

Table 4: Comparison with state-of-the-art methods on DFEW. Bold denotes the best. Underline denotes the second best. TI denotes time interpolation [79, 80]. DS denotes dynamic sampling.

Methods	Sample	Metrics (%)		Comlexity
	Strategies	UAR	WAR	(GFLOPs)
EmotiW-2019 Baseline [15]	n/a	n/a	38.81	n/a
C3D [60]	DS	43.75	46.72	38.57
I3D-RGB [7]	DS	41.86	45.41	6.99
R(2+1)D [61]	DS	42.89	46.19	42.36
3D ResNet18 [24]	DS	42.14	45.67	8.32
ResNet18+LSTM [25, 26]	DS	43.96	48.82	7.78
ResNet18+GRU [11, 25]	DS	45.12	49.34	7.78
Former-DFER (Ours)	DS	**47.42**	**50.92**	9.11

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

and deeper model with our default setting. The experimental results indicate that the performance of the shallower model is weak due to its limited parameters. Moreover, for in-the-wild DFER tasks, due to the limited training data, the deeper network always entails overfitting. Therefore, increasing the model layer will not improve the performance and even though decrease the performance.

4.4 Visualizations

To prove the robustness of the proposed Former-DFER under challenging conditions, we conduct an experiment to visualize the learned facial feature maps shown in Fig. 3, in which occlusion, non-frontal pose, and head motion are all included. For the first facial sequence, we can notice that our method can ignore the occlusion region and pay attention to the non-occlusion part. And for the second sequence, the learned features contain more rich facial information related to emotion even though the non-frontal pose. Moreover, for the large head movement among frames, our method is robust to the interference and can focus on the crucial part.

Furthermore, we also utilize t-SNE [62] to illustrate the feature distribution learned by the baseline and our Former-DFER. The comparison shown in Fig. 4 demonstrates that the proposed Former-DFER can better gather the samples of the same category, proving that our method can learn more discriminative features for in-the-wild DFER.

4.5 Comparison with State-of-the-Arts

In this section, we compare our best results with several state-of-the-art methods on the DFEW and AFEW benchmarks.

Comparison on DFEW: For the DFEW dataset, consistent with the previous work [30], the experiments are conducted under 5-fold cross-validation. Our Former-DFER and baseline models are all trained by dynamic sampling, and we also experiment with some methods using dynamic sampling as a comparison. The comparative performance is shown in Tab. 4. As shown in the table, the proposed method outperforms the compared methods both in UAR and WAR. Specifically, our method shows an improvement of 8.34% and 9.19% in UAR and WAR than the previous state-of-the-art method EC-STFL. Moreover, our Former-DFER shows an improvement of 2.01% and 1.68% in UAR and WAR than our baseline model. From Tab. 4, we can catch that the poor performance on "disgust" and "fear", we think that the insufficient samples result in poor performance. Regarding the DFEW dataset, the proportion of "disgust" and "fear" is 1.22% and 8.14%, respectively.

Comparison on AFEW: We also conduct a further evaluation on AFEW. For the AFEW dataset, in addition to the EmotiW-2019 Baseline, all the methods are first pre-trained on DFEW (fd1) and then fine-tuned on AFEW with the same setting. The comparative performance shown in Tab. 5 demonstrates that our Former-DFER achieves the best results both in UAR and WAR. Moreover, the proposed method improves the UAR and WAR of the baseline by 2.30% and 1.58%, respectively.

5 CONCLUSION

This paper proposes a dynamic facial expression recognition transformer (Former-DFER) for the in-the-wild scenario. Specifically, the proposed Former-DFER mainly consists of two-part: a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former). The CS-Former consists of five convolution blocks and N spatial encoders, which is designed to guide the network to learn occlusion- and pose-robust facial features from the spatial perspective. And the temporal transformer consists of M temporal encoders, which is designed to allow the network to learn contextual facial features from the temporal perspective. The abundant ablation studies have been studied to validate the effectiveness of each part in our Former-DFER. Moreover, the heatmaps of the leaned facial features indicate that the proposed method can handle the issues such as occlusion, non-frontal pose, and head motion. And the visualization of the feature distribution shows that the proposed method can learn more discriminative facial features. Furthermore, the comparison with the previous methods shows that our Former-DFER achieves state-of-the-art results on two popular benchmarks.

6 ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61825601 and in part by the Natural Science Foundation of Jiangsu Province (NSF-JS) under Grant BK20192004B.

REFERENCES

[1] Masih Aminbeidokhti, Marco Pedersoli, Patrick Cardinal, and Eric Granger. 2019. Emotion recognition with spatial attention and temporal softmax pooling. In *ICIAR*. 323–331.

[2] TS Ashwin and Ram Mohana Reddy Guddeti. 2020. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies* 25, 2 (2020), 1387–1415.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. In *NeurIPS*.

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

[5] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, Shizhong Han, Ping Liu, Min Chen, and Yan Tong. 2019. Feature-level and model-level audiovisual fusion for emotion recognition in the wild. In *MIPR*. 443–448.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *ECCV*. 213–229.

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

[8] Joyati Chattopadhyay, Souvik Kundu, Arpita Chakraborty, and Jyoti Sekhar Banerjee. 2018. Facial expression recognition for human computer interaction. In *ICCVBIC*. 1181–1192.

[9] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. 2014. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *ICMI*. 508–513.

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

[11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

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

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

[14] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*. 5203–5212.

[15] Abhinav Dhall. 2019. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *ICMI*. 546–550.

[16] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion recognition in the wild challenge 2013. In *ICMI*. 509–516.

[17] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia* 19, 03 (2012), 34–41.

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

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

[20] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *ICMI*. 467–474.

[21] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018. Video-based emotion recognition using deeply-supervised neural networks. In *ICMI*. 584–588.

[22] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *ICMI*. 445–450.

[23] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video action transformer network. In *CVPR*. 244–253.

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

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

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

[27] Qiaoping Hu, Chuanneng Mei, Fei Jiang, Ruimin Shen, Yitian Zhang, Ce Wang, and Junpeng Zhang. 2020. RFAU: A Database for Facial Action Unit Analysis in Real Classrooms. *IEEE Transactions on Affective Computing* (2020).

[28] Xiaohua Huang, Qiuhai He, Xiaopeng Hong, Guoying Zhao, and Matti Pietikainen. 2014. Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In *ICMI*. 514–520.

[29] Ramin Irani, Kamal Nasrollahi, Marc O Simon, Ciprian A Corneanu, Sergio Escalera, Chris Bahnsen, Dennis H Lundtoft, Thomas B Moeslund, Tanja L Pedersen, Maria-Louise Klitgaard, et al. 2015. Spatiotemporal analysis of RGB-DT facial images for multimodal pain level recognition. In *CVPRW*. 88–95.

[30] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild. In *MM*. 2881–2889.

[31] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing* 30 (2021), 2340–2349.

[32] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in Vision: A Survey. *arXiv preprint arXiv:2101.01169* (2021).

[33] Jean Kossaifi, Antoine Toisoul, Adrian Bulat, Yannis Panagakis, Timothy M Hospedales, and Maja Pantic. 2020. Factorized higher-order CNNs with an application to spatio-temporal emotion estimation. In *CVPR*. 6060–6069.

[34] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. 2020. Two-Stream Aural-Visual Affect Analysis in the Wild. In *FG*. 366–371.

[35] Vikas Kumar, Shivansh Rao, and Li Yu. 2020. Noisy Student Training using Body Language Dataset Improves Facial Expression Recognition. In *ECCV*. 756–773.

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

[37] Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn. 2020. Multi-modal recurrent attention networks for facial expression recognition. *IEEE Transactions on Image Processing* 29 (2020), 6977–6991.

[38] Min Kyu Lee, Dong Yoon Choi, Dae Ha Kim, and Byung Cheol Song. 2019. Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In *FG*. 1–8.

[39] Beibin Li, Sachin Mehta, Deepali Aneja, Claire Foster, Pamela Ventola, Frederick Shic, and Linda Shapiro. 2019. A facial affect analysis system for autism spectrum disorder. In *ICIP*. 4549–4553.

[40] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020).

[41] Sunan Li, Wenming Zheng, Yuan Zong, Cheng Lu, Chuangao Tang, Xingxun Jiang, Jiateng Liu, and Wanchuang Xia. 2019. Bi-modality fusion for emotion recognition in the wild. In *ICMI*. 589–594.

[42] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing* 28, 5 (2018), 2439–2450.

[43] Daizong Liu, Hongting Zhang, and Pan Zhou. 2020. Video-based Facial Expression Recognition using Graph Convolutional Networks. In *ICPR*.

[44] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. 2014. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*. 1749–1756.

[45] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen. 2014. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *ICMI*. 494–501.

[46] Yang Liu, Jinzhao Zhou, Xin Li, Xingming Zhang, and Guoying Zhao. 2021. Graph-based Facial Affect Analysis: A Review of Methods, Applications and Challenges. *arXiv preprint arXiv:2103.15599* (2021).

[47] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. 2018. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In *ICMI*. 646–652.

[48] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*. 94–101.

[49] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. 2019. Frame attention networks for facial expression recognition in videos. In *ICIP*. 3866–3870.

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

[51] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. 2017. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing* 10, 1 (2017), 60–75.

[52] Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *ICMI*. 577–582.

[53] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *ICME*.

[54] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. 2018. BAM: Bottleneck Attention Module. In *BMCV*.

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 8026–8037.

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

[57] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[58] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *CVPR*. 7464–7473.

[59] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2020. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877* (2020).

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

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

[62] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9, 11 (2008), 2579–2605.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[64] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *ICMI*. 569–576.

[65] Shanmin Wang, Hui Shuai, and Qingshan Liu. 2020. Phase Space Reconstruction Driven Spatio-Temporal Feature Learning for Dynamic Facial Expression Recognition. *IEEE Transactions on Affective Computing* (2020).

[66] Yanan Wang, Jianming Wu, and Keiichiro Hoashi. 2019. Multi-attention fusion network for video-based emotion recognition. In *ICMI*. 595–601.

[67] Ziheng Wang, Shangfei Wang, and Qiang Ji. 2013. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *CVPR*. 3422–3429.

[68] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*. 499–515.

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

[70] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. Learning texture transformer network for image super-resolution. In *CVPR*. 5791–5800.

[71] Peng Yang, Qingshan Liu, Xinyi Cui, and Dimitris N Metaxas. 2008. Facial expression recognition using encoded dynamic features. In *CVPR*. 1–8.

[72] Peng Yang, Qingshan Liu, and Dimitris N Metaxas. 2009. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters* 30, 2 (2009), 132–139.

[73] Zhenbo Yu, Guangcan Liu, Qingshan Liu, and Jiankang Deng. 2018. Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing* 317 (2018), 50–57.

[74] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. 2017. Aff-Wild: valence and arousal 'in-the-wild' challenge. In *CVPRW*. 34–41.

[75] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. 2020. M3F: Multi-Modal Continuous Valence-Arousal Estimation in the Wild. In *FG*. 617–621.

[76] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619.

[77] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. 2021. Learning Deep Global Multi-scale and Local Attention Features for Facial Expression Recognition in the Wild. *IEEE Transactions on Image Processing* (2021), 1–1.

[78] Zengqun Zhao, Qingshan Liu, and Feng Zhou. 2021. Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. In *AAAI*, Vol. 35. 3510–3519.

[79] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, and Matti Pietikäinen. 2013. A compact representation of visual speech data using latent variables. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 1 (2013), 1–1.

[80] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. 2011. Towards a practical lipreading system. In *CVPR*. 137–144.

[81] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR*. 1–16.