Spatio-Temporal Transformer for Dynamic Facial Expression Recognition in the Wild

Fuyan Ma
College of Electrical and Information
Engineering,
Hunan University
Key Laboratory of Visual Perception
and Artificial Intelligence of Hunan
Province
Changsha, China
mafuyan@hnu.edu.cn

Bin Sun
College of Electrical and Information
Engineering,
Hunan University
Key Laboratory of Visual Perception
and Artificial Intelligence of Hunan
Province
Changsha, China
sunbin611@hnu.edu.cn

Shutao Li
College of Electrical and Information
Engineering,
Hunan University
Key Laboratory of Visual Perception
and Artificial Intelligence of Hunan
Province
Changsha, China
shutao_li@hnu.edu.cn

ABSTRACT

Previous methods for dynamic facial expression in the wild are mainly based on Convolutional Neural Networks (CNNs), whose local operations ignore the long-range dependencies in videos. To solve this problem, we propose the spatio-temporal Transformer (STT) to capture discriminative features within each frame and model contextual relationships among frames. Spatio-temporal dependencies are captured and integrated by our unified Transformer. Specifically, given an image sequence consisting of multiple frames as input, we utilize the CNN backbone to translate each frame into a visual feature sequence. Subsequently, the spatial attention and the temporal attention within each block are jointly applied for learning spatio-temporal representations at the sequence level. In addition, we propose the compact softmax cross entropy loss to further encourage the learned features have the minimum intra-class distance and the maximum inter-class distance. Experiments on two in-the-wild dynamic facial expression datasets (i.e., DFEW and AFEW) indicate that our method provides an effective way to make use of the spatial and temporal dependencies for dynamic facial expression recognition. The source code and the training logs will be made publicly available.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; *Supervised learning*.

KEYWORDS

Dynamic facial expression recognition; spatio-temporal Transformer; loss function

1 INTRODUCTION

Facial expression is one of the most important ways for human to convey their emotions and communicate with each other[10]. Facial expression recognition (FER) has been an emerging topic, due to its essential real-world applications in driver safety monitoring [56], human robot emotional interaction [39], elderly healthcare[5] and so on. Accurately and automatically recognizing facial expression in videos is an interdisciplinary and integrated research, which spans from psychology to computer science. Previous intensive studies (such as [31, 41, 63]) have been conducted on static facial expression recognition (SFER). However, a facial expression is a
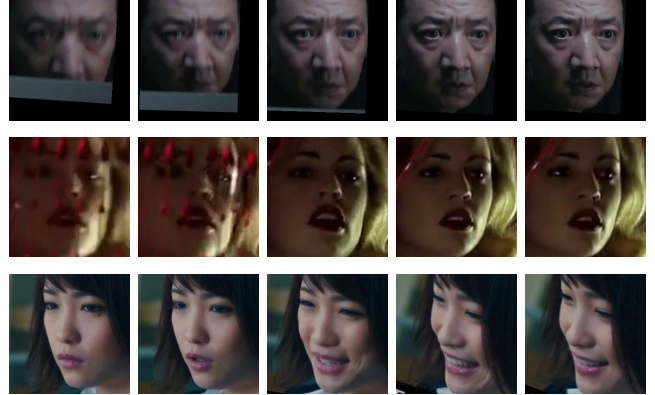

Figure 1: Samples from the DFEW [21] dataset. Variant head poses, poor illumination and occlusions can be seen from these image sequences.

dynamic process, which consists of various facial muscle motions in different facial regions. Temporal facial expression cues are of significance for understanding the emotional state.

Researchers have focused more and more attention on dynamic facial expression recognition (DFER) in recent years. DFER aims to classify a facial image sequence into several discrete expressions (such as happiness, surprise, neutral, sadness, fear, disgust, anger). Constrained by the limited dynamic facial expression datasets, most of DFER methods are designed based on several lab-collected datasets, such as Oulu-CASIA[60], MMI[52] and CK+[40]. The relatively few samples of these lab-collected datasets are not enough for training deep learning-based methods. The lack of large-scale databases seriously hinders the progress of dynamic facial expression recognition. Fortunately, Jiang *et al.* [21] propose a novel and well-annotated dataset named DFEW, with samples shown in Fig. 1. DFEW is the largest in-the-wild dynamic facial expression dataset with respect to the data volume and the sample sources, compared with other datasets (such as Aff-Wild [26], AFEW [11] and CAER [28]), which will definitely push forward the development of in-the-wild DFER.

Previous DFER methods can be mainly divided into two categories (i.e., static frame-based methods and dynamic sequence based methods) [32]. Most of the static frame-based methods utilize

local binary patterns [19], Gabor wavelets [29] and convolutional features[38, 57] to select peak (apex) frames in videos, and further conduct facial expression recognition on these frames. For example, Zhao *et al.* [61] propose to use a sample with peak expression to supervise the peak-piloted deep network to learn from a sample of non-peak expression. In addition, Meng *et al.* [42] propose to use the attention mechanism to aggregate some discriminative frames into a single video representation, achieving superior performance on AFEW and CK+. Although these methods perform well by selecting peak frames, they neglect the temporal dynamics and correlation among facial frames.

Different from static frame-based methods, dynamic sequence based methods usually use 3D convolution neural networks (3DCNN) [2], long-short term memory (LSTM) [54] to learn the spatio-temporal relationships, which can model long-term dependencies and improve the performance of DFER. For instance, Kim *et al.* [23] propose to use a LSTM network to learn the temporal characteristics of the spatial features, which achieves higher recognition rates. Chen *et al.* [8] propose a 3D-Inception-Resnet to directly make the learned features more representative by calculating a spatial-temporal-wise and a channel-wise attention map. Very recently, Li *et al.* [37] utilize a CNN to extract clip-level features and re-weight each clip-based representation to achieve clip-aware dynamic facial expression. Although several approaches have been proposed for in-the-wild DFER, the performances of these methods are still far from being satisfactory, because of occlusions, variant head poses, poor illumination and other unexpected issues in real-world scenes.

It is a challenging task to capture discriminative features in the spatial and temporal domain for in-the-wild DFER. Recent flourishing of Transformer-based methods on computer vision tasks has considerably deepened our understanding about discriminative feature representation and contextual information modeling. Therefore, we propose a simple but effective spatio-temporal Transformer (STT) for in-the-wild DFER, which can exploit and capture the facial appearance information in the spatial domain as well as the evolution information in the temporal domain to enhance the recognition performances. Specifically, we firstly translate each facial frame from a video into a visual feature sequence. Each patch of the feature map can be regard as a token of the feature sequence. We elaborately design the spatio-temporal Transformer for capture discriminative feature tokens and model temporal dependencies among different frames. To further increase the model discriminant ability, we impose the constraint on the prediction distribution by the loss function, following [21, 25, 35, 55]. We propose the compact softmax cross entropy loss to decrease the intra-class distance and increase the inter-class distance. The quantitative results and the visualization results demonstrate the effectiveness of our method for in-the-wild dynamic facial expression recognition. We summarize the contributions of our method as follows:

- We design a simple but effective spatio-temporal Transformer to capture discriminative features and model temporal relationships among facial frames.
- Towards enhancing the intra-class correlation and maximizing the inter-class distance, we propose the compact softmax cross entropy loss to supervise our model in a regularization manner.

- The extensive results and the visualization results on two in-the-wild DFER datasets demonstrate the superiority of our method.

2 RELATED WORK

2.1 Dynamic Facial Expression Recognition in the Wild

With the development of deep learning, especially Convolution Neural Networks, researchers have made significant progress on dynamic facial expression recognition in the wild. An intuitive way to extract the facial expression sequence is to learn the spatial and temporal representations separately. Under this paradigm, several methods [6, 23, 27] have been proposed to use a 2D CNN for extracting spatial features and apply a recurrent neural network (such LSTM) to model temporal correlation. For example, Li *et al.* [33] use a deep network to extract spatial features from each frame and model temporal dynamics by a convolutional network. The discriminative fused features are obtained by a Bi-LSTM network. A global face module is proposed to learn spatial features from the peak frame and a LSTM module is used to extract temporal information [59]. Similarly, Zhang *et al.* propose a novel framework named STRNN to integrate both spatial and temporal information separately. These methods use the spatio-temporal information of image sequence to improve the recognition performances.

Another way to learn the spatial and temporal representation is to model such correlation by the 3D CNN. Liu *et al.* [34] directly apply a 3D CNN with deformable facial action parts constraints to learn part-based representations for dynamic expression recognition. Jung *et al.* [22] also use the 3D CNN to learn spatial features along the time axis and aggregate temporal geometry features to boost the performance. Although these methods employ the 3D CNN to capture the temporal correlation, they only use one or two layers with 3D convolutional filters and fail to model the complicated spatio-temporal correlation. Very recently, Ayral *et al.* [2] propose to re-weight different clips and achieve a clip-aware 3D CNN for dynamic facial expression, which outperforms other methods on AFEW. A novel framework called EC-STFL [21] is also proposed to reduce the inter-class distance and enhance the intra-class correlation. Although these deep-learning based methods can extract spatial and temporal information, they all decouple the spatial features and the temporal features at different stages. We argue that if we can jointly get the spatial and temporal information, the performance of DFER will boost.

2.2 Transformers for Facial Expression Recognition

Transformer-based methods have shown dominant performance on various computer vision tasks, such as object detection [3], image synthesis [13] and pose estimation [58]. ViT [12] directly apply a standard Transformer network for a facial patch sequence and prove its power on image classification. Moreover, the Transformer-based methods [4, 47] have been proposed for video understanding. Regarding the facial expression recognition (FER) task, Ma *et al.* [41] first use Transformers to dynamically modeling the relationships among facial patches, and achieve outstanding performances
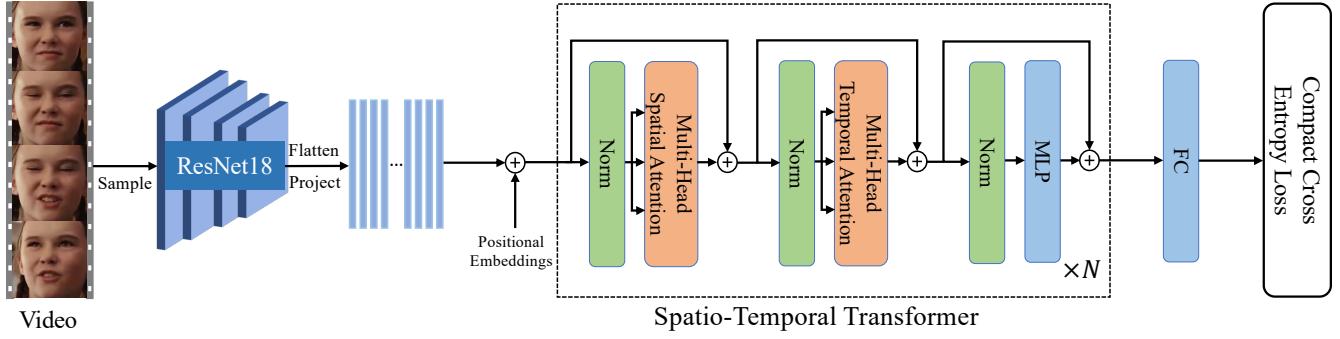
Figure 2: Overview of our method for dynamic facial expression recognition. The facial frames sampled from a video are processed by the CNN backbone to obtain frame-level feature sequences. The spatio-temporal Transformer jointly calculates the spatial attention and the temporal attention for capturing discriminative feature tokens. The compact cross entropy loss is utilized to further supervise our model to learn spatio-temporal representations, which have the close intra-class correlation and the large inter-class distance.

on occlusion and pose-variant FER. Following this idea, Huang *et al.* [20] use the visual Transformer to capture the relationships of different facial regions. Gao *et al.* also achieve occlusion-aware FER by a Transformer architecture. Similarly, Li *et al.* [30] propose a novel mask vision Transformer to focus on discriminative facial parts and filter out the complex backgrounds. Very recently, Former-DFER[62] is proposed to separately extract spatial and temporal correlations by Transformers. Former-DFER achieves current state-of-the-art performances on DFER. The self-attention mechanism of Transformers has the ability to capture long-range interactions among features tokens, which is of significance for the in-the-wild facial expression recognition. Therefore, the applications of Transformers, especially for in-the-wild DFER, are worth exploring.

3 METHOD

3.1 Overview

The proposed framework is shown in Fig. 2, which consists of three parts, i.e., input embedding generation, spatio-temporal Transformer and the compact softmax cross entropy loss. We first use the uniform sampling strategy to obtain the fixed-length facial frame sequence from a video. Then the CNN (i.e., ResNet18) is applied for extracting visual feature maps. We flatten and project the features maps as the input of the spatio-temporal Transformer (STT). The STT utilizes the spatial attention and the temporal attention to jointly learn the discriminative features from the sequence. Subsequently, the compact softmax cross entropy loss is used for minimizing the intra-class distance and maximizing the inter-class distance and further improving the recognition performance.

3.2 Input Embedding Generation

Given an image sequence $X \in \mathbb{R}^{F \times H_0 \times W_0 \times 3}$ with F RGB facial frames of size $H_0 \times W_0$ sampled from the video, we utilize the CNN backbone to extract frame-level features. Specifically, we first divide the video into S segments and uniformly sample T consecutive frames from each segment for training. Finally, we obtain $F = S \times T$ training frames for each video sample. During inference, T frames

in the mid of each segment are heuristically selected, and the length of testing frames is also $F = S \times T$.

A standard CNN backbone (i.e., ResNet18) is used to generate high-level feature maps of size $H \times W$ for each frame. The clip-level features $f_0 \in \mathbb{R}^{F \times H \times W \times C}$ are obtained by concatenating all frame-level feature maps. Then we flatten the spatial dimension of the clip-level feature maps f_0 and project them by a 1×1 convolution, resulting in a new feature sequence $f_1 \in \mathbb{R}^{F \times (H \times W) \times d}$. It is noted that the temporal order of f_1 is in accordance with that of the input X. To supplement the spatio-temporal positional information for the feature sequence f_1, we incorporate the learnable positional embeddings with f_1. Specifically, we add the spatial positional embedding by

$$f_2 = f_1 + e_{pos}^{space}, \tag{1}$$

where $e_{pos}^{space} \in \mathbb{R}^{1 \times (H \times W) \times d}$ is a learnable spatial positional embedding.

We also prepend the classification token to the sequence f_2 at the temporal dimension, which models the global state of the sequence and is further used for recognition. Then we reshape the new $f_2 \in \mathbb{R}^{(F+1) \times (H \times W) \times d}$ into $f_3 \in \mathbb{R}^{(H \times W) \times (F+1) \times d}$, and similarly add the temporal positional embedding by

$$z^0 = f_3 + e_{pos}^{time}, \tag{2}$$

where $e_{pos}^{time} \in \mathbb{R}^{1 \times (F+1) \times d}$ is a learnable temporal positional embedding. Finally, the input embedding z^0 to the spatio-temporal Transformer is obtained.

3.3 Spatio-Temporal Transformer

The spatio-temporal Transformer jointly utilizes the multi-head spatial attention and the multi-head temporal attention to learn a discriminative representation for each frame and exploit the long-range temporal dependencies to perform dynamic facial expression recognition. The spatio-temporal Transformer consists of N blocks and each block is composed with the multi-head spatial attention and the multi-head temporal attention, which iteratively learns the contextual and discriminative spatio-temporal feature representation.
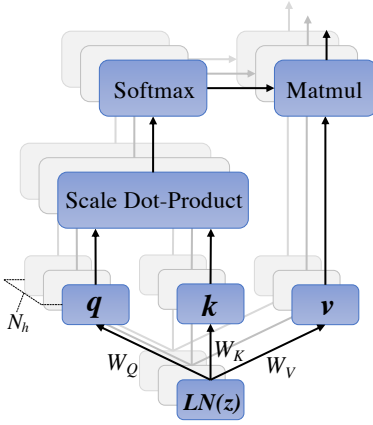
Figure 3: Illustration of the multi-head spatial attention.

Multi-Head Spatial Attention: At the i-th block of the spatio-temporal Transformer, we calculate the query/key/value vector for each patch of the same frame, which can be formatted by

$$q_{(p,t)}^{(i,a)} = W_Q^{(i,a)} LN(z_{(p,t)}^{l-1}),$$
(3)

$$k_{(p,t)}^{(i,a)} = W_K^{(i,a)} LN(z_{(p,t)}^{l-1}),$$
(4)

$$v_{(p,t)}^{(i,a)} = W_V^{(i,a)} LN(z_{(p,t)}^{l-1}),$$
(5)

where a is an index over multiple spatial heads, and $p \in \{1, \ldots, H \times W\}$, $t \in \{1, \ldots, F\}$ are the patch and frame index, respectively. In addition, W_Q, W_K, W_V are the parameters of linear projections. The hidden dimensions of q, k, v for each head is set to $D_h = d/N_h$, and N_h is the number of multiple heads.

The spatial attention is applied between feature patches of the same frame. Therefore, the spatial attention weights are obtained by the scale dot-product by:

$$\alpha_{(p,t)}^{(i,a)space} = SM\left(\frac{q_{(p,t)}^{(i,a)\,\mathrm{T}}}{\sqrt{D_h}} \cdot \left\{k_{(p',t)}^{(i,a)}\right\}_{p'=1,\ldots,H \times W}\right),$$
(6)

where SM represents the softmax activation function. The output of each head is given by

$$s_{(p,t)}^{(i,a)} = \sum_{p'}^{H \times W} \alpha_{(p,t),(p',t)}^{(i,a)space} v_{(p',t)}^{(i,a)}$$
(7)

Subsequently, the outputs are concatenated and projected, using the residual connection:

$$z_{(p,t)}^{\prime(i)} = W_O^\prime \left[s_{(p,t)}^{(i,1)}, \ldots, s_{(p,t)}^{(i,N_h)}\right]^{\mathrm{T}} + z_{(p,t)}^{(i-1)},$$
(8)

where $W_O^\prime$ denote the parameters of a linear projection for the concatenated multi-head features. Although the multi-head spatial attention provides a powerful representation in the spatial dimension, it does not reflect the temporal dependencies across different facial frames.

Multi-Head Temporal Attention: To further capture the temporal correlation within the frame sequence, we apply the multi-head temporal attention on the encoding $z_{(p,t)}^{\prime(i)}$ generated by the

multi-head spatial attention. Towards obtaining the multi-head temporal attention, we make $F + 1$ query-key comparisons, using each patch (p, t) with all patches at the same spatial location in the other frames. The query/key/value vector $(q_{(p,t)}^{\prime(i,a)}, k_{(p,t)}^{\prime(i,a)}, v_{(p,t)}^{\prime(i,a)})$ for each patch (p, t) is generated similarly in Eq. 3, Eq. 4, Eq. 5, respectively. Considering that a classification token is prepended to the frame feature sequence in the temporal dimension, we calculate the temporal attention weights by

$$\alpha_{(p,t)}^{\prime(i,a)time} = SM\left(\frac{q_{(p,t)}^{\prime(i,a)\,\mathrm{T}}}{\sqrt{D_h}} \cdot \left[k_{(0,0)}^{\prime(i,a)} \cdot \left\{k_{(p,t')}^{\prime(i,a)}\right\}_{t'=1,\ldots,F}\right]\right),$$
(9)

where $k_{(p,0)}^{\prime(i,a)}$ is the corresponding key value of the classification token at frame 0. The temporal attention weights $\alpha_{(p,t)}^{\prime(i,a)time}$ are used to sum over the values for each temporal head:

$$s_{(p,t)}^{\prime(i,a)} = \alpha_{(p,0)}^{\prime(i,a)time} v_{(p,0)}^{\prime(i,a)} + \sum_{t'=1}^{F} \alpha_{(p,t')}^{\prime(i,a)time} v_{(p,t')}^{\prime(i,a)}.$$
(10)

The outputs from all temporal heads are concatenated, projected, and passed through a multi-layer perception with a GELU [17] activation function, which can be defined as:

$$z_{(p,t)}^{\prime\prime(i)} = W_O^{\prime\prime} \left[s_{(p,t)}^{\prime(i,1)}, \ldots, s_{(p,t)}^{\prime(i,N_h)}\right]^{\mathrm{T}} + z_{(p,t)}^{\prime(i)},$$
(11)

$$z_{(p,t)}^{(i)} = MLP(LN(z_{(p,t)}^{\prime\prime(i)})) + z_{(p,t)}^{\prime\prime(i)}.$$
(12)

The spatio-temporal encoding $z_{(p,t)}^{(i)}$ of the i-th block serves as the input to the $(i + 1)$-th block. And we apply a single fully connected (FC) layer to the classification token $z_{(0,0)}^{(N)}$ of the final block:

$$p = FC(z_{(0,0)}^{(N)}),$$
(13)

where p is the prediction distribution of C facial expression classes.

3.4 Compact Softmax Cross Entropy Loss

The cross softmax entropy loss with softmax is undoubtedly the most common used supervision loss for the facial expression recognition task. However, learning discriminative spatio-temporal features for in-the-wild DFER requires the loss function to have the ability of maximizing the feature distance between different categories. Therefore, we need a more versatile loss function for DFER.

The label smoothing strategy has been a great success in training deep CNNs, which helps the models overcome the overfitting issue and learn more discriminative features. Given a training sample (x, y) and the prediction $\hat{y}$, the label smoothing strategy performs label regularization by introducing uniform distribution u over C classes on the ground-truth label distribution q:

$$s = q + \lambda u,$$
(14)

where λ is the balancing factor and q can be denoted as $q(\hat{y} = y|x) = 1$ and $q(\hat{y} \neq y|x) = 0$. Therefore, we compute the cross entropy loss by

$$\mathcal{L}_{soft} = -\sum_{k=1}^{C} s_k \log(p_k),$$
(15)

where s_k, p_k are the soft label and the predicted probability for the target category k, respectively. Furthermore, the loss $\mathcal{L}_{soft}$ can be rewritten as:

$$\mathcal{L}_{soft} = -(1-\lambda)\mathcal{L}_{ce} - \lambda\mathcal{L}_{reg}$$

$$= -(1-\lambda)\log(\boldsymbol{p}_y) - \frac{\lambda}{C}\sum_{c=1}^{C}\log(\boldsymbol{p}_c) \qquad (16)$$

$$= -(1-\lambda)\log(\boldsymbol{p}_y) + \lambda\{\mathcal{D}(\boldsymbol{u}||\boldsymbol{p}) + \log(C)\},$$

where L_{ce} and $\mathcal{L}_{reg}$ denotes the standard cross entropy loss and a regularization term. The Kullback-Leibler (KL) divergence $\mathcal{D}(\boldsymbol{u}||\boldsymbol{p})$ measures the difference between the two distributions $\boldsymbol{u}$ and $\boldsymbol{p}$, and suppresses the diversity of $\boldsymbol{p}$ in the loss $\mathcal{L}_{soft}$.

Based on the label smoothing regularization, we design the compact softmax cross entropy loss to decrease the intra-class distance and increase the inter-class distance. Specifically, we leverage the symmetric KL divergence as the regularization term and impose the constraint on the prediction distribution $\boldsymbol{p}'$, with the target prediction excluded. By combining with the standard softmax cross entropy loss, our loss results in

$$\mathcal{L}_{compact} = \mathcal{L}_{ce} + \beta\mathcal{L}'_{reg}$$

$$= \log(\boldsymbol{p}_y) + \frac{\beta}{2}\{\mathcal{D}(\boldsymbol{u}'||\boldsymbol{p}') + \mathcal{D}(\boldsymbol{p}'||\boldsymbol{u}')\}, \qquad (17)$$

of which the KL divergences are given by

$$\mathcal{D}(\boldsymbol{u}'||\boldsymbol{p}') = \sum_{c\neq y}\frac{1}{C-1}\log\left(\frac{1}{(C-1)\boldsymbol{p}'}\right), \qquad (18)$$

$$\mathcal{D}(\boldsymbol{p}'||\boldsymbol{u}') = \sum_{c\neq y}\log\left(\frac{1}{(C-1)\boldsymbol{p}'}\right). \qquad (19)$$

The uniform distribution $\boldsymbol{u}'$ is generated over $C-1$ classes, excluding the target y. Additionally, we calculate the new predicted distribution $\boldsymbol{p}'$ using the softmax function:

$$\boldsymbol{p}' = softmax(\{\hat{y}_c^{logits}\}_{c\neq y}), \qquad (20)$$

where $\{\hat{y}_c^{logits}\}_{c\neq y}$ represents the non-target predicted logits. With the regularization term $\mathcal{L}'_{reg}$, our loss function guides the model to increase the uniformity of the non-target predicted logits. To summarize, our proposed loss function has the ability to induce the smaller intra-class distance and the larger inter-class distance, which is crucial to in-the-wild DFER.

4 EXPERIMENTS

4.1 Datasets

DFEW [21] is the largest publicly available DFER dataset, with 16,372 in-the-wild facial expression video clips. All video clips are extracted from more than 1500 movies, whose themes cover tragedy, comedy, love and so on. Due to the unconstrained conditions, DFEW is a significantly challenging dataset, of which the samples suffer occlusions, head pose changes, extreme illuminations and other problems. Each clip of DFEW is labelled by ten independent professionally trained annotators with one of seven basic facial expressions (i.e., happiness, surprise, neutral, sadness, fear, disgust, anger), indicating that DFEW is a reliable dataset with very few annotation errors. Similar with other FER databases, DFEW has the imbalanced category distribution issue. To better evaluate different methods

on the selected 12,509 clips, five-fold cross-validation protocol is adopted.

AFEW has served as an evaluation dataset for the annual EmotiW from 2013 to 2019. The videos of AFEW are close to real-world data with a wide range of challenges (e.g., complicated backgrounds, different head poses and movements), because the video samples are collected from TV serials and movies. Same with DFEW, each video in AFEW is assigned to seven emotion labels. AFEW consists of 773 videos from 67 movies for training, 383 samples from 33 movies for validation. Since the official test dataset is unavailable, we report the results on the validation dataset to make a fair comparison with other state-of-the-art methods.

4.2 Implementation Details

Data Preprocessing: For DFEW, the processed face region images are publicly available, officially detected and aligned by Megvii Face API [1] and SeetaFace [36]. We directly use the processed data for fair comparison. For AFEW, the FFmpeg toolkit [49] is used to extract frames from the raw videos. We then detect and align faces from the frames by the Dlib toolbox [24]. All the face region images are resized to 112×112 pixels as the inputs to our model.

Training and Testing Settings: We train our model with the PyTorch [43] platform with two NVIDIA GTX 1080Ti GPU cards. Our model is trained with the batch-size of 32 on DFEW and AFEW for 100 epochs and 20 epochs, respectively. The SGD optimizer [45] with an initial learning rate of 0.01 is used to optimize our proposed model on DFEW. The learning rate is divided by 10 every 40 epochs during training. Since AFEW is a relatively tiny dataset, we follow Former-DFER [62] to fine-tune the model on AFEW by the pretrained weights on DFEW (fd1). The learning rate on AFEW is initialized to 0.001. We use the pretrained ResNet18 on MS-Celeb-1M [14] as our CNN backbone. During training, each video clip is divided into $S = 8$ segments and uniformity sample $T = 2$ consecutive frames from each segment. During testing, S and T keep the same settings, and $T = 2$ samples are selected from the mid of each segment, result in 16 facial frames as the input. The number of spatio-temporal Transformer layers N and the number of heads a are empirically assigned to 4 and 8, respectively. In addition, the hidden dimension d is set to 512 and the loss balancing factor β is set to 0.2. The unweighted average recall (UAR) and the weighted average recall (WAR) serves as the evaluation metrics.

4.3 Ablation Study

We conduct the ablation experiments to analyze the impact of each component of our model (i.e., spatio-temporal Transformer and the compact softmax cross entropy loss). All the experiments are conducted on the DFEW dataset with 5-fold cross validation.

Importance of the spatio-temporal Transformer: We first explore the impact of multi-head spatial attention and multi-head spatial attention. The experimental results on DFEW in terms of the UAR and the WAR are shown in Tab. 1. Our proposed spatio-temporal Transformer applies the temporal attention after the spatial attention in one block. To demonstrate things their influences, we replace the spatio-temporal Transformer with a simple LSTM as the baseline. The baseline achieves the UAR and the WAR at 52.02% and 64.48% respectively. The use of multi-head spatial attention
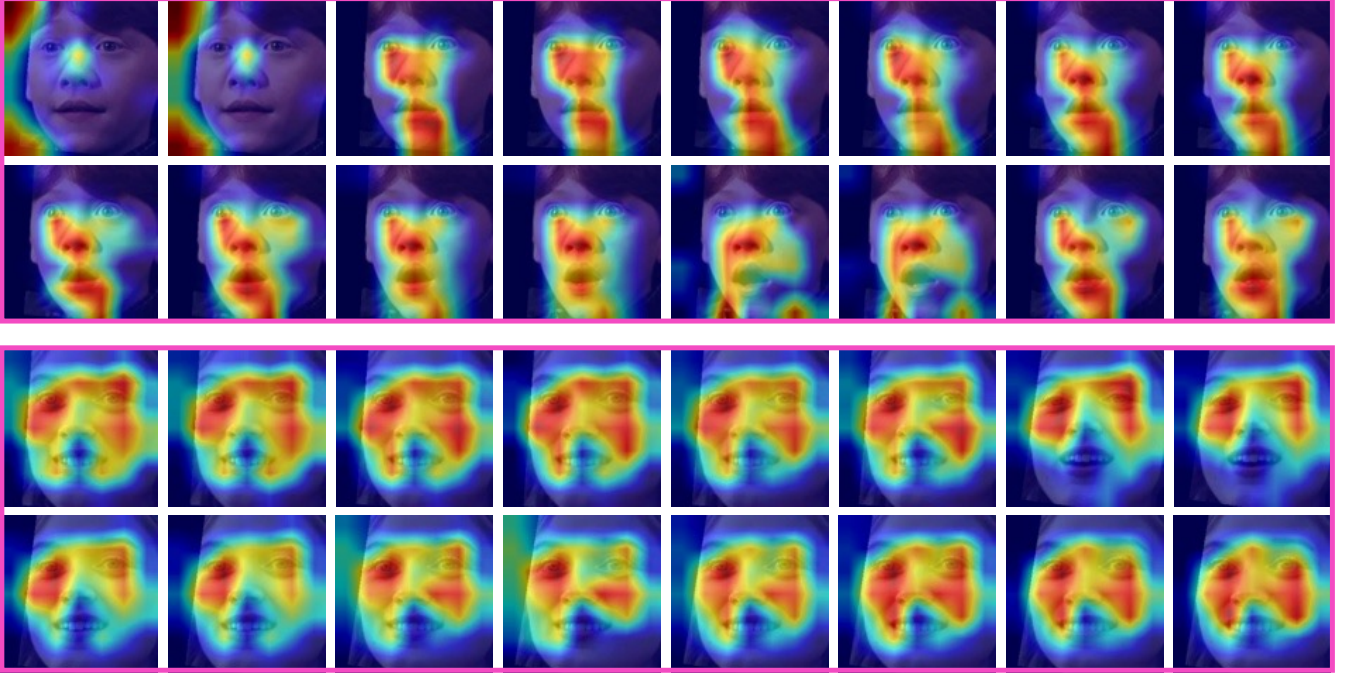
Figure 4: Visualization of the activation maps generated by Grad-CAM [46]. Two frame sequences are presented to demonstrate the discriminative facial regions and the temporal correlations learned by our model.

Table 1: Ablation study w.r.t. multi-head spatial attention, multi-head temporal attention, performed on DFEW with 5-fold cross validation.

Multi-Head Spatial Attention	Multi-head Temporal Attention	Metrics (%)	
		UAR	WAR
✗	✗	52.02	64.48
✔	✗	53.15	65.23
✗	✔	53.30	65.39
✔	✔	54.58	66.65

Table 2: Evaluation of different loss functions on DFEW in terms of the UAR and the WAR.

Loss Function	Metrics (%)	
	UAR	WAR
Cross Entropy Loss	54.15	66.39
Label Smoothing	54.29	66.50
Ours	54.58	66.65

improves the UAR and the WAR by 1.13% and 0.75% over the baseline. And employing the multi-head temporal attention improves the baseline by 1.28% and 0.91% in terms of the UAR and the WAR. Moreover, when the spatial attention and the temporal attention are both applied, the UAR and the WAR are enhanced significantly by 2.56% and 2.71% over the baseline, respectively.

To further demonstrate the impact of the spatial attention and the temporal attention, we visualize the activation maps by Grad-CAM [46] in Fig. 4. As shown in the first frame sequence of Fig. 4, we can find that the first two facial frames are not highlighted on the facial regions. We speculate that our model can learn the temporal correlations among the frames, because the first two frames contribute very less to the target category surprise. Our model can concatenate on the facial regions of the other frames of the first sequence, which indicates the effectiveness of our spatio-temporal Transformer for in-the-wild DFER. The activation maps in the frames from the second sequence also support our conclusion.

Effect of the compact softmax cross entropy loss: We then investigate the influences of different loss functions (i.e., the standard cross entropy loss, label smoothing, our compact loss). The corresponding results are shown in Tab. 2. With the standard cross entropy loss, our model still achieves the UAR of 54.15% and the WAR of 66.39%. Label smoothing and our compact softmax cross entropy loss both improve the performance of our model purely with the cross entropy loss. Specifically, our model trained with label smoothing achieves the gains of 0.14% and 0.11% in terms of the UAR and the WAR. Our compact softmax cross entropy loss achieves the UAR of 54.58% and the WAR of 66.65%, which are 0.43% and 0.26% better than that with the standard softmax cross entropy loss.

To better understand the effect of our proposed loss, we utilize t-SNE [53] to visualize the learned features on DFEW (fd1~fd5) in Fig. 5. The feature distribution on a 2D plane can reflect the intra-class
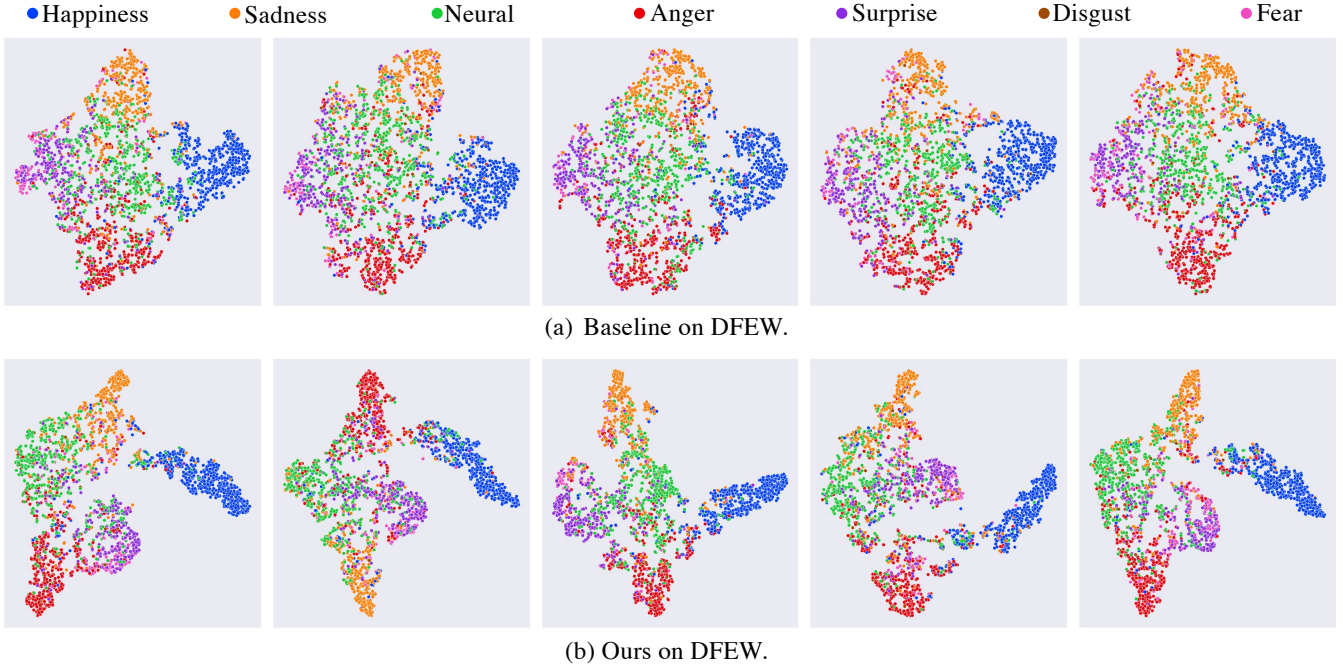
(a) Baseline on DFEW.

(b) Ours on DFEW.

Figure 5: Visualization of the feature distribution generated by t-SNE [53] on DFEW (fd1~fd5). (a) denotes our model with the standard cross entropy loss. (b) represents our model with the compact cross entropy loss.
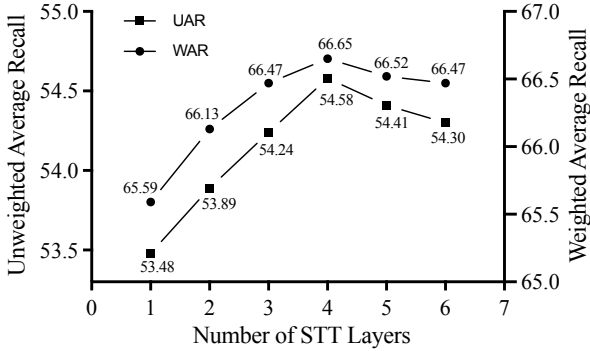

Figure 6: Evaluation of our method with different layers on DFEW.

compactness and the inter-class distance. As shown in Fig. 5, our compact loss enables the learned features have a better aggregation effect and show more clear inter-class boundaries among different expressions, compared with the standard cross entropy loss. Such patterns in Fig. 5 also prove that our model has the ability to learn more discriminative features under the guidance of our proposed loss function.

Impact of Hyper-parameters: Finally, we study the impact of different number of spatio-temporal Transformer layers. We vary the number of layers $N \in \{1, 2, 3, 4, 5, 6\}$ to conduct experiments. N is a trade-off parameter for balancing the recognition performance and the complexity of the model. The experimental results are

shown in Fig. 6. It is apparent that choosing the value of $N = 4$ obtains the best performances of 54.58% and 66.65% in terms of the UAR and the WAR, respectively. The results also indicates that the shallower model's feature representation ability is more limited.

4.4 Comparison with State-of-the Art Methods

We compare our method with other methods on DFEW (5-fold cross validation) and AFEW, with respect to the UAR, the WAR.

Comparison on DFEW: To make a fair comparison, we compare our methods with current state-of-the-art spatio-temporal networks on DFEW, which can be divided into 3D CNNs-based methods and 2D CNN-based methods. The comparison results are shown in Tab. 3. Our method obtains the best results using both metrics. Specifically, Former-DFER is the previous state-of-the-art method with the UAR of 53.69% and the WAR of 66.65%. Our method outperforms Former-DFER by 0.89% and 0.95% in terms of the UAR and the WAR, respectively. Moreover, our method improves EC-STFL by 9.23% and 10.14% in UAR and WAR, which also aims to enhance the intra-class correlation and increase the inter-class distance. As shown in Tab. 3, our method also obtains better results with respect to the category-level accuracy compared with other methods. Specifically, our method achieves the highest accuracies on "happiness", "sadness", "anger" and "fear", and obtains the second best results on "surprise" and "disgust". It is noted that the performances of the expressions "disgust" and "fear" are poor, which are mainly due to insufficient training samples.

Comparison on AFEW: We follow previous methods to conduct transfer learning experiments from DFEW to AFEW. The pre-trained weights on DFEW (fd1) are used to initialize the model

Table 3: Comparison with other state-of-the-art methods on DFEW with 5-fold cross validation. The best results are in bold. Underline represents the second best. TI denotes time interpolation and DS denotes dynamic sampling. The evaluation metrics include the unweighted average recall (UAR) and the weighted average recall (WAR).

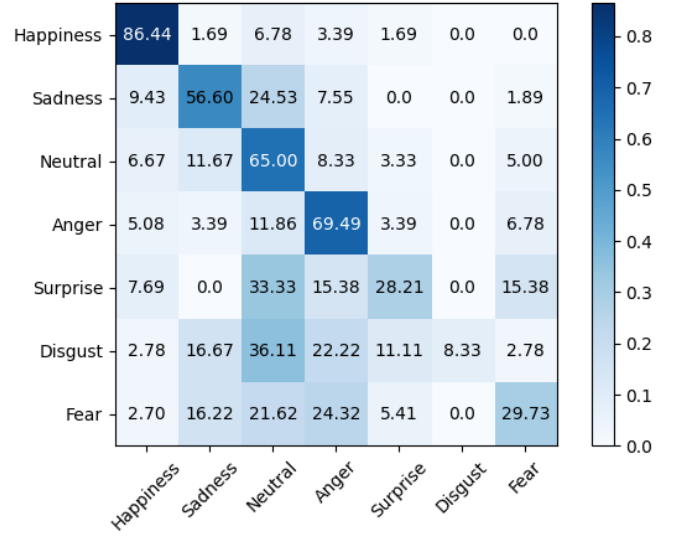| Method	Sample Strategies	Accuracy of Each Emotion (%)							Metrics (%)	
		Happiness	Sadness	Neutral	Anger	Surprise	Disgust	Fear	UAR	WAR
C3D [50]	TI	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
P3D [44]	TI	74.85	43.40	54.18	60.42	50.99	0.69	23.28	43.97	54.47
R(2+1)D18 [51]	TI	79.67	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22
3D Resnet18 [15]	TI	73.13	48.26	50.51	64.75	50.10	0.00	26.39	44.73	54.98
I3D-RGB [7]	TI	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.40	54.27
VGG11+LSTM [18, 48]	TI	76.89	37.65	58.04	60.70	43.70	0.00	19.73	42.39	53.70
ResNet18+LSTM [16, 18]	TI	78.00	40.65	53.77	56.83	45.00	**4.14**	21.62	42.86	53.08
3D R.18+Center Loss [15, 55]	TI	78.49	44.30	54.89	58.40	52.35	0.69	25.28	44.91	55.48
EC-STFL [21]	TI	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
3D Resnet18 [15]	DS	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27
ResNet18+LSTM [16, 18]	DS	83.56	61.56	**68.27**	65.29	51.26	0.00	29.34	51.32	63.85
Resnet18+GRU [9, 16]	DS	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
Former-DFER [62]	DS	84.05	62.57	<u>67.52</u>	70.03	**56.43**	3.45	31.78	53.69	65.70
STT (Ours)	DS	**87.36**	**67.90**	64.97	**71.24**	<u>53.10</u>	<u>3.49</u>	**34.04**	54.58	66.65

Table 4: Comparison with other state-of-the-art methods on AFEW. The best results are in bold.

Methods	Sample Strategies	Metrics (%)	
		UAR	WAR
EmotiW-2019 Baseline [11]	-	-	38.81
C3D [50]	DS	43.75	46.72
I3D-RGB [7]	DS	41.86	45.41
R(2+1)D [51]	DS	42.89	46.19
3D ResNet18 [15]	DS	42.14	45.67
ResNet18+LSTM [16, 18]	DS	43.96	48.82
ResNet18+GRU [9, 16]	DS	45.12	49.34
Former-DFER [62]	DS	47.42	50.92
STT(ours)	DS	**49.11**	**54.23**

Figure 7: Experimental results of confusion matrix on AFEW.

weights for AFEW. The comparison results are shown in Tab. 4. Our method also achieves the best results in UAR and WAR compared with other state-of-the-art methods. Specifically, the performances of our method are 1.69% and 3.31% better than these of Former-DFER. In addition, our method improves the UAR and the WAR of 3D ResNet18 by 6.97% and 8.56%, respectively. We provide the detailed recognition results of different facial expressions by the confusion matrix in Fig. 7. Similar with the results on DFEW, the performances of the expressions "disgust" and "fear" are still far from being satisfactory and can be improved in the future work.

5 CONCLUSION

In this paper, we propose a simple but effective spatio-temporal Transformer and the compact softmax cross entropy loss for in-the-wild dynamic facial expression recognition (DFER). Specifically, the spatio-temporal Transformer is utilized to handle DFER with

the real-world scenario issues (e.g., occlusions, head poses and illuminations) by sequence modeling. We jointly apply the spatial attention and the temporal attention within each block to learn spatio-temporal representations iteratively. The spatial facial cues and the temporal dependencies are captured by a sequence of frame-level feature tokens from a global perspective. To further increase the model discriminant ability, we impose the constraint on the prediction distribution by the compact loss to enhance the intra-class correlation and increase the inter-class distance. The experimental results and the visualization results demonstrate that our method learns discriminative spatio-temporal feature representations and

enhances the classification margin, which result in our favorable performance compared with other state-of-the-art methods.

REFERENCES

[1] MEGVII Face++ Face Detection API. [n.d.]. https://www.faceplusplus.com.
[2] Théo Ayral, Marco Pedersoli, Simon Bacon, and Eric Granger. 2021. Temporal stochastic softmax for 3d cnns: An application in facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3029–3038.
[3] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. 2020. Toward Transformer-Based Object Detection. *arXiv preprint arXiv:2012.09958* (2020).
[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095* 2, 3 (2021), 4.
[5] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. 2022. Impact of Deep Learning Approaches on Facial Expression Recognition in Healthcare Industries. *IEEE Transactions on Industrial Informatics* (2022). doi: 10.1109/TII.2022.3141400.
[6] Youyi Cai, Wenming Zheng, Tong Zhang, Qiang Li, Zhen Cui, and Jiayin Ye. 2016. Video based emotion recognition using CNN and BRNN. In *Chinese Conference on Pattern Recognition*. Springer, 679–691.
[7] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
[8] Weicong Chen, Dong Zhang, Ming Li, and Dah-Jye Lee. 2020. Stcam: Spatial-temporal and channel attention module for dynamic facial expression recognition. *IEEE Transactions on Affective Computing* (2020). 10.1109/TAFFC.2020.3027340.
[9] 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).
[10] Charles Darwin. 2015. *The expression of the emotions in man and animals*. University of Chicago Press.
[11] Abhinav Dhall. 2019. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *International Conference on Multimodal Interaction*. 546–550.
[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12873–12883.
[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Msceleb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*. Springer, 87–102.
[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.
[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[17] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[19] Ming-Wei Huang, Zhe-wei Wang, and Zi-Lu Ying. 2010. A new method for facial expression recognition based on sparse representation plus LBP. In *International Congress on Image and Signal Processing*, Vol. 4. IEEE, 1750–1754.
[20] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. 2021. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences* 580 (2021), 35–54.
[21] 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 *Proceedings of the 28th ACM International Conference on Multimedia*. 2881–2889.
[22] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 2983–2991.
[23] Dae Hoe Kim, Wissam J Baddar, Jinhyeok Jang, and Yong Man Ro. 2017. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing* 10, 2 (2017), 223–236.
[24] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.
[25] Takumi Kobayashi. 2019. Large Margin In Softmax Cross-Entropy Loss.. In *British Machine Vision Conference*. 139.
[26] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* 127, 6 (2019), 907–929.
[27] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. 2018. A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2121–2129.
[28] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10143–10152.
[29] Seung Ho Lee, Wissam J Baddar, and Yong Man Ro. 2016. Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos. *Pattern Recognition* 54 (2016), 52–67.
[30] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. 2021. Mvt: Mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520* (2021).
[31] Shan Li and Weihong Deng. 2018. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* 28, 1 (2018), 356–370.
[32] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020). doi: 10.1109/TAFFC.2020.2981446.
[33] Dandan Liang, Huagang Liang, Zhenbo Yu, and Yipu Zhang. 2020. Deep convolutional BiLSTM fusion network for facial expression recognition. *The Visual Computer* 36, 3 (2020), 499–508.
[34] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. 2014. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer Vision*. Springer, 143–157.
[35] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks.. In *International Conference on Machine Learning*, Vol. 2. 7.
[36] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen. 2017. VIPLFaceNet: an open source deep face recognition SDK. *Frontiers of Computer Science* 11, 2 (2017), 208–218.
[37] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. 2022. Clip-aware Expressive Feature Learning for Video-based Facial Expression Recognition. *Information Sciences* (2022). doi: 10.1016/j.ins.2022.03.062.
[38] Yuanyuan Liu, Xiaohui Yuan, Xi Gong, Zhong Xie, Fang Fang, and Zhongwen Luo. 2018. Conditional convolution neural network enhanced random forest for facial expression recognition. *Pattern Recognition* 84 (2018), 251–261.
[39] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. 2017. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica* 4, 4 (2017), 668–676.
[40] 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 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 94–101.
[41] Fuyan Ma, Bin Sun, and Shutao Li. 2021. Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing* (2021). doi: 10.1109/TAFFC.2021.3122146.
[42] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. 2019. Frame attention networks for facial expression recognition in videos. In *IEEE International Conference on Image Processing*. IEEE, 3866–3870.
[43] 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. *arXiv preprint arXiv:1912.01703* (2019).
[44] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.
[45] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).
[46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.
[47] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. 2021. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915* (2021).
[48] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[49] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux Journal* 2006, 146 (2006), 10.
[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international Conference on Computer Vision*. 4489–4497.

[51] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 6450–6459.

[52] Michel Valstar, Maja Pantic, et al. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proceedings of International Workshop on EMOTION: Corpora for Research on Emotion and Affect.* Paris, France., 65.

[53] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[54] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* 569–576.

[55] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision.* Springer, 499–515.

[56] Torsten Wilhelm. 2019. Towards facial expression analysis in a driver assistance system. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019).* IEEE, 1–4.

[57] Huiyuan Yang, Umur Ciftci, and Lijun Yin. 2018. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2168–2177.

[58] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. 2020. TransPose: Towards Explainable Human Pose Estimation by Transformer. *arXiv preprint arXiv:2012.14214* (2020).

[59] Mingjing Yu, Huicheng Zheng, Zhifeng Peng, Jiayu Dong, and Heran Du. 2020. Facial expression recognition based on a multi-task global-local network. *Pattern Recognition Letters* 131 (2020), 166–171.

[60] 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.

[61] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. 2016. Peak-piloted deep network for facial expression recognition. In *European Conference on Computer Vision.* Springer, 425–442.

[62] Zengqun Zhao and Qingshan Liu. 2021. Former-DFER: Dynamic Facial Expression Recognition Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia.* 1553–1561.

[63] 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* 30 (2021), 6544–6556.