Information Sciences 598 (2022) 182–195

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Clip-aware expressive feature learning for video-based facial expression recognition

Yuanyuan Liu [a], Chuanxu Feng [a], Xiaohui Yuan [b,*], Lin Zhou [a], Wenbin Wang [a], Jie Qin [c], Zhongwen Luo [a]

[a] School of Information Engineering, China University of Geosciences, Wuhan, China
[b] Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA
[c] Wuhan Huazhong Numerical Control Co., Ltd., China

ARTICLE INFO

Article history:
Received 27 September 2021
Received in revised form 11 March 2022
Accepted 19 March 2022
Available online 25 March 2022

Keywords:
Video-based FER
Emotional activation map
Clip-based feature encoder
Clip-aware emotion-rich representation

ABSTRACT

Video-based facial expression recognition (FER) has received increased attention as a result of its widespread applications. However, a video often contains many redundant and irrelevant frames. How to reduce redundancy and complexity of the available information and extract the most relevant information to facial expression in video sequences is a challenging task. In this paper, we divide a video into several short clips for processing and propose a clip-aware emotion-rich feature learning network (CEFLNet) for robust video-based FER. Our proposed CEFLNet identifies the emotional intensity expressed in each short clip in a video and obtains clip-aware emotion-rich representations. Specifically, CEFLNet constructs a clip-based feature encoder (CFE) with two-cascaded self-attention and local–global relation learning, aiming to encode clip-based spatio-temporal features from the clips of a video. An emotional intensity activation network (EIAN) is devised to generate emotional activation maps for locating the salient emotion clips and obtaining clip-aware emotion-rich representations, which are used for expression classification. The effectiveness and robustness of the proposed CEFLNet are evaluated using four public facial expression video datasets, including BU-3DFE, MMI, AFEW, and DFEW. Extensive experiments demonstrate the improved performance of our proposed CEFLNet in comparison with the state-of-the-art methods.

1. Introduction

Video-based Facial Expression Recognition (FER) is an important task for understanding human emotions and behaviors in videos, which classifies a video into several basic emotions such as happiness, anger, disgust, fear, sadness, neutral, and surprise [1,2]. The task faces several challenges such as noise introduced by irrelevant frames, the inherently complex information of subtle facial expressions in videos, the costly computational overhead introduced by heavy models to ensure performance. To address these problems, we introduce a clip-aware, emotion-rich feature learning network to obtain an advanced representation of videos for FER.

* Corresponding author.
E-mail addresses: liuyy@cug.edu.cn (Y. Liu), fcxfcx@cug.edu.cn (C. Feng), xiaohui.yuan@unt.edu (X. Yuan), zhoulin@cug.edu.cn (L. Zhou), wangwenbin@cug.edu.cn (W. Wang), toqingjie@126.com (J. Qin), luozw@cug.edu.cn (Z. Luo).

https://doi.org/10.1016/j.ins.2022.03.062
0020-0255/

Video-based FER methods include static frame-based methods and dynamic sequence-based methods [3]. Most of the static frame-based methods process the manually defined peak (apex) frames, e.g., local binary patterns (LBPs) [4], local phase quantization (LPQ) [5,6], Gabor wavelets [7], convolutional features [8–10], etc. These methods usually neglect the importance of intrinsic relationships between visual information of adjacent frames. In addition, it is labor costly to obtain peak frames via manual annotation.

Recently, more studies focus on the dynamic sequence-based method. Rather than using static frames, methods such as the Long Short-Term Memory (LSTM) [11,12] and C3D network [13], encode the spatio-temporal information by learning from appropriate supervision signals (e.g., video category labels). Modeling long-term dependencies has been widely employed for video-based expression recognition [14,15]. Although the sequence-based methods have shown an improvement for FER, they still face difficulties in two aspects: they usually require overwhelmingly high computation complexity to model video facial expression movements [3,16], and the presence of many frames irrelevant to expressions makes the learned features suboptimal to FER [3].

To address the above limitations, we propose a clip-aware, emotion-rich feature learning network (CEFLNet) that focuses on the most informative frames for FER by identifying the emotional intensities of clips in a video. In particular, we make the CEFLNet automatically locate the most salient frames in a weakly supervised manner without intensity annotations, and thus achieve clip-aware emotion-rich representations for video-based FER. The CEFLNet contains two main components: clip-based feature encoder (CFE) and weakly supervised emotional intensity activation network (EIAN). CFE is used to learn clip-based spatio-temporal features based on inter-frame relations in a clip, exploiting emotional cues between adjacent frames within each clip. EIAN identifies salient clips and obtains clip-aware emotion-rich representations by estimating the emotional activation map.

The contributions of this paper include the following:

- we propose a novel CEFLNet for video-based FER to jointly learn the emotional intensity of clips of a video and recognize facial expressions in a mutually reinforced way. Evaluations on four challenging video-based facial expression datasets demonstrate its advantages over the existing state-of-the-art methods.
- the weakly supervised EIAN is proposed to identify the emotional intensity of each clip and learn clip-aware emotion-rich representation via generating an emotional activation map.
- the CFE is proposed to adaptively aggregate the frame features to form clip-based spatio-temporal features via jointly learning self-attention and local–global relation attention, which fully exploits emotional cues between adjacent frames within each clip.

The remainder of this paper is organized as follows: Section 2 introduces related work in video-based FER. Section 3 presents the proposed CEFLNet for video-based FER in detail. Section 4 discusses the experimental results on four publicly available datasets. Finally, this paper is concluded in Section 5 with a summary and future work.

2. Related work

Video-based FER. Existing video-based FER methods include static frame-based methods and dynamic sequence-based methods. Among the static frame-based methods, we have frame aggregation methods and peak frame extraction methods. The frame aggregation methods strategically combine frame-level features learned from static-based FER networks [16,17] to construct video-level features for FER. The peak frame extraction methods focus on the peak frame of a video and ignore the emotional information from other periods of the video [18,19]. Meng *et.al* [16] proposed the frame attention networks to adaptively aggregate frame features in an end-to-end framework and achieved accuracy of 51.18% on the AFEW 8.0 dataset [20]. To alleviate the influence of redundant and irrelevant frames, Zhao *et al.* [18] proposed a peak-piloted deep network (PPDN) for intensity-invariant expression recognition. This method takes a pair of peak and non-peak expression images with the same expression and subject as input and minimizes the distance between the images with the same expression. Yu *et al.* [19] proposed a deeper cascaded peak-piloted network (DCPN) to enhance the ability of expression representation of the network. These frame-based methods have achieved good results in well-selected peak frames, however, manual selection of peak frames increases labor costs while ignoring other emotional cues existing in adjacent frames.

The dynamic sequence-based method takes the entire video sequence as input and uses the texture information and temporal dependence in the frame sequence to recognize facial expressions [21,13,9,11,3]. Vielzeuf et al. [11] used pre-trained VGG-Face to extract spatial features, then utilized an LSTM layer to encode temporal dependencies in the sequence. Kim *et al.* [13] propose a new spatio-temporal representation learning for FER by integrating C3D and LSTM networks, which is robust to expression intensity variation. In [21], a temporal geometric feature was proposed to improve the discriminative capacity of the learned spatio-temporal appearance features. Although these dynamic-based networks capture spatio-temporal features for FER, they still challenge in describing expression movements in untrimmed videos and require large model capacities to model facial expression changes in videos.

Attention model. Visual attention based networks have been proposed to localize significant regions for many computer vision tasks, including fine-grained recognition [22,23], image captioning [24], person re-identification [25], and object detection [26,27]. Some methods are learned by the aggregating scheme from the internal hidden representations in CNN

[28]. Other methods focus on detecting local regions according to supervised bounding box annotation, e.g., region proposal network (RPN) [26]. Zheng et al. [28] adopted channel grouping sub-network to cluster different convolutional feature maps into groups according to peak responses of maps. Xu et al. [29] proposed an attention shift based on multiple blur levels to avoid occlusions for facial gender classification. SE-Net [23] proposed the Squeeze-and-Excitation (SE) block that re-calibrates channel-wise feature responses by explicitly modeling the inter-dependency between channels. The SE block results in considerable performance improvement for image classification with minor additional computational costs. Meng et al. [16] proposed a frame attention network (FAN) for selecting frames from a video to form a dicriminative video-level representation. Although attention has been successfully employed in many computer vision tasks, it is difficult to directly use it for capturing beneficial expression movements in videos due to the vastly present irrelevant frames and the limited motion variation.

3. Clip-aware emotion-rich feature learning network

3.1. Network Architecture

The architecture of our proposed CEFLNet is shown in Fig. 1(a). CEFLNet consists of CFE and EIAN. Given a video sequence V with facial expression label $Y_V = \{y^e\}$, V is divided into several video clips denoted as $V = \{C_1, C_2, \ldots, C_n\}$, where C_k is the $k-th$ clip. Our learning problem consists of two parts: (1) CFE adaptively encodes frame feature vectors extracted from a clip C_i to form discriminative clip-based features R_i, via jointly self-attention learning and local–global relation learning. (2) After concatenating the clip features, EIAN further focuses on clip-aware emotion-rich representations by generating emotional activation maps in a weakly supervised learning manner, without any peak frames or clip annotation.

3.2. CFE for clip-level representation

The clip-based feature encoder contains two cascaded attention learning modules: self-attention learning (SAL) and local–global relation learning (LRL). Fig. 1(b) shows the detailed structure of the CFE. In practice, SAL models frame-level relation to obtain the self-attention in each clip, and LRL improves the clip-level representation by learning local–global relation attention. Through the two-cascaded attention learning, the CFE exploits the emotional cues of spatio-temporal information in each clip.

Self-attention learning. Self-attention learning models the frame-level relation to obtain spatio-temporal features for clips. Fig. 2 shows the detailed structure of this component. Let $f_{k,i}$ denote the feature vector of the k-th frame in the i-th clip. Note that we use the deep convolutional neural network (DCNN) like a pre-trained ResNet-18 to extract features and consider the global average pooling output of the employed DCNN as $f_{k,i}$. I_i denotes the matrix stacking all the features $f_{k,i}$ of the i-th clip. Given that a clip contains K frames and each $f_{k,i}$ has d dimensions, I_i has a size of $K \times d$. Since we only consider frames of a single clip at this stage, we drop i from the notation for simplicity, *i.e.*, $I = I_i, f_k = f_{k,i}$. Following self-attention learning, we transform I into three different tensors, *i.e.*, a query tensor $I_Q = W^Q I$, a key tensor $I_K = W^K I$, and a value tensor $I_V = W^V I$, where the query/key/value tensor is computed for each visual emotion from the clip feature I. We apply self-attention and obtain feature matrix $f\prime$ that captures visual change patterns of facial expressions:

$$f\prime = softmax\left(\frac{I_Q I_K^T}{\sqrt{d}}\right) I_V. \tag{1}$$

Self-attention learning encodes the spatio-temporal information within a clip. However, it only considers frame-level relations without taking into account the global relation between frames and the clip. To address this limitation, we introduce the local–global relation learning to consider the global information of a clip.

Local–global relation learning Fig. 1(b) shows the structure of the Local–global relation learning. We summarize $f\prime$ into a single clip representation $\hat{f}\prime$ through the pooling operation and compute the local–global relation attention via a sample concatenation and a fully-connected (FC) layer as follows:

$$w_k = \sigma\left(\left[\hat{f}\prime : f\prime_k\right]^T q^0\right) \tag{2}$$

where q^0 is the parameter of the FC layer. $f\prime_k$ is the feature of the k^{th} frame and T is the transpose operation. σ is the sigmoid function. Operator : denotes concatenation that integrates frame features into the clip feature. w_k implies the frames that contain more relevant emotion information in a clip or not. We re-scale and aggregate features of each frame to form the new clip-based representation:

$$R_i = \left[\frac{\Sigma_k w_k f\prime_k}{\Sigma_k w_k}\right] q^1 \tag{3}$$

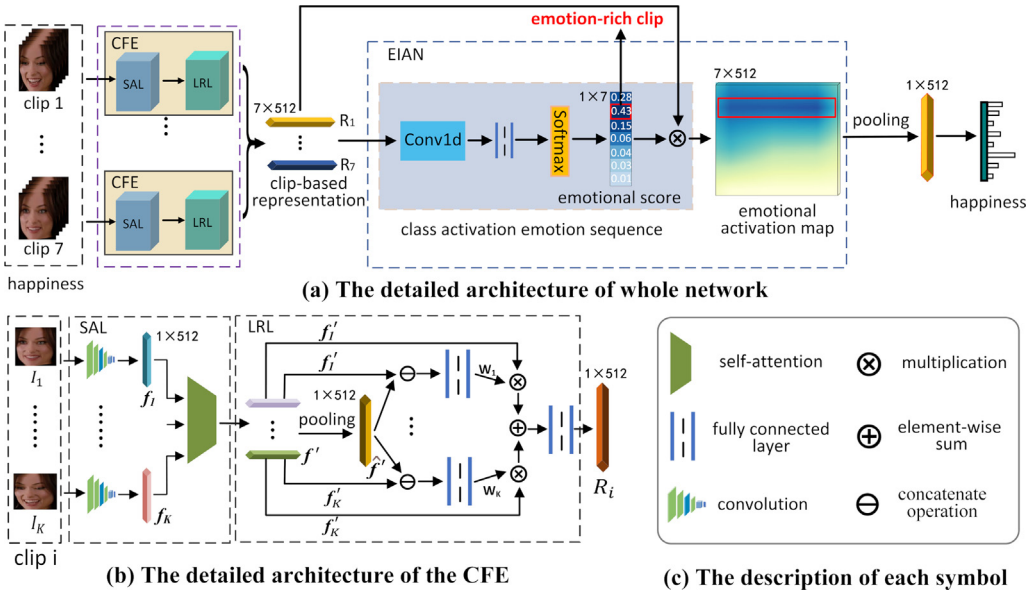

Fig. 1. The overall architecture of our CEFLNet for video-based FER and the structure of CFE..

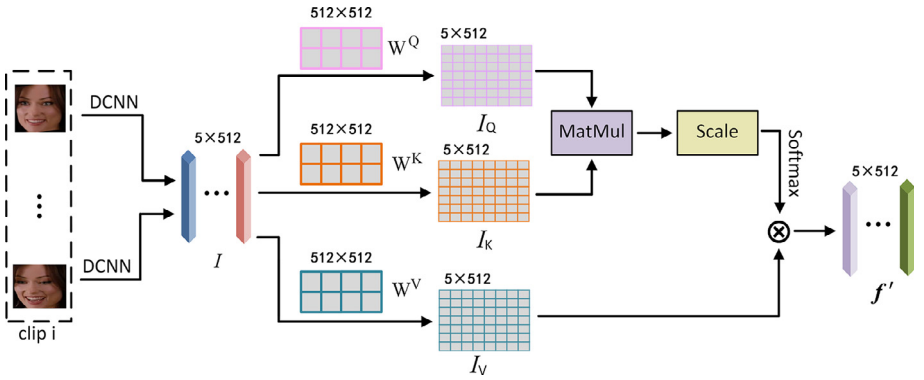

Fig. 2. The structure of the self-attention component. MatMul stands for dot product and Scale stands for scale operation.

where q^1 is the parameter of the FC. The local–global relation attention highlights the more useful visual cues for expression motion in a clip and provides key clip-level features for the following EIAN.

3.3. Weakly supervised EIAN for clip-aware emotion-rich video representation

EIAN identifies the emotional intensity scores of clips and generates emotional activation maps via class activation emotion sequences in a weakly supervised manner. The detailed process for generating the emotional activation map and locating the salient emotion-rich clip is shown in Fig. 3.

The clip-level features are concatenated into a video-level representation V^f

$$V^f = H(R_1, R_2, \ldots, R_n), \tag{4}$$

where $H(\bullet)$ denotes an aggregate operation, n is the number of clips in a video. Inspired by Class Activation Mapping(CAM) [30], we introduce a class activation emotion sequence to generate the emotional activation map by learning the temporal attention of clips. As shown in Fig. 3, the video-level representation V^f is fed to one-dimensional convolutional layers to learn temporal attention. For the attention channels, the results of performing a full-connected layer are $W^T V^f$. Thus, for each video-level expression class y^c, a softmax operation is adopted to identify the emotional intensity scores of clips. The emotional scores A_{y^c} is computed as follows:

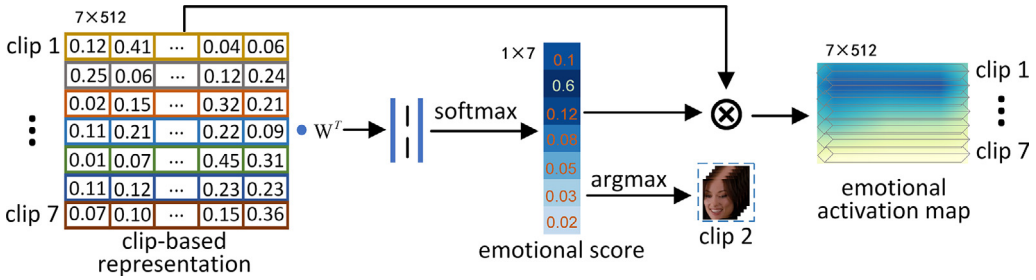

Fig. 3. The detailed process for locating the salient emotion-rich clip and generating an emotional activation map. W^T is a learnable parameter matrix of one-dimensional convolution. Note that darker colors indicate better attention weights, i.e. the current frame contains more emotional information..

$$A_{y^c} = Softmax\left(W^T V^f\right), \tag{5}$$

where W^T is a learnable parameter matrix of one-dimensional convolution.

The emotional scores reflect how much emotional information each clip contains in a video. Unlike the CAM-based bounding box proposals [30], A_{y^c} is a one-dimensional vector of the position of the emotion-rich clips. Hence, we compute the position of the selected emotion-rich clip P_e as follows:

$$P_e = \arg\max\left(A_{y^c}\right), \tag{6}$$

M_c is the emotional activation map of the expression class y^c,

$$M_c = A_{y^c} \cdot V^f, \tag{7}$$

where $\cdot$ represents dot product. M_c gives the importance of the activation at a video temporal sequence leading to the classification of facial expression. The emotion-rich representation $\widehat{V}^f$ of a video is given by:

$$\widehat{V}^f = maxpool(M_c). \tag{8}$$

To classify the emotion-rich representation into facial expression categories, we apply softmax and a fully-connected layer to calculate the probability of facial expressions:

$$p\left(\widehat{Y}_V\right) = softmax\left(\widehat{V}^f q^2\right) \tag{9}$$

$$softmax(Z)_j = \frac{e^{Z_j}}{\sum_{c=1}^{C} e^{Z_c}}, \text{for } j = 1, \ldots, C \tag{10}$$

where $p\left(\widehat{Y}_V\right)$ is the expression category score and q^2 is the parameter vector of the fully-connected layer, Z is the output of the FC layer, C is the number of expression category, and $Softmax(Z)_j$ denotes the probability that the video belongs to the j^{th} expression category.

3.4. Objective function

The objective of CEFLNet has two parts: the CFE guarantees high-quality emotional representations of clips, and EIAN focuses on the emotion-rich features relevant to facial expressions via weakly supervised learning. In our study, only a video-level FER classification loss L_{class} is used to optimize the two objectives of the entire network. Our FER classification loss L_{class} is as follows:

$$L_{class} = -\sum_{\mathcal{V}} Y_{\mathcal{V}} log\left[p\left(\widehat{Y}_{\mathcal{V}}\right)\right] + (1 - Y_{\mathcal{V}}) log\left[1 - p\left(\widehat{Y}_{\mathcal{V}}\right)\right], \tag{11}$$

where $Y_{\mathcal{V}}$denotes the facial expression label for each video, $\mathcal{V}$ indexes a training video, and $p\left(\hat{Y}_{\mathcal{V}}\right)$ denotes the probabilities of facial expressions predicted by the CEFLNet.

4. Experimental Results and Discussion

4.1. Datasets and Implementation Details

To evaluate our method, four video-based face expression datasets were used in our experiments, including BU-3DFE dataset [31], MMI dataset [32], AFEW 8.0 dataset [20], and DFEW dataset [33].

BU-3DFE [31]: The 3D facial expressions are captured at a video rate (25 frames per second). Six emotion labels are included, *i.e.*, anger, disgust, happiness, fear, sadness, and surprise. Each expression sequence contains about 100 frames. BU-3DFE contains 606 3D facial expression sequences captured from 101 subjects, with a total of approximately 60,600 frames. In this study, a 10-fold validation was conducted.

MMI [32]: A total of 205 deliberate expression sequences with frontal faces were collected from 30 subjects. The expression sequences were recorded at a temporal resolution of 24 fps. Each expression sequence of the dataset was labeled with one of the six basic expression classes (*i.e.*, anger, disgust, fear, happiness, sadness, and surprise). The expression sequences were collected such that, the first frame in the sequence was the onset frame and the last frame was the offset frame. In this study, a 10-fold validation was conducted.

AFEW [20]: The AFEW has served as an evaluation platform for the annual EmotiW since 2013. Seven emotion labels are included in AFEW, *i.e.*, anger, disgust, fear, happiness, sadness, surprise, and neutral. AFEW contains videos collected from different movies and TV serials with spontaneous expressions, various head poses, occlusions, and illuminations. AFEW is divided into three splits: Train (738 videos), Val (352 videos), and Test (653 videos). Because we do not have test labels for evaluation, we follow the setting of other compared methods and only used the Training/Val set for experiments.

DFEW [33]: The DFEW is a large-scale unconstrained dynamic facial expression database, containing 16,372 video clips extracted from over 1,500 different movies. It contains 12,059 single-label video clips and also includes seven emotion labels, *i.e.*, anger, disgust, fear, happiness, sadness, surprise, and neutral. DFEW dataset provides five data division methods. Hence, a 5-fold validation was used. Examples of these datasets are shown in Fig. 4.

We kept each video to 105 frames via interpolation and clipping. The face regions are detected using Retinaface [34] and the size of each face is resized to 224×224. A randomly selected frame within the first 30 frames was used as the starting frame and the following 75 consecutive frames were extracted. We split the 75 frames into seven sub-Videos, each of which had 15 frames, with five frames overlapping between each sub-video. To reduce the computation cost, five frames were randomly sampled from each sub-video to form a new expression clip. We conducted a 10-fold validation on BU-3DFE and MMI datasets, a 5-fold validation on the DFEW dataset, and used the training and validation sets for the experiments on the AFEW dataset.

Our method is implemented using Pytorch. The training parameters include initial learning rate (0.0001), cosine annealing schedule to adjust the learning rate, mini-batch size (8), and warm-up. The experiments were conducted on a PC with Intel(R) Xeon(R) Gold 6240C CPU at 2.60 GHz and 128 GB memory, and NVIDIA GeForce RTX 3090 GPU. The key parameters used in training the network are given in Table 1.

4.2. Performance Analysis and Comparison Study

Fig. 5(a) shows the confusion matrix of our method using the BU-3DFE dataset. Among the six expressions, the highest accuracy is 100% (Surprise), while the lowest accuracy is 70.0% (Fear), which has the least amount of facial expression and is difficult to distinguish from the other expressions. The average accuracy of facial expression recognition is 85.33% with a standard deviation of 3.29 for the BU-3DFE dataset. Fig. 5(b) depicts the confusion matrix of our method for processing the MMI dataset. Among the four datasets, our method achieved the best accuracy for predicting facial expressions from the MMI dataset. The proposed method achieved an average accuracy of 91% with a standard deviation of 4.36. For four out of six expressions, including Fear, Happiness, Sadness, and Surprise, we achieved 100% accuracy. There exist a slight confusion between Anger and Disgust expressions and the average accuracy of these two expressions is 83%.

Fig. 5(c) shows the confusion matrix from the AFEW dataset. AFEW is one of the most challenging datasets and great confusion exists among expressions including Disgust, Fear, Sadness, and Surprise. The average accuracy of our method is at 53.98% with a standard deviation of 0.4 and the highest accuracy is 87% for Neutral. The accuracy of Happiness and Anger are 83% and 82%, respectively. Disgust and Fear are the two most confusing expressions in this dataset [35,11]. Fig. 5(d) shows the confusion matrix from the large-scale DFEW dataset. The average accuracy of our method is 65.35% with a standard deviation of 1.13. The highest accuracy is 84% of Happiness followed by Anger and Sadness, the accuracy of which is at 70% and 68%, respectively. Similar to the AEFW dataset, the most confusing expressions include Disgust and Fear. This could be attributed to the extreme imbalance of the category in the DFEW (only occupies 1.22% in the DFEW dataset) [36].

Comparison study (BU-3DFE): We compare our CEFLNet with the state of the arts, including FERAtt + Rep + Cls [37], FAN [16], DeRL [8], C3D [38], ICNP [39], and C3D-LSTM [40]. The dataset used in our comparison study is BU-3DFE. Table 2 report the average accuracy and the feature settings of the methods. The best and second-best results are highlighted with bold font and underscore, respectively. The accuracy of CEFLNet is better than both sequence-based and frame-based methods. Compared to the best sequence-based result, the proposed CEFLNet improved the accuracy by 2.13%. This demonstrates that our method discovers the more informative emotion-related cues by modeling the emotion transition relation in videos.

Fig. 4. Some samples from these four datasets. (a) BU3D, (b) MMI, (c) AFEW, (d) DFEW. The most emotional frames are highlighted with red boxes.

Table 1
The Key parameters in training the network.

Parameters	Settings
Optimizer	ADAM
Init learning rate	0.0001
weight decay	0.0001
Maximum number of iterations	160
Mini-batch size	8
Epoch	120
The number of clips per video	7
The number of frames per clip	5

Comparison study (MMI): In comparison with the state-of-the-art video-based FER methods, Table 3 lists the average accuracy on MMI dataset using frame-based methods (i.e., AUDN [41], DeRL [8], WMDCNN [42] and CER [7], sequence-based methods (i.e., LSTM [13], Deep generative-contrastive networks (DGCN) [9], LPQ-TOP + SRC [6], SAANet [43], and WMCNN-LSTM [42]) and our CEFLNet. The proposed method achieved an average accuracy of 91% with a standard deviation of 4.36, which outperformed existing state-of-the-art FER methods. Compared to the second best method, WMCNN-LSTM [42], the CEFLNet improved the accuracy by 3.9%.

Comparison study (AFEW): Table 4 compares the average accuracy of FER using AFEW dataset. For a fair comparison, we only list these results obtained by the best single models in previous works. Both [44,45] input two LBP maps and a gray image for CNN models. Deeply supervised networks are used in [45,15], which add supervision on intermediate layers. For clip-based methods, [35] uses DenseNet-161 and pre-trains it on both large-scale face datasets and their own Situ emotion video dataset. Additionally, [35] applies complicated post-processing which extracts frame features and computes their mean vector, max-pooling vector, and standard deviation vector. These vectors are then concatenated and finally fed into an SVM classifier. Overall, our CEFLNet improves the baseline (about 2.45%) and achieves performance comparable to that of the best previous single model. It demonstrates that our method achieves the best performance with great robustness, meanwhile, has obvious advantages over other algorithms on the in-the-wild expression dataset.

Comparison study (DFEW): The results in Table 5 show that our method is still far superior to other algorithms. More detailed comparison results can be shown in Table 5. Compared to the state-of-the-art methods reported in [33], the FER accuracy of our CEFLNet achieved significant improvement (over 8.84%) on the challenging large-scale dataset.

4.3. Ablation Study and Analysis

4.3.1. Analysis of Network Components

To analyze the contribution to the learning capability by the components of CEFLNet, Table 6 presents the results of our ablation study that looks into the impact of gradual addition of the self-attention learning, local–global relation learning, and

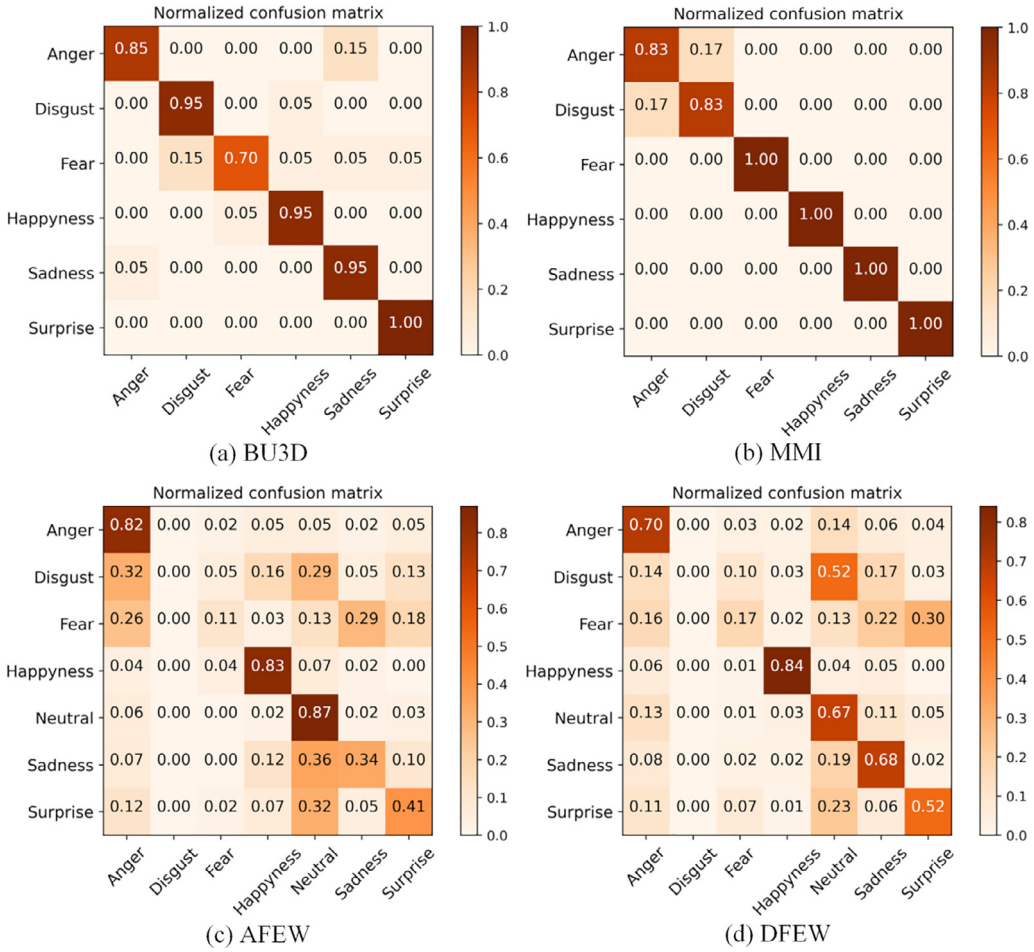

Fig. 5. The confusion matrix of our method using the four datasets.

Table 2
FER accuracy on the BU-3DFE dataset. The best result is highlighted in bold.

Methods	Feature setting	Accuracy(%)
FERAtt + Rep + Cls [37]	frame-based	82.11
FAN [16]	frame-based	84.17
DeRL [8]	peak frame-based	84.17
C3D [38]	sequence-based	75.83
C3D-LSTM [40]	sequence-based	79.17
ICNP [39]	sequence-based	83.20
CEFLNet	clip-based	**85.33**

EIAN training components to the baseline framework (ResNet-18). The training and testing datasets used in this study are BU-3DFE.

ResNet-18 and CNN-LSTM achieved an average accuracy of 62.77% and 79.17%, respectively. In our method, we used SAL to learn frame relation and achieved average recognition accuracy of 84.17%. By adding LRL to the network, the performance was improved by 0.5%, which shows that the local–global relation learning module can better learn the potential relationship between each frame and clip. Note that the integration of EIAN improved the FER accuracy by 0.66%. This demonstrates that the EIAN module learns the emotional intensity from clip-based representations and obtains more distinguishable emotion-rich video features.

Table 3
FER accuracy on the MMI dataset. The best result is highlighted in bold.

Methods	Feature setting	Accuracy(%)
DeRL [8]	frame-based	73.23
WMDCNN [42]	frame-based	78.2
CER [7]	peak frame-based	70.12
AUDN [41]	peak frame-based	75.85
LPQ-TOP + SRC [6]	sequence-based	64.11
LSTM [13]	sequence-based	78.61
DGCN [9]	sequence-based	81.53
WMCNN-LSTM [42]	sequence-based	87.10
SAANet [43]	sequence-based	87.06
CEFLNet	clip-based	**91.00**

Table 4
FER accuracy on AFEW 8.0 dataset. The highest result is highlighted in bold.

Methods	Feature setting	Accuracy(%)
HoloNet [44]	frame-based	44.57
DSN–HoloNet [45]	frame-based	46.47
DSN-VGGFace [15]	frame-based	48.04
FAN [16]	frame-based	51.18
C3D [38]	sequence-based	30.11
VGG16 + TP + SA [46]	sequence-based	49.00
Emotion-BEEU [47]	sequence-based	52.49
DenseNet-161 [35]	clip-based	51.44
CEFLNet	clip-based	**53.98**

Table 5
FER accuracy on DFEW dataset. The highest result is highlighted in bold.

Methods	Feature setting	Accuracy(%)
C3D,EC-STFL [33]	sequence-based	55.50
R3D18,EC-STFL [33]	sequence-based	56.19
VGG11 + LSTM,EC-STFL [33]	sequence-based	56.25
P3D,EC-STFL [33]	sequence-based	56.48
3D ResNet-18,EC-STFL [33]	sequence-based	56.51
CEFLNet	clip-based	**65.35**

Table 6
Ablation study of the proposed CEFLNet. The best results are in bold.

Methods	SAL	LRL	EIAN	Acc(%)
ResNet-18				62.77
CNN-LSTM				79.17
+ SAL	✓			84.17
+ LRL	✓	✓		84.67
+ EIAN	✓	✓	✓	**85.33**

4.3.2. Emotion-rich Clips

Fig. 6 shows the emotional activation maps and clip selection on the four datasets. The orange boxes depict the select emotion-rich clips in videos. It can be seen that the emotion-rich clips have the greatest expression intensity than other clips, which implies that EIAN identifies the salient emotion-rich clip and performs emotional activation according to the emotional intensity of each clip.

In addition, we evaluated the accuracy of emotion-rich clip selection on the four datasets, as shown in Table 7. The proposed MIAN achieved an accuracy of 67.57% on the MMI dataset and achieved an accuracy of 45% on the challenging AFEW dataset. This demonstrates that the EIAN method effectively locates the emotion-rich clip in the untrimmed videos.

We visualized the expression features with different settings in a 2D feature space by using the t-SNE on the four datasets. The visualizations include the following four cases: clip-aware emotion-rich representations by the CEFLNet (see Fig. 7(a)), video attention features extracted by FAN [16] (see Fig. 7(b)), sequence-based video features extracted by LSTM [48] (see

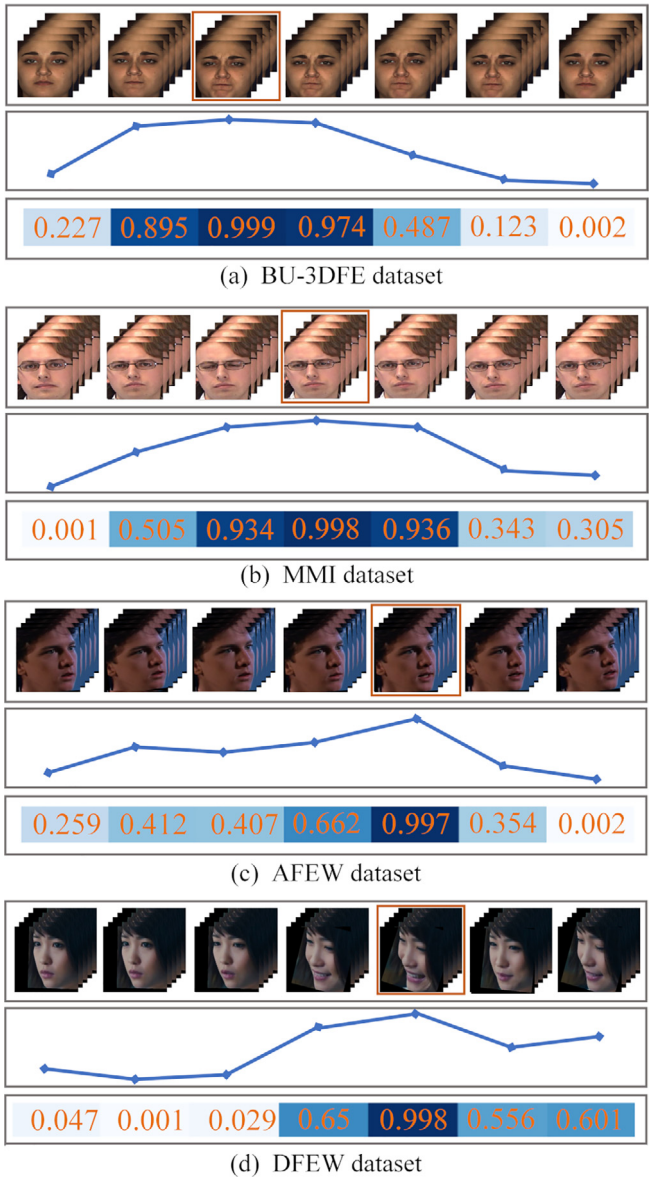

Fig. 6. The emotional activation maps and the located clips (highlighted with orange boxes). Darker colors indicate greater attention weights, i.e., more emotional information..

Table 7
The accuracy of emotion-rich clip locating.

Dataset	BU3D	MMI	AFEW	DFEW
Accuracy(%)	55.83	67.57	45.00	47.65

Fig. 7(c)), frame-based features extracted by ResNet-18[49] (see Fig. 7(d)). Obviously, compared to the features shown in Fig. 7(b), Fig. 7c) and Fig. 7(d), the clip-aware emotion-rich features proposed in this study can significantly be separated according to facial expression categories. It is evident that the proposed CEFLNet can learn more expressive and discriminative representations for video-based FER on the four datasets.

We studied the impact of the number of clips per video and the number of frames per clip on the accuracy of FER. As shown in Fig. 8(a), all four datasets achieved the highest accuracies when the number of clips is 7, and achieved the lowest accuracies when the number of clips is 1. Results show that too many or too few clips are detrimental to the performance of facial expression recognition. As shown in Fig. 8(b), the highest accuracy is achieved when we set the number of frames of

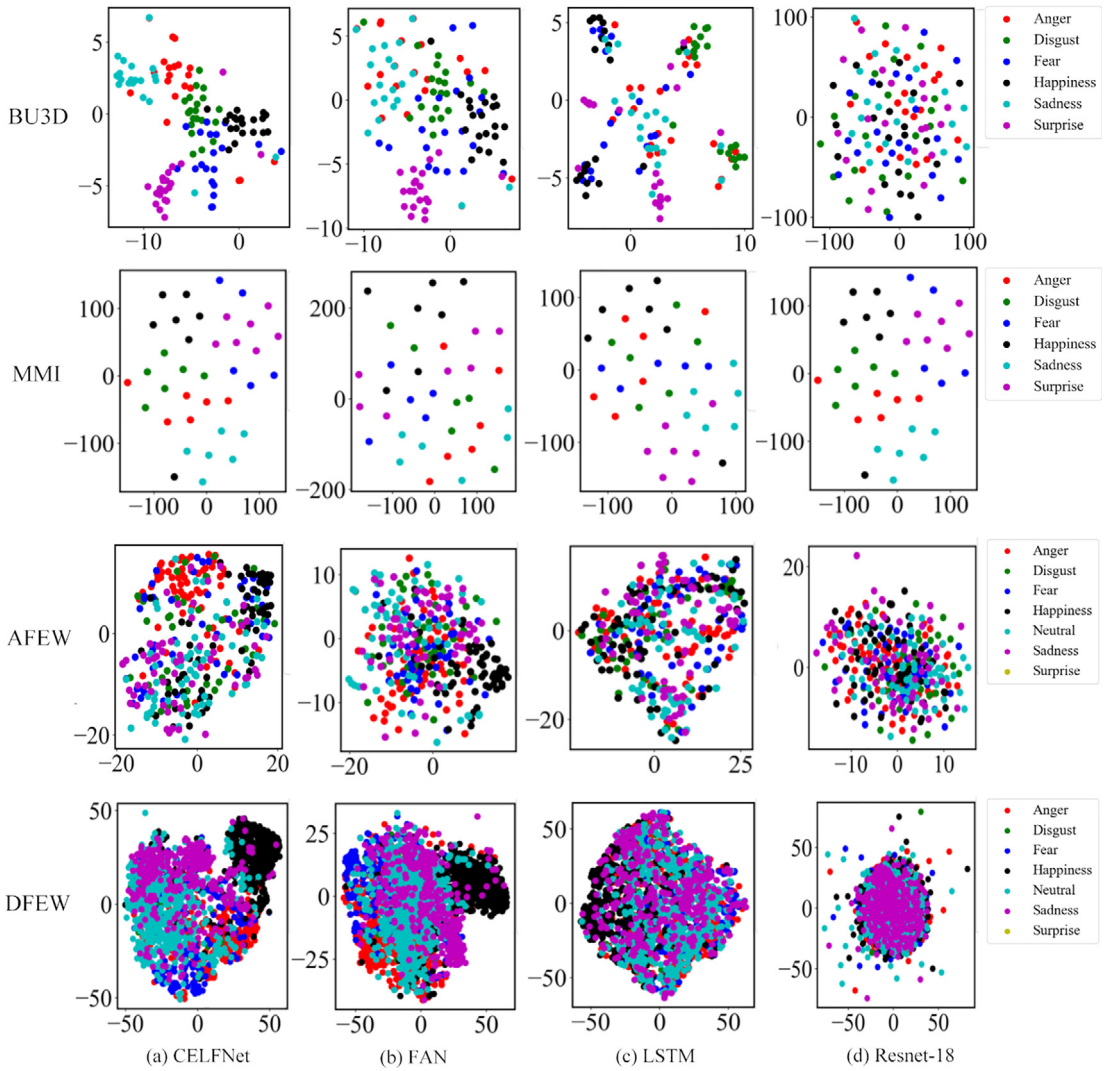

Fig. 7. The t-SNE feature visualization of different representations in 2D space. (a) Clip-aware emotion-rich representations by CEFLNet, (b) video attention features by FAN, (c) sequence-based video features by LSTM, (d) frame-based features by ResNet18.

each clip to 5. When this number is less than 5, the accuracy drops. The performance drop might be a result of emotional information lost. When the number of frames is 15, redundant expressionless frames cause expression inconsistency and hence reduce recognition accuracy. In our experiments, we keep the number of clips of each video to 7 and the number of frames of each clip to 5.

4.4. Computational Complexity

Table 8 reports model parameters and computational cost of the three spatio-temporal learning methods in processing the BU-3DFE dataset. We use Multiply–Accumulate Operations(MACs) [1] to measure the computational cost. Our CEFLNet resulted in the best performance (FER accuracy of 85.33%) with the least computational cost (63.8G) and parameters (12.83 M) among the compared methods, which demonstrates that the proposed method exhibits improved accuracy and efficiency.

Table 9 lists the average accuracy and the computation cost with respect to the number of frames. Clearly, when less number of frames are used, the computational cost is lower. However, the best accuracy is achieved when the number of frames is 5. Hence, to balance speed and accuracy, a five-frame per clip is a proper choice.

[1] https://github.com/sovrasov/flops-counter.pytorch

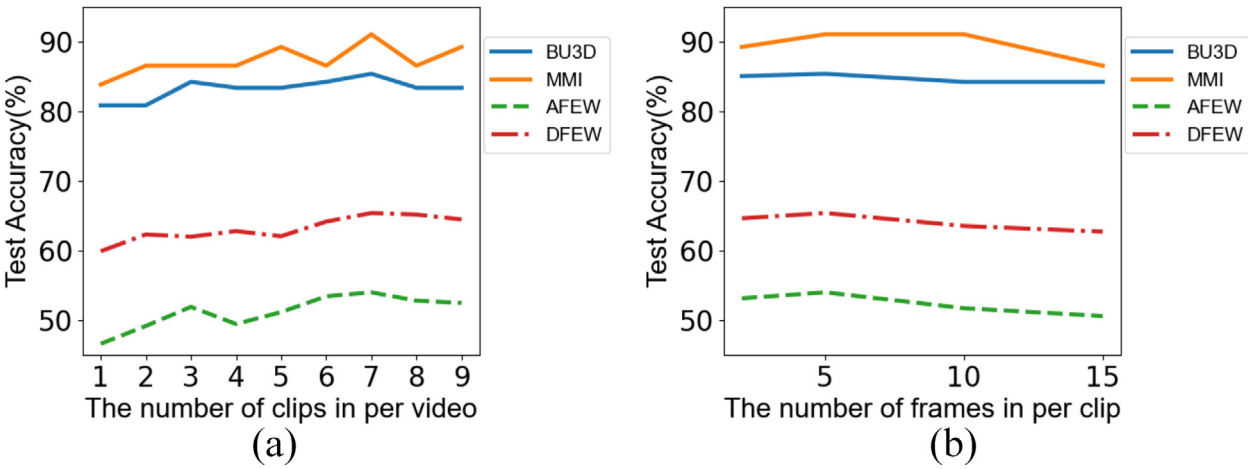

Fig. 8. The accuracy of the number of frames per clip and the number of clips per video for FER on four datasets. (a) The effect of the number of expression clips, (b) the effect of the number of frames..

Table 8
Comparison of model complexity and efficiency.

Method	Backbone	Params(M)	MACs(G)	Acc(%)
C3D	C3D	79.99	326.41	75.83
C3D-LSTM	C3D	110.24	282.26	79.17
CEFLNet	ResNet-18	**12.83**	**63.80**	**85.33**

Table 9
The effect of the number of frames on the computation cost and classification accuracy.

# of frames	MACs(G)	Acc(%)
2	**25.52**	85
5	63.80	**85.33**
10	127.59	84.17
15	191.39	84.17

5. Conclusion and Future Work

In this paper, we propose an effectively clip-aware emotion-rich feature learning network to jointly identify the emotion-rich clips and recognize dynamic facial expressions in a video. CEFLNet decomposes a video into several small video clips and extracts the clip-level spatio-temporal features via two-cascaded self-attention and local–global relation learning within each video clip. Our method generates an emotional activation map that is used to identify salient emotion clips for clip-aware emotion-rich representations. Our proposed method requires no clip-wise or frame-wise annotations for training the model and can be trained in an end-to-end manner.

Experiments were conducted using four public video datasets, namely the BU-3DFE, MMI, AFEW, and DFEW. Due to suppressing the redundancy information from expression-irrelevant clips, the proposed method was found to achieve a much-improved performance for video-based FER, with great robustness and efficiency; the highest accuracy for each of these datasets was 85.33%, 91%, 53.98%, and 65.35%. In our future work, we plan to study self-supervised learning to model the extraction of key information from complex facial video sequences with multiple expressions.

CRediT authorship contribution statement

Yuanyuan Liu: Conceptualization, Methodology, Writing - original draft, Writing. **Chuanxu Feng:** Methodology, Software, Writing - original draft. **Xiaohui Yuan:** Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing. **Lin Zhou:** Data curation, Investigation, Resources. **Wenbin Wang:** Validation, Investigation. **Jie Qin:** Validation, Resources. **Zhongwen Luo:** Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by a National Natural Science Foundation of China grant (62076227) and Wuhan Applied Fundamental Frontier Project under Grant (2020010601012166).

References

[1] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural network-driven feature learning method for multi-view facial expression recognition, IEEE Transactions on Multimedia 18 (12) (2016) 2528–2536.
[2] J. Wu, Z. Lin, W. Zheng, H. Zha, Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition, Neurocomputing 239 (2017) 143–152.
[3] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Transactions on Affective Computing 01 (2020), 1–1.
[4] M.-W. Huang, Z.-w. Wang, Z.-L. Ying, A new method for facial expression recognition based on sparse representation plus lbp, in: 2010 3rd International Congress on Image and Signal Processing, Vol. 4, IEEE, 2010, pp. 1750–1754..
[5] Z. Wang, Z. Ying, Facial expression recognition based on local phase quantization and sparse representation, in: 2012 8th International Conference on Natural Computation, Springer, 2012, pp. 222–225.
[6] B. Jiang, M. Valstar, B. Martinez, M. Pantic, A dynamic appearance descriptor approach to facial actions temporal modeling, IEEE Transactions on Cybernetics 44 (2) (2013) 161–174.
[7] S.H. Lee, W.J. Baddar, Y.M. Ro, 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.
[8] H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2168–2177.
[9] Y. Kim, B. Yoo, Y. Kwak, C. Choi, J. Kim, Deep generative-contrastive networks for facial expression recognition, arXiv preprint arXiv:1703.07140..
[10] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolution neural network enhanced random forest for facial expression recognition, Pattern Recognition 84 (2018) 251–261.
[11] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 569–576.
[12] T. Chen, H. Yin, X. Yuan, Y. Gu, F. Ren, X. Sun, Emotion recognition based on fusion of long short-term memory networks and svms, Digital Signal Processing 117 (2021) 103153.
[13] D.H. Kim, W.J. Baddar, J. Jang, Y.M. Ro, 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.
[14] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 445–450.
[15] Y. Fan, J.C. Lam, V.O. Li, Video-based emotion recognition using deeply-supervised neural networks, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 584–588.
[16] D. Meng, X. Peng, K. Wang, Y. Qiao, Frame attention networks for facial expression recognition in videos, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 3866–3870.
[17] B. Knyazev, R. Shvetsov, N. Efremova, A. Kuharenko, Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video, arXiv preprint arXiv:1711.04598..
[18] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, S. Yan, Peak-piloted deep network for facial expression recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 425–442.
[19] Z. Yu, Q. Liu, G. Liu, Deeper cascaded peak-piloted network for weak expression recognition, The Visual Computer 34 (12) (2018) 1691–1699.
[20] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, 2015, pp. 423–426..
[21] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–2991.
[22] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4438–4446.
[23] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, 2015, pp. 2048–2057..
[25] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3219–3228.
[26] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99..
[27] X. Yuan, Z. Qiao, A. Meyarian, Scale attentive network for scene recognition, Neurocomputing, Dec. 2021, in press.
[28] H. Zheng, J. Fu, T. Mei, J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5209–5217.
[29] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian, M. Savvides, Deepgender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 68–77..
[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
[31] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3d facial expression database for facial behavior research, in: 7th international conference on automatic face and gesture recognition (FGR06), IEEE, 2006, pp. 211–216..
[32] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Paris, France, 2010, p. 65.
[33] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, J. Liu, Dfew: A large-scale database for recognizing dynamic facial expressions in the wild, in: Proceedings of the 28th ACM International Conference on Multimedia (MM), 2020, pp. 2881–2889..
[34] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5203–5212..
[35] C. Liu, T. Tang, K. Lv, M. Wang, Multi-feature based emotion recognition for video clips, ACM ICMI (2018) 630–634..
[36] X. Yuan, M. Abouelenien, M. Elhoseny, A boosting-based decision fusion method for learning from large, imbalanced face data set, in: Quantum Computing: An Environment for Intelligent Large Scale Real Application, Springer, Cham, 2018, pp. 433–448..
[37] P.D. Marrero Fernandez, F.A. Guerrero Pena, T. Ren, A. Cunha, Feratt: Facial expression recognition with attention net, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 837–846.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
[39] Q. Zhen, D. Huang, Y. Wang, L. Chen, Muscular movement model-based automatic 3d/4d facial expression recognition, IEEE Transactions on Multimedia 18 (7) (2016) 1438–1450.
[40] P. Parmar, B. Tran Morris, Learning to score olympic events, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 20–28.
[41] M. Liu, S. Li, S. Shan, X. Chen, Au-inspired deep networks for facial expression feature learning, Neurocomputing 159 (2015) 126–136.
[42] H. Zhang, B. Huang, G. Tian, Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture, Pattern Recognition Letters 131 (2020) 128–134.
[43] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, S. Wen, Saanet: Siamese action-units attention network for improving dynamic facial expression recognition, Neurocomputing 413 (2020) 145–157.
[44] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, Y. Chen, Holonet: towards robust emotion recognition in the wild, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 472–478.
[45] P. Hu, D. Cai, S. Wang, A. Yao, Y. Chen, Learning supervised scoring ensemble for emotion recognition in the wild, in: Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 553–560.
[46] M. Aminbeidokhti, M. Pedersoli, P. Cardinal, E. Granger, Emotion recognition with spatial attention and temporal softmax pooling, in: International Conference on Image Analysis and Recognition, Springer, 2019, pp. 323–331.
[47] V. Kumar, S. Rao, L. Yu, Noisy student training using body language dataset improves facial expression recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2020, pp. 756–773.
[48] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, 2015, pp. 802–810..
[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.