



Clip-aware Expressive Feature Learning for Video-based Facial Expression Recognition

Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, Zhongwen Luo



PII: S0020-0255(22)00281-X

DOI: <https://doi.org/10.1016/j.ins.2022.03.062>

Reference: INS 17405

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

ScienceDirect

To appear in: *Information Sciences*

Received Date: 27 September 2021

Revised Date: 11 March 2022

Accepted Date: 19 March 2022

Please cite this article as: Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, Z. Luo, Clip-aware Expressive Feature Learning for Video-based Facial Expression Recognition, *Information Sciences* (2022), doi: <https://doi.org/10.1016/j.ins.2022.03.062>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## Clip-aware Expressive Feature Learning for Video-based Facial Expression Recognition

Yuanyuan Liu<sup>a</sup>, Chuanxu Feng<sup>a</sup>, Xiaohui Yuan<sup>b,\*</sup>, Lin Zhou<sup>a</sup>, Wenbin Wang<sup>a</sup>, Jie Qin<sup>c</sup>, Zhongwen Luo<sup>a</sup>

<sup>a</sup>School of Information Engineering, China University of Geosciences, Wuhan, China

<sup>b</sup>Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

<sup>c</sup>Wuhan Huazhong Numerical Control Co., Ltd.

### 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 ex-

---

\*Corresponding author

Email addresses: liuyy@cug.edu.cn (Yuanyuan Liu), fcxfc@cug.edu.cn (Chuanxu Feng), xiaohui.yuan@unt.edu (Xiaohui Yuan), zhoulin@cug.edu.cn (Lin Zhou), wangwenbin@cug.edu.cn (Wenbin Wang), toqingjie@126.com (Jie Qin), luozw@cug.edu.cn (Zhongwen Luo)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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.

*Keywords:* Video-based FER, emotional activation map, clip-based feature encoder, clip-aware emotion-rich representation

---

## 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.

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

25 overwhelmingly high computation complexity to model video facial expression  
26 movements [3, 16], and the presence of many frames irrelevant to expressions  
27 makes the learned features suboptimal to FER [3].

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

40 The contributions of this paper include the following:

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

53 The remainder of this paper is organized as follows: Section 2 introduces

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

54 related work in video-based FER. Section 3 presents the proposed CEFLNet for  
55 video-based FER in detail. Section 4 discusses the experimental results on four  
56 publicly available datasets. Finally, this paper is concluded in Section 5 with a  
57 summary and future work.

## 58 2. Related work

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

78 The dynamic sequence-based method takes the entire video sequence as in-  
79 put and uses the texture information and temporal dependence in the frame  
80 sequence to recognize facial expressions [3, 9, 11, 13, 21]. Vielzeuf *et al.* [11]  
81 used pre-trained VGG-Face to extract spatial features, then utilized an LSTM  
82 layer to encode temporal dependencies in the sequence. Kim *et al.* [13] propose

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

83 a new spatio-temporal representation learning for FER by integrating C3D and  
84 LSTM networks, which is robust to expression intensity variation. In [21], a  
85 temporal geometric feature was proposed to improve the discriminative capacity  
86 of the learned spatio-temporal appearance features. Although these dynamic-  
87 based networks capture spatio-temporal features for FER, they still challenge in  
88 describing expression movements in untrimmed videos and require large model  
89 capacities to model facial expression changes in videos.

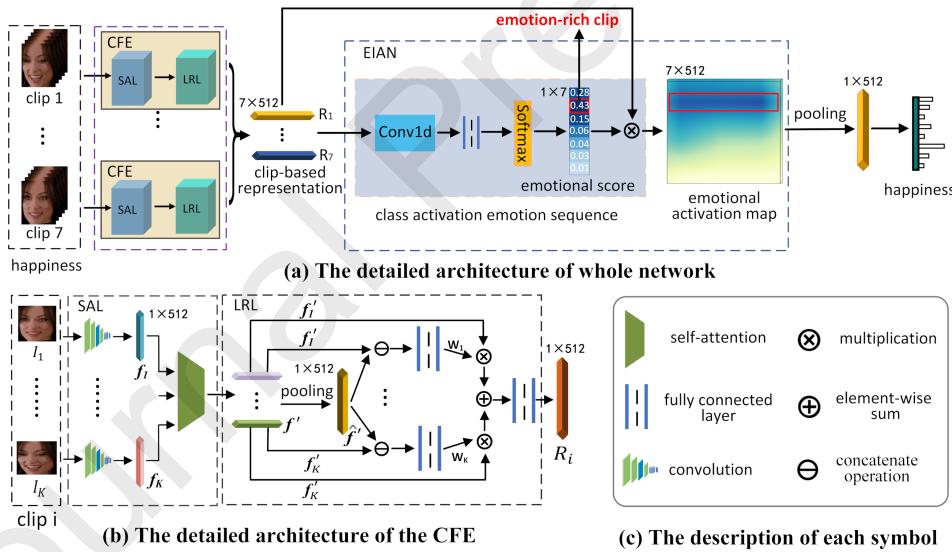
90 **Attention model** Visual attention based networks have been proposed  
91 to localize significant regions for many computer vision tasks, including fine-  
92 grained recognition [22, 23], image captioning [24], person re-identification [25],  
93 and object detection [26, 27]. Some methods are learned by the aggregating  
94 scheme from the internal hidden representations in CNN [28]. Other methods  
95 focus on detecting local regions according to supervised bounding box annota-  
96 tion, e.g., region proposal network (RPN) [26]. Zheng *et al.* [28] adopted channel  
97 grouping sub-network to cluster different convolutional feature maps into groups  
98 according to peak responses of maps. Xu *et al.* [29] proposed an attention shift  
99 based on multiple blur levels to avoid occlusions for facial gender classification.  
100 SE-Net [23] proposed the Squeeze-and-Excitation (SE) block that re-calibrates  
101 channel-wise feature responses by explicitly modeling the inter-dependency be-  
102 tween channels. The SE block results in considerable performance improvement  
103 for image classification with minor additional computational costs. Meng *et*  
104 *al.* [16] proposed a frame attention network (FAN) for selecting frames from  
105 a video to form a discriminative video-level representation. Although attention  
106 has been successfully employed in many computer vision tasks, it is difficult to  
107 directly use it for capturing beneficial expression movements in videos due to  
108 the vastly present irrelevant frames and the limited motion variation.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

### 109 3. Clip-aware emotion-rich feature learning network

#### 110 3.1. Network Architecture

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



#### 121 3.2. CFE for clip-level representation

122 The clip-based feature encoder contains two cascaded attention learning  
 123 modules: self-attention learning (SAL) and local-global relation learning (LRL).

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

124 Figure 1(b) shows the detailed structure of the CFE. In practice, SAL models  
 125 frame-level relation to obtain the self-attention in each clip, and LRL improves  
 126 the clip-level representation by learning local-global relation attention. Through  
 127 the two-cascaded attention learning, the CFE exploits the emotional cues of  
 128 spatio-temporal information in each clip.

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

$$f' = \text{softmax}\left(\frac{I_Q I_K^T}{\sqrt{d}}\right) I_V. \quad (1)$$

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

149     **Local-global relation learning** Fig 1(b) shows the structure of the Local-  
 150 global relation learning. We summarize  $f'$  into a single clip representation  $\hat{f}'$   
 151 through the pooling operation and compute the local-global relation attention

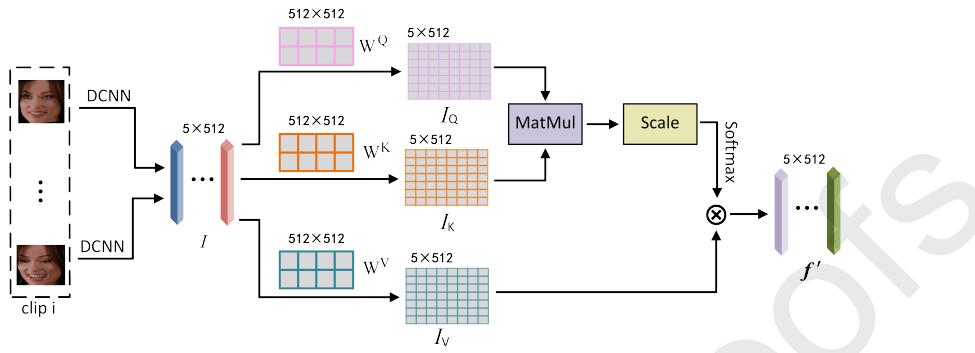


Figure 2: The structure of the self-attention component. MatMul stands for dot product and Scale stands for scale operation.

152 via a sample concatenation and a fully-connected layer (FC) as follows:

$$w_k = \sigma([\hat{f}' : f'_k]^T q^0) \quad (2)$$

30  
31 153 where  $q^0$  is the parameter of the FC.  $f'_k$  is the feature of the  $k^{th}$  frame and  
32 154  $T$  is the transpose operation.  $\sigma$  is the sigmoid function. Operator  $:$  denotes  
33 155 concatenation that integrates frame features into the clip feature.  $w_k$  implies  
34 156 the frames that contain more relevant emotion information in a clip or not.  
35 157 We re-scale and aggregate features of each frame to form the new clip-based  
36 158 representation:

$$R_i = [\frac{\sum_k w_k f'_k}{\sum_k w_k}] q^1 \quad (3)$$

40  
41 159 where  $q^1$  is the parameter of the FC. The local-global relation attention high-  
42 160 lights the more useful visual cues for expression motion in a clip and provides  
43 161 key clip-level features for the following EIAN.

### 44 162 3.3. Weakly supervised EIAN for clip-aware emotion-rich video representation

45 163 EIAN identifies the emotional intensity scores of clips and generates emo-  
46 164 tional activation maps via class activation emotion sequences in a weakly su-  
47 165 pervised manner. The detailed process for generating the emotional activation  
48 166 map and locating the salient emotion-rich clip is shown in Figure 3.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

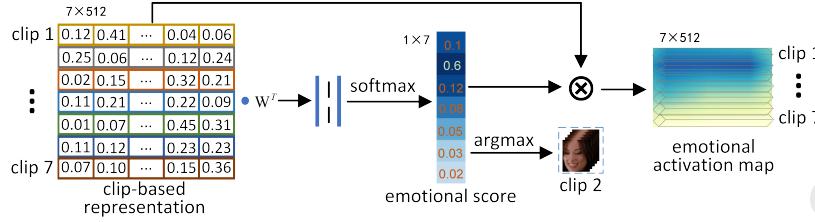


Figure 3: The detailed process for locating the salient emotion-rich clip and generating an emotional activation map.  $W^T$  is a learnable parameter matrix of one-dimensional convolution. Note that darker colors indicate better attention weights, i.e. the current frame contains more emotional information.

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

$$V^f = H(R_1, R_2, \dots, R_n), \quad (4)$$

168 where  $H(\bullet)$  denotes an aggregate operation,  $n$  is the number of clips in a video.

169 Inspired by Class Activation Mapping(CAM) [30], we introduce a class activa-  
170 tion emotion sequence to generate the emotional activation map by learning the  
171 temporal attention of clips. As shown in Figure 3, the video-level representation  
172  $V^f$  is fed to one-dimensional convolutional layers to learn temporal attention.  
173 For the attention channels, the results of performing a full-connected layer are  
174  $W^T V^f$ . Thus, for each video-level expression class  $y^c$ , a softmax operation is  
175 adopted to identify the emotional intensity scores of clips. The emotional scores  
176  $A_{y^c}$  is computed as follows:

$$A_{y^c} = \text{Softmax}(W^T V^f), \quad (5)$$

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

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

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

<sup>182</sup>  $M_c$  is the emotional activation map of the expression class  $y^c$ ,

$$M_c = A_{y^c} \cdot V^f, \quad (7)$$

<sup>183</sup> where  $\cdot$  represents dot product.  $M_c$  gives the importance of the activation at a  
<sup>184</sup> video temporal sequence leading to the classification of facial expression. The  
<sup>185</sup> emotion-rich representation  $\hat{V}^f$  of a video is given by:

$$\hat{V}^f = \text{maxpool}(M_c). \quad (8)$$

<sup>186</sup> To classify the emotion-rich representation into facial expression categories,  
<sup>187</sup> we apply softmax and a fully-connected layer to calculate the probability of  
<sup>188</sup> facial expressions:

$$p(\hat{Y}_V) = \text{softmax}(\hat{V}^f q^2) \quad (9)$$

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

<sup>190</sup> where  $p(\hat{Y}_V)$  is the expression category score and  $q^2$  is the parameter vector of  
<sup>191</sup> the fully-connected layer,  $Z$  is the output of the FC layer,  $C$  is the number of  
<sup>192</sup> expression category, and  $\text{Softmax}(Z)_j$  denotes the probability that the video  
<sup>193</sup> belongs to the  $j^{th}$  expression category.

### <sup>194</sup> 3.4. Objective function

<sup>195</sup> The objective of CEFLNet has two parts: the CFE guarantees high-quality  
<sup>196</sup> emotional representations of clips, and EIAN focuses on the emotion-rich fea-  
<sup>197</sup> tures relevant to facial expressions via weakly supervised learning. In our study,  
<sup>198</sup> only a video-level FER classification loss  $L_{class}$  is used to optimize the two ob-  
<sup>199</sup> jectives of the entire network. Our FER classification loss  $L_{class}$  is as follows:

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

<sup>200</sup> where  $Y_{\mathcal{V}}$  denotes the facial expression label for each video,  $\mathcal{V}$  indexes a training  
<sup>201</sup> video, and  $p(\hat{Y}_{\mathcal{V}})$  denotes the probabilities of facial expressions predicted by the  
<sup>202</sup> CEFLNet.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

#### 203 4. Experimental Results and Discussion

##### 204 4.1. Datasets and Implementation Details

205 To evaluate our method, four video-based face expression datasets were used  
206 in our experiments, including BU-3DFE dataset [31], MMI dataset [32], AFEW  
207 8.0 dataset [20], and DFEW dataset [33].

208 **BU-3DFE [31]:** The 3D facial expressions are captured at a video rate  
209 (25 frames per second). Six emotion labels are included, *i.e.*, anger, disgust,  
210 happiness, fear, sadness, and surprise. Each expression sequence contains about  
211 100 frames. BU-3DFE contains 606 3D facial expression sequences captured  
212 from 101 subjects, with a total of approximately 60,600 frames. In this study,  
213 a 10-fold validation was conducted.

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

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

229 **DFEW [33]:** The DFEW is a large-scale unconstrained dynamic facial  
230 expression database, containing 16,372 video clips extracted from over 1,500  
231 different movies. It contains 12,059 single-label video clips and also includes  
232 seven emotion labels, *i.e.*, anger, disgust, fear, happiness, sadness, surprise, and

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

<sup>233</sup> neutral. DFEW dataset provides five data division methods. Hence, a 5-fold  
<sup>234</sup> validation was used. Examples of these datasets are shown in Fig 4.



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

<sup>235</sup> We kept each video to 105 frames via interpolation and clipping. The face  
<sup>236</sup> regions are detected using Retinaface [34] and the size of each face is resized  
<sup>237</sup> to  $224 \times 224$ . A randomly selected frame within the first 30 frames was used as  
<sup>238</sup> the starting frame and the following 75 consecutive frames were extracted. We  
<sup>239</sup> split the 75 frames into seven sub-Videos, each of which had 15 frames, with  
<sup>240</sup> five frames overlapping between each sub-video. To reduce the computation  
<sup>241</sup> cost, five frames were randomly sampled from each sub-video to form a new  
<sup>242</sup> expression clip. We conducted a 10-fold validation on BU-3DFE and MMI  
<sup>243</sup> datasets, a 5-fold validation on the DFEW dataset, and used the training and  
<sup>244</sup> validation sets for the experiments on the AFEW dataset.

<sup>245</sup> Our method is implemented using Pytorch. The training parameters include  
<sup>246</sup> initial learning rate (0.0001), cosine annealing schedule to adjust the learning  
<sup>247</sup> rate, mini-batch size (8), and warm-up. The experiments were conducted on a  
<sup>248</sup> PC with Intel(R) Xeon(R) Gold 6240C CPU at 2.60GHz and 128GB memory,  
<sup>249</sup> and NVIDIA GeForce RTX 3090 GPU. The key parameters used in training the

network are given in Table 1.

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

250

#### 251 4.2. Performance Analysis and Comparison Study

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

265 Figure 5(c) shows the confusion matrix from the AFEW dataset. AFEW  
 266 is one of the most challenging datasets and great confusion exists among ex-  
 267 pressions including Disgust, Fear, Sadness, and Surprise. The average accuracy  
 268 of our method is at 53.98% with a standard deviation of 0.4 and the highest

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

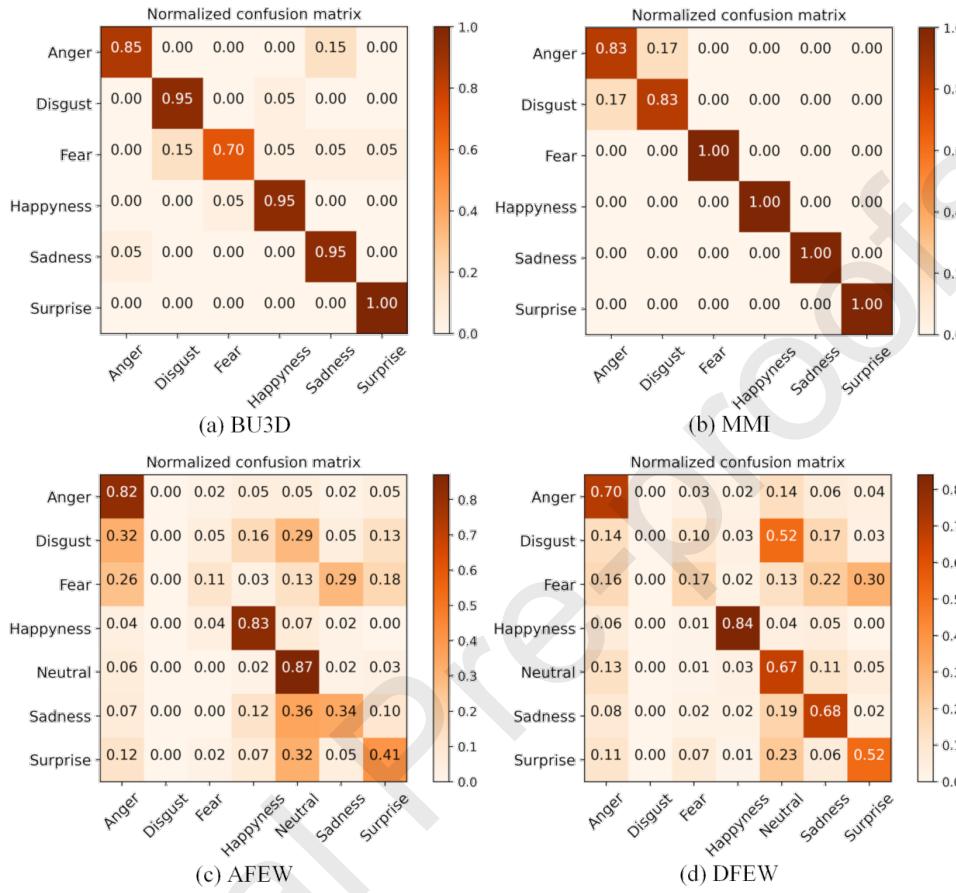


Figure 5: The confusion matrix of our method using the four datasets.

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 [11, 35]. Figure 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].

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

*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.

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	<u>84.17</u>
DeRL [8]	peak frame-based	<u>84.17</u>
C3D [38]	sequence-based	75.83
C3D-LSTM [40]	sequence-based	79.17
ICNP [39]	sequence-based	83.20
<b>CEFLNet</b>	clip-based	<b>85.33</b>

*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

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	<u>87.06</u>
CEFLNet	clip-based	<b>91.00</b>

299 by the best single models in previous works. Both [44] and [45] input two  
300 LBP maps and a gray image for CNN models. Deeply supervised networks  
301 are used in [45] and [15], which add supervision on intermediate layers. For  
302 clip-based methods, [35] uses DenseNet-161 and pre-trains it on both large-  
303 scale face datasets and their own Situ emotion video dataset. Additionally, [35]  
304 applies complicated post-processing which extracts frame features and computes  
305 their mean vector, max-pooling vector, and standard deviation vector. These  
306 vectors are then concatenated and finally fed into an SVM classifier. Overall,  
307 our CEFLNet improves the baseline (about 2.45%) and achieves performance  
308 comparable to that of the best previous single model. It demonstrates that our  
309 method achieves the best performance with great robustness, meanwhile, has  
310 obvious advantages over other algorithms on the in-the-wild expression dataset.

311 *Comparison study (DFEW):* The results in Table 5 show that our method  
312 is still far superior to other algorithms. More detailed comparison results can  
313 be shown in Table 5. Compared to the state-of-the-art methods reported in  
314 [33], the FER accuracy of our CEFLNet achieved significant improvement (over  
315 8.84%) on the challenging large-scale dataset.

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	<u>52.49</u>
DenseNet-161 [35]	clip-based	51.44
CEFLNet	clip-based	<b>53.98</b>

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	<u>56.51</u>
CEFLNet	clip-based	<b>65.35</b>

316     *4.3. Ablation Study and Analysis*

317     *4.3.1. Analysis of Network Components*

318       To analyze the contribution to the learning capability by the components of  
 319       CEFLNet, Table 6 presents the results of our ablation study that looks into  
 320       the impact of gradual addition of the self-attention learning, local-global relation  
 321       learning, and EIAN training components to the baseline framework (ResNet-  
 322       18). The training and testing datasets used in this study are BU-3DFE.

323       ResNet-18 and CNN-LSTM achieved an average accuracy of 62.77% and  
 324       79.17%, respectively. In our method, we used SAL to learn frame relation and  
 325       achieved average recognition accuracy of 84.17%. By adding LRL to the net-  
 326       work, the performance was improved by 0.5%, which shows that the local-global

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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	✓	✓	✓	<b>85.33</b>

<sup>327</sup> relation learning module can better learn the potential relationship between  
<sup>328</sup> each frame and clip. Note that the integration of EIAN improved the FER  
<sup>329</sup> accuracy by 0.66%. This demonstrates that the EIAN module learns the emo-  
<sup>330</sup> tional intensity from clip-based representations and obtains more distinguishable  
<sup>331</sup> emotion-rich video features.

#### <sup>332</sup> 4.3.2. Emotion-rich Clips

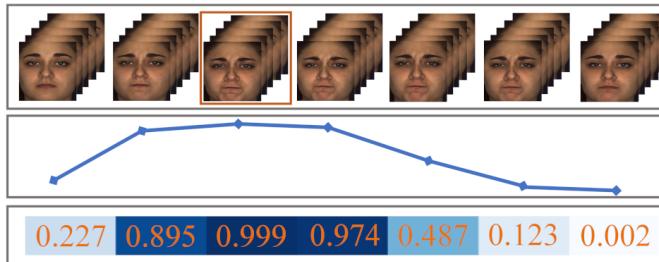
<sup>333</sup> Figure 6 shows the emotional activation maps and clip selection on the four  
<sup>334</sup> datasets. The orange boxes depict the select emotion-rich clips in videos. It can  
<sup>335</sup> be seen that the emotion-rich clips have the greatest expression intensity than  
<sup>336</sup> other clips, which implies that EIAN identifies the salient emotion-rich clip and  
<sup>337</sup> performs emotional activation according to the emotional intensity of each clip.

<sup>338</sup> In addition, we evaluated the accuracy of emotion-rich clip selection on the  
<sup>339</sup> four datasets, as shown in Table 7. The proposed MIAN achieved an accuracy of  
<sup>340</sup> 67.57% on the MMI dataset and achieved an accuracy of 45% on the challenging  
<sup>341</sup> AFEW dataset. This demonstrates that the EIAN method effectively locates  
<sup>342</sup> the emotion-rich clip in the untrimmed videos.

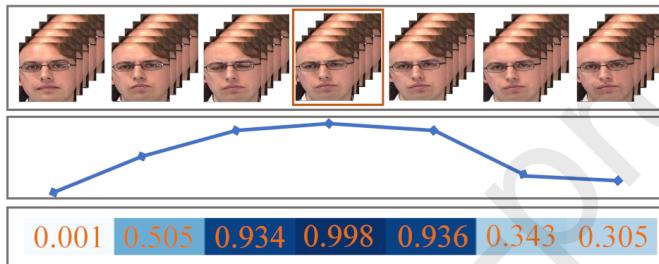
Table 7: The accuracy of emotion-rich clip locating

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

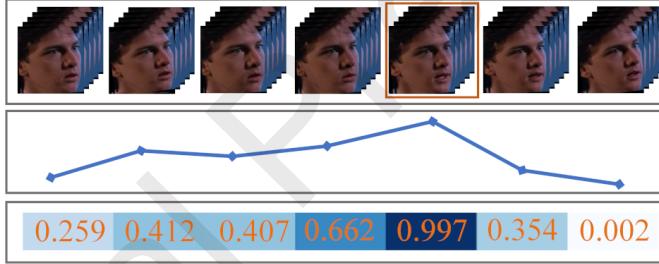
<sup>343</sup> We visualized the expression features with different settings in a 2D fea-  
<sup>344</sup> ture space by using the t-SNE on the four datasets. The visualizations include



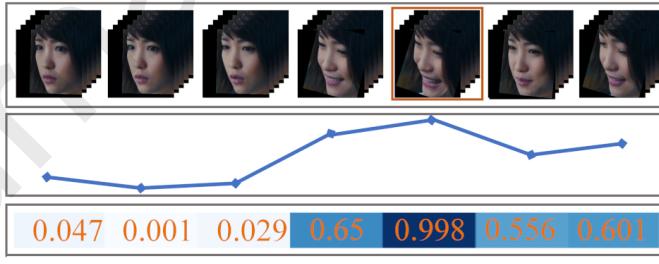
(a) BU-3DFE dataset



(b) MMI dataset



(c) AFEW dataset



(d) DFEW dataset

Figure 6: The emotional activation maps and the located clips (highlighted with orange boxes).

Darker colors indicate greater attention weights, i.e., more emotional information.

<sup>345</sup> the following four cases: clip-aware emotion-rich representations by the CE-  
<sup>346</sup> FLNet (see Figure 7(a)), video attention features extracted by FAN [16] (see

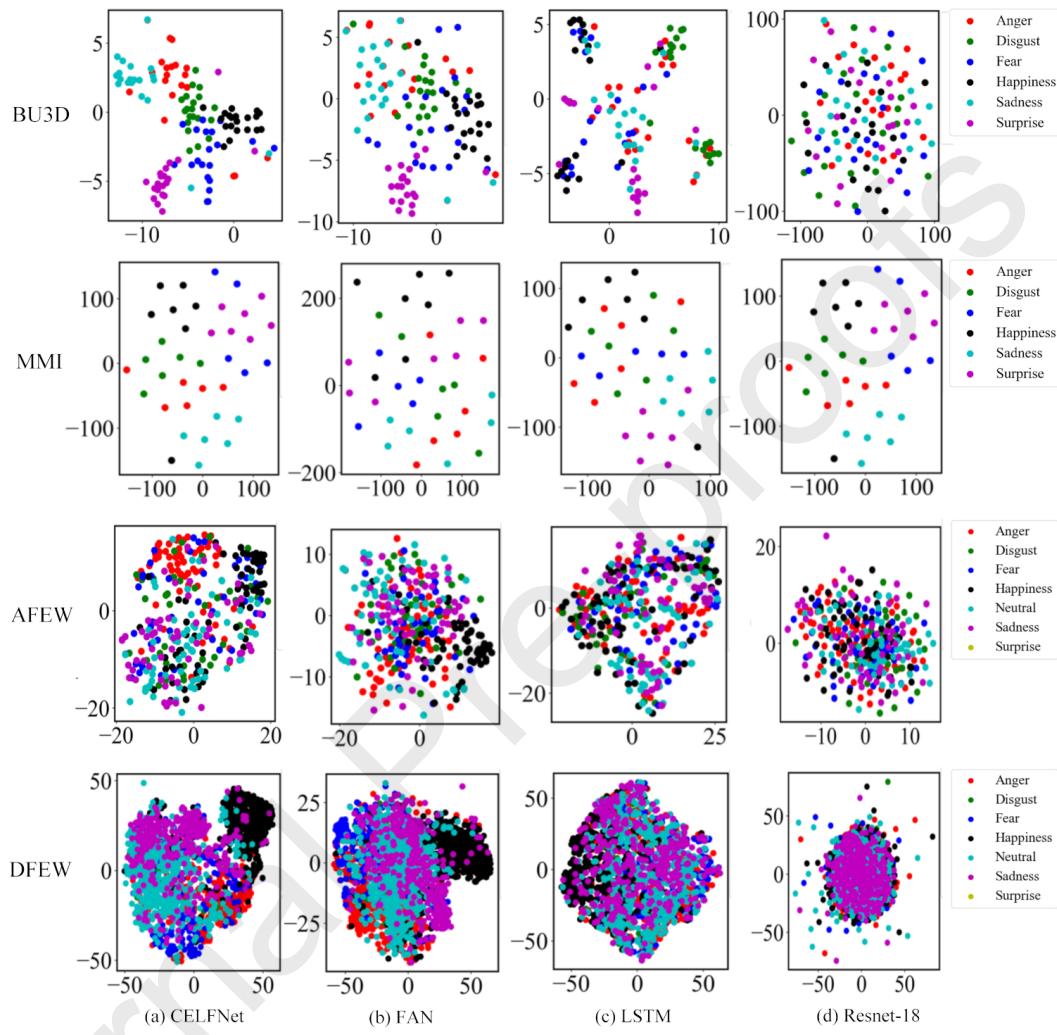


Figure 7: The t-SNE feature visualization of different representations in 2D space. (a) Clip-aware emotion-rich representations by CELFNet, (b) video attention features by FAN, (c) sequence-based video features by LSTM, (d) frame-based features by ResNet18.

<sup>347</sup> Figure 7(b)), sequence-based video features extracted by LSTM [48] (see Figure  
<sup>348</sup> 7(c)), frame-based features extracted by ResNet-18[49] (see Figure 7(d)).  
<sup>349</sup> Obviously, compared to the features shown in Figure 7(b), Figure 7(c) and  
<sup>350</sup> Figure 7(d), the clip-aware emotion-rich features proposed in this study can  
<sup>351</sup> significantly be separated according to facial expression categories. It is evi-

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

352 dent that the proposed CEFLNet can learn more expressive and discriminative  
 353 representations for video-based FER on the four datasets.

354 We studied the impact of the number of clips per video and the number  
 355 of frames per clip on the accuracy of FER. As shown in Figure 8(a), all four  
 356 datasets achieved the highest accuracies when the number of clips is 7, and  
 357 achieved the lowest accuracies when the number of clips is 1. Results show that  
 358 too many or too few clips are detrimental to the performance of facial expression  
 359 recognition. As shown in Figure 8(b), the highest accuracy is achieved when  
 360 we set the number of frames of each clip to 5. When this number is less than  
 361 5, the accuracy drops. The performance drop might be a result of emotional  
 362 information lost. When the number of frames is 15, redundant expressionless  
 363 frames cause expression inconsistency and hence reduce recognition accuracy.  
 364 In our experiments, we keep the number of clips of each video to 7 and the  
 365 number of frames of each clip to 5.

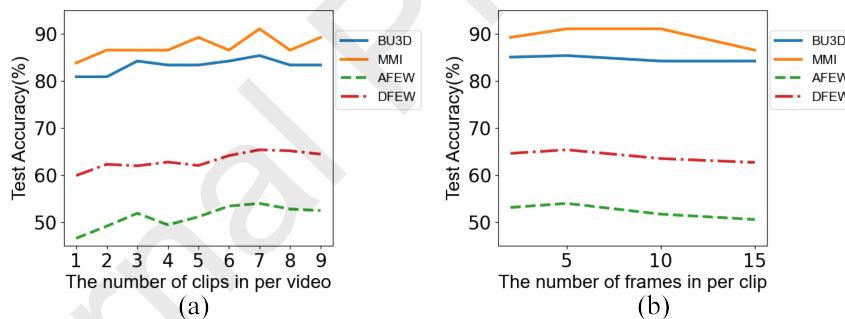


Figure 8: The accuracy of the number of frames per clip and the number of clips per video for FER on four datasets. (a) The effect of the number of expression clips, (b) the effect of the number of frames.

#### 366 4.4. Computational Complexity

367 Table 8 reports model parameters and computational cost of the three spatio-  
 368 temporal learning methods in processing the BU-3DFE dataset. We use Multi-

<sup>369</sup>      ply-Accumulate Operations(MACs)<sup>1</sup> to measure the computational cost. Our  
<sup>370</sup>      CEFLNet resulted in the best performance (FER accuracy of 85.33%) with the  
<sup>371</sup>      least computational cost (63.8G) and parameters (12.83M) among the compared  
<sup>372</sup>      methods, which demonstrates that the proposed method exhibits improved ac-  
<sup>373</sup>      curacy and efficiency.

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	<b>12.83</b>	<b>63.80</b>	<b>85.33</b>

<sup>374</sup>      Table 9 lists the average accuracy and the computation cost with respect  
<sup>375</sup>      to the number of frames. Clearly, when less number of frames are used, the  
<sup>376</sup>      computational cost is lower. However, the best accuracy is achieved when the  
<sup>377</sup>      number of frames is 5. Hence, to balance speed and accuracy, a five-frame per  
<sup>378</sup>      clip is a proper choice.

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

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

## <sup>379</sup> 5. Conclusion and Future Work

<sup>380</sup>      In this paper, we propose an effectively clip-aware emotion-rich feature learn-  
<sup>381</sup>      ing network to jointly identify the emotion-rich clips and recognize dynamic

---

<sup>1</sup><https://github.com/sovrasov/flops-counter.pytorch>

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

382 facial expressions in a video. CEFLNet decomposes a video into several small  
383 video clips and extracts the clip-level spatio-temporal features via two-cascaded  
384 self-attention and local-global relation learning within each video clip. Our  
385 method generates an emotional activation map that is used to identify salient  
386 emotion clips for clip-aware emotion-rich representations. Our proposed method  
387 requires no clip-wise or frame-wise annotations for training the model and can  
388 be trained in an end-to-end manner.

389 Experiments were conducted using four public video datasets, namely the  
390 BU-3DFE, MMI, AFEW, and DFEW. Due to suppressing the redundancy in-  
391 formation from expression-irrelevant clips, the proposed method was found to  
392 achieve a much-improved performance for video-based FER, with great robust-  
393 ness and efficiency; the highest accuracy for each of these datasets was 85.33%,  
394 91%, 53.98%, and 65.35%. In our future work, we plan to study self-supervised  
395 learning to model the extraction of key information from complex facial video  
396 sequences with multiple expressions.

### 397 **Acknowledgments**

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

### 401 **References**

- 402 [1] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural  
403 network-driven feature learning method for multi-view facial expression  
404 recognition, *IEEE Transactions on Multimedia* 18 (12) (2016) 2528–2536.
- 405 [2] J. Wu, Z. Lin, W. Zheng, H. Zha, Locality-constrained linear coding based  
406 bi-layer model for multi-view facial expression recognition, *Neurocomputing*  
407 239 (2017) 143–152.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- [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, IEEE, 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* (2017).
- [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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- [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* (2017).
- [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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- [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 (2021).
- [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. Cherodion, M. Savvides, Deepgender: Occlusion and low resolution robust facial gender classification via progres-

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 491 sively trained convolutional neural networks with attention, in: Proceedings  
492 of the IEEE conference on computer vision and pattern recognition  
493 workshops, 2016, pp. 68–77.
- 494 [30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep  
495 features for discriminative localization, in: Proceedings of the IEEE con-  
496 ference on computer vision and pattern recognition, 2016, pp. 2921–2929.
- 497 [31] L. Yin, X. Wei, Y. Sun, J. Wang, M. J. Rosato, A 3d facial expression  
498 database for facial behavior research, in: 7th international conference on  
499 automatic face and gesture recognition (FGR06), IEEE, 2006, pp. 211–216.
- 500 [32] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition  
501 to the mmi facial expression database, in: Proc. 3rd Intern. Workshop on  
502 EMOTION (satellite of LREC): Corpora for Research on Emotion and  
503 Affect, Paris, France., 2010, p. 65.
- 504 [33] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, J. Liu, Dfew: A  
505 large-scale database for recognizing dynamic facial expressions in the wild,  
506 in: Proceedings of the 28th ACM International Conference on Multimedia  
507 (MM), 2020, pp. 2881–2889.
- 508 [34] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-  
509 shot multi-level face localisation in the wild, in: Proceedings of the IEEE  
510 Conference on Computer Vision and Pattern Recognition (CVPR), 2020,  
511 pp. 5203–5212.
- 512 [35] C. Liu, T. Tang, K. Lv, M. Wang, Multi-feature based emotion recognition  
513 for video clips, ACM ICMI (2018) 630–634.
- 514 [36] X. Yuan, M. Abouelenien, M. Elhoseny, A boosting-based decision fusion  
515 method for learning from large, imbalanced face data set, in: Quantum  
516 Computing: An Environment for Intelligent Large Scale Real Application,  
517 Springer, Cham, 2018, pp. 433–448.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- [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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- [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.