# FRAME ATTENTION NETWORKS FOR FACIAL EXPRESSION RECOGNITION IN VIDEOS

*Debin Meng, Xiaojiang Peng*, Kai Wang, Yu Qiao*

Shenzhen Institutes of Advanced Technology, Chinese Academy of Science, Shenzhen, China
Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen, China
University of Chinese Academy of Sciences, Beijing, China
michaeldbmeng19@outlook.com, {xj.peng, kai.wang, yu.qiao}@siat.ac.cn

## ABSTRACT

The video-based facial expression recognition aims to classify a given video into several basic emotions. How to integrate facial features of individual frames is crucial for this task. In this paper, we propose the Frame Attention Networks (FAN)[1], to automatically highlight some discriminative frames in an end-to-end framework. The network takes a video with a variable number of face images as its input and produces a fixed-dimension representation. The whole network is composed of two modules. The feature embedding module is a deep Convolutional Neural Network (CNN) which embeds face images into feature vectors. The frame attention module learns multiple attention weights which are used to adaptively aggregate the feature vectors to form a single discriminative video representation. We conduct extensive experiments on CK+ and AFEW8.0 datasets. Our proposed FAN shows superior performance compared to other CNN based methods and achieves state-of-the-art performance on CK+.

***Index Terms***— facial expression recognition, audio-video emotion recognition, frame attention networks, CNN, AFEW

## 1. INTRODUCTION

Automatic facial expression recognition (FER) has recently attracted increasing attention in academia and industry due to its wide range of applications such as affective computing, intelligent environments, and multimodal human-computer interface (HCI). Though great progress have been made recently, facial expression recognition in the wild remains a challenging problem due to large head pose, illumination variance, occlusion, motion blur, etc.

Video-based facial expression recognition aims to classify a video into several basic emotions, such as happy, angry, disgust, fear, sad, neutral, and surprise. Given a video, the popular FER pipeline with a visual clue (FER with an audio clue is out of the scope of this paper) mainly includes three steps, namely frame preprocessing, feature extraction, and classification. Especially, frame preprocessing refers to face detection, alignment, illumination normalizing and so on. Feature extraction or video representation is the key part for FER which encodes frames or sequences into compact feature vectors. These feature vectors are subsequently fed into a classifier for prediction.

Feature extraction methods for video-based FER can be roughly divided into three types: static-based methods, spatial-temporal methods, and geometry-based methods.

Static-based feature extraction methods mainly inherit those methods from static image emotion recognition which can be both hand-crafted [1, 2] and learned [3, 4, 5]. For the hand-crafted features, Littlewort *et al.* [1] propose to use a bank of 2D Gabor filters to extract facial features for video-based FER. Shan *et al.* [2] use local binary patterns (LBP) and LBP histogram for facial feature extraction. For the learned features, Tang [3] utilizes deep CNNs for feature extraction, and win the FER2013. Some winners in audio-video emotion recognition task of EmotiW2016 and EmotiW2017 only use static facial features from deep CNNs trained on large face datasets or trained with multi-level supervision [4, 5].

Spatial-temporal methods aim to model the temporal or motion information in videos. The Long Short-Term Memory (LSTM) [6], and C3D [7] are two widely-used spatial-temporal methods for video-based FER. LSTM derives information from sequences by exploiting the fact that feature vectors are connected semantically for successive data. This pipeline is widely-used in the EmotiW challenge, e.g. [8, 9, 10, 11]. C3D, which is originally developed for video action recognition, is also popular in the EmotiW challenge.

Geometry based methods [12, 11] aim to model the motions of key points in faces which only leverage the geometry locations of facial landmarks in every video frames. Jung *et al.* [12] propose a deep temporal appearance-geometry network (DTAGN) which first alternately concatenates the x-coordinates and y-coordinates of the facial landmark points

*Xiaojiang Peng is the corresponding author. Email: xj.peng@siat.ac.cn

This work was supported by the National Natural Science Foundation of China (U1613211, U1713208), Shenzhen Research Program (JCYJ20170818164704758, JSGG20180507182100698), and International Partnership Program of Chinese Academy of Sciences (172644KYSB20150019).

[1] Code is available at https://github.com/Open-Debin/Emotion-FAN

from each frame after normalization and then concatenates these normalized points over time for a one-dimensional trajectory signal of each sequence. Yan *et al.* [11] construct an image-like map by stretching all the normalized facial point trajectories in a sequence together as the input of a CNN.

Among all the above methods, static-based methods are superior to the others according to several winner solutions in EmotiW challenges. To obtain a video-level result with varied frames, a frame aggregation operation is necessary for static-based methods. For frame aggregation, Kahou *et al.* [13] concatenate the $n$-class probability vectors of 10 segments to form a fixed-length video representation by frame averaging or frame expansion. Bargal *et al.* [4] propose a statistical encoding module (STAT) to aggregate frame features which compute the mean, variance, minimum, and maximum of the frame feature vectors.

One limitation of these existing aggregation methods is that they ignore the importance of frames for FER. For example, some faces in Figure 1 are representative for the 'happy' category while the others not. In this paper, inspired by the attention mechanism [14] of machine translation and the neural aggregation networks [15] of video face recognition, we propose the Frame Attention Networks (FAN) to adaptively aggregate frame features. The FAN is designed to learn self-attention kernels and relation-attention kernels for frame importance reasoning in an end-to-end fashion. The self-attention kernels are directly learned from frame features while the relation-attention kernels are learned from the concatenated features of a video-level anchor feature and frame features. We conduct extensive experiments on CK+ and AFEW8.0 (EmotiW2018) datasets. Our proposed FAN shows superior performance compared to other CNN based methods with only facial features and achieves state-of-the-art performance on CK+.

## 2. FRAME ATTENTION NETWORKS

We propose Frame Attention Networks (FAN) for video-based facial expression recognition (FER). Figure 1 illustrates the framework of our proposed FAN. It takes a facial video with a variable number of face images as its input and produces a fixed-dimension feature representation for FER. The whole network consists of two modules: feature embedding module and frame attention module. The feature embedding module is a deep CNN which embeds each face image into a feature vector. The frame attention module learns two-level attention weights, i.e. *self-attention weights* and *relation-attention weights*, which are used to adaptively aggregate the feature vectors to form a single discriminative video representation.

Formally, we denote a video with $n$ frames as $\mathbf{V}$, and its frames as $I_1, I_2, \cdots, I_n$, and the facial frame features are $\{f_1, \cdots, f_n\}$.

**Self-attention weights**. With individual frame features, our FAN first applies a FC layer and a sigmoid function to assign coarse self-attention weights. Mathematically, the self-attention weight of the $i$-th frame is defined by:

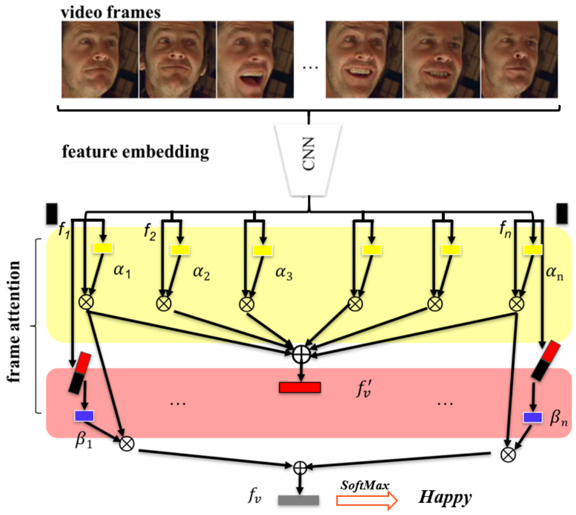

**Fig. 1**. Our proposed frame attention network architecture.

$$\alpha_i = \sigma(f_i^T \mathbf{q}^0) \tag{1}$$

where $\mathbf{q}^0$ is the parameter of FC, $\sigma$ denotes the sigmoid function. With these self-attention weights, we aggregate all the input frame features into a global representation $f'_v$ as follows,

$$f'_v = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i}. \tag{2}$$

We use $f'_v$ as a video-level global anchor for learning further accurate relation-attention weights.

**Relation-attention weights**. We believe that learning weights from both a global feature and local features is more reliable. The self-attention weights are learned with individual frame features and non-linear mapping, which are rather coarse. Since $f'_v$ inherently contains the contents of the whole video, the attention weights can be further refined by modeling the relation between frame features and this global representation $f'_v$.

Inspired by the relation-Net in low-shot learning [16], we use the sample concatenation and another FC layer to estimate new relation-attention weights for frame features. The relation-attention weight of the $i$-th frame is formulated as,

$$\beta_i = \sigma([f_i : f'_v]^T \mathbf{q}^1), \tag{3}$$

where $\mathbf{q}^1$ is the parameter of FC, $\sigma$ denotes the sigmoid function.

Finally, with self-attention and relation-attention weights, our FAN aggregates all the frame features into a new compact feature as,

$$f_v = \frac{\sum_{i=0}^{n} \alpha_i \beta_i [f_i : f'_v]}{\sum_{i=0}^{n} \alpha_i \beta_i}. \tag{4}$$

# 3. EXPERIMENTS

## 3.1. Datasets and Implementation Details

**CK+** [17] contains 593 video sequences from 123 subjects. Among these videos, 327 sequences from 118 subjects are labeled with seven basic expression labels, i.e. anger, contempt, disgust, fear, happiness, sadness, and surprise. Since CK+ does not provide training/testing splits, most of the algorithms evaluated on this database with 10-fold person-independence cross-validation experiments. We constructed 10 subsets by sampling ID in ascending order with a step size of 10 as in several previous works [18, 19], and report the overall accuracy over 10 folds.

**AFEW 8.0** [20] 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 video clips collected from different movies and TV serials with spontaneous expressions, various head poses, occlusions, and illuminations. AFEW 8.0 is divided into three splits: Train (773 samples), Val (383 samples) and Test (653 samples), which ensures data in the three sets belong to mutually exclusive movies and actors. Since the test split is not publicly available, we train our model on training split and report results on validation split.

**Implementation details**. We preprocess video frames by face detection and alignment in the Dlib toolbox We extend the face bounding box with a ratio of 25% and then resize the cropped faces to scale of 224×224. We implement our method by the Pytorch toolbox. By default, for feature embedding, we use the ResNet18 which is pre-trained on MS-Celeb-1M [21] face recognition dataset and FER_Plus expression dataset [22]. For training, on both CK+ and AFEW 8.0, we set a batch to have 48 instances with $K$ frames in each instance. For frame sampling in a video, we first split the video into $K$ segments and then randomly select one frame from each segment. By default, we set $K$ to 3. We use the SGD method for optimization with a momentum of 0.9 and a weight decay of $10^{-4}$. On CK+, we initialize the learning rate (*lr*) to 0.1, and modify it to 0.02 at 30 epochs, and stop training after 60 epochs. On AFEW 8.0, we initialize the *lr* to 4e-6, and modify it to 8e-7 at 60 epochs and 1.6e-7 at 120 epochs, and stop training after 180 epochs.

## 3.2. Evaluation on CK+

We evaluate our FAN on CK+ with comparisons to several state-of-the-art methods in Table 1. On CK+, due to the fact that the videos show a shift from a neutral facial expression to the peak expression, most of the methods conduct data selection manually. Zhang et al [23] propose to combine a spatial CNN model and a temporal network, where the spatial CNN model only uses the last peak frame. Jung et al [12] select a fixed length sequence for each video with a lipreading method [26], and jointly fine-tune a deep temporal appearance-geometry network. Cai et al [24] select the last three frames and the first frame for each video, and train CNN models with a new Island loss function. We argue that *manual data selection is an ad-hoc operation on CK+ and it is impractical since we can not know which is the peak frame beforeahead*. Sikka et al [25] use all frames with a new latent ordinal model which extracts CNN/LBP/SIFT features for sub-event detection and uses multi-instance SVM for expression classification. Our baseline method uses ResNet18 to generate scores for individual frame and applies score fusion (summation) for all frames. It achieves 94.8% which is 2.8% better than [25]. Our proposed FAN with only self-attention gets 99.08% which significantly boosts the baseline by 4.28%. Adding relation-attention weights further improves the accuracy to 99.69% which sets up a new state of the art on CK+.

**Table 1**. Evaluation of our FAN with a comparison to state-of-the-art methods on CK+ database. Note that only those methods evaluated with 7 classes are included.

| Method | Training data | Test data | Acc. |
|---|---|---|---|
| ST network [23] | S: the last frame<br>T: all frames | S: the last frame<br>T: all frames | 98.50 |
| DTAGN [12] | Fixed length | Fixed length | 97.25 |
| CNN+Island loss [24] | The last three frames and the first frame | The last three frames and the first frame | 94.35 |
| LOMo [25] | All frames | All frames | 92.00 |
| Score fusion (baseline) | All frames | All frames | 94.80 |
| FAN(w/o Relation-attention) | All frames | All frames | **99.08** |
| FAN | All frames | All frames | **99.69** |

## 3.3. Evaluation on AFEW 8.0

From the view of performance, AFEW is one of the most challenging videos FER dataset. The EmotiW challenge shares the same data from 2016 to 2018. Table 2 presents the evaluation of our FAN on AFEW with comparisons to recent state-of-the-art methods. For a fair comparison, we only list these results obtained by the best single models in previous works. From the last three rows of Table 2, our proposed FAN improves the baseline by 2.36%. Both [27] and [10] use VGGFace backbone and a recurrent model with long-short-term memory units. These methods aim to capture temporal dynamic information for videos. Most of the methods focus on improving static face based CNN models and

**Table 2**. Evaluation of our FAN with a comparison to state-of-the-art methods on AFEW 8.0 database. It is worth noting that we only compare to the best *single* models of previous works.

| Method | Model type | Accuracy |
|---|---|---|
| CNN-RNN (2016) [27] | Dynamic | 45.43 |
| VGGFace + Undirectional LSTM (2017) [10] | Dynamic | 48.60 |
| HoloNet (2016) [28] | Static | 44.57 |
| DSN-HoloNet (2017) [29] | Static | 46.47 |
| DenseNet-161 (2018) [31] | Static | **51.44** |
| DSN-VGGFace (2018) [30] | Static | 48.04 |
| Score fusion (baseline) | Static | 48.82 |
| FAN w/o Relation-attention | Static | **50.92** |
| FAN | Static | **51.18** |

combine scores for video-level FER. Both [28] and [29] input two LBP maps and a gray image for CNN models. Deeply-supervised networks are used in [29] and [30], which add supervision on intermediate layers. For static methods, [31] gets slightly better performance than ours. However, [31] uses DenseNet-161 and pretrains it on both large-scale face datasets and their own Situ emotion video dataset. Additionally, [31] applies complicated post-processing which extracts frame features and compute their mean vector, max-pooled vector, and standard deviation vector. These vectors are then concatenated and finally fed into an SVM classifier. Overall, our FAN improves the baseline significantly and achieves performance comparable to that of the best previous single model.

### 3.4. Visualization and Hyper-parameters

To better understand the self-attention and relation-attention modules in our FAN, we visualize the attention weights in Figure 2. Figure 2 shows one sequence for each category with blue and orange weight bars, where blue bars represent the self-attention weights (i.e. $\alpha$ in Eq. (1)) of our FAN w/o relation-attention and orange bars the final weights (i.e. $\alpha\beta$ in Eq. (4)) of our FAN. In total, both kinds of weights can reflect the importance of frames. Comparing the blue and orange bars, we find that the final weights of our FAN can always assign higher weights to the more obvious face frames, while self-attention module could assign high weights on some obscure face frames, see the 1st, 2th, and 3rd rows of Figure 2 (left). This explicitly explains why adding relation-attention boost performance.

**Evaluation of Hyper-parameters**. We evaluate two hyper-parameters of our FAN on CK+, i.e. backbone CNN networks and the parameter $K$ mentioned in implementation details, to validate the robustness of our method. For the parameter $K$, besides the default value, we try several other values, i.e. $\{2, 5, 8\}$, and find the performance is not sensitive to $K$. Specifically, our FAN obtains 99.39% with $K$=$\{2, 5\}$. and gets 99.69% with $K$=8. Since the default value, $K$=3 gets 99.69%, we use this default setting in the remainder of this paper.

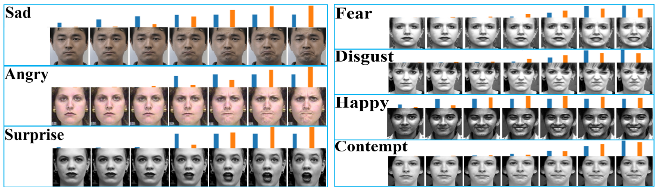

**Fig. 2**. Visualization of the self-attention weights (blue bar) and the final weights of FAN (orange bar) on CK+ dataset.

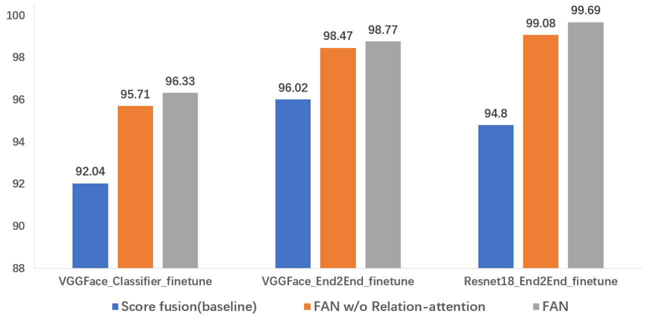

**Fig. 3**. Evaluation of backbone CNN models and training strategies on CK+.

For the backbone CNN model evaluation, we try the VGGFace model which is widely-used in previous works. Similarly, we also pretrain the VGGFace model on the FERPlus dataset. Since [5] shows that it is better to freeze all the feature learning layers after pretrained on FERPlus for VGGFace model, we also conduct the same experiment on CK+ with VGGFace. Figure 3 shows the default comparisons with different backbone CNN models. On CK+, compared with freezing all the feature layers for VGGFace, it gets better results with fine-tuning all layers which may be explained by the domain discrepancy between FERPlus and CK+. Overall, the results are significantly improved by self-attention weights and further improved by the relation-attention weights.

## 4. CONCLUSION

We propose Frame Attention Networks for video-based facial expression recognition. The FAN contains a self-attention module and a relation-attention module. The experiments on CK+ and AFEW show that our FAN with only self-attention improves the baseline significantly and adding relation-attention further boosts performance. With a visualization on CK+, we demonstrate that our FAN can automatically capture the importance of frames. Our single model achieves performance on par with that of state-of-the-art methods on AFEW and obtains state-of-the-art results on CK+.

## 5. REFERENCES

[1] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan, "Dynamics of facial expression extracted automatically from video," *in IVC*, 2006.

[2] Caifeng Shan, Shaogang Gong, and Peter W. Mcowan, "Facial expression recognition based on local binary patterns: A comprehensive study," *in IVC*, 2009.

[3] Yichuan Tang, "Deep learning using linear support vector machines," *in CS*, 2013.

[4] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang, "Emotion recognition in the wild from videos using images," in *ACM ICMI*, 2016.

[5] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," in *ACM ICMI*, 2017.

[6] Sepp Hochreiter and Jrgen Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[7] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[8] Yuanliu Liu, Yuanliu Liu, Yuanliu Liu, and Yuanliu Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *ACM ICMI*, 2016.

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

[10] Valentin Vielzeuf, Stphane Pateux, and Frdric Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *ACM ICMI*, 2017.

[11] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, and Yuan Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, 2018.

[12] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *ICCV*, 2015.

[13] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, and Raul Chandias Ferrari, "Combining modality specific deep neural networks for emotion recognition in video," in *ACM ICMI*, 2013.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[15] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua, "Neural aggregation network for video face recognition.," in *CVPR*, 2017.

[16] Flood Sung Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018.

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

[18] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *CVPR*, 2014.

[19] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis, "A compact deep learning model for robust facial expression recognition," in *CVPRW*, 2018.

[20] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," *arXiv preprint:1808.07773*, 2018.

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016.

[22] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *ACM ICMI*, 2016.

[23] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE TIP*, 2017.

[24] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James OReilly, and Yan Tong, "Island loss for learning discriminative features in facial expression recognition," in *FG*, 2018.

[25] Karan Sikka, Gaurav Sharma, and Marian Bartlett, "Lomo: Latent ordinal model for facial analysis in videos," in *CVPR*, 2016.

[26] Ziheng Zhou, Guoying Zhao, and M. Pietikainen, "Towards a practical lipreading system," in *CVPR*, 2011.

[27] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *ACM ICMI*, 2016.

[28] Anbang Yao, Dongqi Cai, Ping Hu, Shandong Wang, Liang Sha, and Yurong Chen, "Holonet: towards robust emotion recognition in the wild," in *ACM ICMI*, 2016.

[29] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *ACM ICMI*, 2017.

[30] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li, "Video-based emotion recognition using deeply-supervised neural networks," in *ACM ICMI*, 2018.

[31] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang, "Multi-feature based emotion recognition for video clips," in *ACM ICMI*, 2018.