Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition

Hengshun Zhou[*,1], **Debin Meng**[*,2], **Yuanyuan Zhang**[1], **Xiaojiang Peng**[†,2], **Jun Du**[1], **Kai Wang**[2], **Yu Qiao**[2*]
[1]National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP), University of Science and Technology of China, P.R. China
[2]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

ABSTRACT

The audio-video based emotion recognition aims to classify a given video into basic emotions. In this paper, we describe our approaches in EmotiW 2019, which mainly explores emotion features and feature fusion strategies for audio and visual modality. For emotion features, we explore audio feature with both speech-spectrogram and Log Mel-spectrogram and evaluate several facial features with different CNN models and different emotion pretrained strategies. For fusion strategies, we explore intra-modal and cross-modal fusion methods, such as designing attention mechanisms to highlights important emotion feature, exploring feature concatenation and factorized bilinear pooling (FBP) for cross-modal feature fusion. With careful evaluation, we obtain 65.5% on the AFEW validation set and 62.48% on the test set and rank third in the challenge.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Emotion Recognition; Attention Mechanism; Deep learning; Affective Computing; Convolutional Neural Networks

ACM Reference Format:
Hengshun Zhou[*,1], Debin Meng[*,2], Yuanyuan Zhang[1], Xiaojiang Peng[†,2], Jun Du[1], Kai Wang[2], Yu Qiao[2]. 2019. Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340555.3355713

[*]Hengshun Zhou and Debin Meng contributed equally to this research.
† Xiaojiang Peng is the corresponding author.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3355713

1 INTRODUCTION

Emotion recognition(ER) has attracted increasing attention in academia and industry due to its wide range of applications such as human-computer interaction [7], clinical diagnosis [19], and cognitive science [14]. Although great progress in the face and video analysis has been made [4, 23, 26–29], audio-video emotion recognition in the wild remains a challenging problem due to the expression suffers from the large pose, illumination variance, occlusion, motion blur, etc.

Audio-Video emotion recognition can be summarized as a simple pipeline shown in Fig 1, which includes four parts, namely Video preprocessing, Feature Extraction, Feature Fusion, and Classifier. Specifically, video preprocessing refers to extract the spectrogram of the audio, the faces or landmarks of video. Feature extraction and feature fusion respectively extracts emotion features from the audio or visual signal and fuses emotion features into compact feature vectors, which are subsequently fed into a classifier for prediction.

Reviewing the methods of Audio-Video emotion recognition, we find that some methods emphasize feature extraction and other methods emphasize feature fusion. Yao et al [31] construct Holonet as discriminative feature extraction, which combines residual structure [12] and CReLU [22] to increase network depth and maintain efficiency. The EmotiW2017 winner team [13] gets robust feature extraction with Supervised Scoring Ensemble (SSE) which adds supervision to intermediate layers and shallow layers. Since SSE only uses high-level representations, Fan et al[8] further improve SSE by utilizing middle feature maps to provide more discriminative features. These methods mainly use average pooling to obtain video-level representation from frame-level.

Many feature fusion strategies have been used in previous EmotiW challenges. [9, 18, 25] extract CNN-based frame features and use LSTM[10] or BLSTM[11] to fuse them. [1, 15, 17] use Statistical encoding module to aggregate frame features which compute the mean, variance, minimum, and

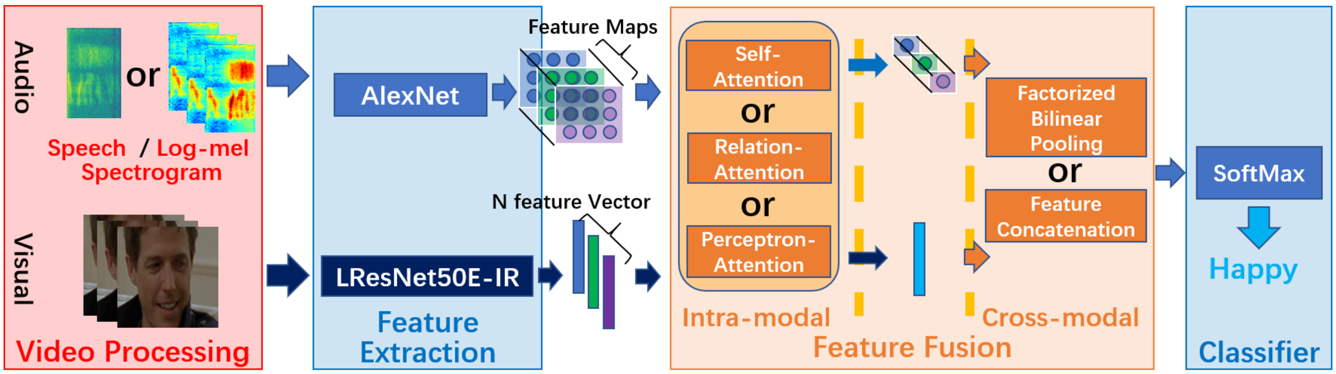

Figure 1: The pipeline of audio-video emotion recognition.

maximum of the frame feature vectors. However, these methods ignore the importance of frames. Besides, all previous methods mainly apply score averaging or feature concatenation for audio-video fusion, which ignores the correlation between the features from different modalities.

In this paper, we exploit three types of intra-modal fusion methods, namely self-attention, relation-attention, and transformer[24]. They are used to learn weights for frame features to highlight important frames. For cross-modal fusion, we explore feature concatenation and factorized bilinear pooling (FBP) [32]. Besides, we evaluate different emotion features, including convolutional neural networks (CNN) for audio information with both speech-spectrogram and Log Mel-spectrogram and several facial features with different CNN models and different emotion pretrained strategies. Finally, we obtain 62.48% and rank third in the challenge.

Our contributions and finds can be summarized as follows.

- We experimentally show that better face recognition CNN models and choosing suitable emotion datasets to further pretrain the face CNN models is important.
- We design three kinds of attention mechanisms for visual and audio feature fusion.
- We apply a Factorized Bilinear Pooling (FBP) for cross-modal feature fusion.

2 THE PROPOSED METHOD

We develop our ER system based on the pipeline of Video preprocessing-Feature Extraction-Feature Fusion-Classifier.

Video preprocessing

Face detection and alignment. We apply face detection and alignment by Dlib toolbox[1]. We extend the face bounding box with a ratio of 30% and then resize the cropped faces to scale of 224×224. We do not apply face detection and alignment for AffectNet dataset, due to the face bounding box had been provided. For AFEW dataset, If no face is detected in the picture, the entire frame is passed to the network.

Audio processing and Spectrogram calculation. For each audio, the speech spectrogram and log Mel-spectrogram extraction process is consistent with [32] and [3] respectively. For speech spectrogram, we use the Hamming window with 40 msec window size and 10 msec shift. Finally, the 200-dimensional low-frequency part of the spectrogram is used as the input to the audio modality. As for log Mel-spectrogram, we calculate its deltas and delta-deltas.

Feature Extraction

Visual Features. We apply three CNN backbones to extract facial emotion features, namely VGGFace, ResNet18, and IR50 [4]. The dimensions are 4096, 512, and 512, respectively.

Audio Feature. We extract the feature maps of the audio from the last Pooling layer of AlexNet. The size of a 3-dimensional feature map is $H \times W \times C$, where the $H(W)$ is the height(width) of the feature map, and C is the number of the channel of the feature map. The feature maps are then split into n vectors($n = H \times W$). Each vector is C-dimensional.

Intra-modal Feature Fusion

We apply the attention-based strategies for intra-modal feature fusion. It converts a variable number of emotion features(from audio or visual modality) into a fixed-dimension feature. We explore three attention methods, namely Self-attention, Relation-attention, and Transformer-attention. Formally, we denote a number of emotion features as $\{f_1, \cdots, f_n\}$.

Self-attention. We apply 1-dimensional Fully-Connected(FC) layer $\mathbf{W}^{\mathbf{0}}_{\mathbf{d}\times\mathbf{1}}$ and a sigmoid function σ for each emotion feature, the weight of the i-th feature f_i^T is defined by:

$$\alpha_i = \sigma(f_i^T \cdot \mathbf{W}^{\mathbf{0}}_{\mathbf{d}\times\mathbf{1}}) \tag{1}$$

[1]http://dlib.net/

With these self-attention weights, we aggregate all the emotion features into a global representation f_s as follows:

$$f_s = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i}. \quad (2)$$

Relation-attention. This attention module was designed to learn weights from the relationship between features. After the self-attention, features are aggregated into a single vector f_s. Since f_s inherently contains global representation of these features, we use the sample concatenation of individual features and global represenation $[f_i : f_s]$ to model the global-local relation. Similar to the Self-attention module, with individual emotion features, we apply 1-dimensional FC layer $\mathbf{W}^1_{\mathbf{d}\times 1}$ and a sigmoid function σ. The relation-attention weight of the i-th feautre $[f_i : f_s]^T$ is formulated as follows:

$$\beta_i = \sigma([f_i : f_s]^T \cdot \mathbf{W}^1_{\mathbf{d}\times 1}), \quad (3)$$

With Self-attention and Relation-attention weights, all the emotion features was convert into a new feature as follows:

$$f_r = \frac{\sum_{i=0}^{n} \alpha_i \beta_i [f_i : f_s]}{\sum_{i=0}^{n} \alpha_i \beta_i}. \quad (4)$$

Transformer-attention. Inspired by the works in[32] and [30], we formulate the attention weight as follows:

$$f_i' = \mathbf{W}^2_{\mathbf{m}\times\mathbf{d}} \cdot f_i + b \quad (5)$$

$$\gamma_i = exp(\mathbf{u^t}_{d\times 1} \cdot tanh(f_i')) \quad (6)$$

To reduce the dimension of the feature f_i, we use a $w \times d$-dimensional FC layer $\mathbf{W}^2_{\mathbf{m}\times\mathbf{d}}$ in Eq.(5). Then the weight of the i-th feautre f_i is processed by a 1-dimensional FC layer $\mathbf{u^t}$, $exp()$ and $tanh()$ function in Eq.(6).

With these transformer-attention weights, we aggregate all the emotion features into a single feature f_t as follows:

$$f_t = \frac{\sum_{i=1}^{n} \gamma_i f_i}{\sum_{i=1}^{n} \gamma_i}. \quad (7)$$

Cross-modal Feature Fusion

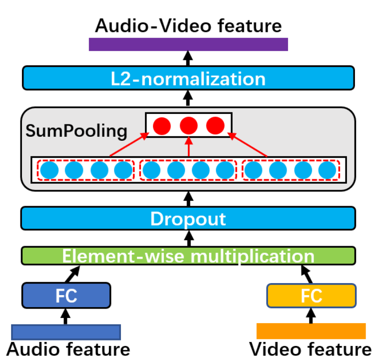

Figure 2: Our factorized bilinear pooling(FBP) module.

We apply **Factorized Bilinear Pooling(FBP)** for cross-modal feature fusion. Given two features in different modalities,i.e. the audio feature vector $\boldsymbol{a} \in \mathbb{R}^m$ for a spectrogram and visual feature $\boldsymbol{v} \in \mathbb{R}^n$ for frame sequence, the simplest cross-modal bilinear model is defined as follows:

$$z_i = \boldsymbol{a}^T \boldsymbol{W}_i \boldsymbol{v} \quad (8)$$

where $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ is a projection matrix, $\boldsymbol{z_i} \in \mathbb{R}$ is the output of the bilinear model. we use the Eq.(9) to obtain the output feature $\boldsymbol{z} = [z_1, \cdots, z_o]$. The formula derivation from formula Eq.(8) to Eq(9) was discribed in the paper[32].

$$\boldsymbol{z} = [z_1, \cdots, z_o] = \text{SumPooling}(\tilde{\boldsymbol{U}}^T \boldsymbol{a} \circ \tilde{\boldsymbol{V}}^T \boldsymbol{v}, k) \quad (9)$$

The implementation of Eq(9) is illustrated in Fig2, where $\tilde{\boldsymbol{U}}^T \boldsymbol{a}$ and $\tilde{\boldsymbol{V}}^T \boldsymbol{v}$ are implemented by feeding feature $\boldsymbol{a}$ and $\boldsymbol{v}$ to FC layers, respectively, and the function SumPooling$(\boldsymbol{x}, k)$ applies sum pooling with non-overlapped windows to $\boldsymbol{x}$. Besides, Dropout is adopted to prevent over-fitting. The $l2$-normalization ($\boldsymbol{z} \longleftarrow \boldsymbol{z}/\|z\|$) is used to normalize the energy of $\boldsymbol{z}$ to avoid the dramatical variation of the output magnitude, due to the introduced element-wise multiplication.

3 EXPERIMENTS

Dataset

In this work we use four emotion datasets to train our models, i.e. AffectNet[20], RAF-DB[16], FER+[2], AFEW[5, 6].

The human-annotated part of AffectNet dataset contains 287,651 training images and 4,000 test images, which are annotated with both emotion labels and arousal valence values. Only emotion labels are used in this task.

The RAF-DB dataset consists of 15,339 images labeled with 7-class basic emotion and 3,954 labeled with 12-class compound emotion. Only images labeled with basic emotion are used in this study.

The FER+ dataset contains 28,709 training, 3,589 validation and 3,589 test images. We combine its training data with validation data for the training split and evaluate the model performance on the test data.

The AFEW contains 773 train, 383 val and 653 test samples, which are collected from movies and TV serials with spontaneous expressions, various poses, and illuminations.

Exploration of Emotion Features

We explore emotion features in two perspectives, namely CNN backbones and pretraining emotion datasets.

For the choice of the CNN model, we compare IR50[4], ResNet18[12], and VGGFace[21] in the Table 1, where the former two models are pretrained on MS-Celeb-1M dataset and the last one on VGGFace dataset. We find that the large CNN, IR50, is superior to the other two models.

We use the well-trained IR50 model to extract features and only train softmax classifier using these features. The IR50 models pre-trained on FER+, RAF-DB, and AffectNet achieve 50.13%, 51.436%, and 53.78%, respectively. Therefore, we choose the IR50 model pretrain on AffectNet as our visual features in the following fusion experiments.

Table 1: Exploration of CNN models and pretrained emotion datasets.

Model	FER+	RAF-DB	AffectNet
VGGFace	88.84%	86.93%	51.425%
ResNet18	88.65%	86.696%	52.075%
IR50	**89.257%**	**89.075%**	**53.925%**

Exploration of Fusion Strategies

We explore three intra-modal attention strategies with the FBP cross-modal fusion. We use speech spectrogram for audio CNN, which obtains 38% on AFEW validation set individally. In the Table 2, we find the FBP improves performance for all the intra-modal fusion methods. Transformer attention for intra-modal fusion is the best for FBP.

Table 2: Evaluation of intra-modal fusion methods.

Audio \ Visual	Self	Relation	Transformer
Self	54.6%	56.9%	60.3%
Relation	54.0%	57.2%	60%
Transformer	54.8%	58%	**61.1%**

We also use log Mel-spectrogram for audio CNN, which obtains a little better performance, but the final results are very similar after intra- and cross-modal fusion. Besides, the concatenation of audio and visual vectors gets 58% accuracy in AFEW validation set with transformer attention. This is 3% lower than FBP which shows the effectiveness of FBP.

Feature Enhancement

In the Table 3, the **Basic Features** means that we only extract one feature vector for each frame. Besides, We apply 5 kinds of feature enhancement strategies as presented in Table 3. Specifically, for feature *F-Mean*, we first obtain 18 transformation frames by using three rotations, three scales, and flipping for a frame. After that, we compute the features of these 18 transformation frames and average these 18 features as the feature *F-Mean*. For the feature *F-MeanStd*, we compute the average feature and feature standard deviation of these 18 features. We then concatenate the average feature and the standard deviation as *F-MeanStd*. For the feature *F-normFFT*, we first compute the Fast Fourier transform(FFT) of the Basic Feature, and then normalize the feature and concatenate the real and imaginary parts as *F-normFFT*. For the feature *F-AR-Mean*, *A* means that the features are extracted by the models pre-trained on Affectnet, and *R* by the models pre-trained on RAF-DB. we concatenate these two mean features of two different pretrained models as *F-AR-Mean*.

Table 3: Evaluation of five feature enhancement strategies. The default setting is Rotation $\in [-2°, 0°, 2°]$, scale $\in [1, 1.03, 1.07]$

Visual Feature	Augmentation details	AFEW Val acc
Basic Feature	–-	**61.1%**
Basic Feature_RAF-DB	–-	58.5%
F-Mean	*default setting*	62.14%
F-MeanStd	*default setting*	**63.7%**
F-MeanStd-2	$Rotation \in [-15°, 0°, 15°]$ $scale \in [0.75, 1, 1.25]$	62.4%
F-NormFFT	Normalized FFT	61.35%
F-AR-Mean	*default setting*	62.92%
FG-Net	–-	59%

Table 3 shows that the five feature enhancement methods further improve the performance of FBP where the feature F-MeanStd achieves the best result on the validation set.

Table 4: Submission results of different model combinations.

Sub	Val	Test	Fusion detail
(1)	–-	**62.481%**	4 FG-Net-1
(2)	–-	59.112%	2 F-MeabStd-2 + 2 F-AR-Mean
(3)	–-	54.518%	4 FG-Net-2
(4)	64.5%	**61.41%**	4 F-MeanStd
(5)	65.5%	**62.328%**	F-Mean + F-MeanStd + F-NormFFT + F-MeanStd-2 + F-AR-Meam

Results On EmotiW2019

In the Table 4. The first three submitted models are trained on the training and validation set of AFEW, and the last two models are trained on the training set of AFEW. We find that it is difficult to choose models and fuse models if combining the validation set with the training set. We adopt class weight in all submissions, which means that we reweight the predicted scores by the square root of the sample numbers([0.15, 0.097, 0.129, 0.185, 0.138, 0.082, 0.215]).

4 CONCLUSIONS

In this paper, we exploit three types of intra-modal fusion methods, namely self-attention, relation-attention, and transformer. They are mainly used to highlight important emotion feature. For the fusion of audio and visual information, we explore feature concatenation and factorized bilinear pooling (FBP). Besides, we evaluate different emotion features, including an audio feature with both speech-spectrogram and Log Mel-spectrogram and several facial features with different CNN models and different emotion pretrained strategies. With careful evaluation, we obtain 62.48% and rank third in the EmotiW 2019 Challenge.

5 ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (U1613211), Shenzhen Basic Research Program (JCYJ20170818164704758), the Joint Lab of CAS-HK.

REFERENCES

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

[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. In *ACM ICMI*.

[3] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 2018. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters* 25, 10 (2018), 1440–1444.

[4] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. 2018. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698* (2018).

[5] Abhinav Dhall, Roland Goecke, Shreya Ghosh, and Tom Gedeon. 2019. EmotiW 2019: Automatic Emotion, Engagement and Cohesion PredictionTasks. In *ACM International Conference on Mutimodal Interaction*.

[6] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* (2012).

[7] Alan Dix. 2009. Human-computer interaction. In *Encyclopedia of database systems*. Springer, 1327–1331.

[8] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018. Video-based Emotion Recognition Using Deeply-Supervised Neural Networks. In *ACM ICMI*.

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

[10] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

[11] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

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

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

[14] Philip Nicholas Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science* 4, 1 (1980), 71–115.

[15] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. 2018. Leveraging large face recognition data for emotion classification. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 692–696.

[16] Shan Li and Weihong Deng. 2019. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE TIP* (2019).

[17] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. 2018. Multi-Feature Based Emotion Recognition for Video Clips. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 630–634.

[18] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. 2018. Multiple Spatio-temporal Feature Learning for Video-based Emotion Recognition in the Wild. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 646–652.

[19] Alex J Mitchell, Amol Vaze, and Sanjay Rao. 2009. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* 374, 9690 (2009), 609–619.

[20] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 1949. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* PP, 99 (1949), 1–1.

[21] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep face recognition.. In *BMVC*, Vol. 1. 6.

[22] Wenling Shang, Diogo Almeida, Diogo Almeida, and Honglak Lee. 2016. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *International Conference on International Conference on Machine Learning*.

[23] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2017. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 549–552.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (2017).

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

[26] Kai Wang, , Xiaoxing Zeng, Jianfei Yang, Debin Meng, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. 2018. Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[27] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. 2019. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *arXiv preprint arXiv:1905.04075* (2019).

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

[29] Jianfei Yang, Kai Wang, Xiaojiang Peng, and Yu Qiao. 2018. Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 594–598.

[30] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.

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

[32] Yuanyuan Zhang, Zi-Rui Wang, and Jun Du. 2019. Deep Fusion: An Attention Guided Factorized Bilinear Pooling for Audio-video Emotion Recognition. *arXiv preprint arXiv:1901.04889* (2019).