Semi-Supervised Multimodal Emotion Recognition with Expression MAE

Zebang Cheng
Shenzhen Technology University
Shenzhen, China
2200411013@stumail.sztu.edu.cn

Yuxiang Lin
Shenzhen Technology University
Shenzhen, China
lin.yuxiang.contact@gmail.com

Zhaoru Chen
Shenzhen Technology University
Shenzhen, China
zr.chen0@gmail.com

Xiang Li
Shenzhen Technology University
Shenzhen, China
KTTRCDL@outlook.com

Shuyi Mao
Shenzhen Technology University
Shenzhen, China
maoshuyi2020@email.szu.edu.cn

Fan Zhang
Shenzhen Technology University
Shenzhen, China
fanzhang@gatech.edu

Daijun Ding
Shenzhen Technology University
Shenzhen, China
ding_dai_jun@outlook.com

Bowen Zhang
Shenzhen Technology University
Shenzhen, China
zhang_bo_wen@foxmail.com

Xiaojiang Peng*
Shenzhen Technology University
Shenzhen, China
pengxiaojiang@sztu.edu.cn

ABSTRACT

The Multimodal Emotion Recognition (MER 2023) challenge aims to recognize emotion with audio, language, and visual signals, facilitating innovative technologies of affective computing. This paper presents our submission approach on the Semi-Supervised Learning Sub-Challenge (MER-SEMI). First, with large-scale unlabeled emotional videos, we train both image-based and video-based Masked Autoencoders to extract visual features, which termed as expression MAE (expMAE) for simplicity. The expMAE features are found to be largely complementary with other official baseline features. Second, since there is only a few labeled data, we use a classifier to generate pseudo labels for unlabeled videos which have high confidence for a certain category. In addition, we also explore several advanced large models for cross-feature extraction like CLIP, and apply factorized bilinear pooling (FBP) for multimodal feature fusion. Our methods finally achieved 88.55% in F1 score on MER-SEMI, ranking second place among all participating teams.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Human-centered computing** → HCI design and evaluation methods.

KEYWORDS

Multimodal Emotion Recognition, Semi-Supervised Learning, Masked Autoencoder

ACM Reference Format:
Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. 2023. Semi-Supervised Multimodal Emotion Recognition with Expression MAE. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581783.3612840

1 INTRODUCTION

Multimodal emotion recognition (MER) has garnered significant attention in recent years due to its potential applications in various domains, including human-computer interaction [12], AI in healthcare [10], and education [8]. Researchers have made notable progress in analyzing facial expressions and visual cues in videos, leveraging techniques such as deep learning and multimodal fusion [1, 3, 13]. MER research typically centers on video data, which is dissected into three modalities. The acoustic modality encompasses voice, speech, and audio. HuBERT [7] introduces a self-supervised approach with an unsupervised clustering step, addressing problems in the acoustic field through masked prediction of hidden units. The lexical modality extracted emotional information through semantic and contextual analysis of the spoken words from the narration in video clips [5]. MacBERT [4] mitigates the gap between the pre-training and fine-tuning stages by masking a word with a similar word. The visual modality contains facial and body language cues that may indicate emotions. DEFR [17] presents a dynamic facial expression recognition transformer, featuring a convolutional spatial transformer and a temporal transformer. These enable robust spatial and temporal facial feature extraction. However, a challenge persists in the effective integration of these multimodal inputs, which is crucial for achieving accurate emotion recognition.

In response to these prevailing challenges, the ACM MM unveiled MER2023 [9]. The MER-SEMI sub-challenge introduces a dataset comprising a large volume of unlabeled videos encouraging the application of semi-supervised learning for enhanced MER. However, the brevity of these video clips and skewed label distribution further compound the challenge of accurately extracting

[*] Corresponding author

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612840

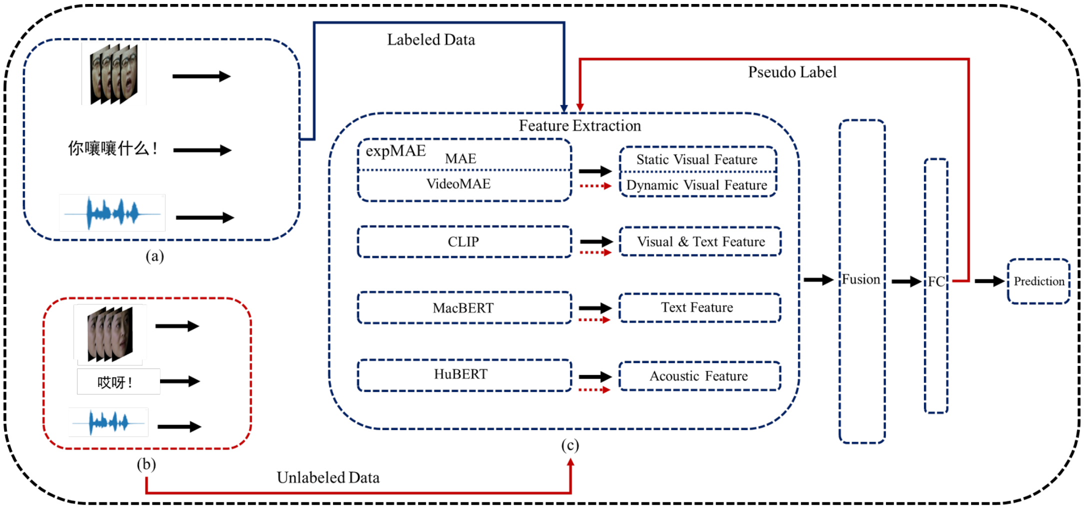

Figure 1: Illustration of the pipeline of our model. (a) Labeled input. (b) Unlabeled input. (c) Feature extraction using multi-modality.

emotional context. In this paper, we present our participation in this challenge by cross-features extraction, a semi-supervised training strategy. An overview of our pipeline is depicted in Figure 1.

In the visual modality, where Masked Autoencoders [6] (MAE) have been showing promising results by pre-training on a large number of unlabeled images. However, MAE is limited when dealing with videos, primarily due to their inability to extract temporal information from the various frames. Furthermore, the videoMAE [14] employs a more appropriate pre-train method, which masks the visual information extend on the video timeline, learning the temporal information. To combine the advantages of static and dynamic features as mentioned above, we train the image-based and video-based MAE from scratch using the unlabeled data, termed expression MAE (expMAE). The result has shown expMAE outperformance in MER compared to single MAE. In terms of multi-modality, we also found that expMAE complements the other official baseline features effectively.

Based on the extracted features mentioned above, we construct an initial model to recognize basic emotions. However, given the scarcity of labeled data and the imbalanced distribution within the training set, we draw inspiration from the classic pseudo-label semi-supervised method [2]. Following this approach, we generate pseudo labels for the unlabeled data and incorporate them into the training set. Besides, we explored several advanced large models for cross-feature extraction and apply factorized bilinear pooling (FBP) [19] for multimodal feature fusion. Recognizing the relatively weaker performance of the lexical modality, we employed visual-text and audio-text cross-feature models, including CLIP [11] and Tacotron[16]. Overall, our method achieved an 88.55% F1 score on MER-SEMI.

2 METHODS

This section focuses on our semi-supervised MER method. Initially, we extract features from various modalities using a small labeled training set. We then perform feature fusion and train a naive emotion classifier. Subsequently, we leverage a dataset of over 70,000 unlabeled samples to generate pseudo-labels through the trained classifier. To address the class imbalance, we employ a strategy that involves selecting a subset of videos with added pseudo-labels and incorporating them into the training set for joint training. The overall framework is shown in Figure 1.

2.1 Expression Masked Autoencoders

MAE containing an Encoder-Decoder structure, are a type of self-supervised learners for computer vision. It segments the input image into patches, with a randomized 75% masking, then the unobscured patches are processed through the Encoder to derive a condensed representation of the original input. Decoder capitalizes on the encoder's output to reconstruct the source image. Once the encoder has been trained, it can be directly reused for downstream tasks. This flexibility enables the formation of various network configurations to meet specific requirements. In dealing with MER-SEMI challenge, we select 16 frames from the video of the dataset, and in each training epoch, we randomly chose one of them to train the model, while in the inference part, we use all of the 16 frames and average the output vectors.

The MAE model's limitations become apparent as it can only glean static features, not accommodating changes in facial expressions during the progression of a video. To tackle this issue, VideoMAE [14] is designed to process video inputs and apply a tube masking strategy to prevent inadvertent feature information leakage. This technique effectively derives dynamic visual features that include temporal information directly from the video.

To further accentuate the expression information within the video, we train expMAE, as mentioned above, to better suit the specific task of emotion recognition. In addition, we applied a preprocessing step involving discarding the background and centering the human face, which allows us to extract the most pertinent and expressive facial features.

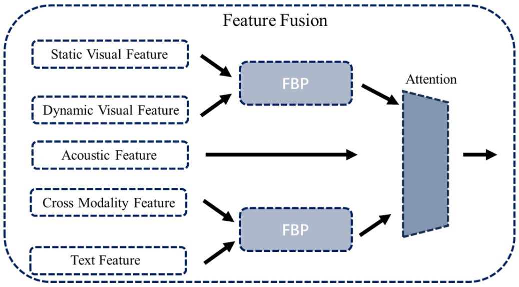

Figure 2: Illustration of the Fusion block in Figure 1.

2.2 Multimodal Fusion for Classification

For acoustic and lexical modalities that also contain emotional information, we first leverage the official baseline features provided by HuBERT and MacBERT. To address the suboptimal performance of the lexical modality, we creatively explore the role of cross-modal features in MER tasks. The Contrastive Language-Image Pre-Training (CLIP) model [11] is selected to handle visual-text features, while a variant of Tacotron [16] is employed for audio-text features. CLIP was pre-trained on a vast text-image corpus, it employs contrastive learning to align textual and image representations in a shared embedding space. Tacotron was a pre-trained speech synthesis model to integrate text and speech features. These cross-modality features enhance the learning capabilities of the classifier, allowing it to better comprehend the transitions between different modalities.

After collecting features from different modalities, we apply the factorized bilinear pooling (FBP) module [19] to fuse each contextually related feature, generating the fused features (as shown in Figure 2). We then follow the official approach of training with the attention module.

2.3 Semi-Supervised Training Strategy

In the proceedings, we observed that the class in label data was skewed, making some classes hard to learn in our model (Figure 3). To address this issue, we employed a semi-supervised strategy to balance the data. The traditional data augmentation techniques such as over-sampling and under-sampling methods are not suitable for this task, as they are prone to overfitting and information loss. we draw inspiration from the classic pseudo-label semi-supervised method [2] by using unlabeled data with pseudo-label.

Specifically, we leverage the pseudo-labeled samples generated from the unlabeled dataset. From the test set, we selected the samples predicted as the surprise class with model confidence (calculated from the output vectors from the softmax function) of 0.8 or higher, as they were likely to have more distinct and reliable features. We followed a similar approach for the worried and sad classes. Ultimately, we included these selected pseudo-labeled samples, along with the original labeled training set, resulting in a final training set from 3373 increase to 4138 samples.

The incorporation of the pseudo-labeled samples with more discernible features aimed to provide the model with a better representation of the underrepresented classes, finally improving its ability to generalize and make accurate predictions.

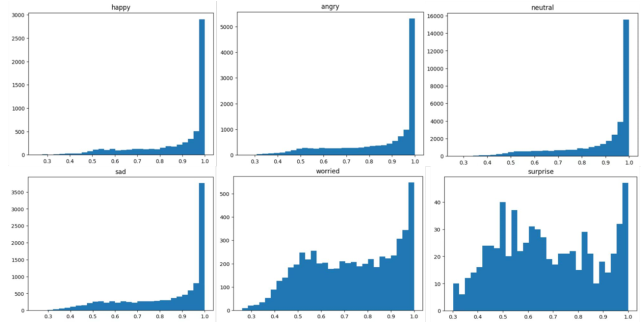

Figure 3: The distributions of the confidence scores from the initial model, from left to right, top to bottom are happy, angry, neutral, sad, worried, and surprise, shows that model is hard to predict the class of surprise, worried, with a lower confidence score.

Table 1: Unimodal results of the baseline.

Feature	Train&Val		
	$metric_e$ (↑)	$metric_v$ (↓)	$metric$ (↑)
Acoustic Modality			
HuBERT-base[7]	60.72	1.53	0.22
HuBERT-large[7]	**65.67**	**1.27**	**0.34**
Lexical Modality			
MacBERT-base[4]	40.96	2.42	-0.19
MacBERT-large[4]	**42.62**	**2.39**	**-0.17**
Visual Modality			
MANet-RAFDB[18]	57.48	1.38	0.23
DFER[17]	43.63	2.02	-0.06
MAE [6]	60.01	1.42	0.25
VideoMAE [14]	61.98	1.33	0.28
expMAE	**62.56**	**1.29**	**0.30**
Cross Modality			
Tacotron-Var [16]	44.01	2.44	-0.17
CLIP [11]	**60.99**	**1.26**	**0.29**

3 EXPERIMENTS AND RESULTS

In this section, we present the experiments conducted and results. We process video into three modalities and two cross-modalities, and experiments were carried out on both of it. Then we investigated different strategies to determine the classifier for emotion recognition tasks. Ultimately, we compare our semi-supervised strategy with different data augmentation techniques.

For emotion classification $metric_e$ is the average F1 score across six classes, and for emotion’s valance regression, $metric_v$ is the mean square error. The combined metric is computed as $metric = metric_e - 0.25 \times metric_v$. Although in MER-SEMI, $metric_v$ is not computed as a score, we include it as part of our metric to boost the model’s ability of emotion recognition.

3.1 Unimodal Comparison

For the acoustic and lexical modalities, we utilize HuBERT and MacBERT respectively, which perform best in the MER2023 paper

Table 2: Performance of the baseline with uni-modality and multi-modality, we select acoustic features from HuBERT-large (HL), lexical features from MacBERT-large (ML), and visual features from MANet-RAFDB (MR), MAE, VideoMae, and expMAE, cross features from CLIP and Tacotron-variant(T-var)."A", "L", "V" and "C" represents the acoustic, lexical, visual, and cross modalities, respectively.

A	L	V	C	Train&Val		
				$metric_e$ ($\uparrow$)	$metric_v$ ($\downarrow$)	$metric$ ($\uparrow$)
				Bimodal Results		
HL	ML	–	–	67.02	1.16	0.38
HL	–	MR	–	72.92	0.86	0.52
HL	–	MAE	–	71.32	0.89	0.49
HL	–	VideoMAE	–	72.90	0.81	0.52
HL	–	expMAE	–	**73.40**	**0.78**	**0.53**
HL	–	–	CLIP	71.32	0.79	0.51
HL	–	–	T-Var	65.24	1.19	0.35
–	ML	MR	–	61.19	1.28	0.29
–	ML	MAE	–	63.66	1.32	0.30
–	ML	VideoMAE	–	66.70	1.12	0.39
–	ML	expMAE	–	**67.13**	**1.10**	**0.39**
–	ML	–	CLIP	64.14	1.12	0.35
–	ML	–	T-Var	53.64	2.19	-0.01
				Trimodal Results		
HL	ML	MR	–	73.39	0.87	0.52
HL	ML	MAE	–	73.8	0.92	0.49
HL	ML	VideoMAE	–	72.42	0.78	0.53
HL	ML	expMAE	–	**74.52**	**0.76**	**0.55**
				Multi Results		
HL	ML	MAE	T-Var	73.07	0.84	0.518
HL	ML	VideoMAE	T-Var	74.52	0.77	0.552
HL	ML	expMAE	T-Var	74.65	0.76	0.556
HL	ML	MAE	CLIP	72.65	0.89	0.502
HL	ML	VideoMAE	CLIP	74.78	0.74	0.561
HL	ML	expMAE	CLIP	**75.01**	**0.66**	**0.585**

[9]. When dealing with the visual modality, we compare the performance with MANet-RAFDB [18], as well as DFER, MAE, VideoMAE, and expMAE. In the cross-modality, we resort to the use of the CLIP model and a Tacotron combined with a GE2E Speaker Encoder [15], namely Tacotron-Var, to effectively extract cross-modal features. The results of these experiments are presented in Table 1.

Among the visual modality, expMAE significantly perform better $metric_e$ and $metric$ values than others, while also demonstrating lower $metric_v$ values.

3.2 Multimodal Comparison

We carried out experiments on multimodal features. These features were derived by leveraging the combined features from visual, acoustic, lexical, and cross modalities. The results of these experiments are presented in Table 2.

By combining the best unimodal feature extractor, we proceeded to test various classifiers for emotion recognition tasks. We employed an ensemble voting Support Vector Machine (SVM) with Radial Basis Function (RBF) kernels, naive attention, Transformer-Encoder with 3 and 6 layers, and bilinear sum pooling methods with FBP module and attention (refer Figure 2). The outcomes of these varied experiments are presented in Table 3.

Table 3: Comparison of different classifiers.

Classifier	Train&Val		
	$metric_e$ ($\uparrow$)	$metric_v$ ($\downarrow$)	$metric$ ($\uparrow$)
SVM	61.70	–	–
Transformer (3 layers)	73.55	0.89	0.512
Transformer (6 layers)	72.14	0.88	0.499
naive attention	75.01	**0.66**	**0.585**
FBP	**75.53**	0.82	0.550

Table 4: Comparison of semi-supervised strategy, MER-SEMI shows the test result on the MER-SEMI dataset.

Augmentation	Train&Val			MER-SEMI
	$metric_e$ ($\uparrow$)	$metric_v$ ($\downarrow$)	$metric$ ($\uparrow$)	$metric_e$ ($\uparrow$)
baseline	75.53	0.82	0.550	0.8799
threshold (reliable label)	76.01	0.75	0.570	0.8759
threshold (weak reinforce)	**78.10**	**0.64**	**0.622**	**0.8855**

3.3 Semi-Supervised Strategy Performance

To mitigate the impact of skewed class distribution on the classifier, we introduced two data augmentation techniques: threshold-based reliable labeling and threshold-based weak reinforcement. Using the highest performing method from Section 3.2, we generate soft pseudo-labels from the unlabeled data. In the reliable labeling strategy, samples with pseudo-label confidence exceeding 0.8 were filtered and added to the training set. In contrast, the weak reinforcement strategy only added the unbalanced classes, since they are hard to learn by the model (Figure 3). The outcomes of these experiments are documented in the following Table 4.

4 CONCLUSION

In our research, we explored multimodal emotion recognition, focusing on the MER-SEMI challenge of the ACM MM 2023 Grand Challenge. expMAE proved instrumental in handling this challenge, effectively extracting both static and dynamic visual expression features from videos and largely complementary with other modality features. Furthermore, we utilized a well-suited semi-supervised learning strategy to address the limitations posed by limited and imbalanced training data, resulting in composite features that significantly enhanced emotion recognition. Our experiments also demonstrated the effectiveness of the BPF module across these modalities.

We would explore two aspects for future direction. Firstly, a stronger semi-supervised training strategy may be utilized in the task, such as Multi-view Learning, Network Embedding. Secondly, it may be interesting to finetune both the encoder (such as expMAE, MacBERT, HuBERT) and the classifier (such as the fusion module) together rather than only train on the decoder.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (62176165), the Natural Science Foundation of Top Talent of SZTU (grant no.GDRC202131), and the Stable Support Projects for Shenzhen Higher Education Institutions (SZWD2021011).

REFERENCES

[1] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends* 2, 02 (2021), 52–58.

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[3] Tanja Bänziger, Didier Grandjean, and Klaus R Scherer. 2009. Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). *Emotion* 9, 5 (2009), 691.

[4] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. *arXiv preprint arXiv:2004.13922* (2020).

[5] Tiquan Gu, Hui Zhao, Zhenzhen He, Min Li, and Di Ying. 2023. Integrating external knowledge into aspect-based sentiment analysis using graph neural network. *Knowledge-Based Systems* 259 (2023), 110025. https://doi.org/10.1016/j.knosys.2022.110025

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. [n. d.]. *Masked autoencoders are scalable vision learners.* IEEE. 15979–15988 pages.

[7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[8] Maryam Imani and Gholam Ali Montazer. 2019. A survey of emotion recognition methods with emphasis on E-Learning environments. *Journal of Network and Computer Applications* 147 (2019), 102423. https://doi.org/10.1016/j.jnca.2019.102423

[9] Zheng Lian, Haiyang Sun, Licai Sun, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, et al. 2023. MER 2023: Multi-label Learning, Modality Robustness, and Semi-Supervised Learning. *arXiv preprint arXiv:2304.08981* (2023).

[10] Prameela Naga, Swamy Das Marri, and Raiza Borreo. 2023. Facial emotion recognition methods, datasets and technologies: A literature survey. *Materials Today: Proceedings* 80 (2023), 2824–2828.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[12] Gaurav Sinha, Rahul Shahi, and Mani Shankar. 2010. Human computer interaction. In *2010 3rd International Conference on Emerging Trends in Engineering and Technology*. IEEE, 1–4.

[13] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2011. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing* 3, 2 (2011), 211–223.

[14] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. [n. d.]. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training.

[15] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4879–4883.

[16] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[17] Zengqun Zhao and Qingshan Liu. 2021. Former-DFER: Dynamic Facial Expression Recognition Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 1553–1561. https://doi.org/10.1145/3474085.3475292

[18] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. [n. d.]. Learning deep global multiscale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing* 30 ([n. d.]), 6544–6556,.

[19] Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. 2019. Exploring emotion features and fusion strategies for audio-video emotion recognition. In *2019 International conference on multimodal interaction*. 562–566.