Facial Expression Recognition Based on Multi-modal Features for Videos in the Wild

Chuanhe Liu[1, *, †], Xinjie Zhang[2, *], Xiaolong Liu[1], Tenggan Zhang[2]
Liyu Meng[1], Yuchen Liu[2], Yuanyuan Deng[1], Wenqiang Jiang[1]
[1] Beijing Seek Truth Data Technology Co.,Ltd.
[2] School of Information, Renmin University of China

Abstract

This paper presents our work to the Expression Classification Challenge of the 5th Affective Behavior Analysis in-the-wild (ABAW) Competition. In our method, the multimodal features are extracted by several different pertained models, which are used to build different combinations to capture more effective emotion information. Specifically, we extracted efficient facial expression features using MAE encoder pre-trained with a large-scale face dataset. For these combinations of visual and audio modal features, we utilize two kinds of temporal encoders to explore the temporal contextual information in the data. In addition, we employ several ensemble strategies for different experimental settings to obtain the most accurate expression recognition results. Our system achieves the average F1 Score of 0.4072 *on the test set of Aff-wild2 ranking 2nd, which proves the effectiveness of our method.*

1. Introduction

Affective computing has an extensive spectrum of application requirements in human-computer interaction, security, robotics manufacturing, automation, medical, and communications. Actively creating machines that can understand the feelings, emotions, and behaviors of humans would help them interact with humans more intimate way and serve effectively [1]. Facial expressions are one of the most powerful, natural, and pervasive signals that humans use to communicate their emotional state and intentions. Machines can analyze human expressions leading to understanding human emotions. The majority of recent research in emotion recognition is based on deep learning, which requires a large quantity of labeled data. Nowadays, there are several datasets, such as Aff-wild [2–4] and Aff-wild2 [3–12], which provides us with large-scale data with high-quality labels, which are convenient for training neural networks and increasing the accuracy of expression recognition.

The facial expressions information can be mainly obtained from the visual modality. Currently, there does not exist a certain expression dataset that is much larger than the others in scale. For extracting more efficient expression features for a particular dataset, the feature extractor can be pre-trained with multiple datasets. Different models as feature extractors do not extract the same features from the facial images and are most likely to obtain complementary sentiment information. Therefore, a richer and more complete visual facial expression representation can be obtained by using multiple visual features from multiple pre-trained models. With the rise of image self-supervised models, such as MAE [13], a large amount of unlabeled data can also be utilized as training data. MAE can improve its ability to encode face images by using a large number of face images, which contributes to the acquisition of emotional information.

Nevertheless, it is well known that audio modality also contains certain emotional information. The information of single modality may be affected by various noises. To obtain a more complete emotional representation, information from multiple modalities can be utilized. The multiple modalities' information can supplement and enhance the unimodality information to a degree, improving the recognition ability, generalization, and robustness of the model.

To improve the performance of emotion recognition, we design a multimodal expression recognition system for videos, which investigates a better combination of features and incorporates multimodal information more efficiently. Our system for the expression recognition contains several key components. First, we examine the officially provided aligned and cropped images and labels, and then supplement the images containing labels but no official images. Second, multiple pre-trained feature extractors are employed to extract visual and audio features. Then, we de-

[*] These authors contributed equally to this work and should be considered co-first authors.

[†] Corresponding Author.

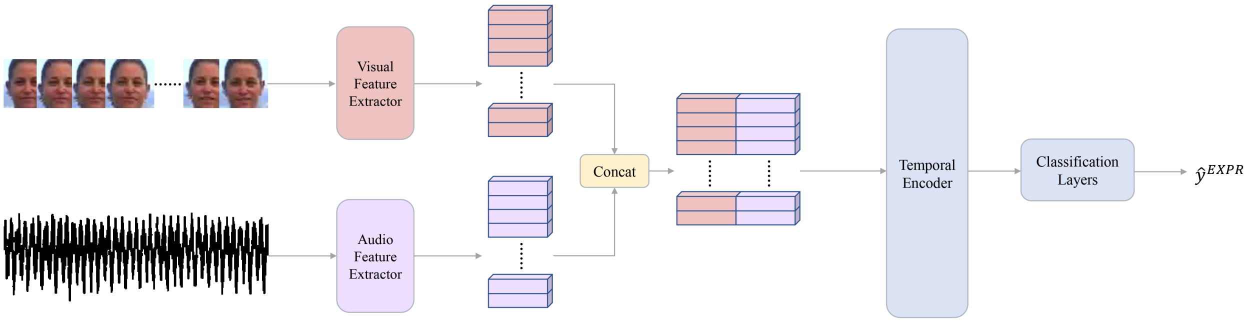

Figure 1. The overall pipeline of our method.

signed multimodal feature combinations and concatenated multiple features into multimodal feature representations. Multimodal features are fed into the temporal encoder. Two different types of temporal encoders, LSTM [14] and Transformer [15], are applied to extract contextual information from the multimodal features. Several techniques are also utilized for optimization. Finally, we adopted several ensemble strategies to ensemble the experimental results for different settings to raise the accuracy of recognition.

2. Related Works

During the ABAW series events, facial expression recognition has consistently been a high-profile task. In this section, we will provide an overview of some relevant approaches on facial expression recognition in ABAW competitions.

[16–18] incorporates both temporal and spatial information to enhance the representational capacity of the network utilizing multi-modal features. [19–21] only utilize visual features to make predictions about facial expressions. Specially, [20] employs a semi-supervised learning strategy to fully exploit the available data and achieve optimal model performance. By leveraging both labeled and unlabeled data, the model is able to better capture the underlying patterns in the data and generalize well to new data.

Overall, the facial expression recognition task remains a challenging problem in the ABAW series competitions. Various techniques and models have been proposed to improve the performance of the architectures, including the use of multi-modal features, temporal and spatial modeling, and semi-supervised learning strategies. Our proposed approach not only utilizes multi-modal features but also incorporates temporal and spatial information, resulting in enhanced representational capacity and competitive performance.

3. Method

For a given video X, it can be separated into two parts, the visual data X^{vis} and the audio data X^{aud}. The visual data can be stated as an image frames sequence$\{F_1, F_2, ..., F_n\}$, and n denotes the number of image frames in X. The goal of the Expression Classification Challenge is to predict the sentiment label for each frame in the video.

The overall pipeline is illustrated in Figure 1. In our approach, the raw data is first preprocessed by converting the video data into a sequence of image frames and audio. Then the features of the sequences are extracted using multiple feature extractors, such as MAE-based [13] model, DenseNet-based [22] model, Wav2Vec 2.0 [23] model and so on. As the phase of multimodal fusion, we combined visual and audio features to a huge feature vector as multimodal feature representation. Multimodal features are encoded by a temporal encoder and then sent to the classification layers to obtain the predicted expression class. The feature extractors and temporal encoders will be described in detail on the following sections. During the training phase, we fed the same feature into the temporal encoder and head twice because of using RDrop [24]. But during the inference phase, we just infer once to get the prediction.

3.1. Pre-processing

Firstly, the officially provided video data is divided into multiple image frames. For each image frame, the face and facial landmarks are recognized by the face detector. The face part is cropped out according to the bounding box to facilitate the extraction of more accurate emotional information later. In order to be consistent with the official labels provided, we use the cropped and aligned face images provided by the competition for the actual processing.

We matched the labels with the images one by one, and found that some of the images corresponding to the labels did not exist in the cropped and aligned set given by the competition, probably because the face images of the corresponding frames were not detected in the video due to the lighting, angle, and other circumstances. For each of these non-existent images, we complement it by finding the nearest frame in the temporal dimension.

3.2. Multimodal Feature Representation

For more efficient use of information from video data, in the feature extraction phase, we extracted various features in both visual and audio modalities by using multiple types of feature extractors. Afterwards, visual and audio features are merged as multimodal feature representation.

3.2.1 Visual Features

For visual modality, many models used for extracting features including MAE-based [13] model , DenseNet-based [22] model, IResNet100-based [25] model, IResNet100-based [25] facial action unit (FAU) detection model and the MobileNet-based [26] model.

MAE-based Feature The first type of visual feature is the feature extracted by pretrained MAE-based model [13]. MAE is a self-supervised model that trains label-free data by masking random patches from the input image and reconstructing the missing patches in the pixel space. We used a face dataset collection of scale 1.2 million, including AFEW [27], DFEW [28], AVEC2019 [29], Emotionet [30], FERV39k [31],MEC2017 [32], to pre-train the MAE encoder. Considering that it is expensive and labor-intensive to obtain a large amount of labeled data, we utilize unlabeled data to train the MAE encoder. past studies have found that action units (au) [33] are closely related to facial expressions. We designed the training task by extracting rough au labels using the OpenFace tool [34] and generating emotion labels by combining the simple correspondence between au and expressions. In this way, we intend to enhance the ability of MAE encoder to mine potential emotion representations through self-supervised pre-training and pseudo-labeled supervised pre-training, and be to be able to generate representations containing a large amount of facial expression information. We denoted this kind of feature as mae, and the dimension of it is 768.

DenseNet-based Feature The second type of visual features is extracted by a pre-trained DenseNet model. Specifically, the DenseNet model is pre-trained on the FER+ and the AffectNet datasets. The dimension of the DenseNet-based visual features is 342. And this kind of feature is denoted as densenet.

IResNet100-based Feature The third type of visual feature is the feature extracted by pretrained IResNet100-based model, and the dimension of IResNet100-based feature is 512. A large-scale facial expression recognition data which consists of FER+ [35], RAF-DB [36] [37] and AffectNet [38] dataset is utilized to pretrain our facial expression recognition IResNet100-based model, which is denoted as ires100. And a commercial authorized facial action unit detection dataset is used to pretrain the other IResNet100-based model, which is denoted as fau.

MobileNet-based Feature The fourth type of visual feature is the feature extracted by pretrained MobileNet-based model, and the dimension of MobileNet-based feature is 512. The MobileNet model is trained on valence-arousal estimation task of AffectNet to further enhance feature representation.

3.2.2 Audio Features

For audio modality, many models used for extracting audio features including eGeMAPS [39], ComParE 2016 [40], VGGish [41], Wav2Vec 2.0 [23], ECAPA-TDNN [42] and HuBERT [43].

Hand-craft Features The first type of audio feature is hand-craft features, which consists of eGeMAPS, ComParE 2016 and fbank. eGeMAPS and ComParE 2016 can be extracted using openSmile, and the dimension of these features are 23 and 130. The dimension of fbank is 80. For convenience, we denotes them as egemaps, compare and fbank.

Deep Features The second type of audio feature is deep features, which consists of Wav2Vec 2.0, ECAPA-TDNN, VGGish and HuBERT. The dimension of Wav2Vec 2.0 feature is 1024, the dimension of ECAPA-TDNN feature is 512, the dimension of VGGish feature is 128 and the dimension of HuBERT is 512. We denotes them as wav2vec, ecapatdnn, vggish and hubert respectively.

3.2.3 Multimodal Fusion

Given the visual features f^v and audio features f^a corresponding to a frame, they are first concatenated and then fed into a fully-connected layer to produce the multimodal features f^m. It can be formulated as follows:

$$f^m = W_f[f^v; f^a] + b_f \tag{1}$$

where W_f and b_f are learnable parameters.

3.3. Temporal Encoder

Due to the limitation of GPU memory, we split the videos into segments at first. Given the segment length l and stride p, a video with n frames would be split into $[n/p] + 1$ segments, where the i-th segment contains frames $\{F_{(i-1)*p+1}, ..., F_{(i-1)*p+l}\}$. With the multimodal features of the i-th segment f_i^m, we employ a temporal encoder to model the temporal context in the video. Specifically, two kinds of structures are utilized as the temporal encoder, including LSTM and Transformer Encoder.

3.3.1 LSTM-based Temporal Encoder

Long and short term memory networks (LSTM) are commonly applied to model sequential dependencies of time sequences. In a practical game, we use LSTM to model the temporal relationships in a sequence of frame images from a video. f_i^m means the input of LSTM are multimodal features for the i-th video segment.

Additionally, to encode the contextual information between two sequential segments, s_{i-1} and s_i, the last hidden state of s_{i-1} is also fed into the LSTM. It can be formulated as follows:

$$g_i, h_i = \text{LSTM}(f_i^m, h_{i-1}) \tag{2}$$

where h_i denotes the hidden states at the end of s_i. There is no overlap between two adjacent segments, in other words, the stride p is the same as the segment length l, which ensures that the last frame of the former segment and the first frame of the latter segment are consecutive.

3.3.2 Transformer-Based Temporal Encoder

We used a Transformer Encoder to model the temporal feature in the video segment, which can be formulated as:

$$g_i = \text{TRMEncoder}(f_i^m) \tag{3}$$

Unlike LSTM, the transformer encoder just models the context in a single segment and ignores the dependencies of frames between segments.

3.4. Classification Layers

We used fully-connected blocks to build the classifer of our method. One fully-connected block is consists of $\{$ FC, Dropout, ReLU $\}$, and the dropout ratio is the same as dropout ratio in temporal encoder. Specially, the last fully-connected block is only consists of a FC layer, and the output dimension is set to 8, which is equal to the number of expressions in expression recognition challenge.

3.5. Loss Function

In the training phase, we utilize the RDrop loss which can be formulated as:

$$\begin{aligned} L^{EXPR} = & \frac{1}{2} * (CE(\hat{y}_1, y) + CE(\hat{y}_2, y)) + \\ & \alpha * \frac{1}{2} * (KL(\hat{y}_1, \hat{y}_2) + KL(\hat{y}_2, \hat{y}_1)) \end{aligned} \tag{4}$$

where CE denotes cross entropy loss, KL denotes Kullback-Leibler divergence loss. y_1 and y_2 denotes the first and the second inference prediction logits, y denotes the label.

4. Experiments

4.1. Dataset

Expression(Expr) Classification Challenge in fifth ABAW competition is based on Aff-Wild2, a large-scale dataset. Aff-Wild2 consists of 548 videos and is annotated with 8 expressions (i.e. neutral, anger, disgust, fear, happiness, sadness, surprise, and other). As for the feature extractors, we used some other dataset for pretraining, which consists of FER+ [35], RAF-DB [36, 37], and AffectNet [38]. In addition, an authorized commercial FAU dataset is also used for pretraining visual feature extractor, which consists of 7K images in 15 face action unit categories(AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU11, AU12, AU15, AU17, AU20, AU24, and AU26). As for the audio feature extractors, we used some different open-source models to extract features. Wav2Vec 2.0 [23], HuBERT [43], and ECAPA-TDNN [42] are the deep open-source model for extracting audio features.

4.2. Evaluation Metric

According to the competition regulations, we use the average F1 score across 8 categories, which can be formulated as:

$$p = \frac{\sum_i^8 F1(\hat{y}_i, y_i)}{8} \tag{5}$$

where $F1$ denotes F1 score, $\hat{y}_i$ and y_i denotes the i-th category of prediction and label respectively.

4.3. Experiment Settings

First, declare that we used Adam [44] optimizer to train models for 25 epochs. As for the Transformer model, the learning rate is 0.0001, the α in equation 3.5 is 5, the affine dimension is 1024, the number of Transformer encoder layers is 4, the attention heads number is 4, the dropout ratio in the Transformer encoder layer is 0.3, the sequence length of one segment is 128. The classifier is consists of 2 fully-connected blocks, and the hidden size of head layers are $\{512, 256\}$.

4.4. Overall Performance on Validation Set

Table 1 shows the results of our method on validation set. Among all the results we post in the table, we utilized the same training settings as we described in experiment settings. As is shown in the table, both Transformer and LSTM models achieve competitive performance and the LSTM model achieves better performance than the Transformer model. Different feature combinations can lead to different result using Transformer and LSTM model.

4.5. Model Ensemble

During the model selection phase, we trained multiple models with varying structures, feature combinations, and hyper-parameters, all of which achieved competitive performance. To further improve the robustness and performance of our system, we employed a vote strategy. The results obtained from the ensemble of different architectures and

Table 1. The performance of our method on the validation set.

Model	Visual Features	Audio Features	F1
Transformer	ires100,mae	ecapatdnn,hubert,wav2vec	0.38362
Transformer	densenet,ires100,mae	ecapatdnn,hubert	0.39112
Transformer	densenet,fau,ires100,mae	ecapatdnn,hubert	0.3938
lstm	ires100,mae	ecapatdnn,hubert,wav2vec	0.37972
lstm	fau,ires100	ecapatdnn,hubert	0.38928
lstm	ires100,mae	wav2vec	0.39385
lstm	densenet,ires100,mae	wav2vec,ecapatdnn,hubert	0.40178
lstm	densenet,ires100,mae	ecapatdnn,hubert	0.41410

Table 2. The results of each single model and the ensemble of them for the expression prediction task on the validation set.

Model	Visual Features	Audio Features	F1
Transformer	ires100,mae	ecapatdnn,hubert,wav2vec	0.38362
Transformer	fau,ires100	ecapatdnn,hubert,wav2vec	0.35119
Transformer	densenet,fau,ires100	ecapatdnn,hubert	0.36087
Transformer	densenet,mae,ires100	ecapatdnn,hubert	0.39380
LSTM	densenet,mae,ires100	ecapatdnn,hubert,wav2vec	0.40178
LSTM	mae,ires100	ecapatdnn,hubert,wav2vec	0.40832
LSTM	fau,ires100	ecapatdnn,hubert	0.38928
LSTM	densenet,mae,ires100	ecapatdnn,hubert	0.40889
LSTM	densenet,mae,ires100	ecapatdnn,hubert	0.41410
Ensemble			**0.45774**

Table 3. The performance of our method on the 5-fold cross-validation. The First 4 folds are from training set, and the last fold is the original validation set.

	F1
Fold 1	0.43697
Fold 2	0.38015
Fold 3	0.35646
Fold 4	0.39170
Fold 5	0.38146
Average	0.38935

feature combinations are presented in Table 2, which indicates that this approach yields significant benefits on the validation set.

4.6. Cross Validation

As the experiment proceeded, we found that the model overfitting was very serious, to address the issue of overfitting, we implemented K-fold cross-validation with K set to 5, because we had 4 times as many training videos as validation videos. We used the original validation set of Aff-Wild2 as the 5th fold. We trained a Transformer encoder model using mae, ires100, wav2vec, ecapatdnn and hubert features. Table 3 presents the results obtained using this approach.

4.7. Ablation Study

We conducted an ablation study to evaluate the impact of different features and feature combinations on the performance of our transformer-based model. Table 4 shows the results obtained on the validation set using the same experimental setup as in the training phase, except for the feature combinations.

Table 4. Ablation study of features on the validation set.

Visual Features	Audio Features	F1
densenet,fau	ecapatdnn	0.35276
densenet,fau	ecapatdnn,fbank	0.33887
densenet,fau	ecapatdnn,hubert	0.34391
densenet,fau,ires100	ecapatdnn,hubert	0.36266
densenet,fau,ires100	ecapatdnn,wav2vec	0.36944
densenet,ires100,mae	ecapatdnn,hubert	0.39380

4.8. Results on the test set

In this competition, we have adopted five submission strategies. We briefly describe each strategy and show the performance on the test set in this section. Table 5 shows the strategies and results for each of our five submissions. The first submission strategy is the ensemble of multiple features and models generated by fusing the test set after training on the officially given training set and selecting the one that performs better on the officially provided validation set. As for the second submission strategy, the best model and feature combination obtained by the ensemble on the officially divided validation set is chosen. For the third and fourth ones, the training and validation sets are combined as a new training set, and the test set results of multiple epochs of multiple model and feature combinations are picked for the ensemble. Furthermore, the fourth strategy also combines the results of the first strategy. The last strategy takes the result of k-fold cross-validation. Here, the 5-fold is divided according to the ratio of the official dataset and the validation set. The above mentioned ensemble uses the vot-

Table 5. The results on the test set of different submissions.

Submit	Strategy	F1
1	Ensemble 1	0.3856
2	Ensemble 2	0.3845
3	Train-Val-Mix	0.4020
4	Train-Val-Mix&Ensemble 1	0.3940
5	5-Fold	0.4072

Table 6. The results of the Expression Classification Challenge.

rank	Teams	F1
1	Netease Fuxi Virtual Human [16]	0.4121
2	**Ours**	**0.4072**
3	CtyunAI [17]	0.3532
4	HFUT-MAC [18]	0.3337
5	HSE-NN-SberAI [19]	0.3292

ing method. Our method comes runner up in this challenge, with a small difference to first place, and the results are shown in the Table 6.

5. Conclusion

In this paper, we propose our framework for the Expression Classification Challenge of the fifth Affective Behavior Analysis in-the-wild (ABAW) Competition. Our approach leverages information from multiple modalities in the spatiotemporal dimension. Various temporal encoders are applied to capture the temporal contextual information in the video. We pre-trained the MAE encoder using a large amount of unlabeled face data to enhance the ability of the model to encode faces and extract expression information. In addition, we design multiple high-quality feature combinations to extract more effective emotional information. Our method achieves a performance of 0.4072 on the test set.

References

[1] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 1

[2] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 26–33, 2017. 1

[3] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 1

[4] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1

[5] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022. 1

[6] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[7] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[8] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[9] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 1

[10] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[11] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 1

[12] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 1

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 2, 3

[14] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014. 2

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[16] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition. *arXiv preprint arXiv:2303.10849*, 2023. 2, 6

[17] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer. *arXiv preprint arXiv:2303.08356*, 2023. 2, 6

[18] Ziyang Zhang, Liuwei An, Zishun Cui, Tengteng Dong, et al. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge. *arXiv preprint arXiv:2303.09158*, 2023. 2, 6

[19] Andrey V Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition. *arXiv preprint arXiv:2303.09162*, 2023. 2, 6

[20] Jun Yu, Zhongpeng Cai, Renda Li, Gongpeng Zhao, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Exploring large-scale unlabeled faces to enhance facial expression recognition. *arXiv preprint arXiv:2303.08617*, 2023. 2

[21] Dang-Khanh Nguyen, Ngoc-Huynh Ho, Sudarshan Pant, and Hyung-Jeong Yang. A transformer-based approach to video frame-level prediction in affective behaviour analysis in-the-wild. *arXiv preprint arXiv:2303.09293*, 2023. 2

[22] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 2, 3

[23] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 2, 3, 4

[24] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021. 2

[25] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE, 2021. 3

[26] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 3

[27] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 3

[28] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889, 2020. 3

[29] Fabien Ringeval, Björn W. Schuller, Michel F. Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Meßner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In Fabien Ringeval, Björn W. Schuller, Michel F. Valstar, Nicholas Cummins, Roddy Cowie, and Maja Pantic, editors, *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, AVEC@MM 2019, Nice, France, October 21-25, 2019*, pages 3–12. ACM, 2019. 3

[30] Carlos Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martínez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA,*

June 27-30, 2016, pages 5562–5570. IEEE Computer Society, 2016. 3

[31] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20890–20899. IEEE, 2022. 3

[32] Ya Li, Jianhua Tao, Björn Schuller, Shiguang Shan, Dongmei Jiang, and Jia Jia. Mec 2017: Multimodal emotion recognition challenge. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–5, 2018. 3

[33] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014. 3

[34] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 3

[35] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016. 3, 4

[36] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. 3, 4

[37] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 3, 4

[38] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 3, 4

[39] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015. 3

[40] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini, et al. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, pages 2001–2005, 2016. 3

[41] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 3

[42] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020. 3, 4

[43] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 3, 4

[44] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4