EVAEF: Ensemble Valence-Arousal Estimation Framework in the Wild

Xiaolong Liu[1,*], Lei Sun[2,*], Wenqiang Jiang[1], Fengyuan Zhang[2]
Yuanyuan Deng[1], Zhaopei Huang[2], Liyu Meng[1], Yuchen Liu[2], and Chuanhe Liu[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 Valence-Arousal Estimation Challenge of the 5th Affective Behavior Analysis in-the-wild (ABAW) competition. We explore the problems in this VA challenge from three aspects: 1) To obtain efficient and robust feature representations, we explore the role of multiple visual and video feature extractors; 2) Based on multimodal feature representations that fuse the visual and video information, we utilize four types of temporal encoders to capture the temporal context information in the video, including the LSTM, GRU, Transformer based encoder and a combined encoder of Transformer and LSTM; 3) five model ensemble strategies are used to combine multiple results with different model settings. Our system achieves the performance in Concordance Correlation Coefficients (CCC) of 0.6193 for valence, 0.6634 for arousal, and a mean CCC of 0.6414 on the test set, which demonstrates the effectiveness of our proposed method and ranks first place in the challenge.

1. Introduction

As a vital component of human-computer interaction, affective computing is widely applicable in scenarios involving education, healthcare, market research, social interaction, and other types of interaction. It also has extremely valuable theoretical implications and real-world practical application value for the realization of humanized communication for intelligent machines. However, emotions usually arise in response to either an internal or external event that has a positive or negative meaning for an individual [35]. Ambiguity or confusion in emotion perception can result from tiny variations in emotional displays when recognizing emotions. Fortunately, with the continuous research in psychology and the rapid development of deep learning, affective computing is gaining more and more attention. For example, Aff-wild [17,22,42] and Aff-wild2 [16,18–25,42] provide us with large-scale annotated datasets, driving the development of affective computing.

Emotion can be described as discrete categorical states or values in a continuous dimensional space. One of the most popular dimensional descriptions of emotion is the circumplex model [33] which uses the two dimensions of valence and arousal to represent emotional states. Valence refers to the positive or negative degree of emotion, while arousal refers to emotional intensity. In this work, we explore methods to estimate the valence and arousal annotations of a person appearing in a video recording. Our system contains four key components. First, for multimodal feature extraction, we use a variety of feature extractors to obtain visual and audio features. Then, based on the above features, to fully capture the temporal context information in the videos, we further apply four types of temporal encoders, including LSTM [34], GRU [4], Transformer [38], and a combination model of Transformer and LSTM. Next, we explore an early-fusion and a late-fusion strategy for feature fusion. Last, five ensemble strategies are used to get better results, which proves to be effective.

2. Related Works

The Valence-Arousal Estimation task of the 5th ABAW competition attracted a lot of attention from researchers, and many brand-new methods were proposed to tackle the challenging task of predicting continuous valence and arousal in videos.

[28–30, 43, 45–47] demonstrate the importance of considering multiple modalities to better capture complex information of human emotions. The integration of both audio and visual features has been shown to provide complementary information, leading to better performance in Valence-Arousal Estimation task. [32, 36, 39, 40, 44, 48] focuse on modeling the emotional content of videos through

*These authors contributed equally to this work and should be considered co-first authors.
†Corresponding Author.

the use of visual features. In particular, [36] leverages multiple training methods to improve the model's ability to represent emotional content from facial expressions. As mentioned by the aforementioned methods, the utilization of multi-modal features enhances performance when compared to using visual features exclusively.

Our proposed approach, which leverages both visual and audio features to model different combinations of features and incorporates temporal and spatial information, achieved competitive performance in the task.

3. Method

Given a video X, it can be divided into the visual input X^{vis} and the audio input X^{aud}, where X^v can be illustrated as a sequence of image frames $\{F_1, F_2, ..., F_n\}$, and n denotes the number of image frames in X. In the Valence-Arousal estimation task, each frame in X is annotated with an emotion label y consisting of a valence label y^v and an arousal label y^a. The task is to predict the emotion label for each frame in the video.

The overall pipeline is illustrated in Fig.1, which consists of four components. First, we extract various frame-level features of the input video in visual and audio modalities. These features are then fed into temporal encoders to model the context information. Afterward, the temporal-aware representations are fed into regressors to acquire predictions of each independent model. Finally, we ensemble several models based on different features or temporal encoders to get a combined prediction as the final one.

3.1. Pre-processing

The videos are first split into image frames, and a face detector is applied to get the face bounding box and facial landmarks in each image. Then, the face in each image is cropped out according to the bounding box, and these cropped images are aligned based on the facial landmarks. Here we simply utilize the cropped and aligned facial images provided by the ABAW5 competition officials.

In addition, there is no valid face in some frames, where faces in these frames are not detected or there is no face in them. For such frames, we simply use their nearest frames with valid faces to represent them. Besides, some frames are annotated with label -5, which means that these annotations are invalid. Such frames are discarded in the pre-processing stage.

3.2. Multimodal Feature Representation

In order to obtain multi-modal representations of the frames in the videos, we employ multiple types of pre-trained models to extract visual features and audio features of the frames. Then, the visual and audio features are fused to get the multi-modal feature representations.

3.2.1 Visual Features

Six kinds of visual feature extractors are employed for the visual feature extraction, which results in five kinds of visual features. The visual feature extractors include the DenseNet-based [13] facial expression model, the IResNet100-based [6] facial expression model, the IResNet100-based [13] facial action unit (FAU) detection model, the MobileNet-based [11] valence-arousal estimation model, the MAE-based [9] facial expression, and action unit model. We will introduce the five kinds of visual features in detail below.

DenseNet-based Features The DenseNet-based features are extracted by a pre-trained DenseNet model, where the DenseNet model is pre-trained on the facial expression datasets, including FER+ [2] and AffectNet [31] datasets. This kind of feature is denoted as *densenet* in Section 4.

IResNet100-based FE Features We make an IResNet100 model pre-trained on the image-level face expression (FE) task and then use it as a feature extractor. The pre-training datasets include FER+ [2], RAF-DB [27] [26], and AffectNet [31] datasets. The extracted features are denoted as *ires100* in Section 4.

IResNet100-based FAU Features Another IResNet100 model is pre-trained on a commercially authorized facial action unit (FAU) detection dataset. The extracted features are denoted as *fau* in Section 4.

MobileNet-based Features The MobileNet-based features are extracted by a MobileNet model pre-trained on the AffectNet [31] dataset for the valence-arousal estimation task. They are denoted as *ms_va* in Section 4.

MAE-based Features MAE-based features are extracted by an MAE model pre-trained on DFEW [14], Emotionet [3], FERV39k [41] datasets. They are denoted as *mae* in Section 4.

3.2.2 Audio Features

For audio modality, we employ seven features including low-level and deep-level types. We will introduce the features at the two levels respectively.

Low-level Features We first employ three manually designed low-level descriptors (LLDs). FBank is a commonly used feature in the field of speech processing. We also utilize it in our work and denote it as *fbank*. Besides, we take advantage of two off-the-shelf feature sets eGeMAPS [7] and ComParE 2016 [37], which proved helpful in some previous emotion estimation works. We denote these two features as *egemaps* and *compare*, respectively.

Deep-level Features We further extract four kinds of features by deep networks. The VGGish model [10] is pre-trained on an audio events dataset AudioSet [8]. We denote the feature extracted by this model as *vggish*. Wav2Vec 2.0 [1] and HuBERT [12] are two recent self-supervised

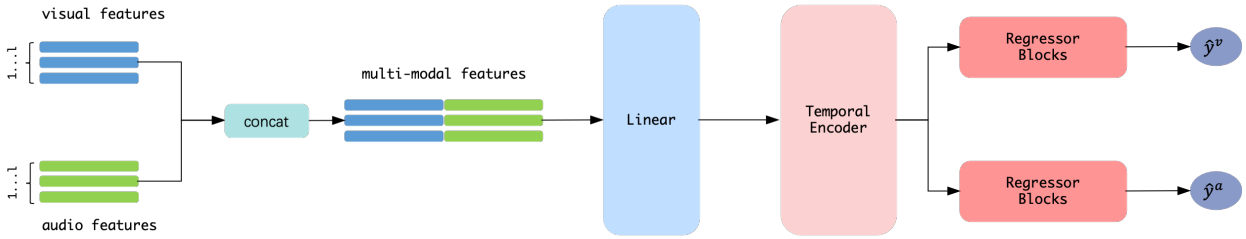

Figure 1. The overall framework of our proposed method.

pre-trained models and performed well on various audio downstream tasks. We explore them in our emotion estimation task and denote the extracted features as *wav2vec* and *hubert*. In addition, we employ a speaker verification model ECAPA-TDNN [5] to provide speaker-related features, which are denoted as *ecapatdnn*.

3.2.3 Multimodal Fusion

In order to fuse the visual and audio features to obtain the multi-modal feature representations, we use two multimodal feature fusion strategies, including early fusion and late fusion.

Early 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 representations 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. Afterward, the multi-modal representations are fed into a temporal encoder for context modeling.

Late fusion With the late fusion strategy, we employ two separate temporal encoders to encode the visual and audio context in the video respectively. Given the visual features f^v and audio features f^a corresponding to a frame, they are first fed into corresponding temporal encoder E_v, E_a and then concatenated to together produce the multimodal representations f^m. It can be formulated as follows:

$$\begin{aligned} \overline{f^v} &= \mathrm{E}_v(f^v) \\ \overline{f^a} &= \mathrm{E}_a(f^a) \\ f^m &= [\overline{f^v}; \overline{f^a}] \end{aligned} \tag{2}$$

where $\overline{f^v}$ and $\overline{f^a}$ are visual and audio context representations respectively. The multimodal representations f^m are directly used for the regression module.

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 single-modal or multi-modal features of the i-th segment, which are uniformly denoted as f_i^m, we employ a temporal encoder to model the temporal context in the video. Specifically, four kinds of structures are utilized as the temporal encoders, including LSTM, GRU, Transformer encoder, and a combined encoder of Transformer and LSTM.

3.3.1 LSTM-based Temporal Encoder

We employ a Long Short-Term Memory Network (LSTM) to model the sequential dependencies in the video. For the i-th video segment s_i, the multimodal features f_i^m are directly fed into the LSTM. In addition, the last hidden states of the previous segment s_{i-1} are also fed into the LSTM to encode the context between two adjacent segments. It can be formulated as follows:

$$g_i, h_i = \mathrm{LSTM}(f_i^m, h_{i-1}) \tag{3}$$

where h_i denotes the hidden states at the end of s_i. h_0 is initialized to be zeros. To ensure that the last frame of s_{i-1} and the first frame of segment s_i are consecutive frames, there is no overlap between two adjacent segments when LSTM is used as the temporal encoder. In other words, the stride p is the same as the segment length l.

3.3.2 GRU-based Temporal Encoder

We use a Gate Recurrent Unit Network (GRU) to encode the temporal information of the image sequence. Segment s_i means the i-th segment, and f_i^m means the input of GRU is the visual features for s_i. Furthermore, the hidden states of the last layer are fed from the previous segment s_{i-1} into the GRU to utilize the information from the last segment.

$$g_i, h_i = \mathrm{GRU}(f_i^m, h_{i-1}) \tag{4}$$

where h_i denotes the hidden states at the end of s_i. h_0 is initialized to be zeros. To ensure that the last frame of s_{i-1} and the first frame of segment s_i are consecutive frames, there is no overlap between the two adjacent segments.

3.3.3 Transformer-Based Temporal Encoder

Besides, we also use a Transformer Encoder to model the temporal information, which can be formulated as:

$$g_i = \text{TRMEncoder}(f_i^m) \tag{5}$$

Using the transformer encoder model, there is no need to feed a full video sequence to the model, while we just split it into a small sequence.

3.3.4 Transformer-LSTM-based Temporal Encoder

Moreover, to make full use of temporal information extracted from both Transformer and LSTM, we combine two encoders together, which can be formulated as follows:

$$f_i^h = \text{TRMEncoder}(f_i^m) \tag{6}$$

$$g_i, h_i = \text{LSTM}(f_i^h, h_{i-1}) \tag{7}$$

where f_i^h denotes the last hidden layer's output of the Transformer encoder. It is believed that with the local temporal features encoded by Transformer, we can further model the global temporal information by feeding these features into an LSTM-based Temporal Encoder.

3.4. Regressor

After the temporal encoder, the features g_i are finally fed into fully-connected layers for regression, which can be formulated as follows:

$$\hat{y}_i = W_p g_i + b_p \tag{8}$$

where W_p and b_p learnable parameters, $\hat{y}_i \in \mathbb{R}^{l\times 2}$are the predictions of the valence and arousal labels of s_i.

3.5. Loss Function

In the training phase, we utilized CCC Loss which can be formulated as:

$$\begin{aligned} L^V &= \frac{1}{N}\sum_{i=1}^{N}(1 - CCC(\hat{y}_i^V, y_i^V)) \\ L^A &= \frac{1}{N}\sum_{i=1}^{N}(1 - CCC(\hat{y}_i^A, y_i^A)) \end{aligned} \tag{9}$$

$$L = \alpha * L^V + \beta * L^A \tag{10}$$

where L^V denotes the valence task loss function, L^A denotes the arousal task loss function, N denotes the number of frames in each batch, $\hat{y}_i^V, y_i^V$ and $\hat{y}_i^A, y_i^A$ respectively denotes the prediction and label of valence and arousal in each batch respectively, L denotes the loss function we utilize, α and β are set to 1 here.

4. Experiments

4.1. Dataset

The 5th ABAW competition includes four challenges: i) Valence-Arousal Estimation ii) Expression Classification, iii) Action Unit Detection, and iv) Emotional Reaction Intensity Estimation Challenge. An augmented version of the Aff-Wild2 database is released in the 5th ABAW competition. This database consists of 594 videos of around 3M frames of 584 subjects annotated in terms of valence and arousal.

To extract visual features, the AffectNet, FER+, and RAF-DB datasets are used for pre-training. AffectNet is a large-scale facial expression recognition dataset that contains around 440K manually annotated images for discrete facial expressions and the intensity of valence and arousal. FER+ is a strict-labeled dataset for facial expression recognition. In order to increase the label accuracy, the label of FER+ is annotated by 10 crowd-sourced taggers. RAF-DB is also a large-scale facial expression recognition dataset that contains 29,672 facial images downloaded from the Internet. During the pre-training phase, we only utilized the discrete basic expression to pre-train our feature extractor.

4.2. Experiment Settings

During the training phase, we use the Adam [15] optimizer to train all our models for 30 epochs. We trained different architectures on Aff-Wild2, as for the transformer-based architecture, the affine dimension is 1024, the number of Transformer encode 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 150, the hidden size of head layers are $\{512, 256\}$, and the dropout ratio of head layers is 0.1.

4.3. Overall Performance on Validation Set

Table 1 shows some results of our method on the validation set of Aff-Wild2. As the results posted in the table, the addition of the densenet features can effectively improve the performance of valence, and the addition of the ires100 feature makes a clear contribution to the performance of arousal. Both Transformer and LSTM models can achieve competitive results using different feature combinations.

4.4. Model Ensemble

The model ensemble is an elegant way to prevent overfitting and enhance the robustness of models. During preparation for the model ensemble, we tried different architectures, hyper-parameters, and combinations of the feature. Moreover, to further improve the variability of the candidate models, we utilized different random seeds and learning rate decay strategies when training these models.

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

Model	Visual Features	Audio Features	Valence	Arousal
Transformer	fau,ires100	ecapatdnn,hubert	0.54859	0.73099
Transformer	fau,densenet	wav2vec,ecapatdnn	0.58029	0.69335
Transformer	affectnet	compare,egemaps,wav2vec	0.60372	0.64089
LSTM	fau,ires100	ecapatdnn,hubert	0.53461	0.70766
LSTM	ires100,mae	compare,wav2vec	0.50706	0.73183

Table 2. The results of every single model and the ensemble of them for the valence task on the validation set.

Model	Features	Valence
Transformer	densenet,compare,wav2vec	0.60372
Transformer	densenet,fau,compare,wav2vec	0.57113
LSTM	ires100,mae,compare,wav2vec	0.58605
LSTM	ires100,ms_va,compare,wav2vec	0.60661
Transformer	fau,densenet,wav2vec,ecapatdnn	0.59588
Transformer	fau,ms_va,wav2vec,fbank,ecapatdnn	0.60750
Transformer	fau,densenet,wav2vec,fbank,ecapatdnn	0.60114
Ensemble		**0.65281**

Table 3. The results of every single model and the ensemble of them for the arousal task on the validation set. Underline indicates the use of late fusion strategy, without underline means the use of early fusion strategy.

Model	Features	Arousal
Transformer	ires100,mae,compare,wav2vec	0.72343
Transformer	mae,ires100,compare,wav2vec	0.71048
Transformer-LSTM	densenet,fau,hubert,ecapatdnn,wav2vec	0.72124
Transformer-LSTM	ires100,mae,fau,ecapatdnn,wav2vec	0.72803
Transformer-LSTM	ires100,mae,compare,wav2vec	0.73845
LSTM	ires100,mae,compare,wav2vec	0.73212
Transformer	ires100,fau,mae,wav2vec,ecapatdnn	0.73142
Transformer	fau,ires100,ecapatdnn,hubert	0.73099
Transformer	ires100,fau,mae,hubert,wav2vec,ecapatdnn	0.73364
Transformer	ires100,fau,mae,wav2vec,ecapatdnn	0.72027
Transformer	ires100,fau,mae,wav2vec,fbank,ecapatdnn	0.72849
GRU	ires100,fau,ms_va,hubert,wav2vec,ecapatdnn	0.72412
Ensemble		**0.76650**

Table 2 and Table 3 show the model ensemble result on the validation set for valence and arousal tasks respectively. According to these results, the ensemble of different models can achieve a better improvement over a single model.

4.5. Cross Validation

To prevent overfitting and improve the robustness of the model, we also utilized a 6-fold cross validation strategy to train our model. For the choice of number 6, we analyzed the data distribution and found that the number of videos in the training set is roughly 5 times the number of videos in the validation set, so we split the training set into 5 segments, which together with the validation set to compose complete 6-fold dataset.

Table 4. The performance of our method on the 6-fold cross-validation. Original means the official validation set. The first five folds are from the training set, and the last fold is the original validation set.

	Valence	Arousal	Mean
Fold 1	0.61800	0.68642	0.65221
Fold 2	0.64664	0.66806	0.65735
Fold 3	0.57118	0.64668	0.60893
Fold 4	0.57468	0.65583	0.61526
Fold 5	0.55084	0.68317	0.61701
Fold 6	0.56636	0.69179	0.62908
Average	0.58795	0.67199	0.62997

As the result shown in table 4, we utilized feature set {fau, ms_va} as the visual feature and {fbank, ecapatdnn} as the audio feature. As the experiment setting we described in 4.2, we utilized a transformer encoder model to train on the 6-fold dataset.

4.6. Ablation Study

Table 5 shows the ablation study performance on the validation set, where all the results are based on the transformer architecture with the same training setting, except for the feature combinations. The addition of ms_va and densenet features leads to better performance on valence than others. With the contributions from ires100, hubert and mae feature, the performance on arousal outperforms the rest of the feature combinations.

Table 5. Ablation study of features on the validation set.

Visual	Audio	Valence	Arousal
fau,densenet	ecapatdnn	0.56295	0.70253
fau,densenet	fbank,ecapatdnn	0.57369	0.69397
fau,densenet	hubert,ecapatdnn	0.57434	0.70275
ires100,fau,densenet	ecapatdnn	0.55862	0.72417
ires100,fau,densenet	hubert,ecapatdnn	0.55997	0.72003
ires100,fau,ms_va	hubert,ecapatdnn	0.54087	0.73136
fau,densenet	wav2vec,fbank,ecapatdnn	0.60114	0.71153
fau,ms_va	wav2vec,fbank,ecapatdnn	0.60750	0.69444
ires100,fau,mae	hubert,wav2vec,ecapatdnn	0.54617	0.73364

4.7. Test Performance

In this section, we briefly describe our five submission strategies and the results on test set. Table 6 shows our strategies and results for each submission. As for the 1st submission, we only train the models on the official training set and choose the models with the best performance on the official validation set, then we ensemble the inference results on the test set from these models. Specifically, we use seven models for the valence estimation task and twelve models for the arousal estimation task.

Table 6. The results on the test set of five submissions.

Submission	Strategy	Valence	Arousal	Mean
1	Ensemble 1	0.6030	0.6501	0.6265
2	Ensemble 2	0.6115	0.6424	0.6269
3	Train-Val-Mix	0.5968	0.6423	0.6195
4	Ensemble 3	0.6095	0.6514	0.6305
5	6-Fold	**0.6193**	**0.6634**	**0.6414**

The model and feature combination of the 1st submission is shown in Table 2 and 3. As for the 2nd submission, we select several models that work better on the validation set from the models used for the first submission to the ensemble. As for the 3rd submission, we mix up the training and validation set and use both of them for training, which is called Train-Val-Mix. To get test set results, we empirically choose the models from 16 to 25 epochs in the training stage for valence and arousal, then we ensemble all these models to get the final test results. As for the 4th submission, we ensemble the test results of submission 1 and submission 3, which proves to be better than the two original results. As for the last submission, we mix up the training and validation set and divide them into six folds. For each time, one fold is used for validation, and the rest five folds are used for training. Since we get the six models, we ensemble all these models to get test results. We call this submission strategy as 6-Fold.

Table 7. The overall results and ranks on the test set.

Teams	Total Score	Valence	Arousal
Ours	**0.6414**	0.6193	**0.6634**
Netease Fuxi Virtual Human [45]	0.6372	**0.6469**	0.6258
CBCR [43]	0.5913	0.5526	0.6299
CtyunAI [47]	0.5666	0.5008	0.6325
HFUT-MAC [46]	0.5342	0.5234	0.5451
HSE-NN-SberAI [36]	0.5048	0.4788	0.5227
ACCC [48]	0.4842	0.4622	0.5062
PRL [39]	0.4661	0.5043	0.4279
SCLAB CNU [32]	0.4640	0.4578	0.4703
USTC-AC [40]	0.2372	0.3007	0.1736
baseline [21]	0.201	0.211	0.191

Finally, Table 7 shows the test results of all the teams in the Valence-Arousal Estimation Challenge, and our proposed method achieves surpasses performance over all the other teams.

5. Conclusion

In this paper, we introduce our method for the Valence-Arousal Estimation Challenge of the 5th Affective Behavior Analysis in-the-wild (ABAW) competition. Our method utilizes multimodal features and uses four different temporal models to obtain sequential information in the videos. We also explore strategies for multimodal feature fusion, including early-fusion and late-fusion. To further improve our predictions, five model ensemble strategies are used to get better final results. The experiment results show that our method achieves 0.6193 ccc for valence, 0.6634 ccc for arousal, and 0.6414 mean ccc on the test set of the expanded Aff-Wild2 dataset, which proves the effectiveness of our proposed method.

References

[1] 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

[2] 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. 2

[3] 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. 2

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 1

[5] 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

[6] 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. 2

[7] 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. 2

[8] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2

[9] 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. 2

[10] 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. 2

[11] 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. 2

[12] 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. 2

[13] 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

[14] 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. 2

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

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

[17] 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

[18] 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, 2020. 1

[19] 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

[20] 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

[21] 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, 6

[22] 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

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

[24] 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

[25] 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

[26] 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. 2

[27] 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. 2

[28] Chuanhe Liu, Wenqiang Jiang, Minghao Wang, and Tianhao Tang. Group level audio-video emotion recognition using hybrid networks. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 807–812, 2020. 1

[29] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 630–634, 2018. 1

[30] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Chuanhe Liu, and Qin Jin. Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2345–2352, June 2022. 1

[31] 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. 2

[32] 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. 1, 6

[33] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 1

[34] 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. 1

[35] Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990. 1

[36] Andrey V Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition. *arXiv preprint arXiv:2303.09162*, 2023. 1, 2, 6

[37] 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. 2

[38] 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. 1

[39] Tu Vu, Van Thong Huynh, and Soo Hyung Kim. Vision transformer for action units detection. *arXiv preprint arXiv:2303.09917*, 2023. 1, 6

[40] Shangfei Wang, Yanan Chang, Yi Wu, Xiangyu Miao, Jiaqiang Wu, Zhouan Zhu, Jiahe Wang, and Yufei Xiao. Facial affective behavior analysis method for 5th abaw competition. *arXiv preprint arXiv:2303.09145*, 2023. 1, 6

[41] 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. 2

[42] 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

[43] Su Zhang, Ziyuan Zhao, and Cuntai Guan. Multimodal continuous emotion recognition: A technical report for abaw5. *arXiv preprint arXiv:2303.10335*, 2023. 1, 6

[44] Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Jinming Zhao, and Qin Jin. Multi-task learning framework for emotion recognition in-the-wild. In *European Conference on Computer Vision*, pages 143–156. Springer, 2023. 1

[45] 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. 1, 6

[46] 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. 1, 6

[47] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer. *arXiv preprint arXiv:2303.08356*, 2023. 1, 6

[48] Peng Zou, Rui Wang, Kehua Wen, Yasi Peng, and Xiao Sun. Spatial-temporal transformer for affective behavior analysis. *arXiv preprint arXiv:2303.10561*, 2023. 1, 6