Valence and Arousal Estimation based on Multimodal Temporal-Aware Features for Videos in the Wild

Liyu Meng[1,*], Yuchen Liu[2,*], Xiaolong Liu[1], Zhaopei Huang[2]
Wenqiang Jiang[1], Tenggan Zhang[2], Chuanhe Liu[1], and Qin Jin[2]
[1] Beijing Seek Truth Data Technology Co.,Ltd.
[2] School of Information, Renmin University of China

Abstract

This paper presents our submission to the Valence-Arousal Estimation Challenge of the 3rd Affective Behavior Analysis in-the-wild (ABAW) competition. Based on multimodal feature representations that fuse the visual and aural information, we utilize two types of temporal encoder to capture the temporal context information in the video, including the transformer based encoder and LSTM based encoder. With the temporal context-aware representations, we employ fully-connected layers to predict the valence and arousal values of the video frames. In addition, smoothing processing is applied to refine the initial predictions, and a model ensemble strategy is used to combine multiple results from different model setups. Our system achieves the performance in Concordance Correlation Coefficients (ccc) of 0.606 *for valence,* 0.602 *for arousal, and mean ccc of* 0.601*, which ranks the first place in the challenge.*

1. Introduction

As a crucial part of human-computer interaction, affective computing can be widely used in medical, market analysis, social and other interaction scenarios, and it has extremely indispensable theoretical significance and practical application value to realize 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 to an individual [34]. When recognizing emotions, subtle differences in emotional expressions can also produce ambiguity or uncertainty in emotion perception. 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 [16,20,39] and Aff-wild2 [15,17–23,39] has provided us with a large-scale dataset of hard labels, driving the development of affective computing.

In the field of single modality emotion recognition, unimodal information is susceptible to various noises and can hardly reflect the complete emotional state. Multimodal emotion recognition can effectively utilize the information contained in multiple modal recognition, capture the complementary information between modalities, and thus improve the recognition ability and generalization ability of the model [1].

Our system for the V&A prediction challenge contains five key components. First, we preprocess the videos into image frames, extract and align the faces in the images. Then, we apply visual and audio feature extractors to extract visual and audio features respectively, which are concatenated to form the multimodal feature representations. Based on such representations, we further apply two types of temporal encoder, including LSTM [33] and Transformer [37], to capture the temporal context information in the video. Next, we feed these temporal-aware features to a regressor with fully-connected layers to predict the valence and arousal values of the video frames. Finally, we conduct a smoothing processing strategy and a model ensemble strategy to further improve the predictions.

2. Related Works

Various solutions have been proposed on the Aff-wild2 dataset for the ABAW Competition [15, 17–23, 39]. We briefly review some of the studies, including deep learning based approaches for face expression analysis. For example: feature fusion, attention mechanisms and iterative self-distillation.

For feature fusion, Mollahosseini et al. [38] propose a temporal fusion approach to combine multimodal features and temporal features. Multimodal representation learning, which aims to narrow the heterogeneity gap among different modalities, plays an indispensable role in the utilization of ubiquitous multimodal data [10]. For instance, works in [28] [27] use both audio and video channel features to

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

analyze emotions in video clips and achieve decent performance.

For the encoding, CAER-Net [24] propose an attention-based mechanism that can be used to assist the emotion recognition using context features. Based on attention mechanism, the role of the context part is more interpretable. However, this may lead to a certain degree of feature redundancy. Farzaneh et al. [8] propose the Discriminant Distribution-Agnostic loss (DDA loss) to regulate the distribution of deep features. With the help of DDA loss, features with rich semantic information for facial expression recognition can increase inter-class separation and decrease intra-class variations, despite training on unbalanced datasets.

3. Method

Given a video X, it can be divided into the visual data X^{vis} and the audio data 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 a sentiment label y consisting of a valence label y^v and an arousal label y^a. The task is to predict the sentiment label for each frame in the video.

The overall pipeline consists of five components. First, all videos are processed to get independent image frames with facial expressions. Secondly, we extract the visual and audio features corresponding to each frame in the videos, and concatenate them to get multimodal features. Thirdly, the multimodal features are fed into a temporal encoder to model the temporal context in the video. Fourthly, with the temporal-aware representations, fully-connected layers are employed to predict the sentiment labels. Finally, some post processors are applied to further improve the predictions. Figure 1 shows the overall framework of our proposed method.

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 ABAW competition officials.

In addition, some of the frames do not contain valid faces because either the faces in them are not detected or there is no face in them. As for an invalid frame, we find the nearest valid frame around it, and replace it with this valid frame.

3.2. Multimodal Feature Representation

We use three pre-trained models to extract the visual features, including the DenseNet-based [12] facial expression model, IResNet100-based [6] facial expression model, and the IResNet100-based Facial Action Unit (FAU) model. We also extract four types of audio features, which are eGeMAPS [7], ComParE 2016 [36], VGGish [11], and wav2vec2.0 [3].

3.2.1 Visual Features

The first 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.

The other kinds of visual features are based on two pre-trained IResNet100 models. The first one is pre-trained on the FER+ [4], RAF-DB [26] [25], and AffectNet [29] datasets. Specifically, in the pre-training stage, the faces in these datasets are aligned by the five face keypoints, and then resized into 112x112. The accuracy of the model in the pre-training stage is 0.8668, 0.8860, and 0.6037 on the FER+, RAF-DB, and AffectNet dataset respectively. The dimension of the visual feature vectors is 512.

The second model is first trained on the Glint360K [2] dataset with the face recognition pre-training task. Then the model is further trained on a authorized commercial FAU dataset. The dimension of the visual feature vector is 512.

3.2.2 Audio Features

The audio features are composed of manually designed low-level descriptors (LLDs) and more semantically informative features extracted by deep learning methods. The LLDs contain the eGeMAPS and the ComParE 2016, where both of them are extracted by the openSmile. The dimensions of these features are 23 and 130 respectively. The high-level features are based on pre-trained wav2vec2.0 and VGGish models. The wav2vec2.0 is a self-supervised model which is pre-trained and fine-tuned on 960 hours of the Librispeech [31]. The dimension of the wav2vec-based features is 768. The VGGish is pre-trained on a large youtube dataset (Audioset [9]). The dimension of VGGish-based features is 128.

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.

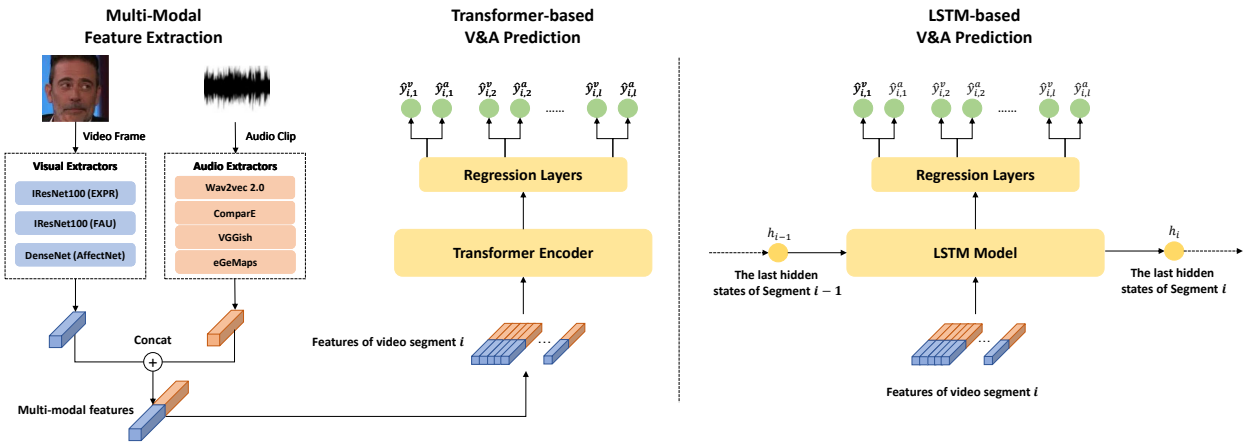

Figure 1. The overall framework of our proposed method.

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

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:

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

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 another word, the stride p is the same as the segment length l.

3.3.2 Transformer-Based Temporal Encoder

We utilize a transformer encoder to model the temporal information in the video segment as well, which can be formulated as follows:

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

Unlike LSTM, the transformer encoder just models the context in a single segment and ignores the dependencies of frames between segments. In order to cover the context of different frames, there can be overlaps between consecutive segments, which means $p \leq l$.

3.4. Training and Inferencing

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

$$\hat{y}_i = W_p g_i + b_p \tag{4}$$

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.

We use the Concordance Correlation Coefficient (CCC) between the predictions and the ground truth labels as the loss function during training, which can be denoted as follows:

$$L = \sum_{c\in\{v,a\}} (1 - CCC(\hat{y}^c, y^c)) \tag{5}$$

where $\hat{y}^v, \hat{y}^a, y^v, y^a$ denotes the predictions and the ground truth labels of valence and arousal in a batch respectively.

3.5. Post-processing

In the testing stage, we apply some additional post processors to the predictions. First, some of the predictions may exceed the range of $[-1, 1]$, and we simply cut these values to -1 or 1.

Secondly, since the sentiment of individuals varies continuously over time, the values of valence and arousal also vary smoothly over time. Thus, we apply a smoothing function to the predictions to make them more temporally smooth. Specifically, given the original prediction of the j-th frame $\hat{y}_j$, the final prediction $\tilde{y}_j$ is set as the average

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

Model	Visual Features	Audio Features	Valence	Arousal
LSTM	DenseNet	wav2vec	0.5544	0.6531
Transformer	DenseNet	wav2vec, ComParE	**0.6050**	0.6416
Transformer	ires100,fau	wav2vec,VGGish,ComParE,eGeMAPS	0.5883	**0.6689**

prediction value of a window with w frames centered on the j-th frame, i.e., $\{\hat{y}_{j-[w/2]}, ..., \hat{y}_{j+[w/2]}\}$.

4. Experiments

4.1. Dataset

The third ABAW competition includes four challenges: i) uni-task Valence-Arousal Estimation, ii) uni-task Expression Classification, iii) uni-task Action Unit Detection, and iv) MultiTask-Learning. All challenges are based on a common benchmark database, Aff-Wild2, a large-scale field database and the first to be annotated according to valence-arousal, expression, and action units. the Aff-Wild2 database extends the Aff-Wild dataset, with more videos and annotations for all behavior tasks. The Valence-Arousal Estimation Challenge contains 567 videos, that have been annotated by four experts using the method proposed in [5].

As for the visual feature extractors, the FER+, RAF-DB, and AffectNet datasets are used for pre-training. The RAF-DB is a large-scale database of facial expressions, which includes about 30,000 images of a wide variety of faces downloaded from the Internet. We use the single-label subset in RAF-DB, including 7 classes of basic emotion. AffectNet dataset contains over one million facial images, collected from the Internet. Approximately half of the retrieved images (approximately 440,000) were manually annotated for the presence of seven discrete facial expressions (classification model) as well as the intensity of value and arousal. In addition, an authorized commercial FAU dataset is also used to pre-train an visual feature extractor. It contains 7K images in 15 face action unit categories(AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU11, AU12, AU15, AU17, AU20, AU24, and AU26).

4.2. Experiment Settings

The models are trained on Nvidia GeForce GTX 1080 Ti GPUs, each with 11GB memory, and with the Adam [14] optimizer. The results reported in the following experiments are based on the average score of 3 random runs. The model is trained for 30 epochs, the batch size is 16 and the dropout rate is 0.3. As for the LSTM model, the learning rate is 0.0003, the dimension of multimodal features and the hidden size are 512, the length of video segments is 100, the number of regression layers is 2 and the hidden sizes are $\{512, 256\}$ respectively.

As for the transformer encoder model, the learning rate is 0.0002, the length of video segments is 250, the stride of segments is 250 or 100, the dimension of multimodal features is 512, the number of encoder layers is 4, the number of attention heads is 4, the dimension of feed-forward layers in the encoder is 1024, the number of regression layers is 2 and the hidden size of regression layers are $\{512, 256\}$ respectively. As for the smooth function in the post-processing stage, the size of the smoothing window is 20 for valence and 50 for arousal.

Table 2. The performance of our method on the 5-fold cross-validation. Original means the official validation set.

	Valence	Arousal	Mean
Fold 1	0.6177	0.6134	0.6156
Fold 2	0.5292	0.6296	0.5794
Fold 3	0.6495	0.6651	0.6573
Fold 4	0.5468	0.6116	0.5792
Fold 5	0.5272	0.6405	0.5839
Average	0.5741	0.6320	0.6031
Original	0.5883	0.6689	0.6286

4.3. Overall Performance on Validation Set

Table 1 shows the experimental results of our proposed method on the validation set of the Aff-Wild2 dataset. The Concordance Correlation Coefficient (CCC) is used as the evolution metric for both valence and arousal prediction tasks. As is shown in the table, our proposed transformer encoder structure achieves the best performance for both valence and arousal, and the LSTM structure achieves competitive performance for arousal as well. It proves the effectiveness of each of our proposed structures.

We also conduct experiments of 5-fold cross-validation, which use the transformer-based structure with the feature set {ires100, fau, wav2vec, VGGish, ComParE, eGeMAPS}. The results are shown in Table 2.

4.4. Model Ensemble

In order to further improve the performance of our proposed models, we apply a model ensemble strategy to these models. We train some models with different basic struc-

Table 3. The results of each single model and the ensemble of them for the valence prediction task on the validation set.

Model	Features	Valence
Transformer	DenseNet,wav2vec,ComParE	0.6089
Transformer	DenseNet,wav2vec,ComparE,VGGish,eGeMAPS	0.6113
Transformer	ires100,fau,wav2vec,VGGish	0.5833
Transformer	ires100,fau,VGGish,ComParE,eGeMAPS	0.5831
Ensemble		**0.6555**

Table 4. The results of each single model and the ensemble of them for the arousal prediction task on the validation set.

Model	Features	Arousal
LSTM	DenseNet,wav2vec	0.6591
Transformer	DenseNet,wav2vec	0.6488
Transformer	DenseNet,wav2vec,ComParE	0.6458
Transformer	DenseNet,wav2vec,VGGish,eGeMAPS	0.6456
Transformer	ires100,fau,wav2vec,VGGish	0.6628
Transformer	ires100,fau,wav2vec,VGGish,ComParE,eGeMAPS	0.6604
Ensemble		**0.7088**

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

Visual	Audio	Valence	Arousal
DenseNet	None	0.5290	0.5969
DenseNet	wav2vec	0.5596	0.6460
ires100, fau	wav2vec	0.5357	0.6412
DenseNet	wav2vec, ComParE	**0.6050**	0.6416
DenseNet	wav2vec, VGGish, ComParE, eGeMAPS	0.5972	0.6370
ires100	wav2vec, VGGish, ComParE, eGeMAPS	0.5055	0.6166
fau	wav2vec, VGGish, ComParE, eGeMAPS	0.5707	0.6168
ires100, fau	wav2vec, VGGish, ComParE, eGeMAPS	0.5883	**0.6689**

tures, hyper-parameters and combinations of features, and get the predictions of them respectively in the testing stage. Then, the average value of the prediction of these models is taken as the final prediction.

Table 3 and Table 4 show the results of model ensembles on the validation set for the valence and arousal prediction task respectively. The results indicate that the model ensemble strategy can combine the strengths of different models and achieve significant improvement over them.

4.5. Ablation Study

In this section, we conduct an ablation analysis of different features to compare their contribution of them. Table 5 shows the results of the ablation study for our proposed visual and audio features. The transformer-based structure is used for the ablation study.

As is shown in the table, each of our proposed features has contributed to the performance. As for the audio features, the ComParE and wav2vec make the most contributions to the valence prediction task, while the VGGish and wav2vec make the most contributions to arousal. As for the visual features, FAU contributes more than ires100 to valence, and DenseNet contributes more than the combination of FAU and ires100 for valence, while less for arousal.

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

Submit	Strategy	Valence	Arousal	Mean
1	Ensemble 1	0.5605	0.5165	0.5385
2	Ensemble 2	0.5779	0.5781	0.5780
3	Train-Val-Mix	**0.6060**	0.5960	**0.6010**
4	Ensemble 3	0.5898	**0.6018**	0.5958
5	5-Fold	0.5929	0.5985	0.5957

4.6. Test Performance

In this section, we briefly introduce our strategies for submissions and show the performance of our proposed method on the test set. Table 6 shows the strategies and results for each of our five submissions. As for the 1st, 2nd and 4th submissions, we apply the simple training and validation strategy, where we only train the models on the official training set and choose the models with the best performance on the official validation set. Specifically, we ensemble only 3 or 4 models to get the predictions for the 1st submission, and ensemble more models with more variations of feature combinations for the 2nd and 4th submission. For example, the model and feature combination of the 2nd submission is shown in Table 3 and 4.

Moreover, as for the 3rd and 5th submissions, we propose two additional training and validation strategies, including Train-Val-Mix and 5-Fold. Specifically, as for the Train-Val-Mix strategy, we mix up the training and validation set, and use both of them for training. In order to choose models with nice and stable performance without data for validation, we empirically choose the models from 16 to 25 epochs in the training stage for Arousal, and from 11 to 16 epochs for Valence. Finally, we ensemble all these models to get test results. As for the 5-Fold strategy, we mix up the training and validation set, and divide them into five folds. For each time, one fold is used for validation, and the rest four folds are used for training. Since we get five models with five folds, we ensemble these models to get test results. As is shown in the table, the Train-Val-Mix strategy achieves the best test performance, and the 5-Fold strategy also achieves competitive performance, which proves the effectiveness of our proposed strategies.

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

5. Conclusion

In this paper, we introduce our method for the Valence-Arousal Estimation Challenge of the 3rd Affective Behav-

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

Method	Valence	Arousal	Mean
Ours	**0.6060**	0.5960	**0.6010**
FlyingPigs [40]	0.5200	**0.6016**	0.5608
PRL [30]	0.4500	0.4448	0.4474
HSE-NN [35]	0.4174	0.4538	0.4356
AU-NO [13]	0.4182	0.4066	0.4124
LIVIA-2022 [32]	0.3742	0.3633	0.3688
Netease Fuxi Virtual Human [41]	0.3005	0.2442	0.2723
baseline	0.1800	0.1700	0.1750

ior Analysis in-the-wild (ABAW) competition. Our method utilizes multimodal information and employs a temporal encoder to model the temporal context in the videos. With the temporal-aware multimodal features, fully-connected layers are applied to get predictions. In addition, a smoothing processing strategy and a model ensemble strategy are used to improve the predictions. The experiment results show that our method achieves 0.606 ccc for valence, 0.602 ccc for arousal and 0.601 mean ccc on the test set of the Aff-Wild2 dataset, which ranks the first place in the challenge.

References

[1] P. A. Abhang, B. W. Gawali, and S. C. Mehrotra. Multimodal emotion recognition - sciencedirect. *Introduction to EEG- and Speech-Based Emotion Recognition*, pages 113–125, 2016. 1

[2] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021. 2

[3] 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

[4] 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

[5] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000. 4

[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] Amir Hossein Farzaneh and Xiaojun Qi. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 406–407, 2020. 2

[9] Jort F Gemmeke, Daniel PW 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. IEEE, 2017. 2

[10] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. 1

[11] 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

[12] 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

[13] Vincent Karas, Mani Kumar Tellamekala, Adria Mallol-Ragolta, Michel Valstar, and Björn W Schuller. Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. *arXiv preprint arXiv:2203.13285*, 2022. 6

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

[15] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

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

[22] 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

[23] 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

[24] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019. 2

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial expression analysis in video. *arXiv preprint arXiv:2203.12891*, 2022. 6

[31] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 2

[32] Gnana Praveen Rajasekar, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro Koerich, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. *arXiv preprint arXiv:2203.14779*, 2022. 6

[33] 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

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

[35] Andrey V Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv preprint arXiv:2203.13436*, 2022. 6

[36] 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

[37] 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

[38] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576, 2017. 1

[39] 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

[40] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. *arXiv preprint arXiv:2203.13031*, 2022. 6

[41] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 6