

Continuous Emotion Recognition with Audio-visual Leader-follower Attentive Fusion

Su Zhang^{1,*}, Yi Ding^{1,*}, Ziquan Wei^{2,*}, Cuntai Guan¹

¹Nanyang Technological University, ²Huazhong University of Science and Technology

sorazcn@gmail.com, ding.yi@ntu.edu.sg, wzquan142857@hust.edu.cn, ctguan@ntu.edu.sg

Abstract

We propose an audio-visual spatial-temporal deep neural network with: (1) a visual block containing a pre-trained 2D-CNN followed by a temporal convolutional network (TCN); (2) an aural block containing several parallel TCNs; and (3) a leader-follower attentive fusion block combining the audio-visual information. The TCN with large history coverage enables our model to exploit spatial-temporal information within a much larger window length (i.e., 300) than that from the baseline and state-of-the-art methods (i.e., 36 or 48). The fusion block emphasizes the visual modality while exploits the noisy aural modality using the inter-modality attention mechanism. To make full use of the data and alleviate over-fitting, the cross-validation is carried out on the training and validation set. The concordance correlation coefficient (CCC) centering is used to merge the results from each fold. On the test (validation) set of the Aff-Wild2 database, the achieved CCC is 0.463 (0.469) for valence and 0.492 (0.649) for arousal, which significantly outperforms the baseline method with the corresponding CCC of 0.200 (0.210) and 0.190 (0.230) for valence and arousal, respectively. The code is available at <https://github.com/sucv/ABAW2>.

1. Introduction

Emotion recognition is the process of identifying human emotion. It plays a crucial role for many human-computer interaction systems. To describe the human state of feeling, psychologists have developed the categorical and the dimensional [1] models. The categorical model is based on several basic emotions. It has been extensively exploited in affective computing largely due to its simplicity and universality. The dimensional model maps the emotion into a continuous space, where the valence and arousal are taken

as the axes. It can describe more complex and subtle emotions. This paper focuses on developing a continuous emotion recognition method based on the dimensional model.

Continuous emotion recognition seeks to automatically predict subject's emotional state in a temporally continuous manner. Given the subject's visual, aural, and physiological data which are temporally sequential and synchronous, the system aims to map all the information onto the dimensional space and produces the valence-arousal prediction. The latter is then evaluated against the expert annotated emotional trace using metrics such as the concordance correlation coefficient (CCC). A number of databases, including SEMAINE [2], RECOLA [3], MAHNOC-HCI [4], SEWA [5], and MuSe [6] have been built for this task. Depending on the subject's context, i.e., controlled or in-the-wild environments, and induced or spontaneous behaviors, the task can be quite challenging due to varied noise levels, illumination, and camera calibration, etc. Recently, Kollias et al. [7–13] build the Aff-Wild2 database, which is by far the largest available in-the-wild database for continuous emotion recognition. The Affective Behavior Analysis in-the-wild (ABAW) competition [14] is later hosted using the Aff-Wild2 database.

This paper investigates one question, i.e., how to appropriately combine features of different modalities to achieve an “1 + 1 > 2” performance. Facial expression is one of the most powerful, natural, and universal signals for human beings to convey or regulate emotional states and intentions [15, 16]. And, voice also serves as a key cue for both the emotion production and emotion perception [17]. Though we human are good at recognizing emotion from multi-modal information, a straightforward feature concatenation may deteriorate an AI emotion recognition system. In a scene where one subject is watching a video or attending a talk show, more than one voice sources can exist, e.g., from the subject him/herself, the video, the anchor, and the audience. It is not trivial to design an appropriate fusion scheme so that the subject's visual information and the complementary aural information are captured.

We propose an audio-visual spatial-temporal deep neural

*They are equal contributors to this work. Cuntai Guan is the corresponding author. This work is partially supported by the RIE2020 AME Programmatic Fund, Singapore (No. A20G8b0102).

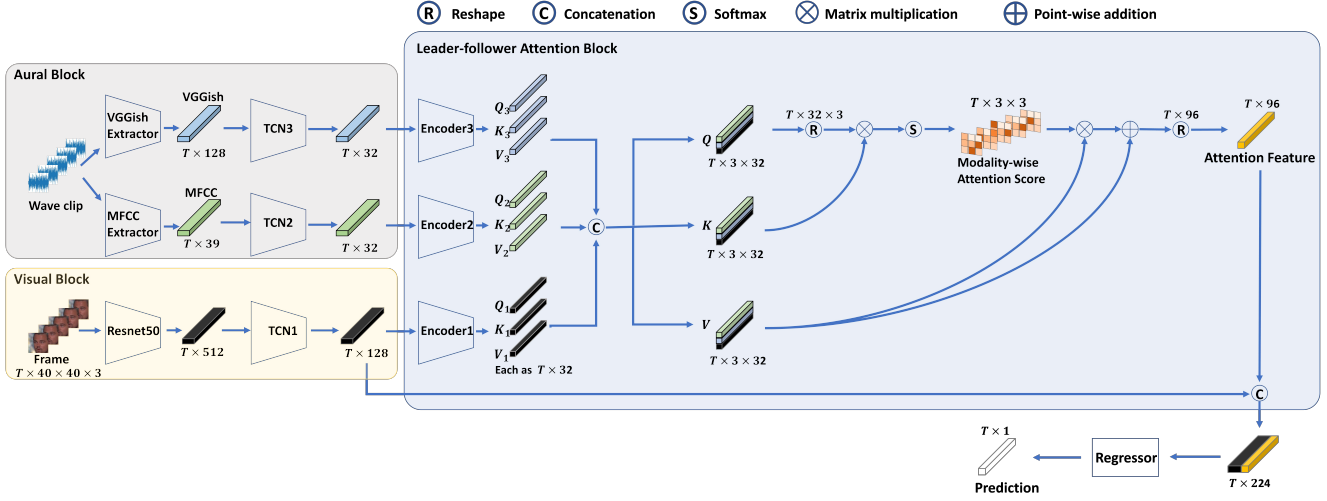


Figure 1. The architecture of our audio-visual spatial-temporal model. The model consists of three components, i.e., the visual, aural, and leader-follower attentive fusion blocks. The visual block has a cascade 2DCNN-TCN structure, and the aural block contains two parallel TCN branches. Together with the visual block, the three branches yield three independent spatial-temporal feature vectors. They are then fed to the attentive fusion block. Three independent attention encoders are used. For the i -th branch, its encoder consists of three independent linear layers, they adjust the dimension of feature vector producing a query Q_i , a key K_i , and a value V_i . They are then regrouped and concatenated to form the cross-modal counterparts. For example, the cross-modal query $Q = [Q_1, Q_2, Q_3]$. An attention score is obtained by Eq. 1. The 3×3 attention score will guide the model to refer to a specific modality at each time step, producing the attention feature. Finally, by concatenating the leading visual features with the attention feature, our model emphasizes the dominant visual modality and makes the inference.

network with an attentive feature fusion scheme for continuous valence-arousal emotion recognition. The network consists of three branches, fed by facial images, mfcc, and VGGish [18] features. A Resnet50 followed by a temporal convolutional network (TCN) [19] is used to extract the spatial-temporal visual feature of the facial images. Two TCNs are employed to extract the spatial-temporal aural feature from the mfcc and VGGish features. The three branches work in parallel and their outputs are sent to the attentive fusion block. To emphasize the dominance of the visual features, which we believe to have the strongest correlation with the label, a leader-follower strategy is designed. The visual feature plays as the leader and owns a skip connection down to the block’s output. The mfcc and VGGish features play as the followers, together with the visual feature they are weighted by an attention score. Finally, the leading visual feature and the weighted attention feature are concatenated and a fully-connected layer is used for the regression. To alleviate over-fitting and exploit the available data, a 6-fold cross-validation is carried out on the combined training and validation set of the Aff-Wild2 database. For each emotion dimension, the final inference is determined by the CCC-centering across the 6 trained models [20].

The remainder of the paper is arranged as follows. Section 2 discusses the related works. Section 3 details the model architecture including the visual, aural, and attentive fusion blocks. Section 4 elaborates the implementation de-

tails including the data pre-processing, training settings, and post-processing. Section 5 provides the continuous emotion recognition results on the Aff-Wild2 database. Section 6 concludes the work.

2. Related Works

2.1. Video-based Emotion Recognition Methods

Along with the thriving deep CNN-based methods comes two fundamental frameworks of neural networks for video-based emotion recognition. The first type possesses a cascade spatial-temporal architecture. The convolutional neural networks (CNN) are used to extract spatial information, from which the temporal information is obtained by using temporal models such as Time-delay, recurrent neural networks (RNN), or long short-term memory networks (LSTM). The second type combines the two separated steps into one and extracts the spatial-temporal feature using 3D-CNN. Our model belongs to the first type.

Two issues hinder the performance of the 3D-CNN based emotion recognition methods. First, they have considerably more parameters than 2D-CNN due to extra kernel dimension. Hence, it is more difficult to train. Second, employing 3D-CNN means to preclude the benefits of large-scale 2D facial image databases (such as MS-CELEB-1M [21] and VGGFace2 [22]). 3D-based emotion recognition databases [23–25] are much fewer. They are mostly based on posed

behavior with limited subjects, diversity, and labels.

In this paper, we employ the cascade CNN-TCN architecture. Systematical comparison [19] demonstrated that TCNs convincingly outperform recurrent architectures across a broad range of sequence modeling tasks. With the dilated and casual convolutional kernel and stacked residual blocks, the TCN is capable of looking very far into the past to make a prediction [19]. Compared to many other methods which utilize smaller window length, e.g., with a sequence length of 70 for AffWildNet [26], or 32 for the ABAW2020 VA-track champion [27], ours of length 300 has achieved promising improvement on the Aff-Wild2 database.

3. Model Architecture

The model architecture is illustrated in Fig. 1. In this section, we detail the proposed audio-visual spatial-temporal model and the leader-follower attentive fusion scheme.

3.1. Visual Block

The visual block consists of a Resnet50 and a TCN. The resnet50 is pre-trained on the MS-CELEB-1M dataset [21] as a facial recognition task, it is then fine-tuned on the FER+ [28] dataset. The Resnet50 is used to extract the independent per-frame features of the video frame sequence, producing the 512-D spatial encodings. The latter is then stacked and fed to a TCN, generating the 128-D spatial-temporal visual features. The TCN utilizes 128×4 as the *channel* \times *layer* setting with a kernel size of 5 and dropout of 0.1. Finally, a fully connected layer is employed to map the extracted features onto a 1-D sequence. Following the labeling scheme of the Aff-Wild2 database where the label frequency equals the video frame rate, each frame of the input video sequence is exactly corresponding to one label point.

3.2. Aural Block

The aural block consists of two parallel branches. The 39-D mfcc and 128-D VGGish [18] features are the inputs, respectively. The mfcc feature is extracted using the OpenS-mile Toolkit [29], and the VGGish feature is obtained from the pre-trained VGGish model [22]. They are fed to two independent TCNs and yield two 32-D spatial-temporal aural features. The TCN utilizes 32×4 as the *channel* \times *layer* setting with the same kernel size and dropout as in the visual block.

3.3. Leader-follower Attention Block

The motivation is two-fold. First, we believe that the representability of the multi-modal information is superior to the unimodal counterpart. In addition to the expressive visual information, the voice usually makes us resonate with

the emotional context. Second, a direct feature concatenation may deteriorate the performance due to the noisy aural information. Voice separation is still a topic of research. When multiple voice sources exist, it is difficult to obtain the voice components relevant to a specific subject.

The block first maps the feature vectors to query, key, and value vectors by the following procedure. For the i -th branch, its encoder consists of three independent linear layers, they adjust the dimension of feature vector producing a query \mathbf{Q}_i , a key \mathbf{K}_i , and a value \mathbf{V}_i . They are then regrouped and concatenated to form the cross-modal counterparts. For example, the cross-modal query $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3]$. After which, the attention feature is calculated as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_K}}) + 1)\mathbf{V}, \quad (1)$$

where $d_K = 32$ is the dimension of the key \mathbf{K} . After which, the *Attention* is normalized and concatenated to the leader feature (i.e., the spatial-temporal visual feature in our case) producing the 224-D leader-follower attention feature. Finally, a fully connected layer is used to yield the inference.

Note that the unimodal version of our model has only the visual block. The inference is obtained based on the 128-D spatial-temporal visual feature.

4. Implementation Details

4.1. Database

Our work is based on Aff-Wild2 database. It consists of 548 videos collected from YouTube. All the video are captured in-the-wild. 545 out of 548 videos contain annotations in terms of valence-arousal. The annotations are provided by four experts using a joystick [30]. The resulted valence and arousal values range continuously in $[-1, 1]$. The final label values are the average of the four raters. The database is split into the training, validation and test sets. The partitioning is done in a subject independent manner, so that every subject's data will present in only one subset. The partitioning produces 346, 68, and 131 videos for the training, validation, and test sets.

4.2. Preprocessing

The visual preprocessing is carried out as follows. The cropped-aligned image data provided by the ABAW2021 challenge are used. All the images are resized to $48 \times 48 \times 3$. Given a trial, the length N is determined by the number of the rows which does not include -5 . A zero matrix \mathbf{B} of size $N \times 48 \times 48 \times 3$ is initialized and then iterated over the rows. For the i -th row of \mathbf{B} , it is assigned as the i -th jpg image if it exists, otherwise doing nothing. After which, the

Table 1. The validation result in CCC using 6-fold cross validation from the unimodal and multi-modal models. The 6-fold cross-validation is used for data expansion and over-fitting prevention, in which the fold 0 is exactly the original data partitioning provided by ABAW2021.

Emotional dimension	Method	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Valence	Baseline	0.210	—	—	—	—	—	—
	Ours-unimodal	0.425	0.614	0.437	0.523	0.526	0.516	0.507
	Ours-multimodal	0.469	0.563	0.428	0.518	0.520	0.503	0.500
Arousal	Baseline	0.230	—	—	—	—	—	—
	Ours-unimodal	0.647	0.651	0.548	0.573	0.587	0.584	0.598
	Ours-multimodal	0.649	0.655	0.559	0.592	0.629	0.603	0.615

matrix **B** is saved in npy format and serves as the visual data of this trial.

The aural preprocessing firstly converts all the videos to mono with a $16K$ sampling rate in wav format. The synchronous mfcc and VGGish features are then extracted, respectively. For the mfcc feature, it is extracted using the OpenSmile Toolbox. The settings are the same as provided by AVEC2019 challenge [31]¹. Since the window and hop lengths of the short-term Fourier transform are fixed to 25ms and 10ms, respectively, the mfcc feature has a fixed frequency of 100 Hz over all trials. Given the varied labeling frequency and the fixed mfcc frequency, the synchronicity is achieved by pairing the i -th label point with the closest-in-time-stamp mfcc feature point. For example, given a label sequence in 30 Hz, the $\{0, 1, 2, 3\}$ -rd label points sampled at $\{0, 0.033, 0.067, 0.100\}$ seconds are paired with the $\{0, 3, 7, 10\}$ -th feature points sampled at $\{0, 0.03, 0.07, 0.10\}$ seconds. For all the label rows that do not contain -5 , the paired mfcc feature points are selected in sequential to form the feature matrix.

For the VGGish feature, it is extracted using the pre-trained VGGish model². First, the log-mel spectrogram is extracted and synchronized with the label points using the same operation above. The log-mel spectrogram matrix is then fed into the pre-trained model to extract the synchronized VGGish features. To ensure that the aural features and the labels have the same length, the feature matrices are repeatedly padded using the last feature points. The aural features are finally saved in npy format.

For the valence-arousal labels, all the rows containing -5 are excluded. The labels are then saved in npy format.

4.3. Data Expansion

The AffWild2 database contains 351 and 71 trials in the training and validation sets, respectively. To clarify, a la-

bel txt file and its corresponding data are taken as a trial. Note that some videos include two subjects, resulting in two separated cropped-aligned image folders and label txt files, with different suffixes. They are each taken as two trials.

To make full use of the available data and alleviate over-fitting, the cross-validation is employed. By evenly splitting the training set into 5 folds, we have 6 folds in total with a roughly equal trial amount, i.e., $70 \times 4 + 71 + 71$ trials. Note that the 0-th fold is exactly the original data partitioning. And there is no subject overlap across different folds. The CCC-centering is employed to merge the inference result on the test set.

Moreover, during training and validation, the resampling window has a 33% overlap, resulting in 33% more data.

4.4. Training

Since we employ the 6-fold cross-validation on 2 emotional dimensions using 2 models, we have 24 training instances to run, each takes about 10Gb VRAM and 1 to 2 days. Multiple GPU cards including Nvidia Titan V and Tesla V100 from various servers are used. To maintain the reproducibility, the same Singularity container is shared over the 24 instances. The code is implemented using Pytorch.

The batch size is 2. For each batch, the resampling window length and hop length are 300 and 200, respectively. I.e., the dataloader loads consecutive 300 feature points to form a minibatch, with a stride of 200. For any trials having feature points smaller than the window length, zero padding is employed. For visual data, the random flip, random crop with a size of 40 are employed for training and only the center crop is employed for validation. The data are then normalized to have 0.5 mean and standard deviation. For aural data, the data is normalized to have 0 mean and unit standard deviation.

The CCC loss is used as the loss function. The Adam optimizer with a weight decay of 0.0001 is employed. The learning rate (LR) and minimal learning rate (MLR) are set to $1e-5$ and $1e-6$, respectively. The *ReduceLROnPlateau*

¹https://github.com/AudioVisualEmotionChallenge/AVEC2019/blob/master/Baseline_features_extraction/Low-Level-Descriptors/extract_audio_LLDS.py

²<https://github.com/tensorflow/models/tree/master/research/audioset/vggish>

scheduler with a patience of 5 and factor of 0.1 is employed based on the validation CCC. The maximal epoch number and early stopping counter are set to 100 and 20, respectively. Two groups of layers for the Resnet50 backbone are manually selected for further fine-tuning, which corresponds to the whole layer4 and the last three blocks of layer3.

The training strategy is as follows. The Resnet50 backbone is initially fixed except the output layer. For each epoch, the training and validation CCC are obtained. If a higher validation CCC appears, then reset the counter to zero, otherwise increase the counter by one. If the counter reaches *Patience*, then reduce *LR* to *MLR*. Next time when the counter reaches *Patience*, release one group of backbone layers (started from the layer4 group) for updating and reset the counter. At the end of each epoch, the current best model state dictionary is loaded. The training is stopped if i) there is no remaining backbone layer group to release, ii) the counter reaches the *Early Stopping Counter*, or iii) the epoch reaches the *Maximal Epoch*. Note that the valence and arousal models are trained separately.

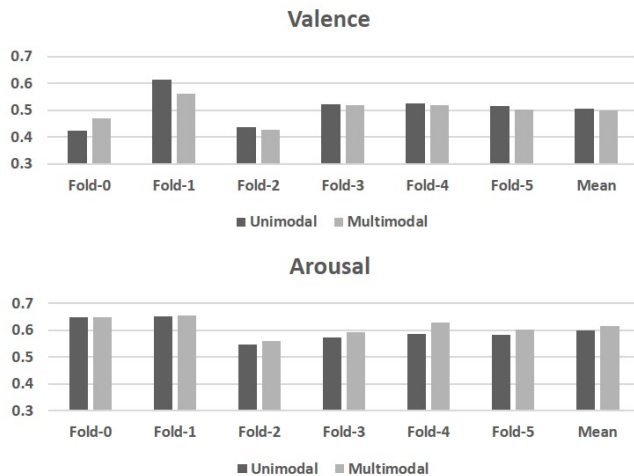


Figure 2. The 6-fold validation result in CCC obtained by our unimodal and multimodal models. Note that the fold 0 is exactly the original data partitioning provided by ABAW2021.

4.5. Post-processing

The post-processing consists of CCC-centering and clipping. Given the predictions from 6-fold cross-validation, the CCC-centering aims to yield the weighted prediction based on the inter-class correlation coefficient (ICC) [31]. This technique has been widely used in many emotion recognition challenges [31–34] to obtain the gold-standard labels from multiple raters, by which the bias and inconsistency among individual raters are compensated. The clipping ensures that the inference is truncated within the interval $[-1, 1]$, i.e., any values larger or smaller than 1 or -1

are set to 1 or -1 , respectively.

In this work, two strategies, i.e., clipping-then-CCC-centering and CCC-centering-then-clipping are utilized. They are called early-clipping and late-clipping.

5. Result

5.1. Validation Result

The validation results of the 6-fold cross-validation on valence and arousal are reported in Table 1. Two types of models are trained. The unimodal model is fed by video frame only, and the multimodal model is fed by video frame, mfcc, and VGGish features.

For fold 0, namely the original data partitioning, both of our unimodal and multimodal models have significantly outperformed the baseline. For other five folds, interestingly, the multimodal information has positive and negative effects on the arousal and valence dimensions, respectively, as shown in Figure 2. We therefore hypothesize that the annotation protocol weighs more on aural perspective for the arousal dimension.

5.2. Test Result

The comparison of our model against the baseline and state-of-the-art methods on the test set are shown in Table 2.

Table 2. The overall test results in CCC. UM and MM denote our unimodal and multimodal models, respectively. CV denotes cross-validation. EC and LC denote early-clipping and late-clipping. The bold fonts indicate the best result from ours and other teams.

Method	Valence	Arousal	Mean
Baseline	0.200	0.190	0.195
ICT-VIPL-VA [35]	0.361	0.408	0.385
NISL2020 [27]	0.440	0.454	0.447
NISL2021 [36]	0.533	0.454	0.494
Netease Fuxi Virtual Human [37]	0.486	0.495	0.491
Morphoboid [38]	0.505	0.475	0.490
STAR [39]	0.478	0.498	0.488
UM	0.267	0.303	0.285
UM-CV-EC	0.264	0.276	0.270
UM-CV-LC	0.265	0.276	0.271
MM	0.455	0.480	0.468
MM-CV-EC	0.462	0.492	0.477
MM-CV-LC	0.463	0.492	0.478

First and foremost, the multimodal model achieves great improvement against the unimodal counterpart, which is up to 70.41% and 58.42% gain (by comparing UM against MM in table 2) over the valence and arousal, respectively. The employment of cross-validation provides incremental improvement on the multimodal result.

We can also see that the three unimodal scenarios and multimodal scenarios have a sharp performance gap, whereas on the validation set the gap is incremental. We hypothesize that the unimodal models, fed only by visual information, suffer from over-fitting and insufficient robustness on the test set. The issue is alleviated by the fusion with aural information. Further investigation will be carried out in our future work using other audio-visual databases where labels of the test set are available.

6. Conclusion

We proposed an audio-visual spatial-temporal deep neural network with an attentive feature fusion scheme for continuous valence-arousal emotion recognition. The model consists of a visual block, an aural block, and a leader-follower attentive fusion block. The latter achieves the cross-modality fusion by emphasizing the leading visual modality while exploiting the noisy aural modality. Experiments are conducted on the Aff-Wild2 database and promising results are achieved. The achieved CCC on test (validation) set is 0.463 (0.469) for valence and 0.492 (0.649) for arousal, which significantly outperforms the baseline method with the corresponding CCC of 0.200 (0.210) and 0.190 (0.230) for valence and arousal, respectively.

References

- [1] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.
- [2] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.
- [3] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.
- [4] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.
- [5] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [6] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *arXiv preprint arXiv:2101.06053*, 2021.
- [7] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal in-the-wild challenge," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1980–1987.
- [8] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 1–23, 2019.
- [9] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," *arXiv preprint arXiv:1910.11111*, 2019.
- [10] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcfac," *arXiv preprint arXiv:1910.04855*, 2019.
- [11] 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)*, pp. 794–800.
- [12] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.
- [13] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution matching for heterogeneous multi-task learning: a large-scale face study," *arXiv preprint arXiv:2105.03790*, 2021.
- [14] D. Kollias, I. Kotsia, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the second abaw2 competition," 2021.
- [15] C. Darwin, *The expression of the emotions in man and animals*. University of Chicago press, 2015.
- [16] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

- [17] K. Scherer, "Component models of emotion can inform the quest for emotional competence. u: G. mathews, m. zeidner, rd roberts (ur.)," *The science of emotional intelligence: knowns and unknowns*, pp. 101–126, 2007.
- [18] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.
- [19] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
- [20] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.
- [21] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Msc-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.
- [22] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.
- [23] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.
- [24] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 2006, pp. 211–216.
- [25] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.
- [26] M. Liu and D. Kollias, "Aff-wild database and affwild-net," *arXiv preprint arXiv:1910.05318*, 2019.
- [27] D. Deng, Z. Chen, and B. E. Shi, "Multitask emotion recognition with incomplete labels," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. IEEE Computer Society, 2020, pp. 828–835.
- [28] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283.
- [29] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
- [30] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.
- [31] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.
- [32] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.
- [33] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3–9.
- [34] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, 2018, pp. 3–13.
- [35] Y.-H. Zhang, R. Huang, J. Zeng, S. Shan, and X. Chen, "M³t: Multi-modal continuous valence-arousal estimation in the wild," *arXiv preprint arXiv:2002.02957*, 2020.

- [36] D. Deng, L. Wu, and B. E. Shi, “Towards better uncertainty: Iterative training of efficient networks for multitask emotion recognition,” *arXiv preprint arXiv:2108.04228*, 2021.
- [37] W. Zhang, Z. Guo, K. Chen, L. Li, Z. Zhang, and Y. Ding, “Prior aided streaming network for multitask affective recognition at the 2nd abaw2 competition,” *arXiv preprint arXiv:2107.03708*, 2021.
- [38] M. T. Vu and M. Beurton-Aimar, “Multitask multi-database emotion recognition,” *arXiv preprint arXiv:2107.04127*, 2021.
- [39] L. Wang and S. Wang, “A multi-task mean teacher for semi-supervised facial affective behavior analysis,” *arXiv preprint arXiv:2107.04225*, 2021.