Proposed Models for Estimating Emotional Reactions Intensity in the Wild

Yang Qian
University of Hawai'i at Mānoa
Hawai'i, USA
qianyang@hawaii.edu

Ali Kargarandehkordi
University of Hawai'i at Mānoa
Hawai'i, USA
kargaran@hawaii.edu

Onur Cezmi Mutlu
Stanford University
California, USA
cezmi@stanford.edu

Saimourya Surabhi
Stanford University
California, USA
mourya.surabhi@stanford.edu

Mohammadmahdi Honarmand
Stanford University
California, USA
mhonar@stanford.edu

*Peter Washington
University of Hawai'i at Mānoa
Hawai'i, USA
pyw@hawaii.edu

*Dennis Paul Wall
Stanford University
California, USA
dpwall@stanford.edu

* Corresponding author

Abstract

Emotions play an essential role in human communication. Investigating an efficient model to recognize emotion automatically has become critical in current human-computer interaction research. As an emerging trend in HCI, awareness of emotional intensity is also imperative in gaining higher efficiency in this communication. It is when the emotion's meaning and level are understood simultaneously by all individuals and machines involved. There are three types of representations when dealing with emotions: Action Units, Valence Arousal (VA), and Categorical Emotions. This paper suggests our submission to a newly introduced challenge called Emotional Reaction Intensity (ERI) Estimation in the 5th CVPR competition for Affective Behavior Analysis in-the-Wild (ABAW). We employed four modified VGG16 and ResNet-50 models and a multimodal to extract a more robust facial representation than the reference baseline network mentioned in the ABAW, Resnet-50 (trained on VGGface2). Our experiments on the Hume-Reaction dataset show the proposed models' effectiveness. In the proposed model combination, the best performance was achieved with ResNet-50 (trained on ImageNet) with 0.4080 on the test set.

1. Introduction

For human beings, natural facial expressions are the most potent, universally recognized signals to convey emotional states and intentions[1, 2]. Mental disease diagnosis, human social/physiological interaction detection, sociable robotics, and many other human-computer interaction systems have been considered a target context for extensive research on automatic emotion recognition [3-8].

Emotional expressions have a crucial role in detecting certain types of developmental disorders. Autism spectrum disorder (ASD) affects almost 1 in 44 people in America [9]. It is the fastest-growing developmental disorder in the United States [10, 11]. Children suffering from autism tend to evoke emotions differently than neurotypical peers, and it is more difficult for them to produce the correct facial expressions [12-14]. In these contexts, to improve social communication, capturing vivid human emotions and making a corresponding reaction is expected to develop an ideal human-computer interaction model [15-23]. Preliminary efforts have been done in this space using various digital and wearable devices that enable families to provide therapy in the comfort of their home setting with the ability to customize the intervention structure to suit their child's needs [24-30].

Developing an efficient model for automatic facial expression detection highly depends on the quality, variability of the facial representations, and diversity of the dataset. Accordingly, there have been various limitations in similar applications in-the-wild to find a compelling dataset. Containing a set of limited human faces, e.g., only one of the three common types of emotional representations: Categorical Emotions (CE), Action Units (AU), and Valence Arousal (VA) is the most common limitation. The existing similarities between some expressions, and in other words, the ambiguity of the labels in the dataset, is another challenge that makes it demanding to distinguish some facial expressions. This ambiguity might originate from the different understanding of the human being of some facial expressions that cause inconsistent labeling. For instance, in many people's points of view, "Sadness" can be pretty similar to "Disgust" and it

might be difficult to distinguish these two facial expressions.

To tackle some of the mentioned limitations above, using video-based datasets has become more prevalent in facial expressions detection tasks. Although these datasets can enhance the model performance and provide more insight into the expression characteristics, dealing with the data, in this case, becomes even more complex. Due to rapid human emotion changes, many frames might not contain reliable information to predict facial expressions and accordingly estimate the emotion's intensity. Labeling the video frame by frame [31] makes it even more complicated. Therefore, the number of video-based datasets in the wild is now quite limited.

In recent years, the widespread usage of deep learning techniques in various interventions has made it easier to access more efficient datasets. For instance, Aff-Wild [32-35] and Aff-Wild2 [36-41] Audio/Visual (A/V) datasets are current shining examples used in both academic and industrial communities that contain all three representation labels mentioned above. Aff-Wild2 comprises 548 videos of around 2.7M frames annotated in terms of the seven primary expressions (i.e., anger, disgust, fear, happiness, sadness, surprise, and neutral).

Hume-React is another large-scale multi-modal database containing user-generated video content. By releasing the Hume-Reaction dataset, Hume is contributing to this year's (CVPR 2023) challenge. The dataset includes more than 75 hours of video recordings consisting of spontaneous reactions of 2,222 individuals to 1841 evocative video elicitors. Each video is annotated by the individuals with seven self-reported emotions at a scale of intensities 1-100 [42].

This work proposes an affect recognition and level estimation model for the Emotional Reaction Intensity (ERI) Estimation task in the ABAW2 Competition [43]. In contrast to the most recent ABAW competitions, where Muti-task Learning was the central theme or among one of the main challenges [44, 45], this year, the focus is on only uni-task solutions for four challenges: Valence-Arousal (VA) Estimation [46, 47], Expression (Expr) Classification [48, 49], Action Unit (AU) Detection [49, 50], Emotional Reaction Intensity (ERI) Estimation. We designed our algorithms to surpass the baseline network performance - resnet50 (pre-trained VGGFACE2) with fixed convolutional weights and employed multiple modifications to enhance the proposed models and achieve better efficiency in detecting emotion labels and estimating their intensity levels. The code will be made publicly available after we are authorized to release the final paper and publish the work.

2. Related works

The 5th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW) introduce four main challenges. Three are on the previously employed challenges and common facial expression representations: 1) Valence-Arousal (VA) Estimation i.e., how positive/negative and active/passive an emotional state is, 2) Expression (Expr) Classification, and 3) Action Unit (AU) Detection (specific movements of facial muscles from Facial Action Coding System). In addition to these challenges, the competition introduced a 4th challenge called Emotional Reaction Intensity (ERI) Estimation, experimenting on a different dataset (Hume Reaction) this year in 2023.

Multi-modal features, including visual, audio, text, and physiological signals, have been well introduced in the research focusing on deep networks. The successful usage of deep learning in computer vision applications and the capability of the networks to learn these multi-modal features has prompted researchers to investigate their applicability to affective behavior analysis. In this field, facial expression as a visual modality is a major component of percept and analyzing emotions. The Facial Action Coding System (FACS) is a widely used affect recognition network that recognizes specific emotions based on facial Action Units (AU)[51]. Gabor wavelet is another emotion recognition tool successfully applied to facial representation [52]. Benefiting from the growing application of deep learning, researchers have discovered that extracting features based on deep learning techniques can achieve higher accuracy. For instance, to extract visual features [53] uses CNN and RNN stack based on a convolutional recurrent neural network. To prove the efficiency of audio/visual networks, [54] proposes the usage of 2D+1D convolutional neural networks. In the AVECs, researchers use other deep learning methods that all perform better than traditional feature extractors [55-57].

Audio modality widely uses prosodic features (nonverbal features such as pitch, loudness, or rate) for emotion recognition tasks. MFCC, LPCC, PLP, RastaPLP, LFPC feature streams, and the nearest class mean classifier are mainly used acoustic features. Like the visual modality, deep neural networks are being applied to extract deep audio features to detect acoustic events [58-60].

The work proposed by Arushi and Vivek [61] has used CNN for human face classification. A 5-layer CNN with fractional Max pooling has been developed in this work. They obtained a validation accuracy of 47% for this model and 38% for a fine-tuned VGG-16 network.

Happy et al. [62] proposes an excellent technique by using selected facial patches for expression recognition on CK+ dataset. Sobel filtering followed by Otsu thresholding, and a Local Binary Patterns (LBP) for feature extraction and classification are used in the proposed model. The final results in this model are pretty significant: 87.8% for Angry to 98.46 for Surprise.

3. Method

For the current competition (CVPR 2023), the visual modality baseline is ResNet50 trained on VGGface2, and for audio is DeepSpectrum [63] - deep feature extraction of spectrograms (e.g., Mel-spectrograms). Its CNN backbone is a DenseNet121 [64] pre-trained on ImageNet [65]. Visual and audio modality baseline networks performance on the validation set is 0.2488 and 0.1087, respectively. For calculating the loss function, this paper uses Mean Squared Error (MSE). To calculate the MSE, the difference between the model's predictions and the ground truth is squared and averaged out across the whole dataset.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (1)$$

For the performance, calculations are based on the Pearson's correlation coefficient formula:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad (2)$$

3.1. Overview

We first employed a VGG16 network to extract the visual modality features of each facial expression type. Due to the extensive training that the VGG16 was going through, its accuracy on the affect recognition tasks tends to be excellent. VGG-16 architecture schematic is illustrated in Figure 1 [66].

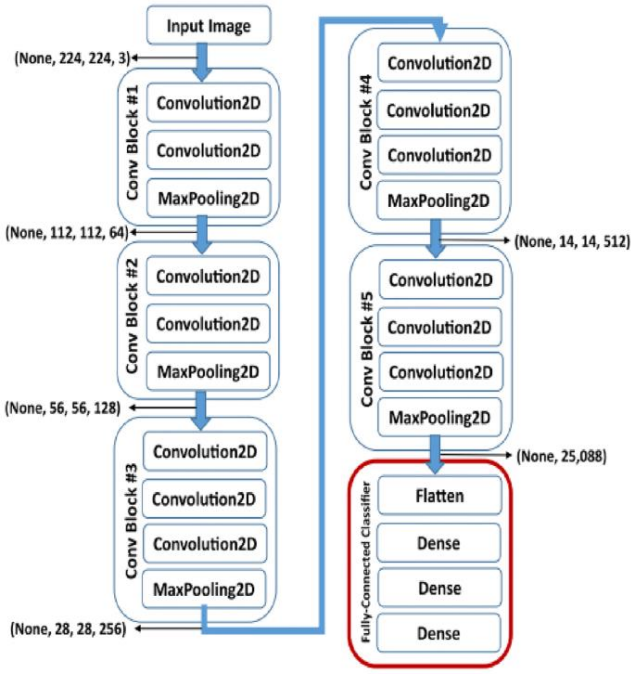

Figure 1: Schematic design of the VGG16.

This network consists of 16 convolution layers with a receptive field of 3 × 3 and a Max pooling layer of size 2 × 2. After the last Max pooling layer, there are three fully connected layers. Softmax classifier is used as the final layer, and for all hidden layers, relu activation is applied.

At this step, with the chunk size 500, and the learning rate 0.0004 the performance was 0.3125 By changing the type and the number of GPU, we tried a larger chunk size with the same network (VGG16), same learning rate and this time the performance increased to 0.3674.

We also employed ResNet-50 pre-trained on ImageNet to compare the results and see if we can achieve a better performance. Regarding the architecture, both VGG-16 and ResNet-50 are comparable, with the exception that Resnet50 has an additional identity mapping capability. This is shown in Figure 2 [67]. By applying ResNet-50 at the learning rate 0.001 and the chunk size of 1000, performance increased to 0.4080.

For audio modality, the signals are processed through a LSTM network. Both modality networks have been trained on ImageNet.

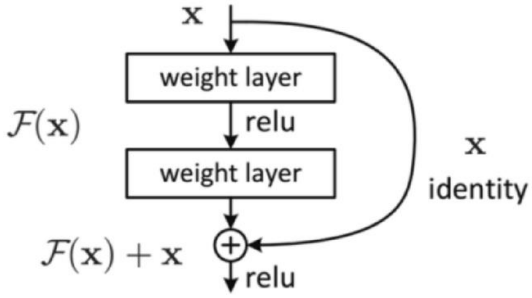

Figure 2: Additional mapping capability of ResNet-50

3.2. Data Preprocessing and Normalization

All frames in the videos have been clipped, and 32 randomly (evenly) distributed frames among all frames have been chosen for each video. All 32 frames will be in a chunk in order. This way, the shape of our training data is a tensor like [x, 32, 224, 224, 3], where x is the total number of chunks. We applied a standard augmentation technique that is commonly used for face analysis and emotion recognition [68]. Brightness increases up to 150% [1, 1.5] of the maximum value. RetinaFace [69] is used as the basic algorithm for face detection. This robust single-level face detection algorithm applies multi-task joint additional supervised learning to perform a pixel-level localization of the faces. Excellent modeling ideas including feature pyramid network, context network and task union are incorporated in this algorithm. For the normalization, the data was normalized by subtracting the mean (μ) of each feature and division by the standard deviation (σ). This way, each feature has a mean of 0 and a standard deviation of 1. This definitely results in faster convergence.

$$x := \frac{x - \mu}{\sigma} \qquad (3)$$

3.3. Model Architecture

Figure 3 provides the overall model architecture. For a sample video with 32 frames (evenly distributed) among all frames, we trained the model YAP-CSM on four different single modal network configurations (table 1) and a multi-modal network (table 2). LSTMs [70] were used to capture temporal relationships. While it helps to avoid vanishing and exploding gradients problems, it is benefiting from the capability of learning long-term dynamics. Instead of a multiplicative status of naive RNN, LSTM uses a summation of memory status instead [71]. Figure 3 represents the YAP-CSM multi-modal that takes Time Distributed layer output extracted from the previous network (e.g. ResNet-50) CNN model and audio signals in a parallel way as inputs. Then, the two parallel LSTM layers with 512 nodes are employed to learn the temporal relationship between the frames. The outputs of the LSTM layers are separately connected to a dense layer. At last, a multi-modal layer connected to a softmax layer predicts the label and the intensity of emotion in a video sequence. The architecture is similar for the other four single modal networks.

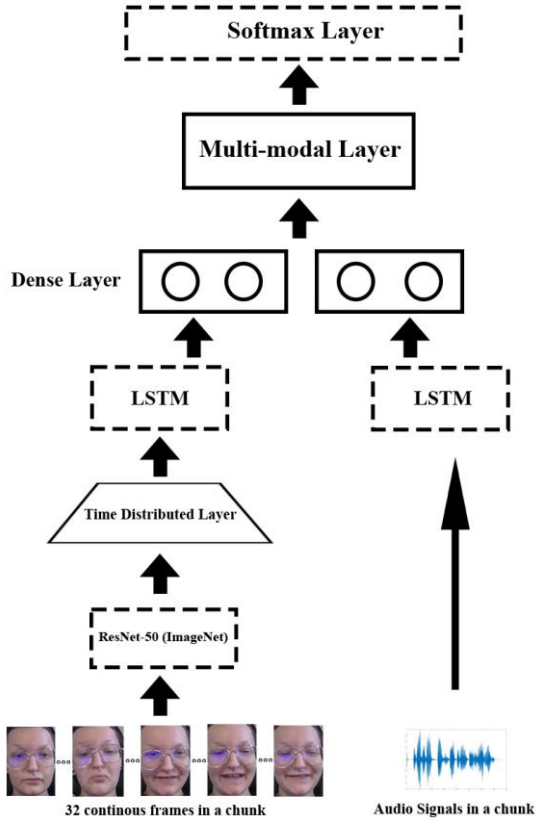

Figure 3: Workflow of the proposed YAP-CSM multi-modal

3.4. Results

Results achieved from the experimented models is presented and discussed within this section.

The details on the performance for each experiment on the Hume-Reaction dataset is provided in table 1. It is clear that YAP-CSM-v4 (ResNet-50) with the chunk size of 1000 (32 frames each video), and learning rate at 0.001, has achieved the best performance among other experiments.

Methodology	Chunk size	No. of Frames	Learning rate	ρ
Baseline Network	-	-	-	0.2488
YAP-CSM-v1 (VGG16)	500	16	0.0004	0.3125
YAP-CSM-v2 (VGG16)	1000	32	0.0004	0.3674
YAP-CSM-v3 (ResNet-50 trained on VGGFace2)	1000	32	0.001	0.3840
YAP-CSM-v4 (ResNet-50 trained on ImageNet)	1000	32	0.001	0.4080

Table 1. Performance comparison between baseline network and our proposed models.

The results above are for the visual modality. We also applied our multi-modal (audio/visual) based on the ResNet50 (pretrained on ImageNet) – Figure 3 for this task. The results for the multi-modal are as follows:

Methodology	Chunk size	No. of Frames	Learning rate	ρ
YAP-CSM-Multimodal (ResNet-50 trained on ImageNet)	1000	32	0.001	0.2980

Table 2. Performance of the proposed YAP-CSM multi-modal.

In the multi-modal experiment, as observed, the performance is lower than the single modal, however, it still benefits from a higher accuracy compared to the baseline network.

4. Conclusion

Our methodology has been tested on the Hume-Reaction dataset using a variety of different networks with modified features were. Among all, we achieved a higher performance by employing YAP-CSM-v4 for estimating of emotional reaction intensity in visual modality. In multimodal setting, the performance tends to be lower possibly due to limitations in the audio quality of the

dataset. Nonetheless, there is potential for further research to improve multimodal models. The preliminary results of all utilized models show a successful implementation of our approach for estimating the emotional reaction compared to the baseline network.

References

[1] C. Darwin and P. Prodger, "The expression of the emotions in man and animals. Oxford University Press, USA," 1998.

[2] 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.

[3] A. Banerjee, P. Washington, C. Mutlu, A. Kline, and D. P. Wall, "Training and profiling a pediatric emotion recognition classifier on mobile devices," *arXiv preprint arXiv:2108.11754*, 2021.

[4] C. Hou, H. Kalantarian, P. Washington, K. Dunlap, and D. P. Wall, "Leveraging video data from a digital smartphone autism therapy to train an emotion detection classifier," *medRxiv*, p. 2021.07. 28.21260646, 2021.

[5] H. Kalantarian *et al.*, "Labeling images with facial emotion and the potential for pediatric healthcare," *Artificial intelligence in medicine*, vol. 98, pp. 77-86, 2019.

[6] H. Kalantarian, K. Jedoui, P. Washington, and D. P. Wall, "A mobile game for automatic emotion-labeling of images," *IEEE transactions on games*, vol. 12, no. 2, pp. 213-218, 2018.

[7] P. Washington *et al.*, "Training an emotion detection classifier using frames from a mobile therapeutic game for children with developmental disorders," *arXiv preprint arXiv:2012.08678*, 2020.

[8] P. Washington *et al.*, "Training affective computer vision models by crowdsourcing soft-target labels," *Cognitive computation*, vol. 13, pp. 1363-1373, 2021.

[9] D. L. Christensen *et al.*, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2012," *MMWR Surveillance Summaries*, vol. 65, no. 13, p. 1, 2018.

[10] K. Ardhanareeswaran and F. Volkmar, "Introduction. Focus: autism spectrum disorders," *The Yale Journal of Biology and Medicine*, vol. 88, no. 1, pp. 3-4, 2015.

[11] E. Gordon-Lipkin, J. Foster, and G. Peacock, "Whittling down the wait time: exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder," *Pediatric Clinics*, vol. 63, no. 5, pp. 851-859, 2016.

[12] J. Manfredonia *et al.*, "Automatic recognition of posed facial expression of emotion in individuals with autism spectrum disorder," *Journal of autism and developmental disorders*, vol. 49, pp. 279-293, 2019.

[13] A. Nag *et al.*, "Toward continuous social phenotyping: analyzing gaze patterns in an emotion recognition task for children with autism through wearable smart glasses," *Journal of medical Internet research*, vol. 22, no. 4, p. e13810, 2020.

[14] A. Lakkapragada *et al.*, "The classification of abnormal hand movement to aid in autism detection: Machine learning study," *JMIR Biomedical Engineering*, vol. 7, no. 1, p. e33771, 2022.

[15] P. Washington *et al.*, "Activity recognition with moving cameras and few training examples: applications for detection of autism-related headbanging," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1-7.

[16] P. Washington *et al.*, "A wearable social interaction aid for children with autism," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, pp. 2348-2354.

[17] N. Haber *et al.*, "A wearable social interaction aid for children with autism," *arXiv preprint arXiv:2004.14281*, 2020.

[18] P. Washington and D. P. Wall, "A Review of and Roadmap for Data Science and Machine Learning for the Neuropsychiatric Phenotype of Autism," *arXiv preprint arXiv:2303.03577*, 2023.

[19] C. Voss, N. Haber, and D. P. Wall, "The potential for machine learning–based wearables to improve socialization in teenagers and adults with autism spectrum disorder—reply," *JAMA pediatrics*, vol. 173, no. 11, pp. 1106-1106, 2019.

[20] J. Daniels *et al.*, "Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism," *NPJ digital medicine*, vol. 1, no. 1, p. 32, 2018.

[21] C. Voss *et al.*, "Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial," *JAMA pediatrics*, vol. 173, no. 5, pp. 446-454, 2019.

[22] H. Kalantarian *et al.*, "The performance of emotion classifiers for children with parent-reported autism: quantitative feasibility study," *JMIR mental health*, vol. 7, no. 4, p. e13174, 2020.

[23] H. Kalantarian, P. Washington, J. Schwartz, J. Daniels, N. Haber, and D. P. Wall, "Guess What? Towards Understanding Autism from Structured Video Using Facial Affect," *Journal of healthcare informatics research*, vol. 3, pp. 43-66, 2019.

[24] A. Kline *et al.*, "Superpower glass," *GetMobile: Mobile Computing and Communications*, vol. 23, no. 2, pp. 35-38, 2019.

[25] C. Voss *et al.*, "Superpower glass: delivering unobtrusive real-time social cues in wearable systems," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp. 1218-1226.

[26] P. Washington *et al.*, "Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 5, no. 8, pp. 759-769, 2020.

[27] N. Haber, C. Voss, and D. Wall, "Making emotions transparent: Google Glass helps autistic kids understand facial expressions through augmented-

reaiity therapy," *IEEE Spectrum*, vol. 57, no. 4, pp. 46-52, 2020.

[28] P. Washington *et al.*, "SuperpowerGlass: a wearable aid for the at-home therapy of children with autism," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 1, no. 3, pp. 1-22, 2017.

[29] J. Daniels *et al.*, "Feasibility testing of a wearable behavioral aid for social learning in children with autism," *Applied clinical informatics*, vol. 9, no. 01, pp. 129-140, 2018.

[30] C. Voss *et al.*, "Designing a holistic at-home learning aid for autism," *arXiv preprint arXiv:2002.04263*, 2020.

[31] A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, "Frame-based facial expression recognition using geometrical features," *Advances in human-computer interaction*, vol. 2014, pp. 4-4, 2014.

[32] 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)*, 2020: IEEE, pp. 637-643.

[33] 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.

[34] D. Kollias *et al.*, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 907-929, 2019.

[35] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: valence and arousal'In-the-Wild'challenge," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 34-41.

[36] D. Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2328-2336.

[37] D. Kollias, "ABAW: learning from synthetic data & multi-task learning challenges," in *European Conference on Computer Vision*, 2023: Springer, pp. 157-172.

[38] 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.

[39] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv preprint arXiv:1910.04855*, 2019.

[40] D. Kollias and S. Zafeiriou, "Analysing affective behavior in the second abaw2 competition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3652-3660.

[41] 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.

[42] L. Christ *et al.*, "The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022, pp. 5-14.

[43] D. Kollias, P. Tzirakis, A. Baird, A. Cowen, and S. Zafeiriou, "ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges," *arXiv preprint arXiv:2303.01498*, 2023.

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

[45] W. Zhang, Z. Guo, K. Chen, L. Li, Z. Zhang, and Y. Ding, "Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition," *arXiv preprint arXiv:2107.03708*, 2021.

[46] L. Meng *et al.*, "Multi-modal emotion estimation for in-the-wild videos," *arXiv preprint arXiv:2203.13032*, 2022.

[47] S. Zhang, R. An, Y. Ding, and C. Guan, "Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2376-2381.

[48] F. Xue, Z. Tan, Y. Zhu, Z. Ma, and G. Guo, "Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2412-2418.

[49] W. Zhang *et al.*, "Transformer-based multimodal information fusion for facial expression analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2428-2437.

[50] W. Jiang, Y. Wu, F. Qiao, L. Meng, Y. Deng, and C. Liu, "Facial action unit recognition with multi-models ensembling," *arXiv preprint arXiv:2203.13046*, 2022.

[51] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[52] Z. Zhang, "Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron," *International journal of pattern recognition and Artificial Intelligence*, vol. 13, no. 06, pp. 893-911, 1999.

[53] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*, 2016: IEEE, pp. 439-448.

[54] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*, 2016: IEEE, pp. 1-10.

[55] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual*

Workshop on Audio/Visual Emotion Challenge, 2017, pp. 19-26.

[56] J. Huang *et al.*, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 11-18.

[57] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[58] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli, "Efficient spatial temporal convolutional features for audiovisual continuous affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 19-26.

[59] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, 2018, pp. 65-72.

[60] J. Li *et al.*, "Hybrid Multimodal Feature Extraction, Mining and Fusion for Sentiment Analysis," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022, pp. 81-88.

[61] A. Raghuvanshi and V. Choksi, "Facial expression recognition with convolutional neural networks," *CS231n Course Projects*, vol. 362, 2016.

[62] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1-12, 2014.

[63] S. Amiriparian *et al.*, "Snore sound classification using image-based deep spectrum features," 2017.

[64] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[65] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211-252, 2015.

[66] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction and building materials*, vol. 157, pp. 322-330, 2017.

[67] D. Theckedath and R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *SN Computer Science*, vol. 1, pp. 1-7, 2020.

[68] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435-442.

[69] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203-5212.

[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[71] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "LSTM for dynamic emotion and group emotion recognition in the wild," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 451-457.