Efficient Feature Extraction and Late Fusion Strategy for Audiovisual Emotional Mimicry Intensity Estimation

Jun Yu[1], Wangyuan Zhu[1]*, Jichao Zhu[1], Zhongpeng Cai[1], Gongpeng Zhao[1],
Zerui Zhang[1], Guochen Xie[1], Zhihong Wei[1], Qingsong Liu[2], Jiaen Liang[2]
[1]University of Science and Technology of China
[2]Unisound AI Technology Co., Ltd.
{harryjun}@ustc.edu.cn
{zhuwangyuan, jichaozhu, zpcai, zgp0531, igodrr, xiegc, weizh588}@mail.ustc.edu.cn
{liuqingsong, liangjiaen}@unisound.com

Abstract

In this paper, we present the solution to the Emotional Mimicry Intensity (EMI) Estimation challenge, which is part of 6th Affective Behavior Analysis in-the-wild (ABAW) 2024. The EMI Estimation challenge task aims to evaluate the emotional intensity of seed videos by assessing them from a set of predefined emotion categories (i.e., "Admiration", "Amusement", "Determination", "Empathic Pain", "Excitement" and "Joy"). To tackle this challenge, we extracted rich dual-channel visual features based on ResNet18 and AUs for the video modality and effective single-channel features based on Wav2Vec2.0 for the audio modality. This allowed us to obtain comprehensive emotional features for the audiovisual modality. Additionally, leveraging a late fusion strategy, we averaged the predictions of the visual and acoustic models, resulting in a more accurate estimation of audiovisual emotional mimicry intensity. Experimental results confirmed the effectiveness of our approach, with the average Pearson's Correlation Coefficient (ρ) of 0.3288 for 6 emotional dimensions in the validation set, and 0.3594 in the test set. Eventually, we achieved third place in the competition.

1. Introduction

In recent years, the study of emotional mimicry intensity estimation has gained significant attention in the fields of affective computing and human-computer interaction. Emotional mimicry refers to the phenomenon where individuals unconsciously imitate the emotional expressions of others, which plays a crucial role in social interaction and empathy[8, 9]. The ability to accurately estimate the intensity of emotional mimicry from audiovisual cues is essential for developing intelligent systems capable of understanding and responding to human emotions effectively[14, 33, 34]. In this paper, we present our comprehensive solution to the EMI Estimation challenge of 6th Affective Behavior Analysis in-the-wild (ABAW) Workshop and Competition[16–21].

The EMI Estimation challenge[21] task aims to investigate emotional mimics through the introduction of a novel and extensive dataset. For this challenge, participants are tasked with employing a multi-output regression approach to predict the intensities of six self-reported emotions: Admiration, Amusement, Determination, Empathic Pain, Excitement and Joy. These emotions are specifically related to decision-making in emotional categories.

In this paper, we propose an efficient feature extraction approach combined with a late fusion strategy for audiovisual emotional mimicry intensity estimation. Our method leverages both auditory and visual modalities to capture comprehensive information about emotional expressions. By extracting discriminative features from audio and video signals, we aim to effectively represent the complex dynamics of emotional mimicry. Furthermore, we introduce a late fusion strategy to integrate the information from both modalities at a later stage of the processing pipeline.

In general, the contributions of our work are as follows:

- we extracted rich dual-channel visual features (ResNet18, AUs) and effective single-channel acoustic features (Wav2Vec2.0). This allowed us to obtain comprehensive emotional features for the audiovisual modality.
- leveraging a late fusion strategy, we averaged the predictions of the visual and acoustic models, resulting in a more accurate estimation of audiovisual emotional mimicry intensity.
- Experimental results validate the effectiveness of our ap-

*Corresponding author

proach, with the average ρ across the 6 emotion dimensionson the test set achieving 0.3594, and securing third place in the EMI Challenge.

The remaining structure of this paper is as follows: Section 2 introduces the related work. Section 3 presents the details of the multimodal features used and the model architecture. Section 4 describes the implementation details of the experiments and provides result analysis. Finally, Section 5 summarizes our work.

2. Related Work

2.1. Video-based facial emotion analysis

Video-based facial emotion analysis is a critical area of research within the field of affective computing, focusing on the automated recognition and interpretation of facial expressions from video data to infer the underlying emotional states of individuals, facial expressions play a vital role in understanding and analyzing emotions. Then a variety of pretrained models for facial expression recognition (FER)[29] or universal image analysis can be employed to extract frame-level visual features. These include ResNet-Affecttet[13, 25], MANet-RAFDB[23, 36], AUs[10], FaceNet[28], ViT[3], among others. Notably, AUs provides an interpretable approach by considering the activation of specific facial muscles to encode facial expressions. ResNet employs skip connections based on identity mappings to facilitate deep neural network training, while MANet combines a global multi-scale module and a local attention module for capturing both local and global information in facial emotion recognition. As for Affectnet datasets, The dataset is a large-scale dataset widely used in research on emotion recognition. It consists of facial images sourced from the internet, annotated with seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. In summary, the AffectNet dataset's extensive size, containing hundreds of thousands of images, makes it a significant resource for studying emotion recognition and sentiment analysis.

2.2. Audio-based emotion analysis

In the realm of audio-based emotion analysis, feature extraction plays a crucial role in capturing discriminative information from acoustic signals to characterize emotional states accurately. Over the years, researchers have explored various feature extraction techniques to represent the complex dynamics of human emotions present in audio recordings. Earlier studies often relied on traditional manual feature extraction methods, such as Mel-frequency cepstral coefficients (MFCC) and extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)[12]. MFCC features capture the spectral characteristics of audio signals by transforming the power spectrum of the signal into the mel frequency domain and applying the discrete cosine transform. MFCC provides a compact representation of the audio signal, capturing crucial spectral information related to mood. The feature can be extracted using the opensmil[11] toolkit. On the other hand, eGeMAPS is specifically designed to capture a wide range of acoustic properties from speech signals, encompassing various acoustic parameters that cover spectrum, prosody, and speech quality characteristics. eGeMAPS has emerged as a popular choice for speech-based emotion recognition and feature extraction in other related fields[32]. However, with the advent of deep learning, pre-trained models have gained widespread adoption in speech feature extraction tasks. DeepSpectrum[1], for instance, leverages pre-trained convolutional neural networks originally designed for image recognition to extract acoustic features. Its effectiveness has been demonstrated in various speech and audio recognition tasks.Furthermore, large models based on BERT[7], Wav2Vec[27] and Wav2Vec2.0[2] have also been applied to speech feature extraction, showcasing the potential of deep learning approaches in capturing complex patterns and dependencies in audio data.

Figure 1. Preprocessed consecutive facial images

3. Approach

In this section, we describe our method in detail, the architecture of the model is shown in Fig.2

3.1. Pre-processing

For the EMI estimation task involving video streams, the primary objective is to detect changes in facial expressions. Therefore, preprocessing of the video data is of paramount importance. We observed varying frame rates in the original videos, resulting in notable discrepancies in the number of extracted video frames. Consequently, we initially utilize the ffmpeg[30] tool to standardize the frame rate of each video sample to 30 fps. For each video sample, we begin by using OpenCV[6] to read the video as a series of consecutive frames. Subsequently, we conduct face detection on these frames, proceeding frame by frame. Frames containing detected faces are retained, while those without faces are discarded. For face detection and facial key point localization, we employ the Multi-task Cascaded Convolutional

Networks (MTCNN)[35] algorithm. Following face detection, we utilize the coordinates of facial key points provided by MTCNN to crop appropriate face regions from the images. These face regions are resized to 112x112 pixels and saved in the corresponding folders based on their sequence order, in preparation for subsequent feature extraction. The consecutive face images after preprocessing are shown in the Fig.1.

3.2. Multimodal feature extraction

3.2.1 Videos features

In the preprocessing stage, we have extracted continuous face images from the video, allowing us to directly extract facial expression features from these images.

Vision Transformer (ViT): As an alternative vision-based approach, we utilize a DINO-trained ViT[3], which has undergone pre-training on the ImageNet-1K dataset using the Label-free Self-distillation (DINO) method. This model has demonstrated effectiveness across a range of image-based tasks, including facial expression emotion recognition [4]. Upon processing the extracted facial images, the model generates a 384-dimensional embedding for each image. No additional pre-training or fine-tuning is performed on the model.

ResNet18: The Convolutional Neural Network (CNN) ResNet18, introduced by [13], is renowned for its exceptional ability to extract features from images. In order to enhance its performance on facial datasets, a ResNet18 network pre-trained on AffectNet[25] is utilized to extract global spatial features from facial images. The features before the final fully connected layer are averaged to obtain a 512-dimensional feature vector. This improvement further enhances the network's capability to accurately extract relevant facial features.

AUs: The Facial Action Coding System (FACS)[10] is a comprehensive method for objectively coding facial expressions. In FACS, Action Units (AUs) correspond to specific facial muscles. Each AU has two dimensions: the first dimension indicates detection, with 0 indicating absence and 1 indicating presence. The second dimension represents intensity, ranging from 0 to 1. We utilize OpenFace2.0[22] to extract the detection and intensity of 17 AUs relevant to facial expressions. Ultimately, each facial image receives a 34-dimensional feature embedding.

3.2.2 Audios features

Before extracting the audio features, we normalized all audio files to -3 dB and converted them to a format of 16 kHz, 16-bit mono.

Wav2Vec2.0: Self-supervised pre-trained Transformer models have garnered considerable attention in the field of computer audition. A prominent example of such a foundational model is Wav2Vec2.0 [2], which is frequently employed for Speech Emotion Recognition (SER) [26]. Given that all subchallenges are emotion-related, we leverage Wav2Vec2.0, specifically a large version fine-tuned on the MSP-Podcast[24] dataset, for continuous emotion recognition. We derive audio signal features by averaging the representations from the final layer of the model, resulting in a 768-dimensional embedding.

3.3. Temporal Encoder

For the temporal encoder input, visual features are dual-channel, while acoustic features are single-channel.For the visual input features $X_v \in \mathbb{R}^{T \times d_v}$, where T represents the temporal dimension, while d_v represents the spatial dimension. The calculation formula is as follows:

$$X_v = \text{Concate}(ResNet18(x), AUs(x)) \tag{1}$$

where $x \in \mathbb{R}^{T \times H \times W \times 3}$ represents the input sequence of facial images, $T = 300$ represents the number of images, where H and W are both 112. "Concate" means to concatenate the two outputs along the spatial dimension.Similarly, acoustic features $X_a \in \mathbb{R}^{T \times d_a}$ represent the feature sequence obtained from the Wav2Vec2.0 model.

Then, the visual features X_v and acoustic features X_a are fed into a Temporal Convolutional Network (TCN)[5] based on one-dimensional causal convolution to gather local temporal context.TCN utilize dilated causal convolutions to capture temporal dependencies over long sequences efficiently. The architecture of TCN typically consists of multiple layers of dilated convolutional filters, followed by activation functions and possibly other layers such as pooling or normalization. The expression for a single layer of a TCN can be represented as follows:

$$y_t = Relu(W * x_{t-d} + b) \tag{2}$$

Where y_t is the output at time step t. x_{t-d} represents the input sequence, with d denoting the dilation factor.The dilation factor d determines the receptive field of the convolutional filter and controls how many past time steps the filter can consider. W is the learnable convolutional filter. b is the bias term. $Relu$ represents the ReLU activation function. After passing through the TCN network, both the visual and acoustic features have the same spatial dimension d_{model}. In addition, behind the visual branch, a Transformer[31] Encoder is cascaded to interact with and integrate different parts of the input sequence, thereby extracting rich feature representations and capturing long-range dependencies in the sequence without introducing recursive structures.

3.4. FFN

Feedforward Neural Network (FFN) consists of two fully connected layers and a non-linear activation function to

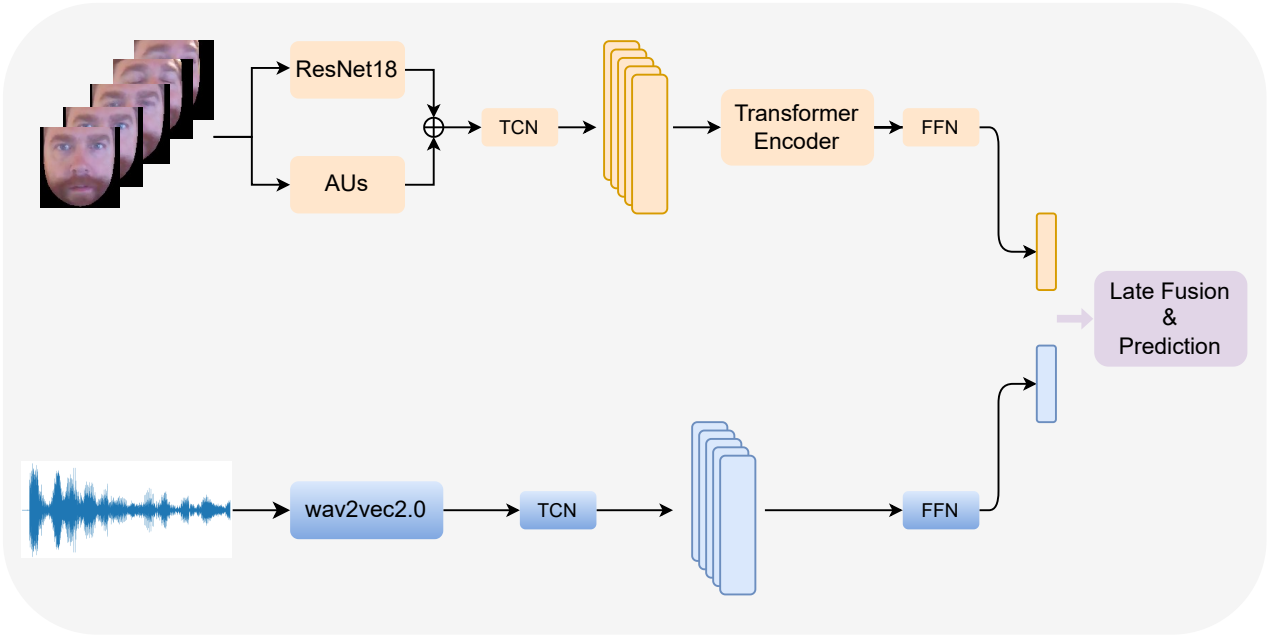

Figure 2. The overall framework we proposed consists of two branches: a video branch with dual-channel visual features based on ResNet18 and AUs, and an audio branch utilizing single-channel acoustic features from Wav2Vec2.0. The predictions from both branches are fused late to obtain the final result. TCN and FFN denote Temporal Convolutional Network and Feedforward Neural Network, respectively, while $\bigoplus$ indicates spatial dimension concatenation of dual-channel features.

achieve non-linear mapping of input features. The input undergoes linear transformation and non-linear activation (ReLU) through the first fully connected layer, followed by linear transformation through the second fully connected layer to obtain the final output. Then the visual and acoustic output can be expressed as:

$$F_v = ReLU(X_v W_1 + b_1)W_2 + b_2 \quad (3)$$

$$F_a = ReLU(X_a W_1 + b_1)W_2 + b_2 \quad (4)$$

where W_1 and b_1 are the weight matrix and bias vector of the first fully connected layer, while W_2 and b_2 are those of the second fully connected layer. $F_v \in \mathbb{R}^6$ and $F_a \in \mathbb{R}^6$ respectively represent the outputs of the visual branch and the acoustic branch, both with a dimensionality of 6.

3.5. Late Fusion and Prediction

Our late fusion approach involves training the visual and acoustic models separately to make individual predictions, and then averaging the audiovisual results to obtain a final prediction from both modalities. The formula is as follows:

$$F_{va} = average(F_v, F_a) \quad (5)$$

where F_{va} represents the final bimodal prediction result.

3.6. Optimisation objective

In this work, we utilize the mean square error (MSE) loss function in our training procedure. Let $y = [y_1, \cdots, y_6]$ and $\hat{y} = [\hat{y}_1, \cdots, \hat{y}_6]$ be the true emotional reaction intensity and the prediction, respectively, then the loss $\mathcal{L}$ can be defined as:

$$\mathcal{L} = \text{MSE}(y, \hat{y}) \quad (6)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

where n represents the number of samples, y_i and $\hat{y}_i$ denote the true label and predicted value of the ith sample, respectively.

4. Experiments

4.1. Dataset

In this work, we employed the multimodal Hume-Vidmimic2[21] dataset, which comprises over 15,000 videos featuring 557 participants and spanning more than 30 hours of audiovisual content. Within this dataset, each participant was instructed to mimic a "seed" video depicting a person expressing a particular emotion. Following the mimicry task, participants were prompted to assess the emotional intensity of the resulting video by selecting from a predefined set of emotional categories. The dataset is speaker-independent and partitioned into training, validation, and test sets. Table 1 presents the dataset statistics for each partition.

Partition	Duration	Samples
Train	15:07:03	8072
Validation	9:12:02	4588
Test	9:04:05	4586
$\sum$	33:23:10	17246

Table 1. Hume-Vidmimic2 partition statistics.

4.2. Implement Details

Evaluation metric Average Pearson's Correlations Coefficient (ρ) is the metric used in intensity estimation, which is a measure of linear correlation between predicted emotional reaction intensity and target, then the metric can be defined as follows:

$$\rho = \sum_{i=1}^{6} \frac{\rho_i}{6} \tag{8}$$

where $\rho_i (i \in \{1, 2 \cdots, 6\})$ for 6 emotions, respectively, and is defined as:

$$\rho_i = \frac{\text{cov}(y_i, \hat{y_i})}{\sqrt{\text{var}(y_i)\text{var}(\hat{y_i})}} \tag{9}$$

where $\text{cov}(y_i, \hat{y_i})$ is the covariance between the predicted value and the target, $\text{var}(y_i)$ and $\text{var}(\hat{y_i})$ are variance respectively.

Training settings The training process is optimized using the Adam optimizer [15]. All experiments were conducted on an NVIDIA RTX 3090 GPU with PyTorch, using an initial learning rate of $3e^{-5}$ and a batch size of 128. Additionally, if the validation set metric does not improve for 10 epochs, the learning rate is halved. The visual branch encoder has a dimension of 546 with 2 encoder blocks and 4 multi-heads. In the temporal encoder, the 1-dimensional convolution kernel size is 3, there are 5 convolution layers, and the feature dimension in attention is 128.

4.3. Results

Unimodal Results. For the EMI estimation challenge, we initially assessed the effectiveness of our unimodal features (video, audio) on the validation set and compared them with the officially provided features. The experimental results are presented in Table 2. It is evident from Table 2 that combining our extracted features with the official ones led to performance improvements over the baseline on the validation set. Concerning the model, retraining the official features with our model resulted in notable enhancements in the visual and audio modalities, with increases of 1% and 4.08%, respectively. Furthermore, our experiments with ResNet18 and AUs for visual feature extraction yielded significant improvements of 0.1236 and 0.1352, respectively. Finally, by integrating the visual dual-channel features (ResNet18, AUs), we achieved a visual result of 0.1479, thus confirming the efficacy of our extracted multi-channel features.

Features	Modality	Mean ρ
ViT(baseline)[21]	V	0.09
ViT(ours)	V	0.10
ResNet18	V	0.1236
AUs	V	0.1352
ResNet18+AUs	V	**0.1479**
Wav2Vec2.0(baseline)[21]	A	0.24
Wav2Vec2.0(ours)	A	**0.2808**

Table 2. The unimodal results on validation set of the EMI Estimation Challenge. We report the Pearson correlation coefficient (ρ) for the average of 6 emotion targets. Where V represents the visual modality, and A represents the acoustic modality.

Multimodal Results. After obtaining the unimodal prediction results, we utilized a late fusion strategy to obtain audiovisual prediction results. The experimental results are presented in Table 3. From Table 3, it is evident that our multimodal results outperform the official ones. Ultimately, by averaging the visual dual-channel features (ResNet18, AUs) and the acoustic single-channel feature (Wav2Vec2.0), we achieved a performance of 0.3288 on the validation set, representing a 7.88 percentage point improvement over the baseline.

Modality	Features	Late Fusion (ρ)
V+A	ViT+Wav2Vec2.0(baseline)	0.25
	ViT+Wav2Vec2.0(ours)	0.2835
	ResNet18+Wav2Vec2.0	0.2983
	AUs+Wav2Vec2.0	0.3026
	ResNet18+AUs+Wav2Vec2.0	**0.3288**

Table 3. Multimodal results of late fusion for mean ρ on the validation set.

Evaluation on the test set. We validated our approach in the Hume-Vidmimic2 test set of the EMI Estimation Challenge. The competition results for all participating teams are shown in Table 4 . The average Pearson's Correlation Coefficient (ρ) across 6 emotional dimensions is 0.3594, achieving third place in the competition.

5. Conclusion

In this paper, we present the solution to the Emotional Mimicry Intensity (EMI) Estimation challenge, which is part of 6th Affective Behavior Analysis in-the-wild (ABAW) Competition. We obtained effective feature representations by extracting visual dual-channel fea-

Team	ρ
Netease Fuxi AI Lab	0.7185
HCAI-VIS	0.5536
USTC-IAT-United (Ours)	**0.3594**
HSEmotion	0.3316

Table 4. The final results on the Hume-Vidmimic2 test set. Numbers are provided by the challenge organizing committee website.

tures (ResNet18, AUs) and acoustic single-channel feature (Wav2Vec2.0). Subsequently, based on a late fusion strategy, we fused the audiovisual results. Experimental results validate the effectiveness of our approach, with the average ρ across the 6 emotion dimensionson the test set achieving 0.3594, and securing third place in the EMI Challenge.

6. Acknowledgments

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B, CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), USTC-IAT Application Sci. & Tech. Achievement Cultivation Program (JL06521001Y), Sci. & Tech. Innovation Special Zone (20-163-14-LZ-001-004-01).

References

[1] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. Snore sound classification using image-based deep spectrum features. 2017. 2

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 2, 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3

[4] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. Vitfer: facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80, 2022. 3

[5] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2020. 3

[6] G Sumanth Naga Deepak, B Rohit, Ch Akhil, D Sai Surya Chandra Bharath, and Kolla Bhanu Prakash. An approach for morse code translation from eye blinks using tree based machine learning algorithms and opencv. In *Journal of Physics: Conference Series*, page 012070. IOP Publishing, 2021. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[8] Chaoyue Ding, Jiakui Li, Martin Zong, and Baoxiang Li. Speed-robust keyword spotting via soft self-attention on multi-scale features. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111. IEEE, 2023. 1

[9] Kevin Ding, Martin Zong, Jiakui Li, and Baoxiang Li. Letr: A lightweight and efficient transformer for keyword spotting. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7987–7991. IEEE, 2022. 1

[10] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2, 3

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010. 2

[12] 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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3

[14] Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng. Multimodal temporal attention in sentiment analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 61–66, 2022. 1

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Dimitrios 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*, pages 2328–2336, 2022. 1

[17] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023.

[18] 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.

[19] 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.

[20] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023.

[21] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 1, 4, 5

[22] Peter A Krause, Christopher A Kay, and Alan H Kawamoto. Automatic motion tracking of lips using digital video and openface 2.0. *Laboratory Phonology*, 11(1), 2020. 3

[23] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 2

[24] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017. 3

[25] 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, 3

[26] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926. IEEE, 2022. 3

[27] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 2

[28] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*, pages 1–5. IEEE, 2020. 2

[29] Yingli Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression recognition. *Handbook of face recognition*, pages 487–519, 2011. 2

[30] Suramya Tomar. Converting video formats with ffmpeg. *Linux journal*, 2006(146):10, 2006. 2

[31] 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. 3

[32] Bogdan Vlasenko, RaviShankar Prasad, and Mathew Magimai.-Doss. Fusion of acoustic and linguistic information using supervised autoencoder for improved emotion recognition. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 51–59. 2021. 2

[33] Jun Yu, Jichao Zhu, Wangyuan Zhu, Zhongpeng Cai, Guochen Xie, Renda Li, Gongpeng Zhao, Qiang Ling, Lei Wang, Cong Wang, et al. A dual branch network for emotional reaction intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5810–5817, 2023. 1

[34] Jun Yu, Wangyuan Zhu, Jichao Zhu, Xiaxin Shen, Jianqing Sun, and Jiaen Liang. Mmt-gd: Multi-modal transformer with graph distillation for cross-cultural humor detection. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, pages 43–49, 2023. 1

[35] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 3

[36] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 2