A Dual Branch Network for Emotional Reaction Intensity Estimation

Jun Yu, Jichao Zhu, Wangyuan Zhu, Zhongpeng Cai, Guochen Xie, Renda Li, Gongpeng Zhao
University of Science and Technology of China

harryjun@ustc.edu.cn,
{jichaozhu, zhuwangyuan, zpcai, xiegc, rdli, zgp0531}@mail.ustc.edu.cn

Abstract

Emotional Reaction Intensity(ERI) estimation is an important task in multimodal scenarios, and has fundamental applications in medicine, safe driving and other fields. In this paper, we propose a solution to the ERI challenge of the fifth Affective Behavior Analysis in-the-wild(ABAW), a dual-branch based multi-output regression model. The spatial attention mechanism is used to better extract visual features, and the Mel-Frequency Cepstral Coefficients technology extracts acoustic features, and a method named modality dropout is added to fusion multimodal features. Our method achieves excellent results on the official validation set.

1. Introduction

With advances in artificial intelligence and deep learning, researchers are increasingly interested in computational methods for human emotional reactions [4]. It can help doctors diagnose whether patients have anxiety, depression, etc. by calculating the intensity of emotional reactions. In addition, it can also be used in scenarios such as education [31], entertainment [1], and driver safety detection [8].

For static facial expression recognition and dynamic facial expression recognition tasks, common emotion description methods include action units (AU), arousal and valence, etc, which have been used in ABAW's challenge [12–16, 18–21, 32].

For traditional methods based on Support Vector Machine-Hidden Markov Model(HMM) [22]. With the development of deep learning, CNN, 3DCNN, and RNN methods have been applied to visual tasks. In recent years, with the excellent performance of Transformer [28] in natural language processing, ViT [5] has been successfully applied in computer vision, and has produced many excellent pre-training models. However, these works mainly classify emotional samples into specific, which is a typical single classification task. Furthermore, datasets collected in laboratory share similar fixed patterns, with emotional expres-

sions having similar intensities. By collecting data on the web and creating datasets to make it more wild. However, in these tasks, static or dynamic expressions are recognized as limited class, while the connection between emotions is ignored [29], which is not enough to finely reflect the emotional state.

In order to promote the development of emotional reaction intensity(ERI), ABAW2023 holds this competition, our goal is to design a model to predict the reaction intensity of different emotions, including: Adoration, Amusement, Anxiety, Disgust, Empathic-Pain, Fear, Surprise, so it's a multi-output regression task. Therefore, in this paper, we propose a video feature extraction model based on CNN and spatial attention. On the basis of fusing local inter-frame information through temporal convolutional network, we use the temporal transformer to obtain the global temporal relationship, and use the MFCC to generate audio branch's feature. Use a timing modeling method similar to video branch to obtain the global timing relationship. Finally, the features of the video and audio are fused into the prediction head to estimate the vector of emotional reaction intensity.

In summary, our contributions can be summarized as follows:

- We propose a dual-branch model for the ABAW ERI Estimation track. It consists of Spatial Encoder with CNN, MFCC and Temporal Encoder.

- We introduce a mechanism based on modality dropout to fuse visual and aduio features.

2. Related Work

In [27] they propose a method called ViPER. Based on the pre-trained Vision Transformer [6], a modality-independent fusion framework is designed to predict people's emotional state, whose input data can be a combination of audio and video frames and text. For dynamic facial expression recognition, unlike static facial expression recognition samples which tend to exhibit consistent high-intensity expressions, [23] designs a global convolution-attention block (GCA) module to weight the feature map
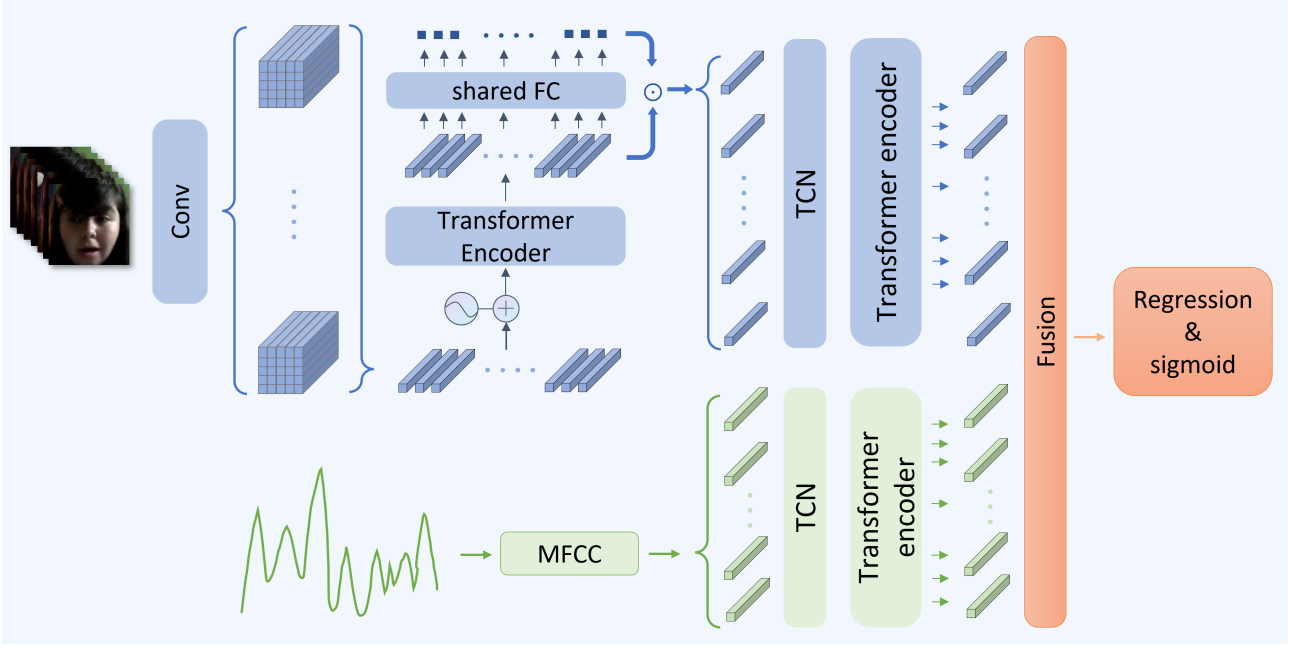
Figure 1. proposed architecture

channels so as not to Correlated features can be suppressed to avoid equal treatment of frames with different expression strengths in a video sequence. Meanwhile, an intensity-aware loss-guided network is designed to distinguish emotion samples with relatively low expression intensity. Feature extraction from video streams affects model performance. [26, 29] both proposed models based on spatial attention mechanism and aggregating different frames through convolutional or linear layers. Among them, [29] is pre-trained on the DFEW dataset and trained on ERI, which improves the effect of processing dynamic emotions. For audio-based ERI estimation, [24, 29] use extended Geneva Minimalistic Acoustic Parameter Set(eGeMAPS) and DeepSpectrum of DenseNet121 pretrained on ImageNet as features, but the performance are lower than video-based methods. [24] also uses ResNet18 [7] as the backbone to extract audio features, achieving the best results based on audio methods.

3. Methodology

In this section, we describe our method in detail, the architecture of the model is shown in Fig. 1

3.1. Pre-processing

It is necessary to preprocess the video streams in the dataset, since the focus of ERI is the facial region, which should be avoided from being disturbed by other factors. The videos are first split into images, and we use the dlib [10] toolkit to detect 68 facial landmarks in each frame to crop the face, and resize to 112×112 as the input size. Besides, interpolation with a window width of 12 frames and frame smoothing are used to handle frames that cannot be detected by dlib. For face samples cannot be detected in the entire video frame, we also use a more robust MTCNN [33] method. Due to the subject's webcam has different fps, the frame number of the video ranges from 50 to 1561, while the duration of the video ranges from 9.9 seconds to 15 seconds. Therefore, in order to facilitate processing and save GPU memory, we uniformly extract 32 frames from each sample as input.

3.2. Visual backbone

CNN. The input x for the visual branch is faces clipped from the original video stream with a linear sampling of $T = 32$ frames, for our model $x \in \mathbb{R}^{T \times H \times W \times 3}$. The first foure convolutional layers of ResNet18 [7] are applied on cropped images to extract low-level features $f \in \mathbb{R}^{T \times c \times h \times w}$.

Spatial-Encoder. In order to input the CNN feature map f of every input frame $t \in \{1, 2, \cdots, T\}$ into a shared spatial-encoder, we first flatten each frame's feature into a two-dimensional shape as $f_t \in \mathbb{R}^{(hw) \times d_{model}}$, and add the spatial positional embedding to f_t. Therefore, the encoder's input can be defined as:

$$f_{t,i} = f_{t,i} + p_i \tag{1}$$

where $p_i(i \in \{1, 2, \cdots, hw\})$ is a learnable location parameter, and an enhanced feature map z_t is obtained after

2

computing by spatial encoder as:

$$z_t = \text{encoder}(f_t) \quad (2)$$

A full-connection layer with softmax is added on z_t to aggregate the information of each position and generate a position weight a_t. Weighted by a_t, from the enhanced feature map z_t, we obtain aggregation feature g_t:

$$a_{t,i} = \text{FC}(z_{t,i}), i \in \{1, 2, \cdots, hw\} \quad (3)$$

$$a_t = \text{softmax}(a_{t,1}, \cdots, a_{t,hw}) \quad (4)$$

$$g_t = \sum_{i=1}^{hw} (a_t z_t) \quad (5)$$

where FC denotes full-connection layer, and g_t is corresponded to a frame. By stacking g_t in the time dimension, we obtain the feature sequence $g = (g_1, \cdots, g_T)$ of a video only containing the position representation.

3.3. Acoustic feature

Mel Frequency Cepstral Coefficient(MFCC) is an audio feature widely used in speech recognition and emotion recognition, which is very close to the human hearing system. We used the Python toolkit Librosa to extract 128-dimensional features, and then combined the adjacent 8-frame features to obtain 1024-dimensional feature vector, which is fed into temporal encoder described in the next.

3.4. Temporal Encoder

In this section, the module proposed by [3] is adopted. The extracted feature vectors from visual or audio backbone are fed into the temporal convolutional network(TCN) based on 1-dimensional causal convolutional with dilation to aggregate local temporal context . Besides, zero padding is also used to ensure that the output of the feature through the convolutional layer has the same length as the input. Causal convolution can ensure that the current time t can only see the information at the previous time, so as to avoid information leakage. Denote ker as the convolution kernel and dilation rate is d, and the input of time dimension x, then the p-th element of output feature y can be calculated by:

$$y_p = \sum_{di+t=p} ker(i) \cdot x_t \quad (6)$$

by using different d, the receptive field of the convolution can be changed dynamically. Finally, add temporal position information to the output of TCN as input of transformer encoder to capture global information.

3.5. Multimodal Fusion and Regression

To efficiently integrate the good features learned in video and audio models, and to avoid the model's excessive dependence on a certain modality during learning, we use a fusion strategy called modality dropout, which is applied at the modal level. With a probability p_m, both audio and video feature are used as input, when only one is used, the video feature is selected with a probability of p_v. Given the audio feature f^a and video feature f^v, the mltimodal feature f^m with modality dropout is:

$$f^{av} = \begin{cases} \text{concat}(f^a, f^v) & \text{with } p_m \\ \text{concat}(0, f^v) & \text{with } (1-p_m)p_v \\ \text{concat}(f^a, 0) & \text{with } (1-p_m)(1-p_v) \end{cases} \quad (7)$$

$$f^m = W_m f^{av} + b_m \quad (8)$$

where W_m and b_m are learnable parameters, concat denotes channel-wise concatenation, and a layer normalization [2] is added after concatenation. Finally, similar aggregation method in spatial encoder is used to estimate the reaction intensity. The difference is that sigmoid is added as the activation function to normalize the value to $(0, 1)$.

3.6. Optimisation objective

In this work, we use mean square error(MSE) loss for our training process. Let $y = [y_1, \cdots, y_7]$ and $\hat{y} = [\hat{y}_1, \cdots, \hat{y}_7]$ be the true emotional reaction intensity and the prediction, respectively, then the loss $\mathcal{L}$ can be defined as:

$$\mathcal{L} = \text{MSE}(y, \hat{y}) = \mathbb{E}[\sum_{i,j} (y_{ij} - \hat{y}_{ij})^2] \quad (9)$$

where i denotes the emotion, j denotes the batch size.

4. Experiment

4.1. Dataset

Hume-Reaction dataset is used for the ERI Estimation Challenge in 5th ABAW. It is a reaction of subjects from two cultures, South Africa and the United States, to emotional video stimuli. It consists of both audio and video parts and is recorded over approximately 75 hours. Corresponding label vector is self-annotated by the subjects and normalized to $[0, 1]$ by its maximum intensity, and 7 elements represent adoration, amusement, anxiety, disgust, empathic-pain, fear, and surprise, respectively.

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:

$$\bar{\rho} = \sum_{i=1}^{7} \frac{\rho_i}{7} \quad (10)$$
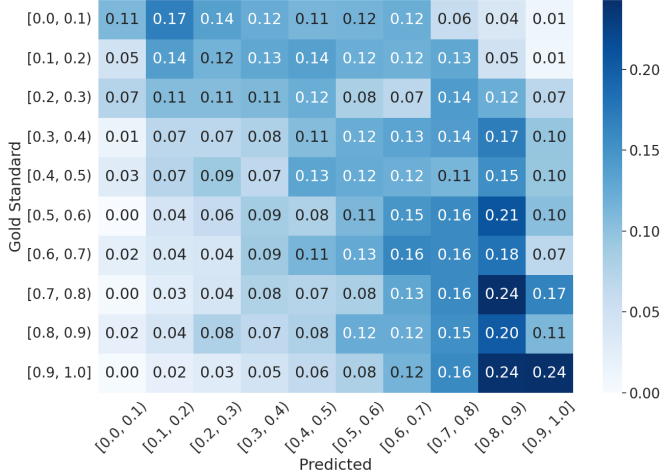
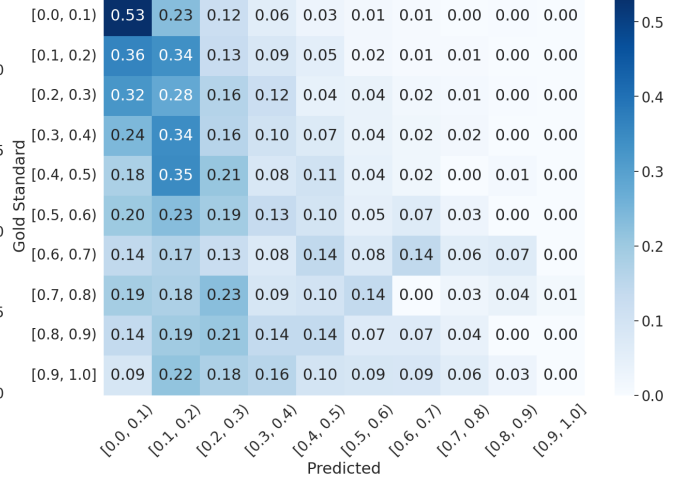Figure 2. Amusement prediction analysis

Figure 3. Amusement prediction analysis

where $\rho_i (i \in \{1, 2 \cdots, 7\})$ for 7 emotions, respectively, and is defined as:

$$\rho_i = \frac{\text{cov}(y_i, \hat{y}_i)}{\sqrt{\text{var}(y_i)\text{var}(\hat{y}_i)}} \quad (11)$$

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 by Adam [11] optimizer. All the experiments are implemented on NVIDIA RTX 3090 with PyTorch, with initial learning rate of e^{-4}, batch size of 64. And when metric on the validation set don't improve for 10 epochs, the learning rate will halved. For visual branch, we use ResNet18 weight from [30] trained on AffectNet [25] as initialization parameter, it can capture effective features in static face representation recognition task. We freeze the unimodal's model parameters with the highest ρ, extract audio and video features and feed them to the fusion module. The dimension of encoder in visual branch is 256, equals to low level feature's channel, and the number of encoder blocks is 4 and number of multi-head is 4. In temporal encoder, the kernel size of 1-dimension convolution is 3, and convolution layer is 5, the dimension of feature in attention is 128. For fusion module, modality dropout and video dropout are set to 0.1 and 0.5, respectively.

4.3. Results

In our unimodal experiment, both the branch based on video and audio, the performance of the model are greatly improved as shown in Tab. 1, and the correlation coefficients are increased to 0.2972, 0.3500, which have a significant improvement compared with baseline.

Ablations We conducted ablation studies on the Hume-Reaction dataset to better understand our proposed model,

Method	Audio	Video	$\bar{\rho}$
Baseline(eGeMAPS) [17]	✓	-	0.0583
Baseline(DeepSpectrum) [17]	✓	-	0.1087
ours	✓	-	0.2972
Baseline(FAU) [17]	-	✓	0.2840
Baseline(VGGFACE2) [17]	-	✓	0.2488
ours	-	✓	0.3500
Baseline [17]	✓	✓	0.2382
ViPER [27]	✓	✓	0.3025
ours	✓	✓	0.4429

Table 1. Results on validation set

and the results are shown in Tab. 2. If p_m is set to 1.0, i.e., the two features from temporal encoder are directly concatenated together in last dimension, ρ is 0.4426, is significantly boosted compared with unimodal, $p_m = 0.9$, ρ has a slight improvement.

Qualitative analysis On the validation set, we calculated the emotional response intensity of different samples, and the average value was 0.44, which indicated that our estimated value could well reflect the emotions of the sample subjects. For the convenience of analysis, we selected amusement and empathic-pain according to the average response intensity of each emotion in the training set, divided them into 10 levels on the basis of intensity, and counted the corresponding number of samples. The results are shown in Fig. 4. The former has more high-intensity samples, while the latter has a majority of low-intensity samples.

Fig. 2 and Fig. 3 show the confusion matrix between the estimated value and the actual intensity of the amusement class and empathic-pain class, respectively. For the amusement class, these values are roughly scattered around
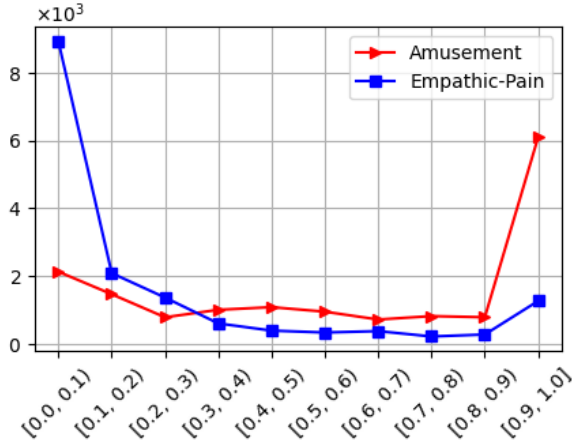
4

Figure 4. Amusement and empathic-pain sample statistics with quantized intensity on train set

Method	p_m	p_v	$\bar{\rho}$
ours	1.0	-	0.4426
ours	0.9	0.5	0.4429
ours	0.8	0.5	0.4424

Table 2. Result with modality dropout on validation set

the diagonal, and perform well for high intensities with 0.28, indicating that the model can calculate the intensity of emotional responses well, while the empathic-pain class is mainly distributed in $[0, 0.3]$. This difference may be caused by the label distribution of the train set, as shown in Fig. 4, the label intensity of the latter is mainly concentrated in $[0, 0.2]$, or consistent with the classification results in [9,23] , the amusement(happy) class is easier to be learned by the model.

5. Conclusion

In this paper, we propose a multimodal based method to improve the performance of emotional response intensity estimation, extracting the global feature information of the face based on spatial attention in the visual branch, and generating features based on MFCC in the acoustic branch. Our method outperforms prior work on the Hume-Reaction dataset.

Acknowledge

References

[1] Moch. Taufik Akbar, Muhamad Nasrul Ilmi, Imanuel V. Rumayar, Jurike V. Moniaga, Tin-Kai Chen, and Andry Chowanda. Enhancing game experience with facial expression recognition as dynamic balancing. *Procedia Computer Science*, 2019. 1

[2] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016. 3

[3] 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, 2021. 3

[4] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Muller, Lukas Stappen, Eva-Maria Messner, Andreas Konig, Alan S. Cowen, E. Cambria, and Björn Wolfgang Schuller. The muse 2022 multimodal sentiment analysis challenge: Humor, emotional reactions, and stress. *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 1

[7] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 2

[8] Mira Jeong and ByoungChul Ko. Driver's facial expression recognition in real-time for safe driving. *Sensors (Basel, Switzerland)*, 18, 2018. 1

[9] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 5

[10] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, 2009. 2

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

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

[13] 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

[14] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 1

[15] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 1

[16] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[17] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan S. Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *ArXiv*, abs/2303.01498, 2023. 4

[18] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 1

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

[20] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[21] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[22] Tarun Krishna, Ayush Rai, Shubham Bansal, Shubham Khandelwal, Shubham Gupta, and Dushyant Goyal. Emotion recognition using facial and audio features. In *International Conference on Multimodal Interaction*, 2013. 1

[23] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. *ArXiv*, abs/2208.10335, 2022. 1, 5

[24] Jia Li, Ziyang Zhang, Jun Lang, Yueqi Jiang, Liuwei An, Peng Zou, Yang Xu, Sheng Gao, Jie Lin, Chunxiao Fan, Xiao Sun, and Meng Wang. Hybrid multimodal feature extraction, mining and fusion for sentiment analysis. *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022. 2

[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*, 10:18–31, 2017. 4

[26] Prajwal K R, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention.

2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5152–5162, 2021. 2

[27] Lorenzo Vaiani, Moreno La Quatra, Luca Cagliero, and Paolo Garza. Viper: Video-based perceiver for emotion recognition. *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022. 1, 4

[28] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 1

[29] Kexin Wang, Zheng Lian, Licai Sun, B. Liu, Jianhua Tao, and Yin Fan. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022. 1, 2

[30] Zhengyao Wen, Wen-Long Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *ArXiv*, abs/2109.07270, 2021. 4

[31] Torsten Wilhelm. Towards facial expression analysis in a driver assistance system. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4, 2019. 1

[32] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1

[33] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 2016. 2