A Dual Branch Network for Emotional Reaction Intensity Estimation

Jun Yu[1], Jichao Zhu [1*], Wangyuan Zhu[1], Zhongpeng Cai[1], Guochen Xie[1], Renda Li[1],
Gongpeng Zhao[1], Qiang Ling[1], Lei Wang[1], Cong Wang[2], Luyu Qiu[2], Wei Zheng[2]
[1]University of Science and Technology of China
[2]Huawei Techologies
{harryjun,qling,wangl}@ustc.edu.cn
{jichaozhu,zhuwangyuan,zpcai,xiegc,rdli,zgp0531}@mail.ustc.edu.cn
{wangcong64,qiuluyu,victor.zhengwei}@huawei.com

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. Temporal Encoder is composed of Temporal Convolutional Network and Transformer Encoder, which is used to capture the temporal relationship between features. And a method named modality dropout is added to fusion multimodal features. Our approach for ERI challenge achieves Pearson's Correlation Coefficient of 0.4439 on the validation set and 0.4380 on the test set, which ranks second in the final leaderboard.

1. Introduction

With advances in artificial intelligence and deep learning, researchers are increasingly interested in computational methods for human emotional reactions [6]. 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 [38], entertainment [1], and driver safety detection [10].

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 [14, 15, 17–19, 21–24, 39].

Some traditional methods(*e.g.* [25]), through support vector machine-hidden Markov model (HMM) model to obtain expression classification. 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 [34] in natural language processing, ViT [7] 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 expressions having similar intensities. By collecting data on the web and creating datasets to make it more wild to some extent. However, in these tasks, static or dynamic expressions are recognized as limited class, while the connection between emotions is ignored [35], which is not enough to finely reflect the emotional state.

To promote the development of emotional reaction analysis, ABAW2023 is organizing a competition to design a model that can predict the intensity of various emotions, including Adoration, Amusement, Anxiety, Disgust, Empathic-Pain, Fear, and Surprise. This is a multi-output regression task, and our objective is to develop a model capable of predicting the intensity of emotional reaction accurately.

In this paper, we introduce a novel video feature extraction model that utilizes Convolutional Neural Networks (CNN) and spatial attention to focus on global facial information. The model extracts features that contain only spatial information for each frame, while local inter-frame information is fused using a temporal causal convolutional network. A temporal transformer is also applied to establish connections between all video frames to capture dynamic emotion information. To generate features for the acoustic branch, we employ Mel-frequency cepstral coefficients (MFCC). Similar to the video branch, we use a timing modeling approach to obtain global timing relationships. Finally, the features from the video and audio branches are

[*] Corresponding author

fused into the prediction head to estimate emotional response strength vectors.

In summary, our contributions can be summarized as follows:

- We propose a dual-branch model for the ABAW ERI Estimation challenge. It consists of Spatial Encoder with CNN, MFCC and Temporal Encoder.
- We introduce a mechanism based on modality dropout to fuse visual and acoustic features, which is better than simple concatenation operation.
- In the 5th ABAW competition, our method exceeded the baseline by a large margin in the test set and ranked second, indicating the effectiveness of our method.

2. Related Work

In [33] they propose a method called ViPER. Based on the pre-trained Vision Transformer [8], 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 where samples often exhibit consistent high performance, [26] designed a Global Convolutional Attention (GCA) module to weight feature map, which can make the features more distinctive and avoid treating frames with different expression intensities equally in video sequences. 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. Both [31, 35] propose models based on spatial attention mechanism and aggregating different frames through convolutional or linear layers. In [16], the FaceRNET consists of two modules, the representation extractor component for extracting different emotion descriptors, and the another utilizing a RNN and a Mask layer for handling input sequences of different lengths.

For audio-based ERI estimation, [28, 35] use extended Geneva Minimalistic Acoustic Parameter Set(eGeMAPS) and DeepSpectrum of DenseNet121 pretrained on ImageNet as features, but the performances are lower than video-based methods. [28] also uses ResNet18 [9] as the backbone to extract audio features, achieving the best results based on audio methods.

At the MuSe-Reaction sub-challenge in MuSe2022 [5], novel and effective approaches were presented based on the Hume Reaction dataset. ViPER, a multi-modal method using Contrastive Language-Image Pre-Training(CLIP) [32], achieved an average Pearson's Correlation Coefficient of 0.2970 on the test set. In addition, the Former-DFER+MLGCN [35] method utilized Transformers to model multi-modal information and established an inter-dependent matrix for emotional reaction text using GCN [4] to fuse visual features, achieving a score of 0.3375 on the test set.

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 firstly split into images, and we use the dlib [12] toolkit to detect 68 facial landmarks in each frame to crop the face, and resize to 112 $\times$ 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 [40] 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

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 four convolutional layers of ResNet18 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 position 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 computing by spatial encoder as:

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

A shared 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:

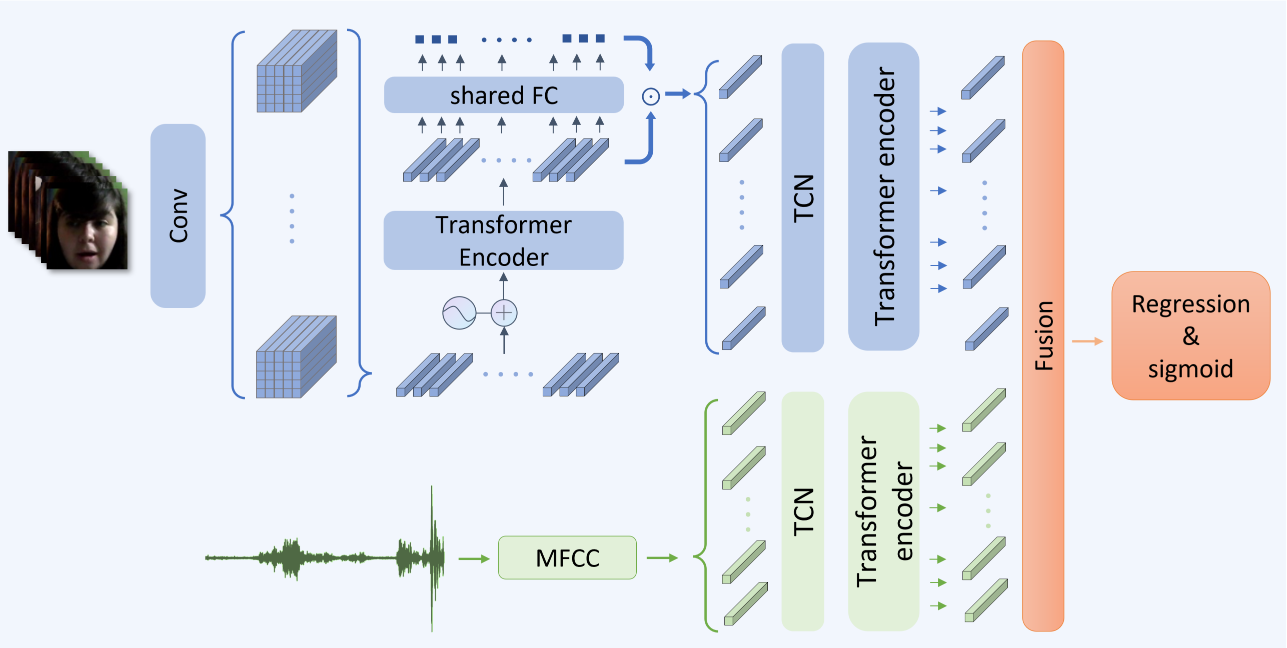

Figure 1. The overall framework of our proposed method. It consists of two branches, uniformly sampled video frames passing through the visual branch (represented by the blue branch in the figure), the acoustic branch (green) based on MFCC features, and finally the multi-modal features fusing through modal dropout and predicting regression values for seven emotional reaction.

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

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

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

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 use the Python toolkit Librosa to extract 128-dimensional features, and then combine 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 from visual or acoustic 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:

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

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 multi-modal feature f^m with modality dropout is:

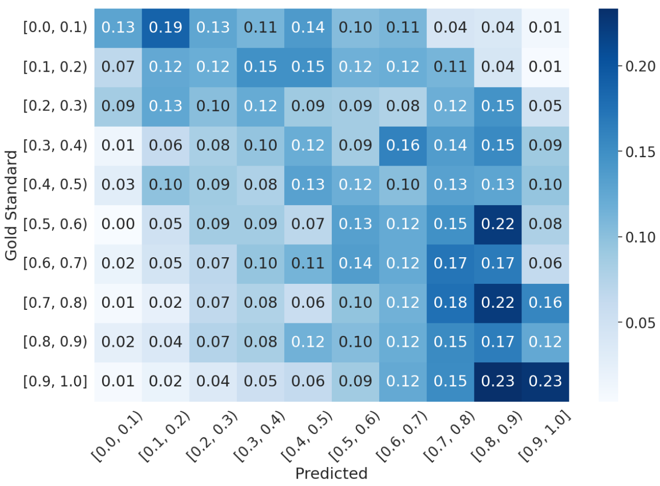

Figure 2. Amusement prediction analysis, $\rho = 0.4620$.

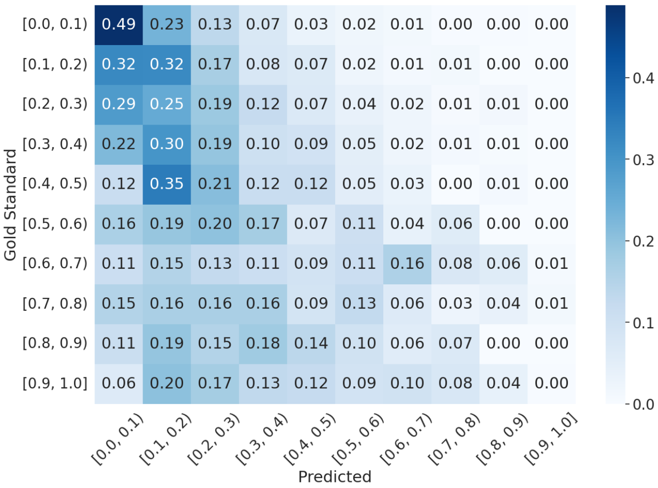

Figure 3. Empathic-pain prediction analysis, $\rho = 0.4266$.

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

$$f^m = W_m f^{av} + b_m \tag{8}$$

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:

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

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:

$$\rho = \sum_{i=1}^{7} \frac{\rho_i}{7} \tag{10}$$

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

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

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 [13] optimizer. All the experiments are implemented on NVIDIA RTX 3090 with PyTorch, with initial learning rate of $1e^{-4}$, 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 [37] trained on AffectNet [30] 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

Method	audio	video	ρ
Baseline(eGeMAPS) [20]	✓	-	0.0583
Baseline(DeepSpectrum) [20]	✓	-	0.1087
ours (MFCC)†	✓	-	0.2972
Baseline(FAU) [20]	-	✓	0.2840
Baseline(VGGFACE2) [20]	-	✓	0.2488
ours (Spatial Transformer)†	-	✓	0.3517
Baseline [20]	✓	✓	0.2382
ours	✓	✓	0.4439

Table 1. Results on validation set. † means that Temporal Encoder is used to obtain the temporal relationship of unimodal features and estimated by the regression layer without fusion.

Method	p_m	p_v	ρ
ours	1.0	-	0.4423
ours	0.9	0.5	0.4439
ours	0.8	0.5	0.4435

Table 2. Results with modality dropout on validation set.

val fold	fold-1	fold-2	fold-3	fold-4	fold-5
ρ	0.5103	0.5186	0.5300	0.5310	0.5168

Table 3. Model trained and tested on different folds (including train and validation set).

dimension of feature in attention is 128. For fusion module, modality dropout and video dropout are set to 0.1 and 0.5, respectively. Experiments are also verified using MindSpore. The code implemented by MindSpore will be open sourced to MindFace [29] (https://github.com/mindspore-lab/mindface).

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, 0.3517, which have a significant improvement compared with baseline.

Ablations We conducted ablation studies on the Hume-Reaction dataset to better understand our proposed model, 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 0.4423, is significantly boosted compared with unimodal,and set p_m =0.9, ρ has a slight improvement. In addition, in order to expand the samples and make full use of the training set and validation set, we mix them together and conduct 5-fold cross-validation experiments. The results are shown in Tab. 3.

Qualitative analysis On the validation set, we calculate the emotional response intensity of different samples, and the average value is 0.4439, which indicates that our estimated value could well reflect the emotions of the sample subjects. For the convenience of analysis, we select amusement and empathic-pain according to the average response intensity of each emotion in the training set, divide them into 10 levels on the basis of intensity, and count the corresponding number of samples. The results are shown in Fig. 4. The former contains a larger proportion of high-intensity samples, whereas the latter mostly consists of low-intensity samples.

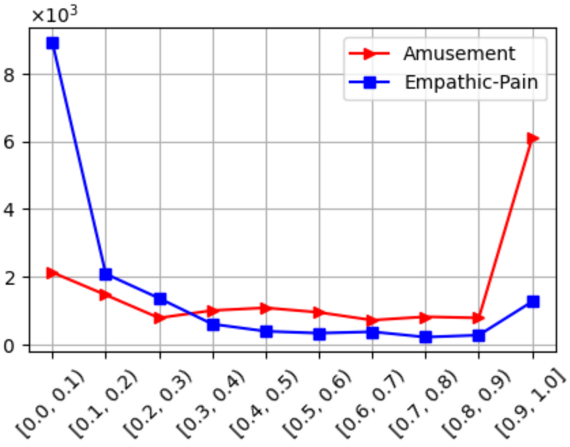

Figure 4. Amusement and empathic-pain sample statistics with quantized intensity on training set.

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 the diagonal, and perform well for high intensities with 0.23, 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 training 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 [11, 26] , the amusement(happy) class is easier to be learned by the model.

Evaluation on the test set We evaluate our proposed method on the test set of the ABAW ERI task and present the performance of each expression class in Tab. 4. The results from all participating teams are shown in Tab. 5. Our approach outperforms the baseline by a significant margin. Specifically, our performance value is 0.0354 lower than first, but it is 0.0334 and 0.0445 higher than the third and fourth ranked teams, respectively.

Emotion	Adoration	Amusement	Anxiety	Disgust	Empathic-Pain	Fear	Surprise
ρ	0.4271	0.4668	0.4501	0.377	0.4091	0.495	0.4405

Table 4. Pearson's Correlations Coefficient (ρ) of 7 emotions on test set.

Team	ρ
HFUT-CVers [27]	0.4734
Netease Fuxi Virtual Human [41]	0.4046
SituTech	0.3935
CASIA-NLPR	0.3865
USTC-AC [36]	0.3730
NISL-2023	0.3667
HFUT-MAC [42]	0.2527
IXLAB	0.1789
USTC-IAT-United($ours$)	0.4380

Table 5. The average Pearson's Correlations Coefficient of different teams on official test set.

5. Conclusion

In this paper, we propose a multimodal based method to improve the performance of emotional reaction intensity estimation, which leverages spatial attention in the visual branch to extract global feature information from face area, while using MFCC to generate features in the acoustic branch. Our method achieves a Pearson's Correlation Coefficient of 0.4380 on the Hume-Reaction test set and wins the runner-up in the CVPR2023 ABAW Challenge 4, which supports the potential of the proposed method in improving the performance of ERI estimation.

Acknowledge

This work is supported by Natural Science Foundation of China (62276242), 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] 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. 4

[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] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5172–5181, 2019. 2

[5] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. 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*, MuSe' 22, page 5–14, New York, NY, USA, 2022. Association for Computing Machinery. 2

[6] 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

[7] 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

[8] 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. 2

[9] 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

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

[11] 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

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

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

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

[15] 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

[16] Dimitrios Kollias, Andreas Psaroudakis, Anastasios Arsenos, and Paraskeui Theofilou. Facernet: a facial expression intensity estimation network. *ArXiv*, abs/2303.00180, 2023. 2

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

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

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

[20] Dimitrios Kollias, Panagiotis Tzirakis, 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. 5

[21] 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

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

[23] 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

[24] 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

[25] 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

[26] 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. 2, 5

[27] Jia Li, Yin Chen, Xuesong Zhang, Jian-Hui Nie, Yang Yu, Zi-Yang Li, M. Wang, and Richang Hong. Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. *ArXiv*, abs/2303.09164, 2023. 6

[28] 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

[29] mindface. mindface:mindface for face recognition and detection. https://github.com/mindspore-lab/mindface/, 2022. 5

[30] 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

[31] 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

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2

[33] 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. 2

[34] 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

[35] 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

[36] Shangfei Wang, Jiaqiang Wu, Feiyi Zheng, Xin Li, Xuewei Li, Su ge Wang, Yi Wu, Yanan Chang, and Xiangyu Miao. Emotional reaction intensity estimation based on multimodal data. *ArXiv*, abs/2303.09167, 2023. 6

[37] 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

[38] 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

[39] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge.

In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1

[40] 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

[41] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition. *ArXiv*, abs/2303.10849, 2023. 6

[42] Ziyang Zhang, Liuwei An, Zishun Cui, Ao xu, Tengteng Dong, Yueqi Jiang, Jingyi Shi, Xin Liu, Xiao Sun, and Meng Wang. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge. *ArXiv*, abs/2303.09158, 2023. 6