Multi-modal Emotion Reaction Intensity Estimation with Temporal Augmentation

Feng Qiu, Bowen Ma, Wei Zhang, Yu Ding*
Virtual Human Group, Netease Fuxi AI Lab
{qiufeng,mabowen01,zhangwei05,dingyu01}@corp.netease.com

Abstract

Emotion reaction intensity (ERI) estimation aims to estimate the emotion intensities of subjects reacting to various video-based stimuli. It plays an important role in human affective behavior analysis. In this paper, we proposed a effective solution for addressing the task of ERI estimation in the fifth Affective Behavior Analysis in the wild (ABAW) competition. Based on multi-modal information, We first extract uni-modal features from images, speeches and texts, respectively and then regress the intensities of 7 emotions. To enhance the model generalization and capture context information, we employ the Temporal Augmentation module to adapt to various video samples and the Temporal SE Block to reweight temporal features adaptively. The extensive experiments conducted on large-scale dataset, Hume-Reaction, demonstrate the effectiveness of our approach. Our method achieves average pearson's correlations coefficient of 0.4160 on the validation set and obtain third place in the ERI Estimation Challenge of ABAW 2023.

1. Introduction

Human affective behavior analysis is a rapidly growing field that has gained great attention in recent years, which involves the analysis and quantify human emotions from various modalities of data. It is capable of endowing machines with emotional intelligence and significantly contributes to human-computer interaction (HCI) [15, 44], such as interactive games and movies, virtual customer service agents [40], and automatic emotion analysis systems.

While deep learning has made remarkable strides in the field of image understanding, predicting human emotions from videos remains a challenging task. This is largely due to the absence of large-scale emotional datasets and the uncertain labels. In order to promote the development of this topic, a sequence of Affective Behavior Analysis in-the-Wild (ABAW) [23, 27, 51] competition is organized. ABAW competition involves collecting high-quality and large-scale emotional datasets [19, 20, 22, 24–26, 28–30], namely Aff-wild1, Aff-wild2 and Hume-Reactions, and exploring more efficient deep learning models and algorithms for emotion analysis. The recent fifth ABAW (ABAW5) competition is split into four challenges, including (1) Valance-Arousal (VA) Estimation Challenge; (2) Expression (Expr) classification Challenge; (3) Action Unit (AU) Detection Challenge and (4) Emotional Reaction Intensity (ERI) Estimation Challenge. This work only focuses on the fourth challenge (ERI) to estimate the emotional intensities of subjects reacting to various video-based stimuli. This task is based on Hume-Reaction dataset and each sample within the dataset has been self-annotated by the subjects themselves for the emotional intensity involving 7 mentioned emotional experiences, including Adoration, Amusement, Anxiety, Disgust, Empathic-Pain, Fear and Surprise.

In this paper, we proposed a multi-modal framework with temporal augmentation module for reaction emotional intensity estimation. Although most information of subject's emotion can be captured from visual modality, the acoustic and text modal characteristics can also be an important way of conveying information and can supplement visual features. Therefore our method consists of three branches to extract uni-modal features from images, audios and texts respectively. These multi-modal features are then integrated for estimating emotional intensity.

For visual branch, we employ vision transformer (ViT) [11], which has achieved great success in image recognition, as our visual encoder to extract spatial features from the facial expression images. Besides, Masked Autoencoder (MAE) [16, 36] has recently been shown to be effective in pre-training Vision Transformers (ViT) for image analysis due to its superior generalization capability. Therefore, we first perform the self-supervised task of reconstructing the original image from only partial image and train the MAE on our private large-scale facial expression dataset. Then, we further finetune the MAE encoder on static expression classification task based on public dataset AffectNet [37]. These spatial features extracted from singe frames in a se-

*Corresponding Author.

quence collectively form the video feature. Due to the varying length of videos and the diversity of video samples, we only focus on 32 frames from each video. These frames are randomly selected by a temporal augmentation module to ensure representative coverage of the video content. After that, we use temporal SE Block [18] and Bidirectional GRUs (BiGRUs) [9] to extract the temporal context information and output final visual features.

For acoustic and text branch, we first employ HuBERT [45] and DeBERTa [17] to extract the acoustic and text features from audios and texts respectively. Then we obtain the global temporal contextual information by the temporal convolutional network and the temporal transformer encoder. Finally, features of images, speeches and texts are fused together by late fusion strategy to estimate the emotional reaction intensity.

To demonstrate the effectiveness of our framework, extensive experiments are conducted on Hume-Reaction dataset. The results show that our model achieves excellent performance and outperforms the baseline method by a significant margin. Our contributions can be summarised as follows:

- This work proposes a multi-modal framework that consists of visual, acoustic and text branches for motion reaction intensity estimation.
- This work introduces a temporal augmentation module in visual branch to improve the generalization capability of the model.

2. Related Works

2.1. Uni-modal features

Visual features. The facial expression in the image plays an important role to analysis human affective behavior. Based on the anatomy, a standardized classification of facial expressions is presented, called Facial Action Coding System (FACS). FACS describes facial expression as combinations of elementary components called Action Units (AU). Thanks to the rapid development of deep learning, model-based features are being utilized by more and more methods [7,21,32] due to its superior performance and powerful generalization ability. These methods train the neural networks on the task of facial expression recognition based on the large-scale dataset. Besides that, another work [36] introduces a robust facial representation model MAE-Face for AU analysis, which can also be used as a visual feature extractor for other tasks of emotion analysis. MAE-Face is trained in self-supervised manner and is capable of learning a high-capacity model from a large-scale face image.

Audio modality. Some traditional acoustic features are widely in human affective analysis, like Mel-frequency Cepstral Coefficients (MFCCs), spectrogram, Linear Predictive Cepstral Coefficients (LPCCs) and Perceptual Linear Prediction (PLP). These features can derive from the audio signals without learning. In recent years, model-based features learned by neural networks become the preferred one since feature extraction is automatic. It is also benefit from the abundance of data and the growth of computation power. Wav2vec 2.0 [4], HuBERT [45] are both large pre-trained language model based on self-supervised approaches for speech representation learning and can be used to extract robust acoustic features from audio signals.

Text modality. Text is also important for human to convey information and communicate with others. In the tasks of Natural Language Processing, words or sentences are usually converted into vectors through embedding models, such as Word2Vec [48] and Glove [39]. The vectors are chosen carefully such that they capture the semantic and syntactic qualities of words. Later, a family of masked-language models, called BERT [10], is published in 2018 by Google. BERT is based on the transformer architecture and pre-trained simultaneously on language modeling and next sentence prediction tasks. After pre-training, BERT learns latent representations of words or sentences in context. Recently, by introducing the disentangled attention mechanism and the enhanced mask decoder, DeBERTa [17] significantly improves the efficiency of model pre-training and the performance of downstream tasks in Natural Language Processing.

2.2. Emotional reaction intensity estimation

Emotional Reaction Intensity (ERI) estimation challenge was first presented in MuSe 2022 [1]. The goal of this task is to estimate 7 emotional intensity of subjects reacting to a wide range of various video-based stimuli, including Adoration, Amusement, Anxiety, Disgust, Empathic-Pain, Fear and Surprise.

The baseline in [8] use GeMAPS [14] and DeepSpectrum [2] to extract audio features, and VGGFACE 2 [7] and Facial Action Units (FAU) [13] to extract visual features. And then a Long Short-Term Memory (LSTM)-RNN is used to aggregate the temporal information. The uni-modal visual features extracted through FAU achieves the best mean Pearson's Correlation Coefficient (ρ) of 0.2801 on the development set of Hume-Reaction. The approaches submitted to MuSe 2022 also involve other visual features, including Resnet-18 [32] trained on AffectNet, ViPER [42] and FaceRNET [21]. And the best results is obtained by Resnet-18 [32], which achieves ρ of 0.3893. Most works show that combining the audio and visual modalities lead to worse results than the uni-modal visual models. However, the TEMMA model [32] who wins the championship in MuSe 2022 improves the metric of ρ by 0.075 through merging visual features and audio features extracted by

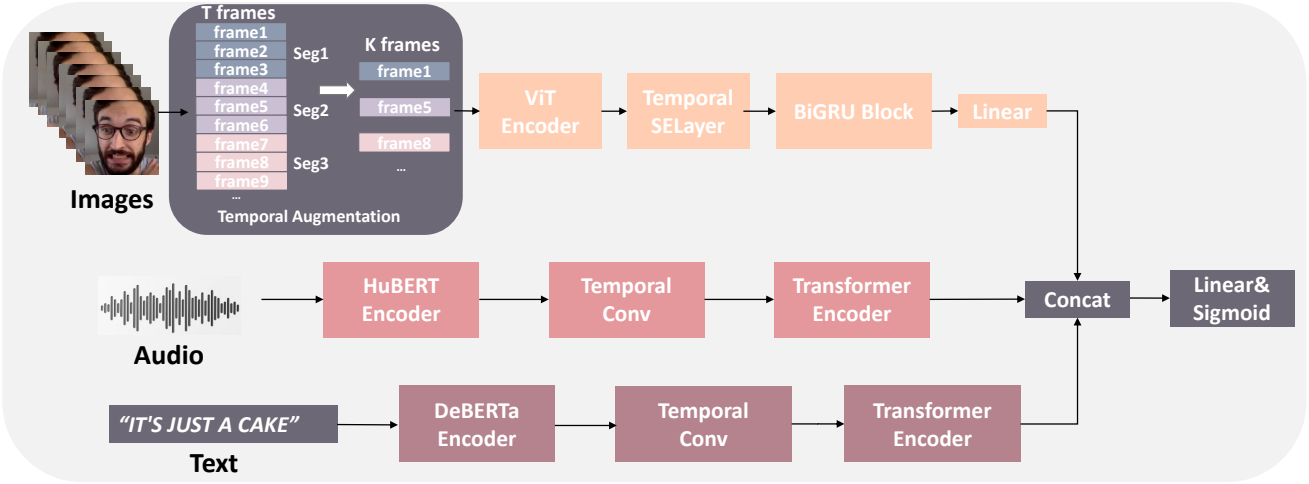

Figure 1. Illustration of our multi-modal framework with for emotional reaction intensity estimation.

ResNet-18 and DeepSpectrum, respectively, in comparison to uni-modal visual model.

3. Method

This section describes our multi-modal framework for estimating the emotional intensities of subjects in a video. As illustrated in Figure 1, our method consists of three branches to extract uni-modal features from images, audios and texts in a video respectively. Then we apply a late fusion strategy for regressing the 7 emotion intensities based on multi-modal information.

3.1. Visual branch

We denote the input face images for visual branch as $\mathcal{X}_v \in \{I_1, I_2, .., I_t, ...I_T\}$, where T is the number of total frames and I_t is the t-th image in a video.

Temporal augmentation. Due to the significant differences in the number of frames T, we adopt a segment-based sampling strategy similar to that used in Temporal Segment Networks (TSN) [47]. This method enables the model to learn temporal feature of the entire video, regardless of its length. Furthermore, it can also be viewed as the process of temporal augmentation by randomly sampling in each segment of the video, which can enhance the temporal generalization ability of the model.

Specifically, for a sequence of images $\mathcal{X}_v$, we divide it into K segments $\{\mathcal{X}_v^1, \mathcal{X}_v^2, ..., \mathcal{X}_v^K\}$ with the same frame number. And then we sample one image from each segment randomly to form new image sequence $\hat{\mathcal{X}}_v \in \{I_1, I_2, .., I_K\}$ with K frames.

ViT encoder. Based on sampling K frames, We use vision transformer (ViT) model as the visual encoder to extract spatial features $f \in \mathbb{R}^{K \times d_{enc}}$ from each frame in $\hat{\mathcal{X}}_v$, where d_{enc} is the feature dimension of the output in ViT encoder. To acquire robust visual features, our ViT encoder is pre-trained with two stages. In the first stage, we employ the masked auto encoding (MAE) method for self-supervised pre-training of ViTs on the large-scale face image datasets, including AffectNet [37], CASIA-WebFace [49], CelebA [33] and IMDB-WIKI [41]. In specific, the model with a ViT encoder and a ViT decoder need to reconstruct raw pixel values by masking a large portion (75%) of the image patches. MAE allows for learning high-capacity models with superior generalization ability. We only preserve the encoder of MAE as our visual encoder.

In the second stage, we finetune the visual encoder combined with two linear layers on the task of expression classification based on AffectNet dataset [37]. This stage aims to further extract the features for the specific task, emotion analysis, which is more related to our final task, emotion intensity estimation. Our model achieves the top-1 accuracy of 69.77% and F1 score of 0.3515 on the test set of AffectNet. After that, we fix the ViT model as our visual encoder to extract spatial features from single frame.

Temporal SE block. To construct informative temporal features, we employ temporal Squeeze-and-Excitation (SE) block similar with SENet [18] along temporal channel to adaptively recalibrate frame-wise feature responses. Temporal SE block first fuses frame-wise spatial information by average global pooling within each image and then produces the per-frame re-weighted features $f \in \mathbb{R}^{K \times d_{se}}$ by a simple self-gating mechanism, where d_{se} is the dimension of the features outputted by temporal SE block.

Subsequently, we employ a BiGRU block to aggregate the feature sequence and extract context information. BiGRU block consists of two BiGRU layers following a layer normalization [3] and a linear layer. To further augment the feature representation capacity, we additionally use a linear layer to produce the final 64-dimensional visual features f_v.

3.2. Audio branch

For audio branch, we use advanced HuBERT [45] as the encoder to extract 768-dimensional acoustic features from audios. HuBERT is a large language model trained by self-supervised approaches for speech representation learning. Different from previous self-supervised model, HuBERT forces the model learn the combined acoustic and language features by applying the prediction loss over the masked regions only. Then, a temporal convolutional network based on causal convolutions [5,38] is used to aggregate local temporal context. Causal convolutions is able to keep the ordering when then model extract temporal informations from the data sequences. After that, we employ a transformer encoder to capture global temporal information and output the final 64-dimensional audio features f_a.

3.3. Text branch

For text branch, we use DeBERTa [17] as our encoder to extract 1024-dimensional features from text. Note that we utilize Speech2Text [46] model to convert the speech into text. Speech2Text is a transformer-based seq2seq model designed for speech recognition and translation proposed by Meta-FAIR. To ensure the accuracy of text extraction, we used the Fuxi Youling Crowdsourcing Platform and Fuxi Agent-Oriented Programmin (AOP) System for data annotation and verification. Because some video clips do not contain any textual information, we use 768-dimensional vectors of zeros to replace them. Finally, similar to the speech branch, we use a temporal convolution block and a transformer encoder to produce 64-dimensional textual features f_t.

3.4. Multi-modal fusion

Signals from different modalities are supplementary and complementary to each other in expressing emotional information. To avoid overfitting to a certain modality during training, we employ a late fusion strategy for the final estimation. In specific, we simply use a concatenation of three features f_v, f_a, f_t extracted by visual, audio and text branch, respectively. To enhance the generalization ability of the model, two fully connected layers following a dropout layer are adopted to estimate the emotional intensities. Because the value range of labels is between 0 and 1, we add a sigmoid activation function to normalize the predicted results to (0,1). The process can be formulated as:

$$\hat{y} = Sigmoid(FC(concat(f_v, f_a, f_t))) \tag{1}$$

where $\hat{y}$ denotes the predicted intensity.

3.5. Loss function

Our experiments perform emotional reaction intensity estimation, which is to regress values from 0-1 for 7 emotions. We take the mean squared error (MSE) loss as follows:

$$\mathcal{L}_{\text{MSE}} = -\frac{1}{N}\sum_{i=0}^{N} \|y_i - \hat{y}_i\|^2 , \tag{2}$$

where N is the mini-batch size and y_i and $\hat{y}_i$ are the ground-truth and predicted intensity for $i\text{-}th$ utterance, respectively.

4. Experiment

4.1. Dataset

We conduct our experiment on the large-scale Hume-Reaction dataset, which consists of more than 70 hours of audios and videos involving 2222 subjects from the United States (1,138) and South Africa (1,084). The age of these subject varies from 18.5 years old to 49.0 years old. Each sample in Hume-Reaction records the reaction of subjects to a wide range of various video-based stimuli. Then these videos are self-annotated by the subjects themselves for the 7 emotional intensities in a range from 1-100, including Adoration, Amusement, Anxiety, Disgust, Empathic-Pain, Fear, Surprise. The total number of videos in Hume-Reaction is 25067. And these videos are split into training, development and test set with 15806, 4657 and 4604 samples.

4.2. Experimental Setting

We extract frames from videos in Hume-Reaction dataset by OpenCV and then crop the face images by the OpenFace [6] detector. All images are resize to 224 $\times$ 224 pixels before fed into the model. We implement our experiment based on PyTorch framework and trained on NVIDIA A100 GPUs. The optimization uses the AdamW optimizer [35] with an initial learning rate of 0.0001. During training process, we adjust the learning rate according to the CosineAnnealing policy [34] with the minimum learning rate of 1e-7 and the number of restart epochs of 5. We set the size of mini-batch to 16. Additionally, The training duration of each model is governed by an early-stopping strategy with the patience of 10 epochs. During training, we employ the Exponential Moving Average (EMA) strategy with a decay rate of 0.999 to enhance the stability of training process. Furthermore, we also apply image augmentation to improve the generalization ability of models, including *RandomRotation*, *RandomHorizontalFlip* and *ColorJitter*.

4.3. Metric

For emotional intensity eatimation, the average Pearson's Correlations Coeffient (ρ) across 7 emotions is used for evaluation, which can be calculated as:

$$\rho = \frac{1}{7}\sum_{c=1}^{7} \frac{\text{Cov}\left(y_c, \hat{y}_c\right)}{\delta_{y_c}\delta_{\hat{y}_c}} \tag{3}$$

Val Set	Adoration	Amusement	Anxiety	Disgust	Empathic-Pain	Fear	Surprise	Average
fold1	0.4142	0.4594	0.4380	0.4147	0.3456	0.4405	0.4508	0.4233
fold2	0.4263	0.4568	0.4299	0.4214	0.3695	0.4509	0.4214	0.4252
fold3	0.4371	0.4919	0.4125	0.4079	0.3674	0.4515	0.4041	0.4246
fold4	0.4024	0.4692	0.4423	0.4046	0.3938	0.4622	0.4280	0.4289
fold5	0.3988	0.4619	0.4235	0.3933	0.3553	0.4343	0.4293	0.4138

Table 1. The results of 5-fold cross-validation.

where $Cov(\cdot)$ represents the covariance, and y_c and $\hat{y}_c$ denotes the ground truth and the predicted results of the c-th emotion intensity, respectively. δ_{y_c} and $\delta_{\hat{y}_c}$ is the standard deviation of y_c and $\hat{y}_c$.

4.4. Experimental results

Comparison on development set. We first train and evaluate our models with multi-modal features on the training and development set of Hume-Reaction, respectively. The results compared with other methods can be seen in Table 2. It is obvious that the visual feature provides the most information about the emotional intensity. Our model used ViT features pre-trained with two stages achieves the best performance and achieves a Pearson's Correlations Coefficient (ρ) of 0.3925 on development set, which is better than the Resnet-18 used in TEMMA by 0.032. And we also try other structures to extract visual features or combine multiple features, like IResNet100 [12] and Inception [43], but these methods all worse than the model with only ViT features. Then we train our model with only audio branch. Previous works use acoustic features of eGeMAPS or DeepSpectrum. Here, we try another advanced acoustic feature HuBERT. The results shows that the model with HuBERT is slightly better than DeepSpectrum by 0.0015.

Although audio provides less information than images, it can be used as a supplement to image features. The combination of visual feature and audio feature leads to a better performance and achieves the ρ of 0.4098.

Some videos in the dataset have textual information while others do not. Nevertheless, the appearance of text in some videos may provide additional information about subjects' reaction emotions. Therefore we also try to extract texts from speeches and then encode the sentence texts by pre-trained large language model. In this work, we use DeBERTa as our text encoder. The experiments show that introducing textual information can indeed further improve performance by 0.052 and achieves 0.4150.

Comparison on test set. In the ERI sub-challenge of ABAW5, we need to predict the labels of the test set of Hume-Reaction dataset. Our method achieves a average ρ of 0.4046 and win the third place in this track.

5-fold cross-validation. To further enhance the generalization ability and test the models' performance, we use 5-fold cross-validation to train multiple models and then ensemble them. We combine the training and development set of Hume-Reaction dataset. And then we split them into 5 folds randomly and train the model on 4 folds of them and take the rest on as the validation set. The results can be found in Table 1

Method	Visual	Audio	Text	ρ
Baseline [8]	FAU	-	-	0.2840
Baseline [8]	VGGFACE 2	-	-	0.2488
FaceRNET [21]	REC	-	-	0.3590
TEMMA [32]	Resnet-18	-	-	0.3893
Ours	ViT	-	-	**0.3925**
Baseline [8]	-	eGeMAPS	-	0.0583
Baseline [8]	-	DeepSpectrum	-	0.1087
TEMMA [32]	-	DeepSpectrum	-	0.1835
Ours	-	HuBERT	-	**0.1850**
Baseline [8]	VGGFACE 2	DeepSpectrum	-	0.2382
TEMMA [32]	Resnet-18	DeepSpectrum	-	0.3968
Ours	ViT	HuBERT	-	**0.4098**
Ours	ViT	HuBERT	DeBERTa	**0.4150**

Table 2. Comparisons of experimental results on the development set of Hume-Reaction dataset.

Team	Rank	p
HFUT-CVers [31]	#1	0.4734
USTC-IAT-United [50]	#2	0.4380
Ours	#3	0.4046
SituTech	#4	0.3935
CASIA-NLPR	#5	0.3865

Table 3. Final competition results on the test set of Hume-Reaction dataset.

Ablation study. As can be seen in Table 4, we also conduct extensive experiments with different settings to further investigate the effectiveness of our used components, including Temporal Augment, Temporal SE, EMA and mixup. We also the select the number of segments used in Temporal Augment by experiments, which is 32 in this work.

5. Conclusion

In this work, we propose a multi-modal framework for emotional intensity estimation. We explore multiple effec-

Method	Temporal Augment	Temporal SE	EMA	Mixup	#frames (K)	ρ
Ours	×	×	×	×	128	0.3942
Ours	×	×	×	×	64	0.3940
Ours	×	×	×	×	32	0.3958
Ours	×	×	×	×	16	0.3850
Ours	✓	×	×	×	32	0.4041
Ours	✓	✓	×	×	32	0.4058
Ours	✓	✓	✓	×	32	0.4115
Ours	✓	✓	✓	✓	32	**0.4150**

Table 4. Comparisons of our methods with different settings on the development set of Hume-Reaction dataset.

tive features for different modalities and incorporate temporal augment module to improve the model's generalization ability. Our method shows superior performance and achieves third place in the ERI sub-challenge of ABAW5.

6. Acknowledgments

The experiments and the data management and storage are supported by NetEase Fuxi Youling platform, based on Fuxi Agent-Oriented Programming (AOP) system that is carefully designed to facilitate task modeling. This work is also supported by the 2022 Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054), and the Key Research and Development Program of Zhejiang Province (No. 2022C01011).

References

[1] Shahin Amiriparian, Lukas Christ, Andreas König, Eva-Maria Meßner, Alan Cowen, Erik Cambria, and Björn W Schuller. Muse 2022 challenge: Multimodal humour, emotional reactions, and stress. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7389–7391, 2022. 2

[2] 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

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[4] 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

[5] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 4

[6] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 4

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[8] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, 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*, pages 5–14, 2022. 2, 5

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[10] 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

[11] 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

[12] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE, 2021. 5

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

[14] 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

[15] Lin Gong and Hongning Wang. *When Sentiment Analysis Meets Social Network: A Holistic User Behavior Modeling in Opinionated Data*, page 1455–1464. Association for Computing Machinery, New York, NY, USA, 2018. 1

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[17] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 2, 4

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2, 3

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

[20] 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

[21] Dimitrios Kollias, Andreas Psaroudakis, Anastasios Arsenos, and Paraskeui Theofilou. Facernet: a facial expression intensity estimation network. *arXiv preprint arXiv:2303.00180*, 2023. 2, 5

[22] 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

[23] 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

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

[25] 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. *arXiv preprint arXiv:2303.01498*, 2023. 1

[26] 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. *arXiv preprint arXiv:2303.01498*, 2023. 1

[27] 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

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

[29] 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

[30] 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

[31] Jia Li, Yin Chen, Xuesong Zhang, Jiantao Nie, Yangchen Yu, Ziqiang Li, Meng Wang, and Richang Hong. Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. *arXiv preprint arXiv:2303.09164*, 2023. 5

[32] Jia Li, Ziyang Zhang, Junjie Lang, Yueqi Jiang, Liuwei An, Peng Zou, Yangyang Xu, Sheng Gao, Jie Lin, Chunxiao Fan, 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*, pages 81–88, 2022. 2, 5

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[36] Bowen Ma, Rudong An, Wei Zhang, Yu Ding, Zeng Zhao, Rongsheng Zhang, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Facial action unit detection and intensity estimation from self-supervised representation. *arXiv preprint arXiv:2210.15878*, 2022. 1, 2

[37] 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. 1, 3

[38] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 4

[39] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[40] AA Rizzo, G Lucas, J Gratch, G Stratou, LP Morency, R Shilling, and S Scherer. Clinical interviewing by a virtual human agent with automatic behavior analysis. In *2016 Proceedings of the international conference on disability, virtual reality and associated technologies*, pages 57–64, 2016. 1

[41] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018. 3

[42] Kai Simon and Georg Lausen. Viper: augmenting automatic information extraction with visual perceptions. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 381–388, 2005. 2

[43] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 5

[44] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. page 1397–1405, New York, NY, USA, 2011. Association for Computing Machinery. 1

[45] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. A

comparison of discrete and soft speech units for improved voice conversion. In *ICASSP*, 2022. 2, 4

[46] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*, 2020. 4

[47] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 3

[48] J Wijffels. word2vec: Distributed representations of words. *R Package Version 0.3*, 4, 2021. 2

[49] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3

[50] Jun Yu, Jichao Zhu, Wangyuan Zhu, Zhongpeng Cai, Guochen Xie, Renda Li, and Gongpeng Zhao. A dual branch network for emotional reaction intensity estimation. *arXiv preprint arXiv:2303.09210*, 2023. 5

[51] 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