

Leveraging TCN and Transformer for effective visual-audio fusion in continuous emotion recognition

Weiwei Zhou*, Jiada Lu*, Zhaolong Xiong, Weifeng Wang
Chinatelecom Cloud

{zhouweiwei, lujiada, xiongzl12, wangweifeng}@chinatelecom.cn

Abstract

Human emotion recognition plays an important role in human-computer interaction. In this paper, we present our approach to the Valence-Arousal (VA) Estimation Challenge, Expression (Expr) Classification Challenge, and Action Unit (AU) Detection Challenge of the 5th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW). Specifically, we propose a novel multi-modal fusion model that leverages Temporal Convolutional Networks (TCN) and Transformer to enhance the performance of continuous emotion recognition. Our model aims to effectively integrate visual and audio information for improved accuracy in recognizing emotions. Our model outperforms the baseline and ranks 3 in the Expression Classification challenge.

1. Introduction

Facial Expression Recognition (FER) can be used in a variety of applications, such as emotion recognition in videos, facial recognition for security purposes, and even in virtual reality applications. Many facial-related tasks have achieved high accuracies, such as face recognition and face attribute recognition. Despite this, the capacity to comprehend the emotions of a person is still not adequate. The subtle distinctions between emotional expressions can lead to ambiguity or uncertainty in the perception of emotions, which makes it harder to assess the emotion of a person. Therefore, the scale of most of the FER datasets are not sufficient to build a robust model.

The appearance of AffWild and AffWild2 dataset and the corresponding challenges [5–12, 12–14, 30] boost the development of affective recognition study. The Aff-Wild2 dataset contains about 600 videos with around 3M frames. The dataset is annotated with three different affect attributes: a) dimensional affect with valence and arousal; b) six basic categorical affect; c) action units of facial mus-

cles. To facilitate the utilization of the Aff-Wild2 dataset, the ABAW5 2023 competition was organized for affective behavior analysis in the wild.

Multi-modal emotion recognition has been proven to be a more effective approach than single-modality emotion recognition, as it can utilize the complementary information between modalities to capture a more complete emotional state while being less susceptible to various noises. This improved recognition ability and generalization ability of the model can lead to more accurate and reliable results.

Considering the fact that visual and audio information contains much emotional information, we propose to use multi-modal features for continuous facial emotion recognition and design a network structure based on TCN and Transformer for feature fusion. Visual and audio features are first fed into their respective TCN modules, then the features are concatenated and fed into the Transformer encoder for learning, and finally, an MLP is used for prediction. Our approach can unify visual and audio features into a temporal model, designing an efficient emotion recognition network with Transformer, thereby improving the evaluation accuracy of Valence-Arousal Estimation, Action Unit Detection, and Expression Classification.

The remaining parts of the paper are presented as follows: Sec 2 describe the study of facial emotion recognition and multi-modal fusion technique. Sec 3 describes our methodology; Sec 4 describes the experiment details and the result; Sec 5 is the conclusion of the paper.

2. Related Work

Many previous studies were focusing on the fusion of visual and audio features for emotion recognition. Juan et al. [19] presented a network that used traditional audio features and visual features extracted with a pre-trained CNN model. Vu et al. [28] built a multi-task model for valence-arousal estimation and facial expressions prediction. The authors applied the distillation knowledge architecture for training and prediction because the dataset does not include labels for all two tasks. One of the approaches using the

*These authors contributed equally to this work.

multi-modal mechanism for facial emotion recognition was proposed by Tzirakis et al. [26], where the visual and audio features are extracted with the CNN module and are concatenated to feed into the LSTM network. Nguyen et al. [18] proposed a network consisting of a two-stream auto-encoder and an LSTM to integrate visual and audio signals for emotion recognition. Zhang et al. [31] proposed a multi-modal multi-feature approach that extracts visual features from 3D-CNN and audio features from a bidirectional recurrent neural network. Srinivas et al. [21] propose a transformer architecture with encoder layers to integrate audio-visual features for expression tracking. Tzirakis et al. [25] use attention-based methods to fuse the visual and audio features.

Previous studies have proposed some useful networks on the Aff-wild2 dataset. Kuhnke et al. [15] combine vision and audio information in the video and construct a two-stream network for emotion recognition and achieving high performance. Yue Jin et al. [4] propose a transformer-based model to merge audio and visual feature.

Temporal Convolutional Network (TCN) was proposed by Colin Lea et al. [16], which hierarchically captures relationships at low-, intermediate-, and high-level time scales. Jin Fan et al. [3] proposed a model with a spatial-temporal attention mechanism to catch dynamic internal correlations with stacked TCN backbones to extract features from different window sizes.

The Transformer mechanism proposed by Vaswani et al. [27] has achieved high performance in many tasks, so many researchers exploit Transformer for affective behavior studies. Zhao et al. [32] proposed a model with spatial and temporal Transformer for facial expression analysis. Jacob et al. [17] proposed a network to learn the relationship between action units with transformer correlation module.

Inspired by the previous work, in this paper we proposed a multi-modal fusion model with TCN and Transformer to enhance the performance of emotion recognition.

3. Methodology

In this section, we describe in detail our proposed method for tackling the three challenging tasks of affective behavior analysis in the wild that are addressed by the 5th ABAW Competition: Valence-Arousal Estimation, EXPR Classification, and AU Detection. We explain how we design our model architecture, data processing, and training strategy for each task and how we leverage multi-modal to improve our performance.

3.1. Preprocessing

We extract the audio stream from the video and preprocess it by converting it to a mono channel with a sample rate of 16,000 Hz. This allows us to reduce the noise and complexity of the audio signal. Some of the video frames

do not contain valid faces, either due to missing or not detected by the face detector. To handle this issue, we replace these frames with the closest frame that has valid face detection. This ensures that we have a consistent sequence of facial images for each video.

3.2. Audio Features

We use Wav2Vec2-emotion [22] to extract the audio features that capture the emotional content of speech.

Wav2Vec2-emotion is a model based on Wav2Vec2-Large-Robust, which is pre-trained on 960 hours of LibriSpeech audio with a sampling rate of 16kHz. The model is then fine-tuned on 284 instances of MSP-Podcast data, which contains emotional speech from different speakers and scenarios. The feature vector dimension is 512, which represents a high-level representation of the acoustic signal.

To align the audio features with the video frames, we resize the features to match the length of each frame using interpolation. This ensures that we have a consistent temporal resolution for both modalities.

3.3. Visual Features

We extract four visual feature vectors using different models that capture various aspects of facial appearance and expression.

The first feature vector is extracted using ArcFace [2] from insightface, which has been pre-trained on the Glint360K dataset [1] for face recognition. This vector encodes the identity and pose of the face with a dimension of 512.

The second feature vector is extracted using EfficientNet-b2 [23, 24], which has been pre-trained on the VGGFace2 dataset [20] for face identification and fine-tuned on the AffectNet8 dataset. This vector captures the facial attributes and expressions with a dimension of 1280.

The third and fourth feature vectors are extracted using a model from DAN [29], pre-trained on MSCeleb, and fine-tuned on RAF-DB and AffectNet8. These vectors represent the global and local features of the face with a dimension of 512 each.

3.4. Split Videos

Videos are first split into segments with a window size w and stride s . Given the segment window w and stride s , a video with n frames would be split into $\lfloor n/s \rfloor + 1$ segments, where the i -th segment contains frames $\{F_{(i-1)*s+1}, \dots, F_{(i-1)*s+w}\}$.

In other words, videos are cut into some overlapping chunks, each with a fixed number of frames. The purpose of doing this is to break down the video into smaller parts that are easier to process and analyze. Each chunk has some

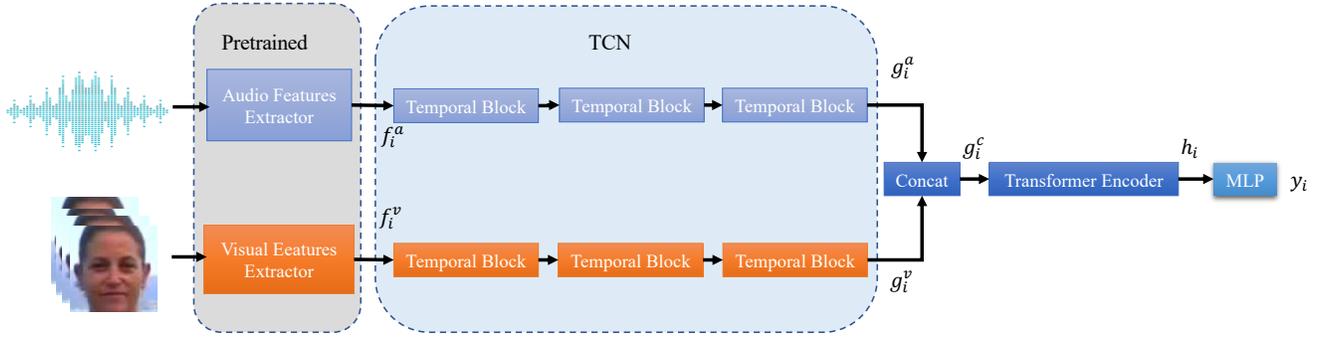


Figure 1. The architecture of our proposed model. The model consists of four components: pre-trained feature extractors for audio and visual features, TCN with three temporal blocks, Transformer encoder, and MLP for final prediction.

degree of overlap with the previous and next ones so that no information in the video is missed.

3.5. Modeling

We denote audio features as f_i^a and visual features as f_i^v corresponding to the i -th segment.

3.5.1 Temporal Convolutional Network

Each feature is fed into a dedicated Temporal Convolutional Network (TCN) for temporal encoding, which can be formulated as follows:

$$g_i^v = \text{TCN}(f_i^v)$$

$$g_i^a = \text{TCN}(f_i^a)$$

where g_i^v denotes visual features, g_i^a denotes audio features. Then, visual features and audio features are concatenated, denoted as g_i^c .

$$g_i^c = [g_i^v, g_i^a]$$

This means that we use a special type of neural network that can capture the temporal patterns and dependencies of the features over time. The TCN takes the input feature vector and applies a series of convolutional layers with different kernel sizes and dilation rates to produce an output feature vector. The output feature vector has the same length as the input feature vector but contains more information about the temporal context. For example, the TCN can learn how the sound and image change over time in each segment of the video. The output feature vectors for both sound and image are then combined together by concatenating them along a dimension. This creates a new feature vector that contains both audio and visual information for each segment of the video.

3.5.2 Temporal Encoder

We utilize a transformer encoder to model the temporal information in the video segment as well, which can be formulated as follows:

$$h_i = \text{TransformerEncoder}(g_i^c).$$

The Transformer encoder only models the context within a single segment, thereby ignoring the dependencies between frames across segments. To account for the context of different frames, overlapping between consecutive segments can be employed, thus enabling the capture of the dependencies between frames across segments, which means $s \leq w$.

We use another type of neural network that can learn the relationships and interactions among the features within each segment. The transformer encoder takes the input feature vector that contains both audio and visual information and applies a series of self-attention layers and feed-forward layers to produce an output feature vector. The output feature vector has more semantic meaning and representation power than the input feature vector. For example, the transformer encoder can learn how different parts of the sound and image relate to each other in each segment of the video. However, the transformer encoder does not consider how different segments of the video are connected or influenced by each other. To solve this problem, we can make some segments overlap with each other so that some frames are shared by two or more segments. This way, we can capture some information about how different segments affect each other. The degree of overlap is controlled by two parameters: s is the length of a segment and w is the sliding window size. If s is smaller than or equal to w , then there will be some overlap between consecutive segments.

3.5.3 Prediction

After the temporal encoder, the features h_i are finally fed into MLP for regression, which can be formulated as fol-

lows:

$$y_i = \text{MLP}(h_i)$$

where y_i are the predictions of i -th segment. For VA challenge, $y_i \in \mathbb{R}^{l \times 2}$. For EXPR challenge, $y_i \in \mathbb{R}^{l \times 8}$. For AU challenge, $y_i \in \mathbb{R}^{l \times 12}$.

The prediction vector contains the values that we want to estimate for each segment. The MLP consists of several layers of neurons that can learn non-linear transformations of the input. The MLP can be trained to minimize the error between the prediction vector and the ground truth vector. The ground truth vector is the actual values that we want to predict for each segment. Depending on what kind of challenge we are solving, we have different types of ground truth vectors and prediction vectors. For the VA challenge, we want to predict two values: valence and arousal. Valence measures how positive or negative an emotion is. Arousal measures how active or passive an emotion is. For the EXPR challenge, we want to predict eight values: one for each basic expression (anger, disgust, fear, happiness, sadness, and surprise) plus neutral and other expressions. For the AU challenge, we want to predict twelve values: one for each action unit (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26).

3.6. Loss Functions

VA challenge: We use the Concordance Correlation Coefficient (CCC) between the predictions and the ground truth labels as the measure, which is defined as in Eq 1. It measures the correlation between two sequences x and y and ranges between -1 and 1, where -1 means perfect anti-correlation, 0 means no correlation, and 1 means perfect correlation. The loss is calculated as Eq 2.

$$CCC(x, y) = \frac{2 * \text{cov}(x, y)}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

$$\text{where } \text{cov}(x, y) = \sum (x - \mu_x) * (y - \mu_y)$$

$$\mathcal{L}_{VA} = 1 - CCC \quad (2)$$

EXPR challenge: We use the cross-entropy loss as the loss function, which is defined as in Eq 3.

$$\mathcal{L}_{EXPR} = -\frac{1}{N} \sum_i \sum_{c=1}^M y_{ic} \log(p_{ic}) \quad (3)$$

where y_{ic} is a binary indicator (0 or 1) if class c is the correct classification for observation i . p_{ic} is the predicted probability of observation i being in class c , M is the number of classes. The multiclass cross entropy loss function measures how well a model predicts the true probabilities of each class for a given observation. It penalizes wrong

predictions by taking the logarithm of the predicted probabilities. The lower the loss, the better the model.

AU challenge: We employ BCEWithLogitsLoss as the loss function, which integrates a sigmoid layer and binary cross-entropy, which is defined as in Eq 4.

$$\mathcal{L}_{AU} = -\frac{1}{N} \sum_i [y_i \cdot \log(\sigma(x_i)) + (1 - y_i) \cdot \log(1 - \sigma(x_i))] \quad (4)$$

where N is the number of samples, y_i is the target label for sample i , x_i is the input logits for sample i , σ is the sigmoid function. The advantage of using BCEWithLogitsLoss over BCELoss with sigmoid is that it can avoid numerical instability and improve performance.

4. Experiments and Results

4.1. Experiments Settings

All models are trained on an Nvidia GeForce GTX 3090 GPU which has 24GB of memory. We use AdamW optimizer and cosine learning rate schedule with the first epoch warmup. The learning rate is $3e - 5$, the weight decay is $1e - 5$, the dropout prob is 0.3, and the batch size is 32.

For VA Challenge, we use Wav2Vec2-emotion, Eff, RAF-DB, and AffectNet8 as the input features.

For EXPR Challenge, we use two types of input features: Eff and AffectNet8 as described above.

For AU Challenge, we use three types of input features: Eff, RAF-DB, and AffectNet8 as described above.

For all three challenges, we split videos using a segment window $w = 300$ and a stride $s = 200$. This means we divide each video into segments of 300 frames with an overlap of 100 frames between consecutive segments. This helps us capture the temporal dynamics of facial expressions and emotions.

4.2. Overall Results

Table 1 displays the experimental results of our proposed method on the validation set of the VA, EXPR, and AU Challenge, where the Concordance Correlation Coefficient (CCC) is utilized as the evaluation metric for both valence and arousal prediction, and F1-score is used to evaluate the result of EXPR and AU challenge. As demonstrated in the table, our proposed method outperforms the baseline significantly. These results show that our proposed approach using TCN and Transformer-based model effectively integrates visual and audio information for improved accuracy in recognizing emotions on this dataset.

Table 2, Table 3, and Table 4 display the overall test results on the three challenges. Notably, Netease Fuxi and SituTech achieved the first and second highest scores in all three challenges, surpassing other teams significantly, indicating their exceptional performance in these challenges.

Experiment	Feature	Valence	Arousal	F1-score
VA	Eff, AffectNet8, RAF-DB, Wav2Vec2-emotion	0.5505	0.6809	-
EXPR	Eff, AffectNet8	-	-	0.4138
AU	Eff, AffectNet8, RAF-DB	-	-	0.5248

Table 1. Performance of our method on the validation dataset of three experiments

Teams	Total Score	CCC-V	CCC-A
SituTech	0.6414	0.6193	0.6634
Netease Fuxi	0.6372	0.6486	0.6258
CBCR	0.5913	0.5526	0.6299
Ours	0.5666	0.5008	0.6325
HFUT-MAC	0.5342	0.5234	0.5451
HSE-NN-SberAI	0.5048	0.4818	0.5279
ACCC	0.4842	0.4622	0.5062
PRL	0.4661	0.5043	0.4279
SCLAB CNU	0.4640	0.4578	0.4703
USTC-AC	0.2783	0.3245	0.2321
baseline	0.201	0.211	0.191

Table 2. The overall test results on VA challenge. The bold fonts indicate the best results.

Teams	F1
Netease Fuxi	0.4121
SituTech	0.4072
Ours	0.3532
HFUT-MAC	0.3337
HSE-NN-SberAI	0.3292
AlphaAff	0.3218
USTC-IAT-United	0.3075
SSSIHL DMACS	0.3047
SCLAB CNU	0.2949
Wall Lab	0.2913
ACCC	0.2846
RT IAI	0.2834
DGU-IPL	0.2278
baseline	0.2050

Table 3. The overall test results on EXPR challenge. The bold fonts indicate the best results.

Our team ranks fourth in the VA challenge, third in the EXPR challenge, and sixth in the AU challenge. Our team’s performance demonstrates our competitive standing in the challenges, with notable achievements in the VA, EXPR, and AU challenges.

Teams	F1
Netease Fuxi	0.5549
SituTech	0.5422
USTC-IAT-United	0.5144
SZFaceU	0.5128
PRL	0.5101
Ours	0.4887
HSE-NN-SberAI	0.4878
USTC-AC	0.4811
HFUT-MAC	0.4752
SCLAB CNU	0.4563
USC IHP	0.4292
ACCC	0.3776
baseline	0.365

Table 4. The overall test results on AU challenge. The bold fonts indicate the best results.

4.3. Ablation Study

In this section, we perform several ablation studies on these three experiments to compare the contribution of different features. From Table 6, it can be seen that almost every feature contributes to the VA prediction task, and the combination of 4 visual features: Eff, ArcFace, AffectNet8, RAF-DB, and the audio features: Wav2Vec2-emotion reach the highest CCC score on VA experiment. Table 7 shows that the use of Eff and AffectNet8 can reach the highest F1-score in the EXPR experiments. Table 8 shows that Eff, AffectNet8, and RAF-DB can reach the highest F1-score in the EXPR and AU experiments. The cross-validation result of the VA, EXPR, and AU experiments are reported in Table 5. Fold 0 is exactly the original data from the ABAW dataset.

5. Conclusion

Our proposed approach utilizes a combination of a Temporal Convolutional Network (TCN) and a Transformer-based model to integrate visual and audio information for improved accuracy in recognizing emotions. The TCN captures relationships at low-, intermediate-, and high-level time scales, while the Transformer mechanism merges audio and visual features. We conducted our experiment on the Aff-Wild2 dataset, which is a widely used benchmark

Task	Evaluation Metric	Partition	Method	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Valence	CCC	Validation	Ours	0.5505	0.6455	0.5889	0.5394	0.5406
			Baseline	0.24	-	-	-	-
		Test	Ours	0.5504	0.4979	0.5008	0.4979	0.4875
			Baseline	0.211	-	-	-	-
Arousal	CCC	Validation	Ours	0.6809	0.6259	0.6539	0.6468	0.6591
			Baseline	0.20	-	-	-	-
		Test	Ours	0.5805	0.5396	0.6325	0.5037	0.5569
			Baseline	0.191	-	-	-	-
EXPR	F1-score	Validation	Ours	0.4138	0.4350	0.3614	0.3959	0.4234
			Baseline	0.23	-	-	-	-
		Test	Ours	0.3406	0.2979	0.3532	0.3293	0.3427
			Baseline	0.2050	-	-	-	-
AU	F1-score	Validation	Ours	0.5248	0.5524	0.5000	0.5060	0.5393
			Baseline	0.39	-	-	-	-
		Test	Ours	0.4735	0.4822	0.4887	0.4818	0.4720
			Baseline	0.365	-	-	-	-

Table 5. Results for the five folds of three tasks

Visual Features	Audio Features	Valence	Arousal
ArcFace	None	0.5013	0.6054
AffectNet8	None	0.5392	0.6629
RAF-DB	None	0.5109	0.6579
Eff	None	0.5208	0.6467
Eff, ArcFace	None	0.5216	0.6519
Eff, ArcFace, AffectNet8	None	0.5345	0.6532
Eff, ArcFace, AffectNet8, RAF-DB	None	0.5429	0.6613
Eff, ArcFace, AffectNet8, RAF-DB	Wav2Vec2-emotion	0.5505	0.6809

Table 6. Ablation study of features on the validation dataset of VA experiment.

Visual Features	Audio Features	F1-score	Visual Features	Audio Features	F1-score
ArcFace	None	0.3512	ArcFace	None	0.4598
AffectNet8	None	0.3937	AffectNet8	None	0.4894
RAF-DB	None	0.3928	RAF-DB	None	0.4915
Eff	None	0.4018	Eff	None	0.5118
Eff, ArcFace	None	0.4015	Eff, ArcFace	None	0.5042
Eff, AffectNet8	None	0.4138	Eff, AffectNet8	None	0.5215
Eff, RAF-DB	None	0.4012	Eff, RAF-DB	None	0.5155
Eff, AffectNet8, ArcFace	None	0.4093	Eff, AffectNet8, ArcFace	None	0.5109
Eff, AffectNet8, RAF-DB	None	0.4087	Eff, AffectNet8, RAF-DB	None	0.5248
Eff, AffectNet8	Wav2Vec2-emotion	0.4028	Eff, AffectNet8, RAF-DB	Wav2Vec2-emotion	0.5134

Table 7. Ablation study of features on the validation dataset of EXPR experiment.

Table 8. Ablation study of features on the validation dataset of AU experiment.

dataset for emotion recognition. Our results show that our method significantly outperforms the baseline. Finally, our

team ranks fourth in the VA challenge, third in the EXPR challenge, and sixth in the AU challenge.

References

- [1] Xiang An, Jiangkang Deng, Jia Guo, Ziyong Feng, Xuhan Zhu, Yang Jing, and Liu Tongliang. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2
- [3] Jin Fan, Ke Zhang, Yipan Huang, Yifei Zhu, and Baiping Chen. Parallel spatio-temporal attention-based tcn for multi-variate time series prediction. *Neural Computing and Applications*, pages 1–10, 2021. 2
- [4] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021. 2
- [5] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022. 1
- [6] 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.
- [7] 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, 2020.
- [8] 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.
- [9] 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.
- [10] 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, 2023.
- [11] 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.
- [12] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1
- [13] 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.
- [14] 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
- [15] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020. 2
- [16] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision—ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*, pages 47–54. Springer, 2016. 2
- [17] Geethu Miriam Jacob and Björn Stenger. Facial action unit detection with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7676–7685, 2021. 2
- [18] Dung Nguyen, Duc Thanh Nguyen, Rui Zeng, Thanh Thi Nguyen, Son N Tran, Thin Nguyen, Sridha Sridharan, and Clinton Fookes. Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. *IEEE Transactions on Multimedia*, 24:1313–1324, 2021. 2
- [19] Juan DS Ortega, Patrick Cardinal, and Alessandro L Koerich. Emotion recognition using fusion of audio and video features. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3847–3852. IEEE, 2019. 1
- [20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 2
- [21] Srinivas Parthasarathy and Shiva Sundaram. Detecting expressions with multimodal transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 636–643. IEEE, 2021. 2
- [22] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021. 2
- [23] Andrey V. Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using efficient-nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2359–2366, June 2022. 2
- [24] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 2
- [25] Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53, 2021. 2
- [26] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309, 2017. 2
- [27] 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. 2
- [28] Manh Tu Vu, Marie Beurton-Aimar, and Serge Marchand. Multitask multi-database emotion recognition. In *Proceed-*

- ings of the IEEE/CVF International Conference on Computer Vision*, pages 3637–3644, 2021. 1
- [29] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021. 2
- [30] 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
- [31] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 632–636. IEEE, 2020. 2
- [32] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 2