Local Region Perception and Relationship Learning Combined with Feature Fusion for Facial Action Unit Detection

Jun Yu[1], Renda Li[1], Zhongpeng Cai[1*], Gongpeng Zhao[1], Guochen Xie[1], Jichao Zhu[1], Wangyuan Zhu[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
{zpcai,rdli,zgp0531,xiegc,jichaozhu,zhuwangyuan}@mail.ustc.edu.cn
{wangcong64, qiuluyu, victor.zhengwei}@huawei.com

Abstract

Human affective behavior analysis plays a vital role in human-computer interaction (HCI) systems. In this paper, we introduce our submission to the CVPR 2023 Competition on Affective Behavior Analysis in-the-wild (ABAW). We propose a single-stage trained AU detection framework. Specifically, in order to effectively extract facial local region features related to AU detection, we use a local region perception module to effectively extract features of different AUs. Meanwhile, we use a graph neural network-based relational learning module to capture the relationship between AUs. In addition, considering the role of the overall feature of the target face on AU detection, we also use the feature fusion module to fuse the feature information extracted by the backbone network and the AU feature information extracted by the relationship learning module. We also adopted some sampling methods, data augmentation techniques and post-processing strategies to further improve the performance of the model. On the official test set, our method ranks third in the AU detection track. This result and subsequent ablation experiments prove the effectiveness of our proposed method.

1. Introduction

The affective behavior analysis in-the-wild (ABAW) [11, 15] is a major targeted characteristic of human-computer interaction (HCI) systems used in real life applications. The target is to create machines and robots that are capable of understanding people's feelings, emotions and behaviors. Thus, being able to interact in a 'human-centered', and effectively serving them as their digital assistants. Human affective behavior analysis plays a significant role in HCI systems.

Different action units (AU) combinations in Facial Action Coding System (FACS) [7] can represent different expressions. The FACS defines a set of facial action units from the perspective of face anatomy, which is used to accurately characterize the facial expression changes. Each facial action unit describes a set of apparent changes generated by facial muscle movements, the combination of which can express arbitrary face expressions. Facial action units (AUs) relate to specific local facial regions based on the FACS, so how to effectively extract the local features associated with the corresponding AU is particularly important. Traditional methods [3,4,8,21,33] use handcrafted methods to represent facial local regions. With the development of deep learning, deep neural networks have been used to improve the accuracy of AU detection, and use face landmarks or divide aligned faces into different patches to locate facial local areas. Obviously, the above two methods fixedly extract the local facial region, which is not accurate enough and cannot adapt to the posture changes of the face. Recent work [27] has emerged that uses a three-stage training strategy to adaptively enable the encoder to extract features that perceive facial local regions. However, this method still needs to use the extra annotations related to the face landmarks, and use multi-task learning to train the model. Because the activation status of AUs are not independent of each other, the activation status of one AU is often correlated with the status of other AUs. Therefore, the relationship between AUs should be taken into consideration when performing AU detection. A recent work [22] uses a graph neural network to obtain the relationship between AUs, and through a two-stage training strategy, obtains multi-dimensional edge features. However, in order to obtain the relationship between AU nodes, using a two-stage training strategy makes

[*] Corresponding author

the training process complicated, therefore, a more effective module is used in our proposed method to extract the features of AU nodes.

In this paper, we propose an single-stage trained method for AU detection. It can more effectively and adaptively consider the facial local regions related to AU detection, so that it only needs to be trained once to extract the desired relationship between AU nodes and output the relation graph of AU. Considering that the overall characteristics of the target face (whether the expression is happy or sad) also plays a certain role in AU detection, the influence of the overall representation of the face will eventually be integrated. Specifically, our method consists of three modules: (i) the **Local region perception (LRP) module** effectively extracts the parts relevant to AU detection in the output of backbone; and the (ii) **AU relationship learning (ARL) module** learns the relation representation between AUs; the (iii) **feature fusion (FF) module** fuses the mutual information between AUs and the overall representation of the target face.

In this work, we focus on Action Unit Detection. Our contributions in this paper are summarized as:

- We use the LRP module to make the model better capture the facial local region related to AU detection without using additional facial landmark annotation and multi-stage training process.
- The ARL module based on the graph neural network is used to learn multiple relationship graphs between AUs.
- In order to better fuse the overall features of the target face and the relationship features between the AUs, We propose a feature fusion module based on self-attention operations, which achieves the best result on the official validation set, but on some other cross-validation sets, simple fixed weight fusion can achieve better result.

2. Related works

In this section, we shortly summarize some works related to the problem of AU detection in the challenge.

2.1. Competition on ABAW

The competition on affective behavior analysis in-the-wild is dedicated to solving the problem of computer analysis of human emotion behavior in natural situations, and thus improving the scenario application capability of human-computer interaction systems.

In the previous challenge [10], many effective approaches offered by some scholars. For example, Zhang *et al.* [40] propose a unified transformer-based multimodal framework for Action Unit detection. Wang *et al.* [30] propose a transfomer [28] based model to detect facial action unit (FAU) in video. They propose a action units correlation module to learn relationships between each action unit labels and refine action unit detection result. We noted that applying a multi-label detection transformer [28] that leverage multi-head attention to learn which part of the face image is the most relevant to predict each AU, which is a very effective solution. These methods have given a great boost to the development of AU detection task.

2.2. AU detection

For the task of AU detection, the limited identity of commonly used datasets and the inability to extract local features relevant to each AU detection are major challenges. Due to the complexity of AU labeling, traditional methods represent local areas of the face by handcrafted has significant limitations. In order to solve the above problems, many recent works focus on using additional facial landmarks annotations to extract important facial local features, and use multi-task learning to improve the performance of AU detection models [2, 9, 26]. In SEV-Net [35], text descriptions of local information are used to generate local region attention maps.

In order to enable the model to adaptively learn features that carry key facial local region information, Tang *et al.* [27] adopts a three-stage training strategy, using face landmarks information in a way similar to multi-task learning, so that the model can pay attention to those important facial local regions. However, this method requires the use of additional face landmarks annotations, and does not pay attention to the possible associations between AUs. Luo *et al.* [22] adopted a two-stage training strategy based on a graph neural network to obtain the associated state relationship between AUs. However, this method only uses a simple fully connected layer to obtain the characteristics of each AU node and doesn't use additional face key landmarks annotation as an auxiliary, so an additional first stage of training is required to enable the fully connected layer to obtain node information well. In short, these methods try to introduce additional tasks or annotations to help the model learn important facial local region features. Thus, The proposed method uses a local region perception (LRP) module to optimize the global features output from the backbone network, making it easier for the fully connected layers to learn the features of each AU node. And we use a more simplified graph neural network than in [22] and a strategy of fusing with the global features output by the backbone network. Next, we will describe our work more specifically.

3. Method

The architecture of the proposed AU detection framework is shown in Fig. 1. The entire framework includes a

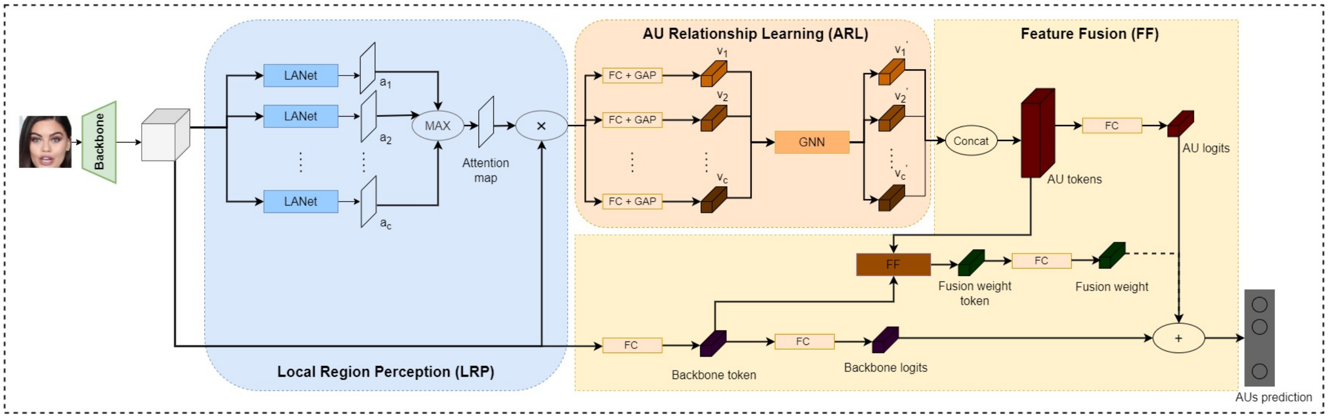

Figure 1. Our proposed framework for AU detection. The backbone uses the input target face to extract the overall features of the image. The local region perception module helps the graph neural network to more effectively extract the relationship between AU nodes, the relationship learning module learns the correlation between different AUs, and the feature fusion module considers the overall information of the target face thereby helping AU detection. In addition, we only draw the feature fusion module based on self-attention, and do not draw the simple fixed weight fusion.

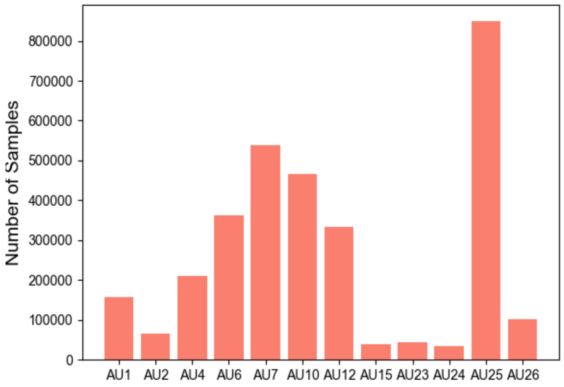

Figure 2. The distribution of the number of AUs in each category in the training set.

feature extractor (IResnet100 [6]) pre-trained on Glint360K [1], a local region perception module, a AU relationship learning module and a feature fusion module. The feature extractor extracts the overall representation of the input image, and then the output of the backbone is input to two branches. One branch effectively extracts the part feature related to AU detection in the output of the backbone through the local attention module, and the optimized output is passed through AU relationship learning module based on graph neural network acquires relational representations between AUs. The other branch is input in parallel to a fully connected layer and a feature fusion module, and the fully connected layer is used to obtain the logits output by the backbone network. We finally chose two types of feature fusion modules, which achieved the best results on different cross-validation sets. One is based on self-attention operation, considering the relationship between the overall features of the target face and the AUs feature graph. This type of feature fusion module obtained the best result on the official verification set; the other is to simply fuse the logits output by the backbone with the logits output by the graph neural network with a fixed weight, and this type of feature fusion module has achieved the best results in some cross-validation sets.

3.1. Local region perception module

In order to help the graph neural network extract mutual information between AU nodes effectively, we propose the Local Region Perception (LRP) module. This module consists of several LANets [31], which effectively notice the local region of the face associated with AU detection. LANet consists of two 1x1 convolutional layers. Assuming that the feature dimension output by the backbone is (c, h, w), after the first 1x1 convolution, the number of channels will be reduced to c/r, where r represents the channel compression rate; after the second 1x1 convolution, the number of channels will be reduced to 1. Several feature maps output by LANets will be stacked together in the channel dimension, and also the maximum value will be calculated in the channel dimension, and finally the attention score map will be obtained through the sigmoid activation function. The attention score map finally output by the local region perception module and the output of the backbone are elementwise multiplied to obtain the required feature map related to AUs.

3.2. AU relationship learning module

The AU Relationship Learning (ARL) module uses the FGG network proposed in [22] to obtain the association information between AUs. The optimized backbone network output obtained from the previous module will first go through the number of categories of fully connected layers and global average pooling to obtain the features $\mathcal{V} = \{v_1, v_2, ..., v_c\}$ of each AU node. Then the AU node features are input into the graph neural network to obtain the relationship information between AUs. The graph neural network generates a specific topology for each target face, and allows multiple relationships between nodes, and the number of relationships is determined by the hyperparameter k. The graph neural network will output the features $\mathcal{V}^{'} = \{v_1^{'}, v_2^{'}, ..., v_c^{'}\}$ of each AU node, which already contains the mutual information between the node and k other AU nodes, and finally stack the features of all AU nodes together through the fully connected layer to get AU logits.

3.3. Feature fusion module

In order to consider the overall information of the target face at the same time, the proposed Feature Fusion (FF) module can add the backbone logits and the AU logits according to a certain fusion weight. In the competition, we used two methods to obtain fusion weights. One method simply uses fixed weights to add AU logits and Backbone logits. Another method based on the self-attention operation is shown in Fig. 3. The AU tokens and Backbone token obtained by passing the output of the backbone network through a fully connected layer are stacked together. Similar to the approach in Vision Transformer [5], we add at the beginning of the sequence a Weight token, which can get the fusion weight through the fully connected layer. Finally, the fused logits pass through the sigmoid activation function to obtain the predicted probabilities of each category.

3.4. Training

Some traning details and tricks are introduced in this part.

3.4.1 Resampling

Since the dataset is composed of continuous video frames, the category changes and image changes of adjacent video frames are small. Therefore, in order to reduce the training time and avoid the model from quickly overfitting to the training set in a few epochs, we uniformly sample one-tenth of the pictures as the training set. In addition, due to the serious category imbalance in the dataset shown in Fig. 2, we use a uniform sampling strategy of one-fifth of the five categories of AU2, AU15, AU23, AU24, and AU26.

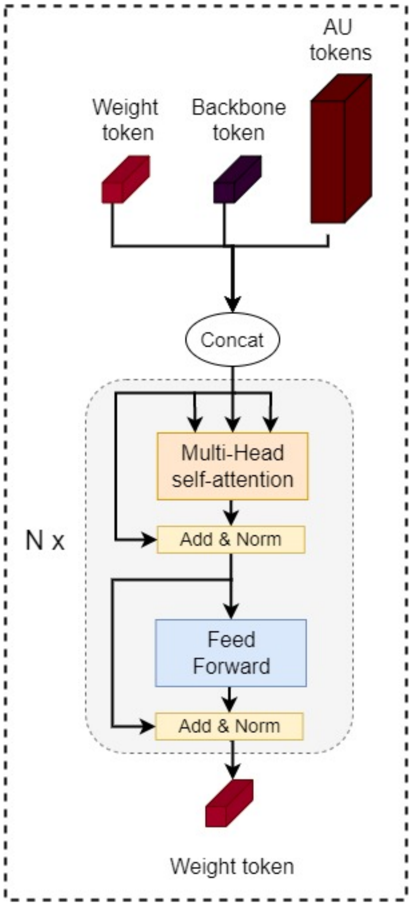

Figure 3. Feature fusion module based on self-attention operation. The weight token is similar to the class token in Vision Transformer [5], which is a learnable vector. The weight token, backbone token and AU tokens concat together as q, k, v are input into the encoder layer to obtain the fusion weight token used to output the fusion weight. In the experiment, the number of encoder layer is 2.

3.4.2 Loss function

AU detection is a multi-label classification task. We use two loss functions, namely binary cross-entropy (bce) loss and circle loss for multi-label classification. Its calculation formula is as follows:

$$\mathcal{L}_{bce} = -\frac{1}{12}\sum_{j=1}^{12}[\, y_j log\hat{y}_j + (1-y_j)log(1-\hat{y}_j)\,] \quad (1)$$

$$\mathcal{L}_{circle} = log(1 + \sum_{i\in\Omega_{neg}} e^{s_i}) + log(1 + \sum_{j\in\Omega_{pos}} e^{-s_j}) \quad (2)$$

$$\Omega_{neg} = \{\, i \mid if\ y_i = 0\}$$

$$\Omega_{pos} = \{\, j \mid if\ y_j = 1\}$$

Val Set	**AU1**	**AU2**	**AU4**	**AU6**	**AU7**	**AU10**	**AU12**	**AU15**	**AU23**	**AU24**	**AU25**	**AU26**	**F1 score**
Official	60.41	54.94	58.46	64.03	75.00	75.61	74.30	34.41	**18.19**	19.26	**84.56**	41.52	55.06
Fold-1	61.95	**57.48**	**51.45**	64.67	**72.42**	**76.07**	74.28	**43.94**	**30.66**	13.37	84.63	**49.28**	57.02
Fold-2	**52.50**	35.26	66.38	**69.15**	**75.97**	74.71	**79.39**	42.85	26.72	**26.50**	**86.34**	34.90	55.97
Fold-3	52.84	**31.50**	**67.52**	64.63	73.31	**73.02**	74.03	36.04	25.34	23.41	85.24	**31.76**	53.05
Fold-4	**63.04**	45.75	53.89	**62.23**	73.90	73.71	**70.63**	**21.43**	29.04	**13.69**	85.70	44.74	53.15

Table 1. The AU F1 scores of models that are trained and tested on different folds (including the official training/validation set). The highest and lowest scores are both indicated in bold.

Probability smoothing **k/F1 score**	**Label smoothing** **k/F1 score**	**Probability/Label smoothing** **k/k/F1 score**
1/55.719	1/55.600	8/1/56.244
2/55.936	2/55.781	8/2/56.249
3/56.059	4/55.913	8/3/56.256
4/56.146	**4/56.006**	8/56.261
5/56.194	5/56.002	8/5/56.265
6/56.204	6/55.981	**8/6/56.270**
7/56.217	7/55.974	8/7/56.261
8/56.243	8/55.963	8/8/56.260
9/56.204	9/55.958	8/9/56.238
10/56.187	10/55.91	8/10/56.227

Table 2. Validation set results for parameter k in probability smoothing and label smoothing in postprocessing. The first column uses only probability smoothing, the second column only uses label smoothing, and the last column has a k value of 8 for fixed probability smoothing and an adjusted k value for label smoothing. The highest score is indicated in bold.

Finally, add bce loss and circle loss together.

$$\mathcal{L}_{total} = \mathcal{L}_{bce} + \mathcal{L}_{circle} \tag{3}$$

3.4.3 Post Process

Considering that the prediction needs to be finally made on consecutive video frames, we use a sliding window based method to smooth the prediction results. We propose two methods for smoothing predictions: one is probabilistic smoothing and the other is label smoothing.

For the category probability predicted by the model for each frame, we will calculate the average predicted probability of the first k frames and the next k frames (including the current frame) of this frame, and use this average as the predicted probability of the frame.

Label smoothing is similar to probabilistic smoothing. For the predicted label of the current frame, we will count the predicted labels of the previous k frames and the next k frames. For a certain category, if it appears in the 2k+1 frame (the predicted label value is 1) more than the number of times it does not appear (the predicted label value is 0), then its predicted label is set to 1, otherwise it is 0.

Obviously for this post-processing method, k is a very important hyperparameter. In our experiments, the best validation set results are obtained when k is 8 for probability smoothing and 6 for label smoothing. Our experiments on the parameter k are shown in Table 2.

IResNet100	Circle loss	Glint360K pre-train	ARL	LRP	FF	F1 score (Official)
✓						50.82
✓	✓					51.78
✓	✓	✓				53.06
✓	✓	✓	✓			53.56
✓	✓	✓	✓	✓		54.19
✓	✓	✓	✓	✓	✓	55.25

Table 3. Ablation experimental results of our proposed method and modules on the official validation set.

4. Experiments

In this part, we will first introduce the dataset used in this competition. We then present our implementation details and result on the validation set. Finally, to demonstrate the effectiveness of the aforementioned modules, we present the results of ablation experiments.

4.1. Datasets

For this Challenge, the Aff-Wild2 dataset will be used. The Aff-wild2 [10, 14, 17–20] was extended from Aff-wild1 [12, 13, 16, 37]. Aff-wild2 expand the number of videos with 567 videos annotated by valence-arousal, 548 videos annotated by 8 expression categories, 547 videos annotated by 12 AUs, and 172,360 images are used that contain annotations of valence-arousal; 6 basic expressions, plus the neutral state, plus the 'other' category; 12 action units. The Action Unit Detection task includes 548 videos annotating the 6 basic expressions, plus the neutral state, plus a category 'other' that denotes expressions/affective states other than the six basic ones. Approximately 2.6 million frames, with 431 participants (265 males and 166 females), have been annotated by seven experts. Therefore, the Aff-Wild2 [10, 14, 17–20] show human spontaneous affective behaviors in the wild, pushing the affective analysis to fit with the real-world scenarios.

4.2. Implementation details

We used IResnet100 pre-trained on Glint360k as the backbone, and the other modules of the model were trained from scratch. The whole training process consists of 15 epochs, the initial learning rate is 0.001, the stochastic gradient descent algorithm is used and the batch size is 256. The learning rate is reduced to one-tenth of the original at the 4th, 6th, and 8th steps. For data enhancement, we only use commonly used weak data enhancements, such as horizontal flipping and color jitter, instead of strong data enhancements such as MixUp [38], because it will conflict with the loss function we use. Experiments are also verified using MindSpore. The code implemented by MindSpore will be open sourced to MindFace [23] (https://github.com/mindspore-lab/mindface).

4.3. Metric

For AU detection chanllenge, we use the average F1 score (F1) of all categories to evaluate the predicted results. See the specific formula below:

$$\mathcal{F}_1^c = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

$$\mathcal{F}_1 = \frac{1}{N}\sum_{c=1}^{N} F_1^c \tag{5}$$

Where N represents the number of classes and c means c-th class.

4.4. Results on validation set

In order to make full use of the official datasets provided. In addition to the official validation set, we also did 4-fold cross-validation. The results of these five partitioned validation sets are shown in Table 1. Some of the best results on the validation set were obtained with a fixed-weight feature fusion module, which we denote in red. Finally, we use the five models to vote on the test set as the final prediction.

4.5. Results on test set

We show the final leaderboard results of the AU detection track on the official test set in Table 4. The F1 score of our method reached 51.44% and ranked third.

Many of the methods of other participating teams are based on multi-modality. Zhang *et al.* [39] used two modalities, audio and vision, and used private data sets for pretraining to achieve excellent generalization of the model; Zhou *et al.* [42] and Zhang *et al.* [41] also used the two modalities of audio and vision use a transformer-based fusion architecture to fuse the features of different modalities. Yin *et al.* [36] used the three modalities of text, audio and vision. The features of the three modes are simply concat together, and the effect is not satisfactory. Wang *et al.* [32] is similar to our method, focusing on the relationship modeling between AUs, but we also consider how to integrate the overall features of the target face. Our method achieves competitive results, which demonstrate the effectiveness of our method.

Teams	F1 score
Netease Fuxi Virtual Human [39]	55.49
SituTech	54.22
USTC-IAT-United (*ours*)	51.44
SZFaceU [34]	51.28
PRL [29]	51.01
CtyunAI [42]	48.87
HSE-NN-SberAI [25]	48.78
USTC-AC [32]	48.11
HFUT-MAC [41]	47.52
SCLAB CNU [24]	45.63
USC IHP [36]	42.92
ACCC	37.76

Table 4. The leaderboard of the AU detection track on the official test set.

4.6. Ablation study

In order to prove the effectiveness of the proposed method and module, we have done detailed ablation experiments shown in table 3, and the results of all experiments are carried out on the official validation set. Our baseline is obtained by adding a classification head after IResNet100. When using Circle loss, the F1 score increased by 0.96%; when IResNet100 was loaded with pre-trained weights on glint360k dataset, the F1 score increased by 1.28%; after using the ARL module, the model learned the relationship between AU, the F1 score has increased by 0.50%; after adding the LRP module, the model can better extract key local area features, so the score has increased by 0.63%; after finally integrating the overall features of the target face output by the backbone network, the F1 score has increased by 1.06%.

5. Conclusion

In this paper, we introduce the proposed single-stage trained AU detection framework for the AU detection challenge in the ABAW5 competition. Aiming at the extraction of local region features of AU detection tasks and the relationship learning problem between AUs, we adopted a Local Region Perception (LRP) module based on LANet and a AU Relationship Learning (ARL) module based on graph neural network. In addition, in order to integrate the overall features of the target face and the relationship features between the AUs, we employ a Feature Fusion (FF) module based on self-attention operations. The results of ablation experiments show that the modules we use can improve the performance of AU detection model, and in the ABAW5 competition, we won the third place in the AU detection track, and the results of the competition proved the effectiveness of our proposed method.

6. Acknowledgments

This work was 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] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021. 3

[2] Carlos Fabian Benitez-Quiroz, Yan Wang, Aleix M Martinez, et al. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, pages 3990–3999, 2017. 2

[3] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3515–3522, 2013. 1

[4] Xiaoyu Ding, Wen-Sheng Chu, Fernando De la Torre, Jeffery F Cohn, and Qiao Wang. Facial action unit event detection by cascade of tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 2400–2407, 2013. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[6] 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. 3

[7] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1

[8] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE international conference on computer vision*, pages 3792–3800, 2015. 1

[9] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

[10] 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. 2, 5

[11] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 1

[12] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020. 5

[13] 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. 5

[14] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 5

[15] 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

[16] 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. 5

[17] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 5

[18] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 5

[19] 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. 5

[20] 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. 5

[21] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013. 1

[22] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-

based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 1, 2, 4

[23] mindface. mindface:mindface for face recognition and detection. https://github.com/mindspore-lab/mindface/, 2022. 6

[24] Dang-Khanh Nguyen, Ngoc-Huynh Ho, Sudarshan Pant, and Hyung-Jeong Yang. A transformer-based approach to video frame-level prediction in affective behaviour analysis in-the-wild. *arXiv preprint arXiv:2303.09293*, 2023. 6

[25] Andrey V Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition. *arXiv preprint arXiv:2303.09162*, 2023. 6

[26] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021. 2

[27] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 1, 2

[28] 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

[29] Tu Vu, Van Thong Huynh, and Soo Hyung Kim. Vision transformer for action units detection. *arXiv preprint arXiv:2303.09917*, 2023. 6

[30] Lingfeng Wang, Shisen Wang, and Jin Qi. Multi-modal multi-label facial action unit detection with transformer. *arXiv preprint arXiv:2203.13301*, 2022. 2

[31] Qiangchang Wang and Guodong Guo. Ls-cnn: Characterizing local patches at multiple scales for face recognition. *IEEE Transactions on Information Forensics and Security*, 15:1640–1653, 2019. 3

[32] Shangfei Wang, Yanan Chang, Yi Wu, Xiangyu Miao, Jiaqiang Wu, Zhouan Zhu, Jiahe Wang, and Yufei Xiao. Facial affective behavior analysis method for 5th abaw competition. *arXiv preprint arXiv:2303.09145*, 2023. 6

[33] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013. 1

[34] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Weicheng Xie, Linlin Shen, et al. Spatio-temporal au relational graph representation learning for facial action units detection. *arXiv preprint arXiv:2303.10644*, 2023. 6

[35] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2021. 2

[36] Yufeng Yin, Minh Tran, Di Chang, Xinrui Wang, and Mohammad Soleymani. Multi-modal facial action unit detection with large pre-trained models for the 5th competition on affective behavior analysis in-the-wild. *arXiv preprint arXiv:2303.10590*, 2023. 6

[37] 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. 5

[38] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

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

[40] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437, 2022. 2

[41] Ziyang Zhang, Liuwei An, Zishun Cui, Tengteng Dong, et al. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge. *arXiv preprint arXiv:2303.09158*, 2023. 6

[42] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer. *arXiv preprint arXiv:2303.08356*, 2023. 6