Coarse-to-Fine Cascaded Networks with Smooth Predicting for Video Facial Expression Recognition

Fanglei Xue[1,2]*, Zichang Tan[3,4], Yu Zhu[3,4], Zhongsong Ma[1,2], Guodong Guo[3,4†]

[1]University of Chinese Academy of Sciences, Beijing, China

[2]Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China

[3]Institute of Deep Learning, Baidu Research, Beijing, China

[4]National Engineering Laboratory for Deep Learning Technology and Application, Beijing, China

xuefanglei19@mails.ucas.ac.cn mazhongsong@csu.ac.cn {tanzichang, zhuyu05, guoguodong01}@baidu.com

Abstract

Facial expression recognition plays an important role in human-computer interaction. In this paper, we propose the Coarse-to-Fine Cascaded network with Smooth Predicting (CFC-SP) to improve the performance of facial expression recognition. CFC-SP contains two core components, namely Coarse-to-Fine Cascaded networks (CFC) and Smooth Predicting (SP). For CFC, it first groups several similar emotions to form a rough category, and then employs a network to conduct a coarse but accurate classification. Later, an additional network for these grouped emotions is further used to obtain fine-grained predictions. For SP, it improves the recognition capability of the model by capturing both universal and unique expression features. To be specific, the universal features denote the general characteristic of facial emotions within a period and the unique features denote the specific characteristic at this moment. Experiments on Aff-Wild2 show the effectiveness of the proposed CFSP. We achieved 3rd place in the Expression Classification Challenge of the 3rd Competition on Affective Behavior Analysis in-the-wild. The code will be released at https://github.com/BR-IDL/PaddleViT.

1. Introduction

In the past years, facial expression recognition has attracted increasing attention in the field of computer vision and psychology science [6, 12, 25, 28, 35, 36, 38, 40]. Facial expression can be regarded as the external signals to reflect people's psychological activities and intentions. Therefore, facial expression recognition is particularly important in

human-computer interaction. Although this topic has been studied for many years, it is still difficult to achieve accurate emotional analysis nowadays.

In order to promote the development of facial expression recognition, Kollias et al. [14–21, 37] organize the competition on Affective Behavior Analysis in-the-wild (ABAW). The competition has been successfully held two times, in conjunction with IEEE FG 2020 and ICCV 2021. The third competition (ABAW3) will be held in conjunction with CVPR 2022. Those competitions provide several large-scale in-the-wild benchmarks for affective behavior analysis, e.g., Aff-Wild2 [19]. Aff-Wild2 contains 558 videos with around 2.8 million frames with rich annotations. Besides the common annotations of facial expressions and Action Units (AUs), it also provides the annotations of valence and arousal to study continuous emotions.

The goal of our work is to study facial expression recognition based on Aff-Wild2. One main challenge in facial expression recognition is the problem of label ambiguity, which may be caused by two aspects: one is the ambiguity of the expression itself, where some expressions are similar and difficult to distinguish. The other is the labeling ambiguity caused by different people having different understandings of different facial expressions, resulting in inconsistent labeling. For example, 'Sadness' is very similar to 'Disgust', and it is hard to distinguish them. To address this problem, we propose the coarse-to-fine cascaded networks to obtain more reliable predictions in a coarse-to-fine manner. Specifically, we first group several similar emotions as a coarse category (including 'Anger', 'Disgust', 'Fear' and 'Sadness'), and this coarse category will be combined with other emotions to train a coarse model at the first stage. In the second stage, an additional network is further employed to distinguish those grouped emotions and then obtain fine predictions. Moreover, inspired by the previous work [32],
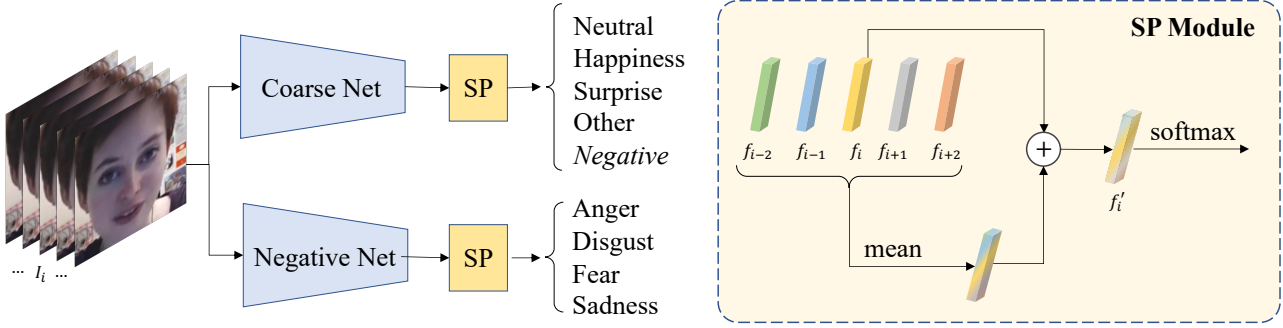

Figure 1. The architecture of our model. Given the input frames $I_{i-2}, ... I_i, ..., I_{i+2}$, we first extract the features of these frames by the Coarse Net and get the coarse results. If the predicted coarse result is *Negative*, the Negative Net is utilized to predict the negative classes. The smooth predicting (SP) module is designed to take multiple frame features (denote as f_i) as input. And every feature is calculated by adding themselves to the mean values of their adjacent features. After this, the predicted result will be more smooth and stable.

the feature embedding of facial expression can be divided into two parts, namely universal and unique parts. The universal part describes the general characteristic of facial emotions, and the unique part describes the specific features of the expressive component. The unique part can be denoted by the features of each specific image. For the universal part, it is captured by using a proposed Smooth Predicting (SP), which collects the features of nearby iterations and outputs the averaging features. The final feature used for prediction is the sum of the universal and unique parts. Furthermore, we train multiple models based on various architectures (e.g., ResNet [10], Swin-Transformer [27]) and improve the performance by using an ensemble of them.

The main contributions of our work can be summarized as follows:

- We propose the Coarse-to-Fine Cascaded networks (CFC) to address the label ambiguity problem in facial expression recognition.

- We propose a Smooth Predicting (SP) to capture both universal and unique features, which can effectively improve the performance of facial expression recognition.

2. Related Works

In this section, we briefly review the related works in facial expression recognition and affective behavior analysis.

Facial Expression Recognition. Facial expression recognition [6, 12, 25, 28, 35, 36, 38, 40], which is a classic research topic in computer vision and psychology science, has been studied for many years. Most of the early methods were based on hand-crafted features [7, 42, 43]. Recently, deep learning has greatly advanced the development of facial expression recognition [6, 12, 31, 35, 36, 38, 40]. For example, Weng et al. [35] propose a multi-branch network to capture both global and local features for facial expression recognition. Xue et al. [36] propose a hybrid architecture with combined CNN and transformer, which achieves state-of-the-art performance on multiple benchmarks. Some researchers [24, 26, 34] propose to extract discriminative features by attention mechanism, which is proven to be robust for occlusions.

Affective Behavior Analysis. This subsection mainly reviews the related works in previous competitions of ABAW. Kuhnke et al. [22] propose a two-stream network to utilize both vision and audio information for affective behavior analysis. Considering the problem of incomplete labels, Deng et al. [1] propose to learn from unlabeled data by a teacher-student scheme. Moreover, Zhang et al. [40] a Prior Aided Streaming Network (PASN) for multi-task expression recognition, including Categorical Emotions (CE), Action Units (AU), and Valence Arousal (VA). In PASN, three tasks are placed in a serial manner (AU→CE→VA), where the basic up-streaming task would aid the recognition of down-streaming tasks. Jin et al. [12] propose a multimodal multi-task method for facial action unit and expression Recognition.

3. Method

3.1. Overview

The overall architecture of our proposed method is illustrated in Fig. 1. It mainly consists of two parts: the Coarse-to-Fine Cascaded networks (CFC) on the left and the Smooth Predicting (SP) module on the right. In the following, we will introduce these two modules in detail.
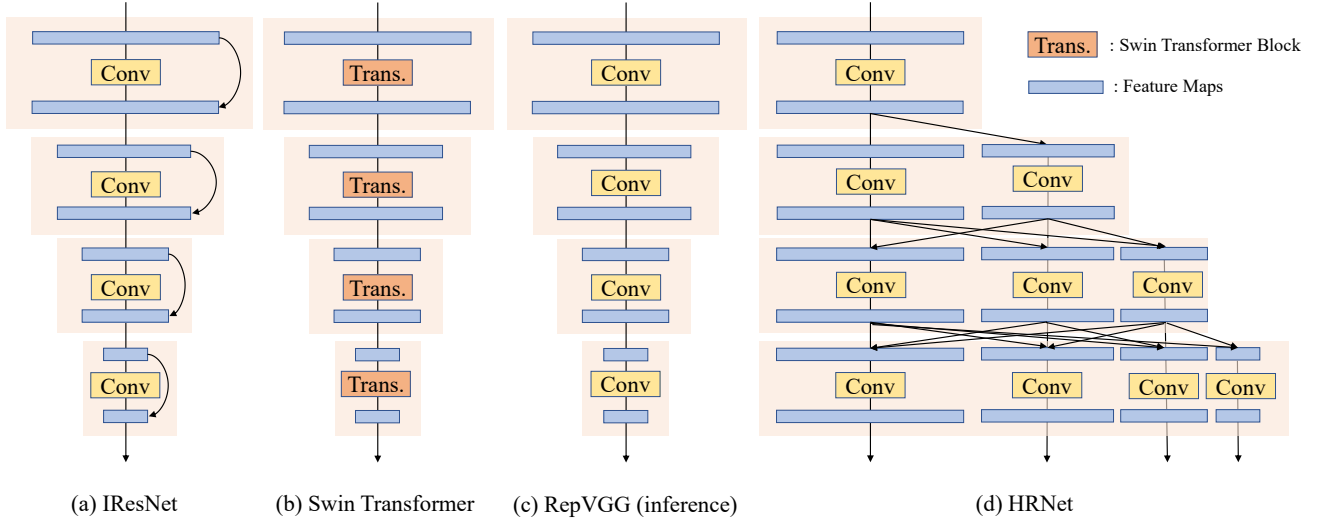
Figure 2. Sketch of different backbone architectures. They have various feature extract methods (convolution or self-attention), various lateral connections (with or without skip connections), and various resolution streams (gradually decrease or keep the same). We believe different architecture could learn different features. Thus, we ensemble these four backbones in the coarse net to generate robust prediction results and prevent over-fitting.

3.2. Coarse-to-Fine Cascaded Networks

Facial expressions are generated by muscle movement in faces. Prior studies [4] split expressions into seven basic categories to represent the common emotions shared by all human beings. The EXPR task in ABAW3 aims to categorize facial images into these seven expressions plus an "other" category. But these seven categories are not equally distributed, some expressions are very similar to others, especially negative expressions. Specifically, four negative expressions (anger, disgust, fear, and sadness) have very similar muscle movements in the same facial areas [5]. Compared with other expressions, like happiness and surprise, these expressions are more difficult to predict correctly. On the other hand, most in-the-wild facial expression datasets are collected from the Internet, where most people share their happy times. Negative emotions are hard to obtain in real scenes [23], making existing FER datasets imbalance. To address these two issues, we adopted the coarse-fine strategy in [23].

To be specific, we utilize two stages to predict coarse and negative labels, separately. Following [23], we combine four negative expressions (anger, disgust, fear, and sadness) to one category (negative) in the first stage, resulting in five classes in the coarse stage. If the coarse net predicts the facial image into the negative class, the image will be fed into the negative net to predict the negative expression. Otherwise, the coarse label will be returned directly. In this manner, the low-frequency negative expression samples are combined in the negative, making the coarse dataset more balance. And the negative net only aims to distinguish negative expressions, which share a similar difficulty.

In order to further enhance the robustness of our framework, we adopt the model ensemble strategy to our coarse net. Specifically, IResNet-152 [2], Swin-S [27], RepVGG [3] and HRNet [33] are adopted as backbone networks. As illustrated in Fig. 3, there are significant differences among these architectures. IResNet (a), Swin Transformer (b), and RepVGG (c) utilize pooling or path merging layers to decrease the spatial resolution of feature maps to reduce the computation cost and enlarge the receptive field, while HRNet (d) remains the high-resolution representations and exchange features across resolutions to obtain rich semantic features. (b) utilizes the self-attention mechanism with shifted windows to explore relationships with different areas while the other three models adopt the traditional convolution method. (a) utilizes residual connection within stages while (b) and (c) do not (c) has skip-connection in the training stage. (d) adopted more connections among different resolutions to extract rich semantic features. These different architectural designs could help different models learn different features, and prevent the whole framework overfit to some noisy features.

3.3. Smooth Predicting

Benefit from the timeline in videos, we can find the process of muscle movement, *e.g.* the process from neural to happiness. So the expression change in videos should be very smooth. But current 2D FER methods only consider
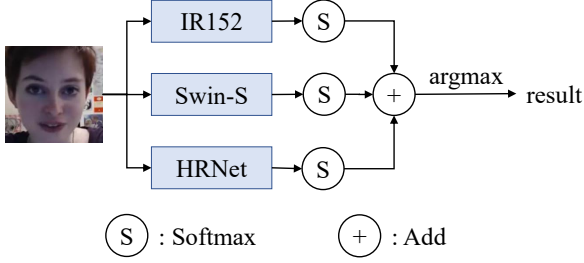
Figure 3. Illustration of our ensemble process. Feature logits are extracted from each backbone network separately and produce the per-class confidential scores by the softmax function. The scores are then summed up and generate the final result with the argmax function.

the spatial information in facial images and ignore the relative information among frames. They predict every frame independently, which may lead to different results in the same expression clip. Videos in the ABAW3 dataset have a high frame rate (30 fps), thus it does not make sense for expressions changing very frequently in adjacent frames.

To address this issue, we propose a smooth predicting module to capture both universal and unique features of adjacent frames. As illustrated in the right part of Fig. 1, the SP module takes a series of feature logits as input and enhance every feature by adding the average information among a specific window (5 in Fig. 1 as an example) to the original feature.

Mathematically, given n adjacent frame images, $I_0, I_1, ...I_{i-1}, I_i, I_{i+1}, ...I_{n-1}$, and the windows size ($w$), we first utilize the network to extract the features from every images independently. After this, n image features will be obtained, named $f_0, ...f_i, ...f_{n-1}$. Then, our proposed SP module will update every frame feature by:

$$f_i^{'} = f_i + \frac{\sum_{j=i-w/2}^{i+w/2} f_j}{w} \quad (1)$$

where the $f_i^{'}$ denotes the updated feature. The updated feature consists of two-part: the unique current frame feature and the universal feature among the given window size. We find that this will improve the smoothness of model outputs as well as the performance.

4. Experiments

In this section, we will first introduce some training details and then give some experiment results based on the official validation set of the Aff-Wild2 [19].

Methods	F1 (%)
Baseline	23
Ours baseline	40
Ours baseline + C-F	42
Ours baseline + SP	43
Ours	**46**

Table 1. The ablation studies to our method with and without C-F strategy and SP module on the validation set of Aff-Wild2.

4.1. Details

We only used the Aff-Wild2 [18] to train our models. All videos are processed into frame images by FFmpeg, and the MTCNN [39] is utilized to detect and align facial images. All the facial images are cropped into 112×112 except for Swin Transformer, which is 224 × 224. To alleviate the over-fitting problem, we utilized several data augmentation methods following [36]: random rotate and crop, random flip, colorjit, blur, and random erasing [44]. The over-sample method proposed in [9] is also adopted to reduce the influence of the class imbalance problem.

An ensemble of IR152 [2], Swin-S [27], and HRNet [33] is adopted as our coarse net, and the RepVGG [3] is adopted as the negative nets. All these backbone nets are pre-trained on the Ms-Celeb-1M [8].

The cross-entropy loss is applied for training. And the f1-score (%) on the validation set is reported by default. The detail of hyperparameters could be found on the corresponding config file in our released codes. All of the experiments are trained on one NVIDIA V100 graphic card.

4.2. Results

Tab. 1 shows the ablation studies of our method on the validation set of Aff-Wild2. By using a more powerful backbone and training the whole model in an end-to-end manner, the baseline performance increased from 23 to 40. The C-F strategy brings another 2 performance increment, indicating that the C-F strategy could reduce the influence of the class imbalance problem and distinguish negative expressions efficiently. The performance further increased to 46 by applying the SP module to help the model capture both universal and unique features along the timeline.

To illustrate the impact of different backbone architectures, we show the coarse and negative results with different backbones in Fig. 3. As we can see, Swin-S achieves the best performance (51.52) in the coarse task and a comparable result (46.70) in the negative task, indicating the great impact achieved by the self-attention mechanism. The self-attention allows the network to use different patterns to extract different content even if they are in the same positions which are very useful to classify highly varied samples. The

Coarse				Negative		
Swin-S	IR152	HRNet	RepVGG	Swin-S	RepVGG	F1 (%)
✓					✓	43.78
	✓				✓	42.74
		✓			✓	43.78
			✓		✓	40.38
✓	✓				✓	44.25
✓	✓			✓		43.28
✓	✓			✓	✓	43.17
✓		✓			✓	43.28
✓	✓	✓			✓	**46.15**
✓	✓	✓		✓	✓	44.62

Table 2. Results of different ensemble strategies on the validation set of the Aff-Wild2. The Swin-S, IR152, and HRNet ensembled as the coarse net, and the RepVGG as the negative net performs the best.

Backbone	Coarse	Negative
IR152 [2]	50.82	44.99
Swin-S [27]	**51.52**	46.70
HRNet [33]	50.67	45.60
RepVGG [3]	48.40	**49.82**

Table 3. Performance on coarse and negative tasks with different backbones on the validation set of Aff-Wild2.

w	F1 (%)
0	42.28
32	43.29
64	43.47
128	43.69
256	**43.78**
512	43.68

Table 4. Evaluation of different windows size w in our method on the validation set of the Aff-Wild2.

Rank	Teams	F1 (%)
1	Netease Fuxi Virtual Human [41]	**35.87**
2	IXLAB [11]	33.77
3	AlphaAff (Ours)	32.17
4	HSE-NN [30]	30.25
5	PRL [29]	28.6
6	dgu [13]	27.2
7	USTC-NELSLIP	21.91
	baseline [14]	20.50

Table 5. Results of the test set of the Expression Classification Challenge.

RepVGG achieves the best performance (49.82) in the negative task, but its performance on the coarse task is the worst (48.40). The performances on different architectures various on these two tasks, indicating they could obtain different features. Therefore, it is intuitive to ensemble them to produce robust results.

Performance with different backbone ensemble strategies is illustrated in Fig. 2. The IR152 and HRNet perform badly in the negative task, so they are not considered in this experiment. The RepVGG performs worse in the coarse task, resulting in bad performance (40.38) in the overall performance. To our surprise, the Swin-S performs badly when ensemble in the negative net. When ensemble with Swin-S, the overall performance will decrease from 0.11 to 1.53. So the Swin-S, IR152, and HRNet are ensembled as the final coarse net, and the RepVGG is utilized solely as the negative net.

We also conduct experiments with different w in the SP module to evaluate the impact of windows size. As is illustrated in Tab. 4, as the window size becomes larger, the model could capture more universal features, and benefit from this, it could generate smooth and stable predictions. The best result (43.78) is obtained when the size of the window is 256, which is about 8 seconds in time. When the w becomes larger than 256, the performance becomes worse. This is because the expression may change in such a long period.

The final results on the test set are illustrated in Tab. 5, our method achieved 3rd place on the leader board. Different from other methods, which utilize other modalities (*e.g.* voice, spoken words), we only adopt frame images as our input. We believe with the help of modality fusion modules proposed in other methods, the performance could increase further.

5. Conclusion

In this paper, we have proposed the Coarse-to-Fine Cascaded Networks with Smooth Predicting (CFC-SP) for video facial expression recognition. To address the problem of label ambiguity, we first group several similar emotions to form a coarse category. Then, the cascaded networks were employed to first conduct a coarse classification and then perform a fine classification. Moreover, we proposed a Smooth Predicting (SP) to further improve the performance. Experiments on Aff-Wild2 have verified the effectiveness of the proposed method.

References

[1] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020. 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. 3, 4, 5

[3] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. RepVGG: Making VGG-style ConvNets Great Again. 3, 4, 5

[4] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. 3

[5] Paul Ekman and Wallace V. Friesen. Facial action coding system: a technique for the measurement of facial movement. 1978. 3

[6] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2402–2411, 2021. 1, 2

[7] Guodong Guo and Charles R Dyer. Learning from examples in the small sample case: face expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):477–488, 2005. 2

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 4

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:5351–5359, aug 2019. 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[11] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Yuchul Jung, and Jin-Woo Jeong. Facial expression recognition based on multi-head cross attention network. *arXiv preprint arXiv:2203.13235*, 2022. 5

[12] 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. 1, 2

[13] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer. *arXiv preprint arXiv:2203.13472*, 2022. 5

[14] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1, 5

[15] 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. 1

[16] 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

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

[18] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019. 1, 4

[19] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1, 4

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

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

[22] 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

[23] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021. 3

[24] Jing Li, Kan Jin, Dalin Zhou, Naoyuki Kubota, and Zhaojie Ju. Attention mechanism-based cnn for facial expression recognition. *Neurocomputing*, 411:340–350, 2020. 2

[25] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 28, pages 2584–2593. IEEE, July 2017. 1, 2

[26] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018. 2

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 4, 5

[28] 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, 2

[29] Kim Ngan Phan, Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. Expression classification using concatenation of deep neural network for the 3rd abaw3 competition. *arXiv preprint arXiv:2203.12899*, 2022. 5

[30] Andrey V Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv preprint arXiv:2203.13436*, 2022. 5

[31] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6248–6257, June 2021. 2

[32] Jiawei Shi, Songhao Zhu, and Zhiwei Liang. Learning to amend facial expression representation via de-albino and affinity. *arXiv preprint arXiv:2103.10189*, 2021. 1

[33] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition. 43(10):3349–3364. 3, 4, 5

[34] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 2

[35] Jun Weng, Yang Yang, Zichang Tan, and Zhen Lei. Attentive hybrid feature with two-step fusion for facial expression recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6410–6416. IEEE, 2021. 1, 2

[36] Fanglei Xue, Qiangchang Wang, and Guodong Guo. TransFER: Learning Relation-aware Facial Expression Representations with Transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, Mar. 2021. 1, 2, 4

[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. 1

[38] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 1, 2

[39] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 4

[40] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 1, 2

[41] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 5

[42] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 2

[43] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007. 2

[44] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 4