Exploring Large-scale Unlabeled Faces to Enhance Facial Expression Recognition

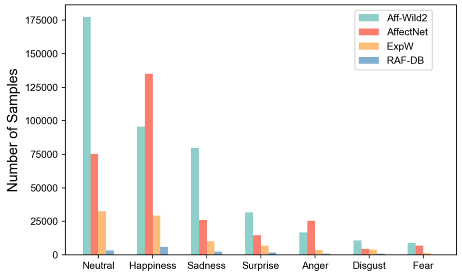
Jun Yu[1], Zhongpeng Cai[1], Renda Li [1*], Gongpeng Zhao[1], Guochen Xie[1], Jichao Zhu[1],
Wangyuan Zhu[1], Qiang Ling[1], Lei Wang[1], Cong Wang[2], Luyu Qiu[2], Wei Zheng[2]
[1]University of Science and Technology of China
[2]Huawei Techologies
{harryjun, qling, wangl}@ustc.edu.cn
{zpcai,rdli,zgp0531,xiegc,jichaozhu,zhuwangyuan}@mail.ustc.edu.cn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

Abstract

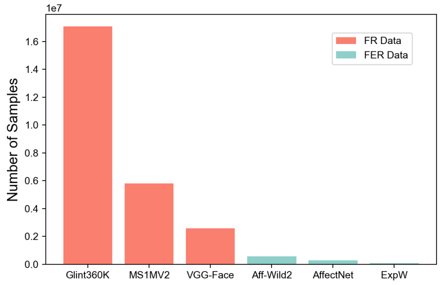
Facial Expression Recognition (FER) is an important task in computer vision and has wide applications in many fields. In this paper, we introduce our approach to the fifth Affective Behavior Analysis in-the-wild (ABAW) Competition which will be held in CVPR20223. For facial expression recognition task, there is an urgent need to solve the problem that the limited size of FER datasets limits the generalization ability of expression recognition models, resulting in ineffective performance. To address this problem, we propose a semi-supervised learning framework that utilizes unlabeled face data to train expression recognition models effectively. Our method uses a dynamic threshold module (DTM) that can adaptively adjust the confidence threshold to fully utilize the face recognition (FR) data to generate pseudo-labels, thus improving the model's ability to model facial expressions. In the 5th ABAW Expression Classification Challenge, our method achieves good results on the Aff-Wild2 validation and test sets, demonstrating that large scale unlabeled faces can indeed improve the performance of face expression recognition.

1. Introduction

According to psychology Research [30] by scientist A.Mehrabia, in human daily communication, the information transmitted through language only accounts for 7% of the total information, while the information transmitted through facial expressions reaches 55% of the total information. Therefore, it is significant to build a robust Facial Expression Recognition (FER) System. In recent years, many FER methods [3, 25, 26, 29, 36, 40, 49, 54] achieved state-of-the-art performance on several benchmark datasets (*e.g.* RAF-DB [27], SFEW [47] and AffectNet [32]).



(a)



(b)

Figure 1. (a) The class distribution of FER datasets. (b) The number of samples of FER data and FR data.

In order to address research questions that are of interest to affective computing, machine learning and multimodal signal processing communities and encourage a fusion of their disciplines. Kollias et al. [13–22, 48] organize the competition on Affective Behavior Analysis in-the-wild (ABAW). The 5th Workshop and Competition on Affec-

[*]Corresponding author

tive Behavior Analysis in-the-wild (ABAW), will be held in conjunction with the IEEE Computer Vision and Pattern Recognition Conference (CVPR), 2023.

In traditional fully supervised face expression recognition methods [3, 4, 39, 46], the accuracy of model predictions relies heavily on a large amount of high-quality labeled data. As shown in Figure 1 (a), existing FER training datasets are biased towards some majority classes, which leads to poor test accuracy for the minority classes. Sometimes the number of minority classes is less than 10% of the number of minority classes. This seriously affects the overall performance of the model. As we all know, it is expensive to obtain a large scale labeled FER data, which makes it difficult to expand the FER training datasets. However, as shown in 1 (b), the scale of Face Recognition (FR) data is much larger than that of FER data. Thus, how to remove the inconsistent data distribution between FR data and FER data becomes an urgent problem of utilizing the (FR) data to enhance Facial Expression Recognition models.

In this paper, we adopted a semi-supervised approach to obtain pseudo-labels for unlabeled data, in order to obtain sufficient training data to help the model to extract facial expressions. At the same time, to alleviate the problem of class-imbalanced dataset, we uniformly sampled the labeled facial expression samples to correct the bias learned by the model from the unlabeled data. We consider that a fixed threshold cannot fully utilize the data and cannot adapt to the class-imbalanced data. Moreover, considering that the discriminative ability of the model can be significantly improved with the increase of training steps, we designed a dynamic threshold module (**DTM**) that can adjust the confidence threshold with different classes and training steps to fully utilize the data.

To sum up, our contributions can be summarized as:

- We propose a semi-supervised learning framework for the task of facial expression recognition. It can apply the unlabeled faces data to the task of facial expression recognition through the use of pseudo labels, which greatly alleviates the problem of small-scale facial expression datasets.

- We design a dynamic threshold module (**DTM**) for the Semi-Supervised Learning method. It can dynamically adjust the confidence threshold for different stages of training and different expression categories, to fully utilize the unlabeled faces to generate pseudo labels.

- In the 5th ABAW competition, our method achieves great performance on official validation and test sets, which proves the effectiveness of our approach.

2. Related Work

2.1. Facial Expression Recognition

The facial expression recognition task is a classic task in the field of pattern recognition, and the approach [3, 4, 36, 40, 46, 54] of making full use of fully supervised data once gained very great progress in the field of FER. In recent years, attention has been focused on extending the dataset to obtain larger scale datasets. To solve the label confusion problem between different expression recognition datasets. IPA2LT [50] is the first work to address the annotation inconsistency in different facial expression datasets. They proposed LTNet embedded with a scheme of discovering the latent truth from multiple inconsistent labels and the input images. Ada-CM [26] is the first solution to explore the dynamic confidence margin in Semi-Supervised Deep Facial Expression Recognition. They designed an adaptive confidence margin to dynamically learn on all unlabeled data for the model's training and conducted a feature-level contrastive objective to learn effective features by applying the InfoNCE [41] loss. Face2Exp [49] proposed the Meta-Face2Exp framework to extract de-biased knowledge from auxiliary FR data through the meta optimization framework. utilized unlabeled face data to enhance expression recognition through the meta optimization framework.

2.2. Learning with Unlabeled Data

An important direction for learning methods using unlabeled data is semi-supervised learning. A popular class of Semi-Supervised Learning methods is to generate an artificial label for an unlabeled image and train the model to predict that artificial label on the input of an unlabeled image [24, 42]. Similarly, consistent regularization [2, 23, 34] uses the predictive distribution of the model to obtain an artificial label after randomly modifying the input or model function. An artificial label is obtained using the predictive distribution of the model after randomly modifying the input or the model function. Fixmatch [37] combines the advantages of these two methods, which use weak data augmentation and strong data augmentation of samples to achieve consistent regularization and obtain pseudo-labeled data by samples with confidence levels greater than a threshold. But the problem of this approach is that his modeling ability is limited in the early stage of model training because the threshold value is fixed. To solve this problem, Flexmatch [51] proposed Curriculum Pseudo Labeling (CPL), a curriculum learning approach to leverage unlabeled data according to the model's learning status. It can flexibly adjust thresholds for different classes at each time step to let pass informative unlabeled data and their pseudo labels. In addition, Dash [43] performs selection by retaining only samples with losses less than a given threshold in each update iteration, which is dynamically adjusted through iterations.

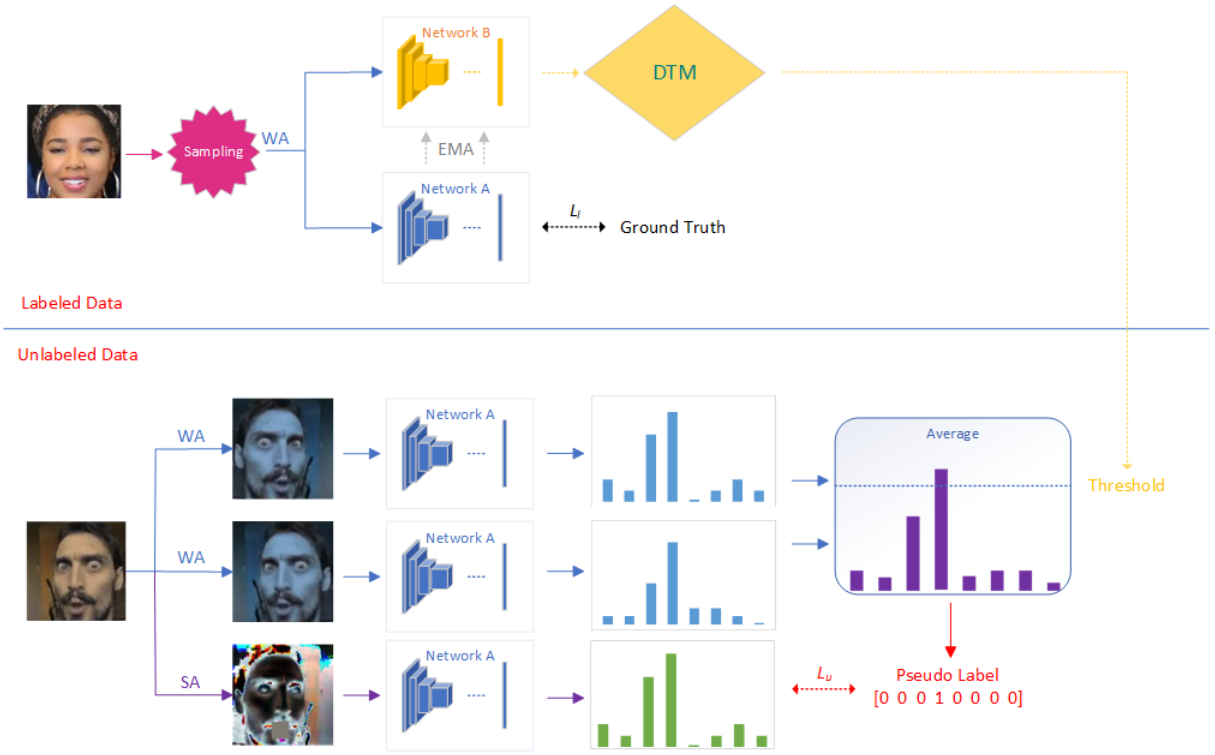


Figure 2. Illustration of the framework. Our approach can be divided into two parts. (1) For the learning of labeled data, in which the network can learn the balanced data distribution and use DTM to dynamically adjust the threshold. (2) For the learning of unlabeled data, The network is optimized by learning the consistency of weakly-augmented (WA) images and strongly-augmented (SA) images.

2.3. Affective Behavior Analysis in-the-wild

Zhang *et al.* [53] utilized the multimodal information from the images, audio and text and proposed a unified multimodal framework to fully use the emotion information, which achieved the best performance in ABAW3 competition. Jeong *et al.* [10] proposed a multi-head cross attention networks and pretrained on Glint360K [1] and some private commercial datasets. Xue *et al.* [45] proposed the Coarse-to-Fine Cascaded networks (CFC) to address the label ambiguity problem and used smooth predicting method to post-process the extracted features. In the 5th ABAW competition, perhaps inspired by [53], the high scoring methods [28, 52, 55, 57] all coincidentally used multi-modal information fusion to improve their scores and achieve great improvement. Our approach aims to leverage information from visual modality and explore the enhancement of large-scale unlabeled faces for FER task.

3. Method

In this section, we will describe our proposed approach in detail. As shown in Figure 2, the labeled data samples are weakly augmented and fed into the Network A to learn the balanced expression features. The unlabeled samples with confidence greater than threshold τ is used to generate pseudo labels, which are fed into the network to learn more facial expression features. The Dynamic Threshold Module (DTM) is introduced to adjust the confidence threshold for each class dynamically in each epoch.

3.1. Data Pre-process

Sampling. We sample the labeled FER data to ensure that each class of expressions has the same number of samples to ensure class balance. In this way, the model learns more class-balanced features that contribute to the de-bias learning of FR data.

Augment. We conduct two kinds of data augment operations. For Weakly-Augment (WA), we mainly use horizontal flip, color jitter, etc., while we choose RandAugment [5] as Strongly-Augment (SA) operation. The unlabeled samples remain semantic consistency after these two data augment operations.

3.2. Semi-Supervised Training

For the labeled data (x_l, y_l), we sample them and then apply weak data augmentation to obtain a probability distribution p_l predicted by neural network A:

$$P_l = Network_A(WA(sampling(x_l)); \theta_A) \tag{1}$$

For unlabeled data x_u, we first generate the weakly-augmented samples x_u^w and strongly-argumented samples x_u^s, then we utilize the network A to extract features and probability distributions:

$$P_u^w = Network_A(\mathit{WA}(x_u); \theta_A) \tag{2}$$

$$P_u^s = Network_A(\mathit{SA}(x_u); \theta_A) \tag{3}$$

Then we obtain the average probability distribution $\hat{P}_u$:

$$\widetilde{P}_u = \frac{1}{2}(P_u^w + P_u^s) \tag{4}$$

If $\mathit{argmax}(\widetilde{P}_u)$ is greater than the confidence threshold τ, we get the pseudo label:

$$\hat{y}_u = \mathit{argmax}(\widetilde{P}_u) \tag{5}$$

3.3. Dynamic Threshold Module

Fixed threshold limited the modeling ability in the early stage. Inspired by [26], we introduce a Dynamic Threshold Module (**DTM**) to adjust the threshold τ during different training stages. To enhance the robustness of the model, the network B is obtained from network A using the exponential moving average (EMA) technique with a decay rate of 0.999. We utilize the network B to extract features and probability distributions of the labeled data ((x_l, y_l)):

$$\check{P}_l = Network_B(x_l; \theta_B) \tag{6}$$

Then, we calculate the average confidence score of all correctly predicted samples in the labeled data for each class:

$$\tau_c = \frac{1}{N_c}\sum_{i=1}^{N_c} p_i^c \tag{7}$$

where N_c is the total number of correctly predicted samples of c-th class, p_i^c is the predicted confidence score of the correct class c for the i-th sample, and τ_c is the threshold of c^th class.

As the number of training epochs increases, the discriminative ability of the model for the trained data significantly increases. Therefore, we perform a weighted average on the threshold to prevent the threshold from increasing too quickly. Therefore, the final confidence threshold τ_c^t for class i at each epoch is:

$$\tau_c^t = \mu\tau_c^{t-1} + (1-\mu)\tau_c \tag{8}$$

where μ is a hyper-parameter.

3.4. Loss Function

We employ the cross-entropy loss function as the objective function for training our model.

$$L_{CE} = -\sum_{i=1}^{8} y_i \log(\hat{y}_i) \tag{9}$$

where y_i represents the label for the i-th class, and $\hat{y}_i$ represents the predicted probability of the i-th class.

For the labeled data, the objective function can be expressed as:

$$L_l = CE(y_l, P_l) \tag{10}$$

For the unlabeled data, the objective function can be expressed as:

$$L_u = CE(\hat{y}_u, \widetilde{P}_u) \tag{11}$$

The whole network minimizes the following loss function:

$$L_{total} = \lambda_1 L_l + \lambda_2 L_u \tag{12}$$

λ_1, λ_2 are hyper-parameters to balance each term's intensity.

3.5. Post-Process

Since Aff-Wild2 [13–22, 48] dataset is derived from all frames of the videos, and an expression takes some time to be generated. So an obvious conclusion is that there will not be a rapid change of expressions within a few adjacent frames. So we set a sliding window to post-process the prediction results for the purpose of smoothing the prediction labels. We first count the number of all predicted labels within a window, and then consider the expression label with the most occurrences as the expression recognition result for all images within this window. Finally, we achieve the predicted expressions smoothing for the whole dataset by means of a sliding window.

4. Experiment

In this section, we will provide a detailed description of the used datasets, the experiment setup, and the experimental results.

4.1. Datasets

FER Datasets. The 5th Workshop and Competition on Affective Behavior Analysis in-the-wild provides the Aff-wild2 [13–22, 48] database as the official datasets. For EXPR Classification Challenge, This database is audiovisual (A/V) and in total consists of 548 videos of around 2.7M frames that are annotated in terms of the 6 basic expressions (i.e., anger, disgust, fear, happiness, sadness, surprise), plus the neutral state, plus a category 'other' that denotes expressions/affective states other than the 6 basic ones. In addition, we used external facial expression

Table 1. Ablation study results on the official validation set, the highest score is indicated in bold.

Method	**Aff-Wild2**	**AffectNet and ExpW**	**MS1MV2**	**Post-Process**	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Other	**F1 Score (%)**
baseline	✓				-	-	-	-	-	-	-	-	23.00
EfficientNet-B7	✓	✓			60.99	15.42	6.25	6.02	40.18	43.21	21.01	45	30.88
SSL	✓	✓			59.63	31.90	21.85	12.29	48.79	41.50	29.28	52.88	37.27
SSL + DTM	✓	✓			56.59	34.89	28.64	9.99	49.99	44.62	31.95	52.82	38.69
SSL + DTM	✓	✓	✓		60.81	39.82	30.76	11.69	**50.26**	42.20	34.37	52.90	40.35
SSL + DTM	✓	✓	✓	✓	**61.58**	**52.24**	**43.97**	**15.86**	46.15	**48.95**	**36.27**	**54.42**	**44.93**

Table 2. The average F1 scores (in %) of different teams on the official Aff-wild2 validation set and test set. Our results are indicated in bold. The last line represent our best results in the post challenge evaluation phase.

Teams	**F1 on Validation Set**	**F1 on Test Set**
Netease Fuxi Virtual Human [52]	49.52	41.21
SituTech [28]	45.77	40.72
CtyunAI [57]	37.67	35.32
HFUT-MAC [55]	40.55	33.37
HSE-NN-SberAI [35]	43.3	32.92
AlphaAff [44]	37.57	32.18
USTC-IAT-United (Ours)	**44.93**	**30.75**
SSSIHL DMACS	-	30.47
SCLAB CNU [33]	47.75	29.49
Wall Lab	-	29.13
ACCC	-	28.46
RT_IAI	-	28.34
DGU-IPL [11]	27.77	22.78
baseline [18]	23	20.50
USTC-IAT-United (Ours Best)	**43.36**	**35.34**

databases, such as AffectNet [32] and ExpW [56]. AffectNet contains about 1M facial images collected from the Internet, it provides eleven emotion and non-emotion categorical labels (Neutral, Happiness, Sadness, Surprise, Fear, Disgust, Anger, Contempt, None, Uncertain, No-Face) and we only used the first 7 categories of images. The Expression in-the-Wild Database (ExpW) contains 91,793 faces downloaded using Google image search. Each of the face images was manually annotated as one of the seven basic expression categories. In our paper, we obtain 8,000 labeled images for each category from the Aff-Wild2 dataset through uniform sampling. Additionally, to increase the diversity of our data and improve the generalization of our model, we randomly sample 8,000 images for each category from the merged dataset of AffecNet and ExpW. For the ”other” category, we use the images from the Aff-Wild2 dataset since there are no such images in the other two datasets. Finally, we include the remaining images of these datasets as the unlabeled samples.

FR Datasets. For Face Recognition Datasets, We use MS1MV2 [9] as the unlabeled data. The MS1MV2 dataset is a semi-automatic refined version of the MS-Celeb-1M dataset [8] proposed by ArcFace [6], which includes 85k ids and 5.8m images. The unlabeled data used in our experiments consists of a subset of the InsightFace [7] MS1MV2 dataset, obtain by uniformly sampling 1/3 of its images. This subset comprises a total of 1.94 million images.

4.2. Setup

All training face images are resized to 224×224 pixels, our proposed method is implemented with the PyTorch toolbox on eight NVIDIA Tesla V100 GPUs. By default, we use Efficient-B7 [38] as the backbone. Adam optimizer [12] is used with the fixed learning rate of 5×10^{-4}. The batch size of labeled and unlabeled data is 32 and we conduct 1000 steps for a epoch for training. The hyper-parameters μ, λ_1, λ_2 are set as 0.9, 1 and 0.8, respectively. Experiments are also verified using MindSpore. The code implemented by MindSpore will be open sourced to MindFace [31] (https://github.com/mindspore-lab/mindface).

4.3. Metrics

According to the requirements of the competition, the evaluation metrics we use is the average F1 score, which is not affected by the class frequency and is more suitable for the imbalanced class distribution . It can be calculated as fallows:

$$F_1^c = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

$$F1 = \frac{1}{N} \sum_{c=1}^{N} F_1^c \tag{14}$$

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

4.4. Results

Validation Set Results. The average F1 scores (in %) of different teams on the official Aff-wild2 validation set are shown in Table. 2. Our method achieves good performance (44.93%) on the official validation set. In fact, this score is also quite high among all the participating teams, indicating to some extent the good potential of our approach. In addition, more discussion of the validation set results can be found in Sec. 4.5.

Test Set Results. However, on the test set, the average F1 score of our method falls very much (from 44.93% to 30.75%), while other methods drop less. We realize that our training is overfitting. Therefore, we try to submit the prediction results of the model training relatively early in the post challenge evaluation phase. Finally, we use the model with less serious overfitting and obtain the best F1 score of 35.34% on the test set. This proves that our approach has great potential and can even surpass some multi-modal approaches. It is worth mentioning that, our method achieves good results with only static images and less than 10% of the annotated data.

4.5. Ablation Study

To prove the effectiveness of the semi-supervised learning method and the dynamic threshold module, we conduct ablation study by comparing the models trained without the corresponding modules and datasets. We present the results for expressioln recognition in Table. 1. Where the dataset is used as shown in Sec. 4.1. As shown in the table, we pretrain the EfficientNet-B7 on the balanced FER dataset achieve the average F1 score of 30.88%. Than we use the SSL method to finetune the backbone and improve the F1 score to 37.27%. By introducing the DTM, the average F1 score can be raised to 38.69%, which proves the effictiveness of the dynamic threshold module. When the MS1MV2 dataset is added as unlabeled data, the F1 score increased to 40.35%. This suggests that unlabelled face data can indeed improve the performance of face expression recognition tasks. By using a sliding window with a window size of 350, our post-processing method ended up with a score of 44.93%.

5. Conclusion

In this paper, we propose a semi-supervised learning approach to improve the performance of face expression recognition task using unlabeled face data. To take full advantage of unlabeled data, we design a dynamic threshold module to leverage confidence thresholds for different training stages and different expression categories to generate more accurate pseudo-labels and alleviate the dataset imbalance problem. Our method achieves good results on the Aff-Wild2 validation and test sets, demonstrating that large scale unlabeled faces can indeed improve the performance of face expression recognition.

Acknowledgement

This work is supported by Natural Science Foundation of China (62276242), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B, CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), USTC-IAT Application Sci. & Tech. Achievement Cultivation Program (JL06521001Y), Sci.&Tech. Innovation Special Zone (20-163-14-LZ-001-004-01).

References

[1] 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] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 2

[3] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9680–9689, 2021. 1, 2

[4] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021. 2

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 3

[6] 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. 5

[7] Jia Guo. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, Accessed on Month Day, Year. 5

[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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 5

[9] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[10] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Jin-Woo Jeong, Yuchul Jung, and Sang-Ho Kim. Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2358, 2022. 3

[11] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Multimodal facial expression recognition with transformer-based fusion networks and dynamic sampling. *arXiv preprint arXiv:2303.08419*, 2023. 5

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[13] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task

learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1, 4

[14] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 1, 4

[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, 4

[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, 4

[17] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1, 4

[18] 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, 4, 5

[19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 1, 4

[20] 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

[21] 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, 4

[22] 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, 4

[23] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2

[24] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 2

[25] Hangyu Li, Nannan Wang, Xi Yang, and Xinbo Gao. Crs-cont: A well-trained general encoder for facial expression analysis. *IEEE Transactions on Image Processing*, 31:4637–4650, 2022. 1

[26] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4166–4175, 2022. 1, 2, 4

[27] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 1

[28] Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, and Wenqiang Jiang. Multi-modal expression recognition with ensemble method. *arXiv preprint arXiv:2303.10033*, 2023. 3, 5

[29] Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 648–665. Springer, 2022. 1

[30] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974. 1

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

[32] 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, 5

[33] 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. 5

[34] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 2

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

[36] 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*, pages 6248–6257, 2021. 1, 2

[37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[38] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[39] Yingli Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression recognition. *Handbook of face recognition*, pages 487–519, 2011. 2

[40] 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. 1, 2

[41] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*, 2021. 2

[42] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[43] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021. 2

[44] Fanglei Xue, Yifan Sun, and Yi Yang. Exploring expression-related self-supervised learning for affective behaviour analysis. *arXiv preprint arXiv:2303.10511*, 2023. 5

[45] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2412–2418, 2022. 3

[46] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 2

[47] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015. 1

[48] 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, 4

[49] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20291–20300, 2022. 1, 2

[50] 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. 2

[51] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 2

[52] 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. 3, 5

[53] 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. 3

[54] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 418–434. Springer, 2022. 1, 2

[55] 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. 3, 5

[56] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126:550–569, 2018. 5

[57] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer. *arXiv preprint arXiv:2303.08356*, 2023. 3, 5