Facial Expression Recognition with Supervised Contrastive Learning and Uncertainty Estimation

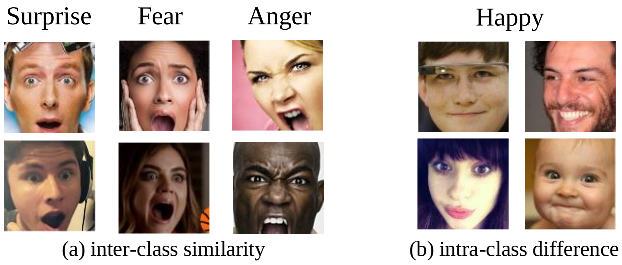
Lin Hu, Yujie Yang, Chen Zu, Qixiang Zhang, Wu Xi, Jiliu Zhou, Yan Wang, *Member, IEEE*

***Abstract*—Facial expression recognition (FER), which aims to recognize expressions of facial images, plays a vital role in human-computer interaction. To achieve automatic FER, several approaches based on deep learning have been proposed. However, most of them lack for the extraction of the discriminative expression semantic information and suffer from the problem of annotation ambiguity. In this letter, we propose an elaborately designed end-to-end recognition network with contrastive learning and uncertainty estimation, to recognize facial expressions efficiently and accurately, as well as to alleviate the impact of annotation ambiguity. Specifically, a supervised contrastive loss is introduced to promote inter-class separability and intra-class compactness, thus helping the network extract fine-grained expression features. As for the annotation ambiguity problem, we propose a Contrastive Uncertainty Estimation Module (CUEM) to estimate the uncertainty of each sample and relabel the unreliable ones. In addition, to deal with the padding erosion problem, we embed an Amending Representation Module (ARM) into the recognition network. Experimental results on three public benchmarks demonstrate the superiority of our method in comparison with the state-of-the-art methods.**

***Index Terms*—Facial expression recognition, Deep learning, Supervised contrastive learning, Uncertainty estimation.**

I. Introduction

FACIAL expression is one of the most essential means of non-verbal communication to reflect emotion and inner states by the movement of facial muscles [1, 2]. Automatic facial expression recognition (FER), which is an important and fundamental task in human-computer interaction, has attracted much attention and is of great importance in many real-word applications such as digital entertainment, intelligent robot system, driver monitoring and so on [3, 4, 5].



(a) inter-class similarity (b) intra-class difference

Fig. 1. Facial images from AffectNet [17] with (a) inter-class similarity and (b) intra-class difference.

Recently, with the rapid development of deep learning (DL), a great deal of DL-based methods have been proposed for the recognition of facial expressions [6, 7, 8, 9, 10]. Despite the impressive performance achieved by these methods, the FER task still remains challenging owing to the following perspectives. First, shared information across different expressions (inter-class similarity) and non-uniform information for the same expression (intra-class difference) interfere the discriminant of facial expressions. In particular, as shown in Fig. 1(a), the images of different expressions show high inter-class similarity due to the habitual facial movement like raised eyebrows, glaring eyes and opening mouth. For the images in Fig. 1(b), the four images from the same expression category (i.e., happy) exhibit great intra-class difference due to variations in gender, identity and age. Obviously, inter-class similarity and intra-class difference make it difficult to extract discriminative expression features, which results in performance decrease. Second, some disturbing factors such as obscuration, illumination and subjectiveness of annotators make it impossible to guarantee that all samples are annotated reasonably, thus involving the annotation ambiguity into the FER dataset. In addition, though convolution padding in DL helps to capture edge information in facial images, it also causes the erosion of the feature maps simultaneously and consequently weakens the representation of expressions, leading to the problem called padding erosion [11].

Manuscript received September 2, 2021. This work is supported by National Natural Science Foundation of China (NSFC 62071314) and Sichuan Science and Technology Program (2021YFG0326, 2020YFG0079). Corresponding author: Yan Wang.

Lin Hu, Yujie Yang, Qixiang Zhang and Yan Wang are with School of Computer Science, Sichuan University, Chengdu, China. (e-mail: 1327722084@qq.com; 2629557339@qq.com; ericZhang5915@gmail.com; wangyanscu@hotmail.com). Chen Zu is with Department of Risk Controlling Research, JD.com, China (e-mail: chenzu@outlook.com). Xi Wu and Jiliu Zhou are with School of Computer Science, Chengdu University of Information Technology, China (e-mail: wuxi@cuit.edu.cn; zhoujiliu@cuit.edu.cn).

The first two authors have equal contribution to this work.

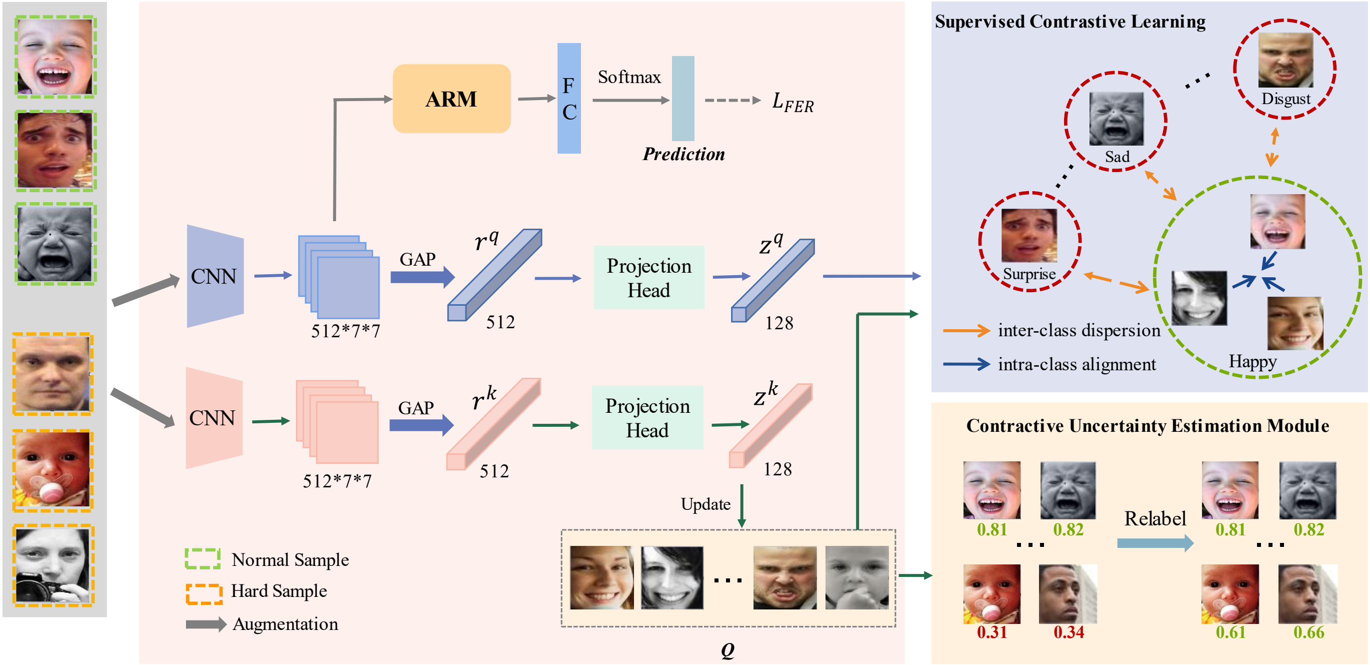


Fig. 2. Overview of the proposed network architecture, including main branch embedded with ARM, supervised contrastive learning and CUEM. Hard sample denotes the facial image difficult to recognize owing to the disturbing factors (i.e., obscuration, pose, illumination, etc) and Normal sample denotes the facial image unaffected by the disturbing factors.

Currently, several works have attempted to solve the above problems [11-14]. For instance, Ruan et al. [12] proposed a feature decomposition and reconstruction learning (FDRL) method to explicitly model inter-class similarity and intra-class difference, enabling the extraction of more discriminative features. Considering the annotation ambiguity existing in FER datasets, Wang et al. [13] developed a self-cure network (SCN) to suppress the ambiguity while preventing deep networks from over-fitting uncertain facial images. She et al. [14] explored a learning framework with latent distribution mining and pairwise uncertainty estimation, which better mines the latent distribution in the label space and fully exploits the pairwise relationship of semantic features between instances to estimate the ambiguity extent in the instance space. Moreover, Shi et al. [11] presented a plug-and-play module named Amending Representation Module (ARM) as a substitute for the pooling layer, which efficiently copes with the padding erosion problem. Although these methods have achieved remarkable progress, most of them only concentrate on one of the problems but neglecting the others. Thus, the FER task still has great potential for the improvement of the recognition performance.

In this letter, we propose an elaborately designed end-to-end recognition network with contrastive learning and uncertainty estimation to jointly resolve the challenges in FER task. The major contributions of this work are as follows. 1) We introduce the supervised contrastive learning to promote inter-class separability and intra-class compactness, thus helping the network extract fine-grained expression features. 2) To alleviate the annotation ambiguity problem in FER, we propose a Contrastive Uncertainty Estimation Module (CUEM) to assess the uncertainty of samples and relabel the unreliable ones. 3) Different from traditional supervised contrastive learning which usually follows the pre-training and transfer learning pipeline, we inject the spirit of multi-task learning into our framework, making it can deal with the FER task in an end-to-end manner. 4) Experimental results on three public benchmarks demonstrate the advancement of our proposed method.

II. Methodology

The overview of the network architecture is illustrated in Fig. 2, which mainly consists of three branches. Considering ARM can effectively alleviate the impact of padding erosion problem [11], we embed it into the top branch which is responsible for the final prediction of facial expressions. The remaining two branches utilize supervised contrastive learning to mine fine-grained expression semantic information for more accurate recognition. In addition, CUEM is introduced to relabel samples with high uncertainty to alleviate the annotation ambiguity problem. The details of the network and objective function will be introduced in the following sub-sections.

A. *Supervised Contrastive Learning*

As shown in Fig. 2, the network employs MoCo framework [15] as the backbone of our supervised contrastive learning. With the guidance of the supervised contrastive loss, the network has the capability to study similar representations for the same expression and distinct representations across different expressions. Specifically, given a batch of randomly sampled facial images including both normal and hard samples, two stochastic-data-augmented views of each image are respectively fed into two branches consisting of a convolutional neural networks (CNN) and global average pooling (GAP) to obtain image representations r^k and $r^q \in \mathbb{R}^{512\times1\times1}$. Since a compacted vector is more suitable for contrastive learning, r^k and r^q are later mapped into vector representations z^k and

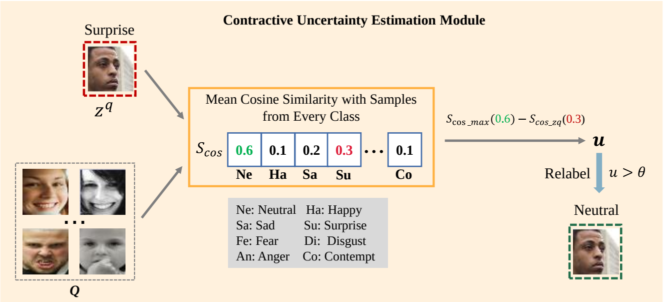


Fig. 3. Overview of Contrastive Uncertainty Estimation Module (CUEM). For each image in a batch, S_{cos} denotes a set of mean cosine similarities calculated with representations in Q. S_{cos_max} and S_{cos_zq} respectively represent the maximum mean cosine similarity and the mean cosine similarity of its corresponding expression label. u denotes the uncertainty obtained by subtracting S_{cos_zq} from S_{cos_max}.

$z^q \in \mathbb{R}^{128\times1\times1}$ by projection heads. Here, we implement the projection head with a three-layer nonlinear multiple-layer perceptron (MLP). Then, z^k and the corresponding annotation are enqueued to a queue $Q = [z_1^k, z_2^k, \ldots, z_K^k]$ with K (set to 8192) slots. z^q and elements in Q form a positive pair if they are representations of facial images with the same label, and otherwise form a negative pair. The supervised contrastive loss is jointly calculated to reduce the intra-class distance, expand the inter-class distance, and finally helps to learn fine-grained discriminative representations for each facial expression. Benefit from the learned feature representations via contrastive learning, the top branch responsible for the final prediction can achieve better FER performance.

B. Contrastive Uncertain Estimation Module (CUEM)

Aiming to alleviate the impact of annotation ambiguity, we design a CUEM to assess the uncertainty and relabel the samples with unreliable annotations, as shown in Fig. 3. In particular, for each facial feature representation z^q in the batch, a set of cosine similarities are calculated with the feature representations in Q mentioned above. By averaging the similarity values for feature representations in each facial expression class, we can obtain the mean cosine similarity S_{cos}. The uncertainty u is then obtained by subtracting the cosine similarity of its corresponding expression label S_{cos_zq} from the maximum mean cosine similarity S_{cos_max}. If u is higher than a preset threshold θ (set to 0.1), we consider the original annotation is unreliable and should be relabeled. With the guidance of the uncertainty u, we can filter out the relatively unreliable samples and select the corresponding label of the maximum mean cosine similarity S_{cos_max} as the new label. In this manner, the relabeled samples can provide more accurate clues for facial expression recognition.

C. Objective Function

The objective function consists of two parts: FER loss and supervised contrastive loss. To be specific, the FER loss can be expressed as:

$$L_{FER} = \frac{1}{N}\sum_i L_c(Y_i, Y_i^*), \qquad (1)$$

where L_c adopts the cross entropy (CE) loss and N is the total number of facial images. Y_i and Y_i^* respectively denote the ground truth and the predicted result.

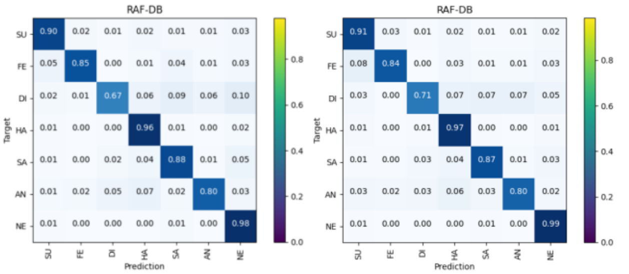


Fig. 4. Confusion matrices of the baseline (left) and our proposed method (right) on RAF-DB.

To extract fine-grained expression semantic information, the supervised contrastive loss (SCL) is introduced to help the network learn useful representations, which can be formulated as:

$$L_{sup} = \sum_{i=1}^{N} \frac{-1}{|P|} \sum_{p\in P} \log \frac{exp\left(z_i^k \cdot z_p^q / \tau\right)}{exp\left(z_i^k \cdot z_p^q / \tau\right) + \sum_{a\in A} exp\left(z_i^k \cdot z_a^q / \tau\right)}, \qquad (2)$$

where P and A represent the index sets of representations of positive samples ($Y_p = Y_i$) and negative samples ($Y_a \neq Y_i$) in Q, respectively. τ is a temperature hyper-parameter set to 0.1.

The total objective function is defined as:

$$L_{total} = L_{FER} + \lambda L_{sup}, \qquad (3)$$

where λ is a hyper-parameter to balance these two items. Herein, we design a learning strategy which gradually adjusts the value of λ in a downward trend according to the number of current epoch and the descent parameters α and β, as follows:

$$\begin{cases} \lambda = 1 - (epoch\ /\ \alpha\ ) & 1 \le epoch \le \alpha \\ \lambda = \beta & \alpha < epoch \le total\ epoch \end{cases}. \qquad (4)$$

III. Experiments

In this section, we first study the effectiveness of our network on three public benchmarks (i.e., RAF-DB [16], AffectNet [17] and FREPlus [18]) by comparing with other state-of-the-art FER methods. Then we evaluate the contribution of key components of the proposed method, including the supervised contrastive learning and the CUEM. In addition, we perform experiments to select appropriate values for key parameters in Eq. (4).

A. Dataset

RAF-DB [16] consists of totally 15,339 facial images with seven basic expressions (i.e. neutral, happiness, surprise, sadness, anger, disgust, and fear). Concretely, 12,271 images are used as training set and the remaining 3068 as testing set. AffectNet [17] is currently the largest FER dataset that includes 440,000 facial images, of which 280,000 images are selected for training and 4,000 for testing. In addition, FREPlus [18] including 28,709 training images and 3,589 testing images is also used in our experiment.

B. Experiment Settings

The proposed network is implemented by Pytorch framework on a NVIDIA RTX 3090 with 24GB memory. Specifically, we utilize ResNet-18 pre-trained on ImageNet as our backbone and take Adaptive moment estimation (Adam) optimizer with weight decay of 1e-4 to optimize the network. The network is trained for 100 epochs and the batch size is

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS. RES DENOTES RESNET. # MEANS REIMPLEMENTATION AND * MEANS USING EXTRA DISTRIBUTION DATA INSTEAD OF SINGLE CATEGORY ANNOTATION.

(a) Comparison on RAF-DB.

Method	Acc.
SCN [13]	87.03
DDL [7]	87.71
DMUE [14]	89.42
FDRL [12]	89.47
Res18+ARM [11]	92.05
Proposed	**92.41**

(b) Comparison on AffectNet.

Method	Acc.
Upsample [17]	47.01
Weighted loss [17]	58.00
Res18+ARM# [11]	59.45
RAN [19]	59.50
SCN [13]	60.23
Proposed	**60.94**

(c) Comparison on FERPlus.

Method	Acc.
PLD* [18]	85.10
Res+VGG [20]	87.40
Res18+ARM# [11]	87.50
SCN [13]	88.01
RAN [19]	88.55
Proposed	**88.56**

TABLE II
ABLATION STUDIES FOR KEY COMPOMENTS OF THE PROPOSED NETWOEK ON AFFECTNET. √ STANDS FOR THE ADDITION OF CORRESPONDING MODULE.

baseline	CUME	SCL	Acc.
√			59.45
√	√		60.21
√		√	60.44
√	√	√	**60.94**

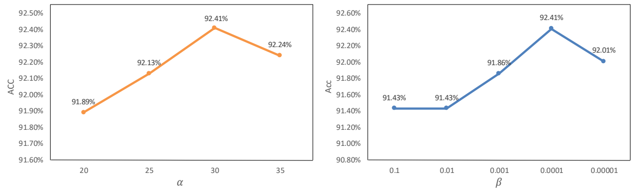
experimentally set to 256. The initial learning rate is set to 0.001 and gradually decreases by 10% after each epoch.

C. *Comparison with the state-of-the-art FER Methods*

To validate the superiority of our method, we implemented several state-of-the-art (SOTA) FER methods for comparison， including deep disturbance-disentangled learning method (DDL) [7], feature decomposition and reconstruction learning method (FDRL) [12], Res18+ARM [11], self-cure network (SCN) [13], latent distribution mining and pairwise uncertainty estimation based method (DMUE) [14] and region attention network (RAN) [19]. For fair comparison, the competing methods utilize pre-trained ResNet-18 as the network backbone as well. In addition, SCN and RAN are pre-trained on MS-Celeb-1M according to their original papers. The experimental results are shown in Table I. As observed, our method achieves the highest accuracy with 92.41%, 60.94% and 88.56% respectively on RAF-DB, AffectNet and FERPlus, demonstrating that our method can extract the discriminative features more efficiently for facial expression.

D. *Ablation Studies of Key Components*

In order to evaluate the contributions of the key components in our method, we conduct the ablation studies in a progressive way by utilizing the top branch embedded with ARM as our baseline. To be specific, our experimental settings include: (1) baseline, (2) baseline + CUEM, (3) baseline + SCL, (4) baseline + CUEM + SCL (proposed). The quantitative ablation results are summarized as Table II. By comparing the first and second rows, we can find that incorporating the CUEM into the baseline improves the recognition accuracy from 59.45% to 60.21%, demonstrating the effectiveness of CUEM of our method. Similarly, after further employing the supervised contrastive loss, the accuracy increases to 60.94%.



(a) evaluation of α (b) evaluation of β

Fig. 5. Evaluation of key parameters α and β on RAF-DB.

The confusion matrices of the baseline and our method on RAF-DB are displayed in Fig. 4, which also demonstrates the effectiveness of key components we proposed. Particularly, we can observe that our method achieves higher accuracy on four expressions (i.e., surprise, disgust, happy and neutral) while maintaining the same level of accuracy with the baseline for the other expressions.

E. *Evaluation of Key Parameters*

Furthermore, aiming to verify the learning strategy in Equation (4), we implement extensive experiments with the different value of α and β on RAF-DB. Specifically, we first fix β = 0.0001 and set α from 20 to 35. Experimental results are given in Fig. 5 (a). We can observe that our method achieves the best performance when α is set to 30, reaching 92.41%. Fig. 5 (b) shows the performance when α is set to 30 and β varies from 0.00001 to 0.1. We can see that when β is set to 0.0001, our method achieves the top performance. Thus, we finally set the values of α and β to 30 and 0.0001, respectively.

IV. Conclusion

In this letter, to tackle the challenges in the FER task including discriminative feature extraction and annotation ambiguity, we propose an end-to-end recognition network with contrastive learning and uncertainty estimation. Concretely, a supervised contrastive loss is introduced to instruct the network to study similar representations for the same expression and distinct representations across different expressions, thus helping the extraction of fine-grained discriminative features. To alleviate the annotation ambiguity problem, we propose a contrastive uncertainty estimation module (CUEM) to access the uncertainty and relabel the unreliable samples. Experimental results on public benchmarks display the advancement of our method in comparison with the state-of-the-art methods.

References

[1] Charles Darwin and Phillip Prodger. "The expression of theemotions in man and animals," Oxford University Press, USA, 1998.

[2] Y. I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97-115, 2001.

[3] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4011-4018, 2019.

[4] F. Zhang, T. Zhang, Q. Mao, L. Duan, and C. Xu, “Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach,” In *Proceedings of the 26th ACM international conference on Multimedia, 2018,* pp. 126-135.

[5] F. Zhang, T. Zhang, Q. Mao, and C. Xu, “Joint pose and expression modeling for facial expression recognition,” In *Proceedings of the IEEE conference on computer vision and pattern recognition, 2018,* pp. 3359-3368.

[6] T. Rao, J. Li, X. Wang, Y. Sun, and H. Chen, “Facial Expression Recognition With Multi-sale Graph Convolutional Networks,” *IEEE MultiMedia*, 2021.

[7] D. Ruan, Y. Yan, S. Chen, J. H. Xue, and H. Wang, "Deep disturbance-disentangled learning for facial expression recognition," In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2833-2841.

[8] Y. Tian, J. Cheng, Y. Li, and S. Wang, “Secondary information aware facial expression recognition,” *IEEE Signal Processing Letters*, vol. 26, no.12, pp. 1753-1757, 2019.

[9] P. Jiang, G. Liu, Q. Wang, and J. Wu, “Accurate and reliable facial expression recognition using advanced softmax loss with fixed weights,” *IEEE Signal Processing Letters*, 2020.

[10] P. Jiang, B. Wan, Q. Wang, and J. Wu, “Fast and efficient facial expression recognition using a Gabor convolutional network,” *IEEE Signal Processing Letters*, vol. *27,* pp. 1954-1958, 2020.

[11] J. Shi, and S. Zhu, “Learning to amend facial expression representation via de-albino and affinity,” arXiv preprint, arXiv:2103.10189.

[12] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, “Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition,” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7660-7669.

[13] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, “Suppressing uncertainties for large-scale facial expression recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6897-6906.

[14] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. 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*, 2021, pp. 6248-6257.

[15] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” arXiv preprint, arXiv:2003.04297.

[16] S. Li, W. Deng, and J. 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*, 2017, pp. 2852-2861.

[17] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing.*, vol. 10, no. 1, pp. 18-31, 2017.

[18] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution.,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279-283.

[19] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region attention networks for pose and occlusion robust facial expression recognition,” *IEEE Transactions on Image Processing.*, vol. 29, pp. 4057-4069, 2020.

[20] C. Huang, “Combining convolutional neural networks for emotion recognition,” in *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2017, pp. 1-4.