MITIGATING LABEL-NOISE FOR FACIAL EXPRESSION RECOGNITION IN THE WILD

Huan Yan[1,2], *Yu Gu*[1,2,3,*], *Xiang Zhang*[1,2], *Yantong Wang*[1,2], *Yusheng Ji*[4] *and Fuji Ren*[5]

[1]S[2]AC Lab, School of Computer and Information, Hefei University of Technology
[2]Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology
[3]Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation
[4]National Institute of Informatics [5]Tokushima University

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859818

ABSTRACT

Label-noise constitutes a major challenge for facial expression recognition in the wild due to the ambiguity of facial expressions worsened by low-quality images. To deal with this problem, we propose a simple but effective *Label-noise Robust Network (LRN)* which explores the inter-class correlations for mitigating ambiguity that usually happens between morphologically similar classes. Specifically, LRN leverages a multivariate normal distribution to model such correlations at the final hidden layer of the neural network to suppress the heteroscedastic uncertainty caused by inter-class label noise. Furthermore, LRN utilizes a confidence-based label-free loss to extract compact intra-class feature representations under label noise while preserving the intrinsic inter-class relationships. Experiments on three in-the-wild facial expression datasets demonstrates the superiority of our method.

***Index Terms*—** Facial expression recognition, aleatoric uncertainty, deep learning, input-dependent

1. INTRODUCTION

Facial expression is a vital nonverbal communication signal used to express one's inner state and intention. With the large number of applications of Facial Expression Recognition (FER) in fields such as lie detection, driver fatigue monitoring and psychological diagnosis, it has received widespread attention in computer vision.

In the past decades, Deep Neural Networks (DNN) trained on large datasets have been successfully applied to different visual recognition tasks, showing good abilities. With the emergence of a large number of laboratory-controlled or in-the-wild datasets (e.g., RAF-DB [1], AffectNet [2], etc), FER based on deep learning has made great progress.

*Corresponding author: Yu Gu. Email:yugu.bruce@ieee.org. This work was supported by the Key Research and Development Plan of Anhui Province (202004b11020018), Fundamental Research Fund of Chinese Academy of Medical Sciences (2020-JKCS-002), Open Fund of Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation (CSSAE-2021-009).

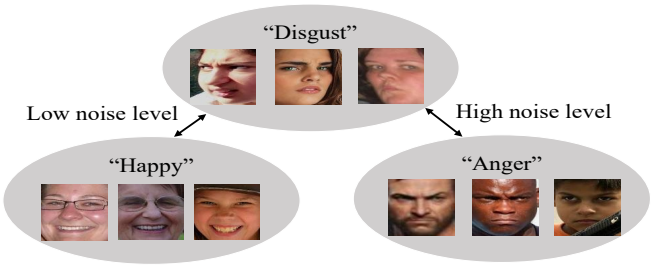

Fig. 1. An illustration of heteroscedastic uncertainty. Some images of *disgust* expression from the RAF dataset are visually similar to *anger* and therefore have high label noise. On the contrary, these *disgust* images are not easy to be mislabeled as *happy* due to the large visual distinction.

However, there are factors such as facial expression ambiguity or low-quality images collected on the Internet, which lead to label noise in in-the-wild datasets. As a result, the learned model cannot well capture useful facial expression features. As mentioned in [3], we call the uncertainty caused by the label noise as aleatoric uncertainty (also known as data uncertainty). According to the correlation of the input space, it can be divided into homoscedastic uncertainty and heteroscedastic uncertainty. The aleatoric uncertainty of the former is constant in the input space, while the latter is variable. Due to the different facial appearances between different facial expressions, an expression image is more likely to be mislabeled to certain type of expression with similar appearances, rather than being randomly assigned to other types. An example is shown in Figure 1. Some images of *disgust* expressions from RAF-DB [1] are more likely to be misjudged as *anger* rather than *happy*. Because morphologically similar categories can cause the annotator to make incorrect annotations and produce noisy labels.

To address the heteroscedastic uncertainty problem in FER, we propose a simple but effective Label-noise Robust Network (LRN) which explores the inter-class correlations for mitigating ambiguity of facial expression. LRN mainly includes two parts: Heteroscedastic Uncertainty Mod-

eling (HUM) and Confidence-based Label-free Discriminative (CLD) loss. For the former, we first assume that the logit of the deep learning model consists of an ideal item independent of noise as well as a noise item. Then a multivariate normal distribution is used to model this inter-class noise correlation to suppress the heteroscedastic uncertainty. Moreover, for FER datasets with fewer categories, the input-dependent covariance matrix can be directly calculated without low-rank approximation. For the latter, due to the subtle differences between expressions and the presence of label noise, it is necessary to enhance the discriminative ability of the deep features learned through the network. Thus, we design a confidence-based label-free discriminative loss to extract compact intra-class feature representations under label noise while preserving the intrinsic inter-class relationships. We verify the effectiveness of our method on several public in-the-wild facial expression datasets (i.e., RAF-DB, AffectNet and FERPlus). Similar to the state-of-the-art noise-tolerant FER methods, we also evaluate our method on the synthetic noisy facial expression datasets, which proves the superiority of our method.

In summary, our main contributions are summarized as follows:

- We propose a novel Label-noise Robust Network (LRN) which explores the inter-class correlations for mitigating ambiguity that usually happens between semantically similar classes.
- We design a confidence-based label-free discriminative loss, which extracts compact intra-class feature representations under label noise while preserving the intrinsic inter-class relationships.
- We verify LRN on several public facial expression datasets (i.e., i.e., RAF-DB, AffectNet and FERPlus). Our LRN achieves a high accuracy of 88.91% on RAF-DB, 60.84% on AffectNet and 89.53% on FERPlus.

2. RELATED WORKS

2.1. Facial Expression Recognition

In recent years, most FER methods based on Convolutional Neural Networks (CNN) have achieved promising performance [4, 5, 6, 7, 8, 9, 10]. For example, DeRL [4] recognizes facial expressions by extracting the information of expressive components through the de-expression learning process. Wang et al. [5] proposed a novel Region Attention Network (RAN) for the robust FER problem to pose and occlusion in the real world, which adaptively captures the importance of facial regions. RAN aggregates and embeds different numbers of regional features generated by the backbone network to form a compact fixed-length representation. Finally, the region biased loss is proposed to encourage the weight of high attention to the most important regions. DDL [11] used multi-task learning and adversarial transfer learning to focus on decomposing a variety of interference factors and encoding information related to expressions to realize expression recognition. Wen et al. [6] integrated the covariance pooling layer and residual network unit into the deep convolutional neural network for better dynamic target learning. Since the facial Action Units (AU) describes the movement of facial muscles, it helps to understand expressions and emotions [7]. Therefore, Pu et al. [12] explored the correlation between AU and facial expressions, designed a corresponding network framework to learn AU representations without AU annotations, and adaptively used AU representations to promote facial expression recognition. Different from previous approaches, our model solves the problem of input-dependent (heteroscedastic) label noise instead of homoscedastic label noise in facial expression recognition (the aleatoric uncertainty is constant across the input space).

2.2. Noisy Label

Due to various factors, such as the cost of the labeling process or the difficulty of correctly classifying data, label noise is a common problem in real-world datasets. An intuitive solution is to identify and correct suspicious labels to their corresponding real classes. Most methods currently used for label noise include sample importance weighting [13, 14], etc. For example, MentorNet [13] provided a curriculum (sample weighting scheme) for StudentNet to focus on samples that may be labeled correctly. Unlike existing courses that are usually predefined by human experts, MentorNet used StudentNet to dynamically learn a data-driven curriculum. Han et al. [15] assumed that samples with small loss are more likely to have clean labels, and train two neural networks at the same time. Each network predicts a small batch, and then feeds samples with small loss to another network for learning. Moreover, some recent research efforts aim to model aleatoric uncertainty [16], which is very important in real-world applications of machine learning (e.g., autonomous driving). Different from [16], we further enhance the ability to distinguish facial features under label noise to facilitate heteroscedastic uncertainty modeling. Moreover, for FER datasets with fewer categories, we can directly calculate the input-dependent covariance matrix without low-rank approximation. As far as we know, this is the first work to model the heteroscedastic uncertainty in the field of FER.

3. PROPOSED METHOD

The graphical illustration of our proposed model is depicted in Figure 2, which mainly consists of the Heteroscedastic Uncertainty Modeling (HUM) and Confidence-based Label-free Discriminative (CLD) loss. Given an in-the-wild dataset $\mathcal{D} = \{(x_i, \hat{y}_i)\}_{i=1}^{N}$, where $x_i \in R^d$ and $\hat{y}_i \in \{1, 2, ..., C\}$ denote the image and label space. Here, $\hat{y}_i$ refers to the label

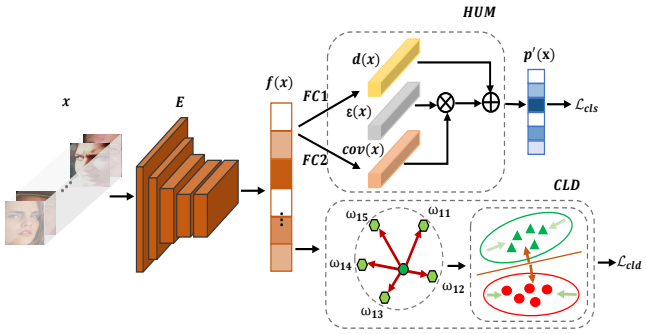

Fig. 2. Overview of our model, mainly including Heteroscedasticity Uncertainty Modeling (HUM) to model the heteroscedasticity uncertainty generated in the facial expression dataset, and a Confidence-based Label-free Discriminative (CLD) loss to encourage features from the same class to be compactly clustered together while preserving the intrinsic inter-class relationships. Best viewed in color. Zoom in for better view.

noise existing in the facial expression dataset. Our goal is to model the uncertainty in the dataset so that the model can extract useful and robust facial features, to correctly classify the expression images in the testset. The details are as follows.

3.1. Heteroscedastic Uncertainty Modeling

Generally speaking, we assume that when x is input, the network output logit $p(x)$ before using the *softmax* function is composed of an ideal term $d(x)$ and a noise term $\epsilon(x)$ [16]. If each noise term $\epsilon(x)$ is selected from the G(0,1) distribution (i.e., Gumbel), then solving the network output probability p_k becomes the classic softmax entropy model solution. This implicitly assumes that the noise components in the dataset are independent and identical, which is impractical for FER in the case of label noise. It is a common sense that the correlation between different expressions in-the-wild datasets is different and the noise level is also different. For example, given a *disgust* image in the RAF-DB dataset shown in Figure 1, some *anger* images may have high levels of noise, but we may have high confidence that they are not *happy*. Moreover, intuitively, it is obvious that the expression of *disgust* and *anger* are more similar than *happy*.

Therefore, we need to model the heteroscedastic uncertainty in-the-wild datasets. Specifically, we assume that the random component comes from a multivariate normal distribution, $\epsilon(x) \sim \mathcal{N}(0, cov(x))$ and model the label noise between classes by calculating the covariance matrix. Then the logit of the final network output can be expressed as $p(x) \sim \mathcal{N}(d(x), cov(x))$. However, the prediction probability of the network does not have a closed-form solution, so Monte Carlo (MC) sampling is used to approximate the prediction probability. In addition, consistent with [16], we use softmax function with temperature T to approximate the prediction probability. Therefore, the network predicted probability is expressed as:

$$\begin{aligned} p'_k(x) &\approx \frac{1}{S}\sum_{s=1}^{S}(softmax_T\, p_s(x))_k \\ &= \frac{1}{S}\sum_{s=1}^{S}\frac{exp((d_k(x)+cov_k(x)\epsilon_k^s)/T)}{\sum_{c=1}^{C}exp((d_c(x)+cov_c(x)\epsilon_c^s)/T)}, \\ \epsilon_1^s, ..., \epsilon_c^s &\sim \mathcal{N}(0_C, I_{C\times C}), \end{aligned} \tag{1}$$

where p_k represents the k-th output probability of the network. S is the number of MC samplings. Note that for the gradient calculation, the reparametrization trick is to rewrite $p(x)$ as $d(x) + cov(x)\epsilon(x)$ (i.e., the affine transformation about $\epsilon(x)$). In the actual operation, we use the linear function of the feature vector $f(x)$ to calculate $d(x)$ and $cov(x)$ (i.e., fully connected layers $FC1$ and $FC2$ after $f(x)$). Therefore, the label-noise robust classification loss $\mathcal{L}_{cls}$ is formulated as:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N}\sum_{c=1}^{C}\Pi_{[i=\hat{y}_{ic}]}log(p'_c(x_i)), \tag{2}$$

3.2. Confidence-based Label-free Discriminative Loss

Due to the subtle differences between expressions and the presence of label noise, it is necessary to enhance the discriminative ability of the features learned through the network. Although the conventional center loss [17] can strongly adjust the distribution of deep features, it is often assumed that the label is real and cannot be applied to the presence of label noise. To achieve this, as depicted in Figure 2, a confidence-based label-free discriminative loss $\mathcal{L}_{cld}$ is designed to extract compact intra-class feature representations under label noise while preserving the intrinsic inter-class relationships. To be specific, given a sample x_i, we choose feature $f(x_i)$ and category center μ_c as the confidence of whether to narrow the feature and category center, expressed as follows:

$$\omega_{ic} = \frac{exp(d(f(x_i), \mu_c))}{\sum_{c=1}^{C}exp(d(f(x_i), \mu_c))}, \tag{3}$$

where d is the metric function. Therefore, ω_{ic} represents the confidence that the i-th sample belongs to category c and the confidence-based label-free discriminative loss $\mathcal{L}_{cld}$ is expressed as follows:

$$\mathcal{L}_{cld} = \frac{\sum_{i=1}^{M}\sum_{c=1}^{C}\omega_{ic}\| f(x_i) - \mu_c \|_2^2}{\sigma_k^2 MC}, \tag{4}$$

where σ_k represents the standard deviation among class centers.

In summary, the overall optimization objective to optimize network parameters is formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{cld}, \tag{5}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{cld}$ represent the label-noise robust classification loss and confidence-based label-free discriminative loss, respectively. λ_1 denotes the regularization parameters. By optimizing this joint loss, LRN is able to extract fine-grained label-noise robust expression features for FER.

4. EXPERIMENTAL RESULTS

4.1. Datasets

We verify the robustness of our proposed method on two synthetic datasets (i.e., synthetic RAF-DB [1], synthetic AffectNet [2]) and three in-the-wild datasets (i.e., RAF-DB [1], AffectNet [2], FERPlus [18]). To make a fair comparison, we use seven categories of expression samples on RAF-DB, and eight categories of expression samples on AffectNet and FERPlus, which are consistent with our state-of-the-art rivals.

RAF-DB is the first in-the-wild dataset containing basic or compound expressions, including approximately 30,000 facial images. Each image is labeled about 40 times, and then the unreliable labels are filtered out, which is a good experimental attribute. We only select images with six basic expressions (i.e., neutral, happy, surprise, sad, anger, disgust, fear) as well as neutral expression for the experiment, including 12,271 images for training and 3,068 images for testing.

AffectNet is the lagest and most challenging FER dataset to date which contains about 1M images. It mainly uses 1,250 emotion-related keywords in six different languages to search for images on the major search engines, of which 450,000 images are manually annotated with eight basic expressions. Similarly to [14], we used 280K training images and 4K testing images (500 for each category), which contains eight basic expressions.

FERPlus is an extension of the FER2013 [19] dataset, which is collected in Google image search through APIs and used in the *ICML 2013 Challenges*. It is divided into 28,709 training, 3,589 validation and 3,589 test images. Each image has a size of 48×48 pixels and is annotated with eight emotion categories (six basic expressions plus neutral and contempt).

4.2. Implementation Details

We use ResNet18 pre-trained on the MS-Celeb-1M [20] dataset as the backbone network to extract features. We use MTCNN [21] to detect faces in images and further resize then to 112×112 pixels, augmented by being horizontally flipped with a probability of 0.5. For the RAF-DB and FERPlus datasets, the Stochastic Gradient Descent (SGD) algorithm with a momentum of 0.9, weight decay of 0.0001 and initial learning rate of 0.0001 is used as the optimizer. For the AffectNet dataset, we use the ADAM algorithm with initial learning rate of 0.0001 to train our model. The learning rate is decayed by a factor of 10× at every 10 epochs. 1,000 MC samples are used at train time. The batch size is 64. The default softmax temperature T is 0.7, and hyperparameter λ_1 is set to 1. Besides, because the AffectNet dataset is extremely unbalanced, the average sampling strategy is used for training to ensure that each class of sample in each batch can be included. We implement our method with Pytorch toolbox and conduct all the experiments on a single NVIDIA RTX 2080Ti.

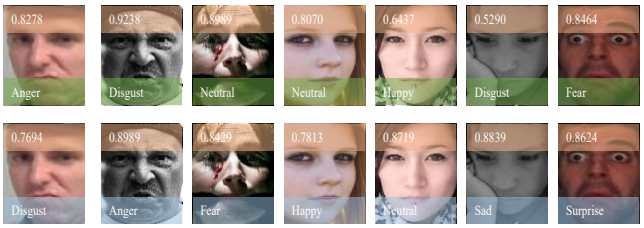

Fig. 3. From top to bottom are the test set prediction results of the baseline model and the LRN model trained on the RAF-DB dataset with 30% synthetic noisy labels. The upper and lower parts of each picture are marked with its maximum prediction score and label in the corresponding model. The true label is the same as the LRN predicted label. Best viewed in color. Zoom in for better view.

4.3. Experimental Results

To prove the effectiveness of the proposed LRN method in suppressing label noise, we conduct experiments on the synthetic RAF-DB and AffectNet datasets and compare LRN with several state-of-the-art noise-tolerant FER methods. Specifically, we randomly select 10%, 20% and 30% of each category in the training sample of the RAF-DB and AffectNet datasets, and the labels in them are flipped to other random categories in order to be consistent with the experimental settings in SCN [14]. Note that for different levels of label noise, we also present the corresponding baseline performance.

The expression classification results of RAF-DB and AffectNet with synthetic noisy annotations are presented in Table 1 with the best results highlighted in bold. Overall, our LRN model strikingly outperforms all other models, surpassing the corresponding baseline model with significant gains of 4.72%, 4.64% and 5.74% on RAF-DB, respectively. There are also performance gains of 2.26%, 1.69% and 1.48% compared to the baseline model on AffectNet. As the noise level increases, our method can still achieve the best performance. Compared with SCE-ALD [23], when the noise ratio on RAF-DB dataset is 10%, 20%, and 30%, LRN increases by 3.18%, 3.03%, and 4.23%, respectively. We have also seen that no matter how high the noise level is, LRN still maintains a good performance gain. To further study the effectiveness of our LRN under noisy annotations, we also visualize the prediction results of the baseline model and LRN model trained on the RAF-DB dataset with 30% synthetic noisy labels. In Figure 3, we can see that the trained baseline model falls easy to misjudge the ambiguous images in the RAF-DB test set, and

Table 1. Accuracies (%) of our approach with baseline, SCN, IPA2LT, and SCE-ALD methods on RAF-DB and AffectNet with synthetic noisy annotations.

Method	Noise=10%		Noise=20%		Noise=30%	
	RAF-DB	AffectNet	RAF-DB	AffectNet	RAF-DB	AffectNet
Baseline	80.73	57.23	78.77	56.52	76.10	55.61
IPA2LT [22]	80.25	57.42	78.26	56.60	72.26	54.29
SCN [14]	82.18	58.85	80.10	57.25	77.46	55.05
SCE-ALD [23]	82.27	58.97	80.38	57.60	77.61	55.54
Our Method	**85.45**	**59.49**	**83.41**	**58.21**	**81.84**	**57.09**

our LRN model can classify these images correctly.

Table 2. Ablation study on different objectives (%).

Structures	Loss	RAF-DB	AffectNet
Baseline	$\mathcal{L}_{ce}$	86.48	58.19
Baseline	$\mathcal{L}_{ce} + \mathcal{L}_{cld}$	87.64	59.31
LRN	$\mathcal{L}_{cls}$	88.35	60.59
LRN	$\mathcal{L}_{cls} + \mathcal{L}_{cld}$	**88.91**	**60.83**

4.4. Ablation Studies

This subsection introduces the effectiveness of each component in LRN via ablation studies on RAF-DB and AffectNet datasets, as shown in Table 2. $\mathcal{L}_{ce}$, $\mathcal{L}_{cls}$ and $\mathcal{L}_{cld}$ represent the baseline cross-entropy loss, heteroscedastic uncertainty modeling loss and confidence-based label-free discriminative loss, respectively. We can see that LRN is improved by 2.43% and 2.64% respectively on the RAF-DB and AffectNet datasets compared to the baseline method. Moreover, some observations can be found: (1) Heteroscedastic uncertainty modeling is more important than others. When only one of the components is used, it is 0.71% and 1.28% higher than confidence-based label-free discriminative loss on RAF-DB and AffectNet respectively. It proves the effectiveness of suppressing heteroscedastic uncertainty caused by label noise between classes. (2) When combined with confidence-based label-free discriminative loss, compared to only using heteroscedastic uncertainty modeling, we have achieved a performance improvement of 0.56% and 0.24% on RAF-DB and AffectNet. It can be seen that when confidence-based label-free discriminative loss is used, the model further extracts compact intra-class feature representation under label noise, while retaining the inherent inter-class relationship, and thus enhancing feature learning. Through the result analysis, all modules in LRN play a vital role in the final result.

4.5. Comparison with the State-of-the-Art

We compare the proposed LRN method with seven state-of-the-art rivals existing approaches: DLP-CNN [1], gACNN [24], IPA2LT [22], SCN [14], Lo et al. [25], EfficientFace [26] and SCE-ALD [23] on the RAF-DB, AffectNet and FERPlus datasets without manually adding label noise.

Table 3. Performance comparison on RAF-DB, AffectNet and FERPlus (%).† means that the model is trained on the RAF-DB and AffectNet training sets.

Methods	RAF-DB	AffectNet	FERPlus
DLP-CNN [1]	80.89	54.47	-
gACNN [24]	85.07	58.78	-
IPA2LT† [22]	86.77	55.71	-
RAN [5]	86.90	59.50	88.55
SCN [14]	87.03	60.23	89.35
Lo et al. [25]	87.35	-	87.45
EfficientFace [26]	88.36	59.89	-
SCE-ALD [23]	87.19	60.31	88.59
Our Method	**88.91**	**60.83**	**89.53**

Performance on RAF-DB and AffectNet RAF-DB and AffectNet are well-known and challenging in-the-wild datasets. We compare with the current most advanced competitors, including the latest solutions to suppress uncertainty in FER (i.e., SCN [14], SCE-ALD [23]). The experimental results are listed in Table 3. Note that since [25] was not evaluated on the AffectNet dataset, we did not give the results. Table 3 shows that our method is better than most existing methods. Compared with the SCE-ALD method, our LRN achieves gains of 1.72% and 0.52% on the RAF-DB and AffectNet datasets. Moreover, comparing with the best results achieved on RAF-DB of the EfficientFace [26] method, our method also has a 0.55% improvement.

Performance on FERPlus FERPlus is an in-the-wild dataset dedicated to FER competitions, and there are some experiments on this dataset. We compare with the latest method of suppressing uncertainty on the FERPlus dataset (i.e. SCN [14], SCE-ALD [23]). We also compare with the existing FER works (RAN [5], Lo et al. [25]) to evaluate the FER performance on the FERPlus dataset. As shown in Table 3, our method achieves the best accuracy, which is 0.94% better than SCE-ALD [23] (the latest result in the comparisons). These results clearly proves that the proposed method is highly competitive.

5. CONCLUSION

In this paper, we propose a simple but effective Label-noise Robust Network (LRN) which explores the inter-class correlations for mitigating ambiguity that usually happens between morphologically similar classes of facial expression. On one hand, a multivariate normal distribution is used to model the noise in the last hidden layer of the neural network to suppress the heteroscedasticity uncertainty caused by the inter-class label noise. On the other hand, a confidence-based label-free loss further encourages the model to extract compact inter-class feature representations under label noise. Experiments on three in-the-wild facial expression datasets demonstrates the superiority of our method.

6. REFERENCES

[1] L. Shan *et al.*, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2852–2861.

[2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2017.

[3] M. Collier, B. Mustafa, E. Kokiopoulou, R. Jenatton, and J. Berent, "A simple probabilistic method for deep classification under input-dependent label noise," *arXiv preprint arXiv:2003.06778*, 2020.

[4] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2168–2177.

[5] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process*, vol. 29, pp. 4057–4069, 2020.

[6] G. Wen, T. Chang, H. Li, and L. Jiang, "Dynamic objectives learning for facial expression recognition," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2914–2925, 2020.

[7] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, pp. 5, 1978.

[8] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "Wife: Wifi and vision based intelligent facial-gesture emotion recognition," *arXiv preprint arXiv:2004.09889*, 2020.

[9] Y. Gu, H. Yan, X. Zhang, Z. Liu, and F. Ren, "3-d facial expression recognition via attention-based multichannel data fusion network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[10] J. She *et al.*, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 6248–6257.

[11] D. Ruan, Y. Yan, S. Chen, J. Xue, and H. Wang, "Deep disturbance-disentangled learning for facial expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2833–2841.

[12] T. Pu, T. Chen, Y. Xie, and H. Wu, "Au-expression knowledge constrained representation learning for facial expression recognition," in *Proc. Int. Conf. Robot. Autom.*, 2021, pp. 11154–11161.

[13] L. Jiang, Z. Zhou, T. Leung, L. Li, and F. Li, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2304–2313.

[14] K. Wang *et al.*, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 6897–6906.

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. neural inf. proces. syst.*, 2018.

[16] M. Collier, B. Mustafa, E. Kokiopoulou, R. Jenatton, and J. Berent, "Correlated input-dependent label noise in large-scale image classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 1551–1560.

[17] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Computer Vision*, 2016, pp. 499–515.

[18] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2016, pp. 279–283.

[19] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. conf. neural inf. proces.*, 2013, pp. 117–124.

[20] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Computer Vision*, 2016, pp. 87–102.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[22] J. Zeng *et al.*, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Computer Vision*, 2018, pp. 222–237.

[23] S. Mao, G. Shi, L. Jiao, S. Gou, Y. Li, L. Xiong, and B. Shi, "Label distribution amendment with emotional semantic correlations for facial expression recognition," *arXiv preprint arXiv:2107.11061*, 2021.

[24] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Trans. Image Process*, vol. 28, no. 5, pp. 2439–2450, 2018.

[25] L. Lo, H. Xie, H. Shuai, and W. Cheng, "Facial chirality: Using self-face reflection to learn discriminative features for facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[26] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3510–3519.