Uncertainty-aware Label Distribution Learning for Facial Expression Recognition

Nhat Le*[1, 2, 3], Khanh Nguyen*[1, 2, 5], Quang Tran[3], Erman Tjiputra[3], Bac Le[1, 2], and Anh Nguyen[4]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]AIOZ, Singapore
[4]Department of Computer Science, University of Liverpool, Liverpool, UK
[5]FPT Software AI Center, Vietnam

Abstract

Despite significant progress over the past few years, ambiguity is still a key challenge in Facial Expression Recognition (FER). It can lead to noisy and inconsistent annotation, which hinders the performance of deep learning models in real-world scenarios. In this paper, we propose a new uncertainty-aware label distribution learning method to improve the robustness of deep models against uncertainty and ambiguity. We leverage neighborhood information in the valence-arousal space to adaptively construct emotiona distributions for training samples. We also consider the uncertainty of provided labels when incorporating them into the label distributions. Our method can be easily integrated into a deep network to obtain more training supervision and improve recognition accuracy. Intensive experiments on several datasets under various noisy and ambiguous settings show that our method achieves competitive results and outperforms recent state-of-the-art approaches. Our code and models are available at `https://github.com/minhnhatvt/label-distribution-learning-fer-tf`.

1. Introduction

Facial expression recognition (FER) plays an important role in understanding people's feelings and interactions between humans. Recently, automatic emotion recognition has gained a lot of attention from the research community [43] due to its applications in healthcare [35], surveillance [7], or human-robot interaction [8]. Most recent FER methods utilize deep learning [28] and achieve better results than handcrafted features approaches [9, 44]. The success of deep networks can be attributed to large-scale FER datasets such as AffectNet [37], EmotioNet [3], and RAF-DB [33]. Some datasets describe emotion in terms of Action Units (AUs) following the Facial Action Coding System [6] or quantify affection over continuous scales, such as valence and arousal [41], while most of them classify facial expressions into basic universal emotions [12, 36] and the neutral state.

Unfortunately, large-scale FER datasets often suffer from the problem of label uncertainty and annotation ambiguity [58, 5, 45]. People with different backgrounds might perceive and interpret facial expressions differently, which can lead to inconsistent and uncertain labels [58, 45]. In addition, real-life facial expressions usually manifest a mixture of feelings [67, 5] rather than a single exaggerated emotion often found in the lab-controlled setting. For example, Figure 1 shows that people may have different opinions about the expressed emotion, particularly in ambiguous images. Consequently, a distribution over emotion categories is better than a single label because it takes all sentiment classes into account and can cover various interpretations, thus mitigating the effect of ambiguity [16]. However, most current large-scale FER datasets only provide a single label for each sample instead of a label distribution, which means we do not have a comprehensive description for each facial expression. This can lead to insufficient supervision during training and pose a big challenge for many FER systems.

To overcome annotation ambiguity in FER, this paper proposes a new uncertainty-aware label distribution learning method that constructs emotion distributions for training samples. Specifically, for each instance, we leverage valence-arousal information to identify a set of neighbors and calculate their corresponding contributions using our adaptive similarity mechanism. We then aggregate neighborhood information with the provided single label, ad-

*Equal contribution

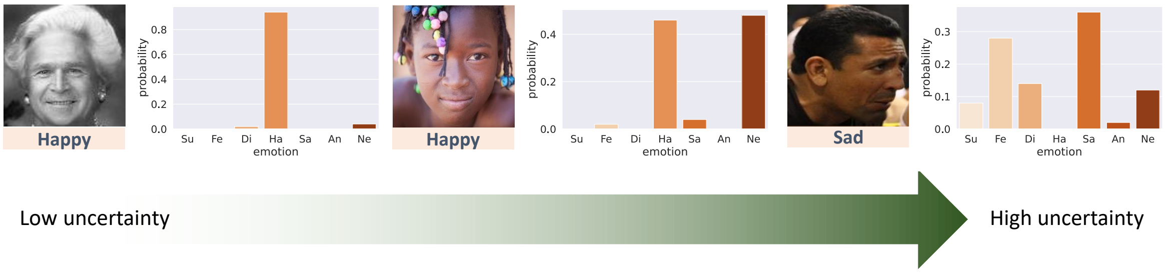

Figure 1: User study results by 50 volunteers on three random images from RAF-DB dataset. The expression on the right image are more ambiguous, which leads to high uncertainty in the emotion label. Labels at the bottom denote the provided annotation from the dataset. (Su=`Surprise`, Fe=`Fear`, Di=`Disgust`, Ha=`Happy`, Sa=`Sad`, An=`Angry`, Ne=`Neutral`)

justed by its learnable uncertainty factor, to generate the target label distribution. Finally, we use the constructed distribution as supervision signals to optimize the model via label distribution learning. We also introduce a discriminative loss that reduces intra-class variations and encourages inter-class differences to improve the model's robustness against ambiguous features. Note that the distribution construction only occurs during training while the inference process remains intact. In summary, our contributions are as follows:

1. We propose a new method, namely **L**abel **D**istribution **L**earning with **V**alence-**A**rousal (LDLVA), for FER with ambiguous annotation by exploiting neighborhood information in the valence-arousal space.

2. Our uncertainty-aware label distribution construction provide more accurate and richer supervision for training deep FER networks, allowing them to learn from ambiguous data effectively in an end-to-end manner.

3. We perform extensive experiments under various synthetic and real-world ambiguity settings and achieve state-of-the-art results on RAF-DB, AffectNet, and SFEW datasets.

2. Related Work

Most recent methods [57, 25, 64, 15, 5, 52, 58, 31, 61, 50, 30] categorize facial expression into discrete classes corresponding to basic universal emotions [12, 36], which is easy to interpret and intuitive to humans. Other approaches [62, 39] attempted to represent human emotion using Action Units (AUs) [6] or continuous scales such as valence and arousal [41]. In this work, we leverage the auxiliary information of continuous scales to mitigate the effect of uncertainty and ambiguity in existing FER datasets when predicting the discrete emotion of a given facial expression.

One challenging problem of FER is that ambiguous facial expressions can make it difficult to correctly identify the expressed emotions, which might lead to noisy and uncertain annotations [5, 45]. Empirical studies also show that neural networks are sensitive to noise and can easily overfit noisy data [1, 48, 29]. To overcome this challenge, previous approaches model the noise by a transition matrix [47, 20, 40, 58]. In [2], precise image features are extracted from a pre-trained model to regulate the learning process with noisy labels. The authors in [63, 54] use noise-tolerant loss functions to increase noise robustness. Other methods [59, 52] measure the uncertainty of each sample and utilize a sample-weighting strategy to help the network tolerate noisy samples. Recently, Zhang *et al.* [60] propose to quantify the uncertainties from the relative difficulty of samples by feature mixup. However, these methods only focus on improving the accuracy on mislabelled data and do not handle the ambiguous nature of facial expressions.

An alternative approach to address label noise and ambiguity is label distribution learning (LDL) [17]. In other domains, previous works [19, 18, 16] leverage prior knowledge to transfer logical labels into discretized bivariate Gaussian label distribution. The authors in [48, 29] utilize network's predictions as label distributions to correct noise, which can be unstable and hard to optimize. Instead, our method not only adaptively utilizes the model's predictions but also exploits domain knowledge of the valence-arousal space to construct target distributions. In FER literature, Zhou *et al.* [67] introduces a framework to map a facial expression to multiple emotions with corresponding intensities. Jia *et al.* [23] proposes to learn emotion distributions by exploiting label correlations at a local level. Zhao *et al.* [65] uses a pre-trained label distribution generator to produce emotion distribution. Other works create the label distribution by computing the membership degrees to the labels [13, 24, 46, 34, 22]. Recently, Chen *et al.* [5] leverages the topology in facial landmarks and action units spaces to acquire more information for label distribution learning. She *et al.* [45] proposed to leverage multiple branches to obtain the latent distribution. However, these methods either rely heavily on good features with local linearity to work properly [13, 11, 24, 46, 34, 38, 22] or only use the mined label distributions to regularize the model's training process instead of directly learning from them [5, 45].

Unlike previous works, our method constructs emotion distributions for training instances and directly uses them as supervision information, thus reducing the effects of annotation ambiguity. We do not need to be provided with label distributions to train the network since they can be accurately estimated using our adaptive similarity mechanism and learnable uncertainty factors. We experimentally show that our approach is more effective as the network is trained end-to-end with label distributions, which brings more meaningful information to the training process.

3. Methodology

We first introduce a list of notations that will be used throughout this paper. Let $\boldsymbol{x} \in \mathcal{X}$ be the instance variable in the input space $\mathcal{X}$ and $\boldsymbol{x}^i$ be the particular i-th instance. The label set is denoted as $\mathcal{Y} = \{y_1, y_2, ..., y_m\}$ where m is the number of classes and y_j is the label value of the j-th class. The logical label vector of $\boldsymbol{x}^i$ is indicated by $\boldsymbol{l}^i$ = $(l^i_{y_1}, l^i_{y_2}, ..., l^i_{y_m})$ with ${}^{i}_{y_j} \in \{0, 1\}$ and $\|\boldsymbol{l}\|_1 = 1$. We define the label distribution of $\boldsymbol{x}^i$ as $\boldsymbol{d}^i$ = $(d^i_{y_1}, d^i_{y_2}, ..., d^i_{y_m})$ with $\|\boldsymbol{d}\|_1 = 1$ and $d^i_{y_j} \in [0, 1]$ representing the relative degree that $\boldsymbol{x}^i$ belongs to the class y_j. A neural network with parameters θ followed by a softmax layer is denoted as $f(\boldsymbol{x}; \theta)$. The corresponding feature vector of $\boldsymbol{x}^i$ extracted by a CNN backbone model is indicated by $\boldsymbol{v}^i \in \mathbb{R}^V$.

3.1. Overview

Most existing FER datasets assign only a single class or equivalently, a logical label $\boldsymbol{l}^i$ for each training sample $\boldsymbol{x}^i$. In particular, the given training dataset is a collection of n samples with logical labels D_l = $\{(\boldsymbol{x}^i, \boldsymbol{l}^i) | 1 \leq i \leq n\}$. However, as depicted in Figure 1, a label distribution $\boldsymbol{d}^i$ is a more comprehensive and suitable annotation for the image than a single label. Inspired by the recent success of label distribution learning (LDL) in addressing label ambiguity [16], we aim to construct an emotion distribution $\boldsymbol{d}^i$ for each training sample $\boldsymbol{x}^i$, thus transform the *training* set D_l into D_d = $\{(\boldsymbol{x}^i, \boldsymbol{d}^i) | 1 \leq i \leq n\}$, which can provide richer supervision information and help mitigate the ambiguity issue. Consequently, our goal is to optimize the parameters θ of the neural network $f(\boldsymbol{x}; \theta)$ such that it can learn an appropriate mapping function for the instance $\boldsymbol{x}^i$ from the input space to the target label distribution $\boldsymbol{d}^i$. Mathematically, we use cross-entropy to measure the discrepancy between the model's prediction and the constructed target distribution [16]. Hence, the solution can be obtained by minimizing the following classification loss:

$$\mathcal{L}_{cls} = \sum_{i=1}^{n} \text{CE}\left(\boldsymbol{d}^i, f(\boldsymbol{x}^i; \theta)\right) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{d}^i_j \log f_j(\boldsymbol{x}^i; \theta). \tag{1}$$

An overview of our method is presented in Figure 2. To construct the *label distribution* for each training instance $\boldsymbol{x}^i$, we leverage its neighborhood information in the valence-arousal space. Particularly, we identify K neighbor instances for each training sample $\boldsymbol{x}^i$ and utilize our *adaptive similarity mechanism* to determine their contribution degrees to the target distribution $\boldsymbol{d}^i$. Then, we combine the neighbors' predictions and their corresponding contribution degrees with the provided label $\boldsymbol{l}^i$ and $\boldsymbol{l}^i$'s uncertainty factor to obtain the label distribution $\boldsymbol{d}^i$. The constructed distribution $\boldsymbol{d}^i$ will be used as supervision information to train the model via label distribution learning. It is worth noting that these steps occur only during training, thus no extra costs are introduced at inference time.

3.2. Adaptive Similarity Measuring

As in previous works [68, 56, 5], we assume that facial images should have similar emotions to their neighbors in an auxiliary or supporting space. Therefore, the label distribution of an instance can be constructed using the information of its neighbors. Since our goal is to reconstruct the target label distribution with high fidelity, the chosen supporting space should highly correlate with the emotion space to transfer as much information as possible. Although information such as facial landmarks and action units can be utilized as the supporting space, we find that valence-arousal values are more closely associated with discrete emotions and thus particularly suitable to be the auxiliary space. In practice, the valence-arousal has been widely used to represent the human emotional spectrum, with valence describing how positive or negative an expression is and arousal indicating the intensity or activation degree of the expression [42].

Similar to the smoothness assumption [68], we assume that the label distribution of the main instance $\boldsymbol{x}^i$ can be computed as a linear combination of its neighbors' distributions. To determine the contribution of each neighbor, we propose an adaptive similarity mechanism that not only leverages the relationships between $\boldsymbol{x}^i$ and its neighbors in the auxiliary space but also utilizes their feature vectors extracted from the backbone. In particular, we first use the K-Nearest Neighbor algorithm to identify K closest points for each training sample $\boldsymbol{x}^i$, denoted as $N(i)$, based on the distance between training instances in the valence-arousal space. We then compute a *local similarity score* between $\boldsymbol{x}^i$ and each of its K neighbors using the following formula:

$$s^i_k = \exp\left(-\frac{\|\boldsymbol{a}^i - \boldsymbol{a}^k\|^2_2}{\delta^2}\right), \quad \forall \boldsymbol{x}^k \in N(i), \tag{2}$$

where $\boldsymbol{a}$ is the corresponding auxiliary valence-arousal vector of $\boldsymbol{x}$, and δ is a hyperparameter controlling similarity measurement. Intuitively, the higher s^i_k is, the more $\boldsymbol{x}^k$ contributes to the label distribution of $\boldsymbol{x}^i$.

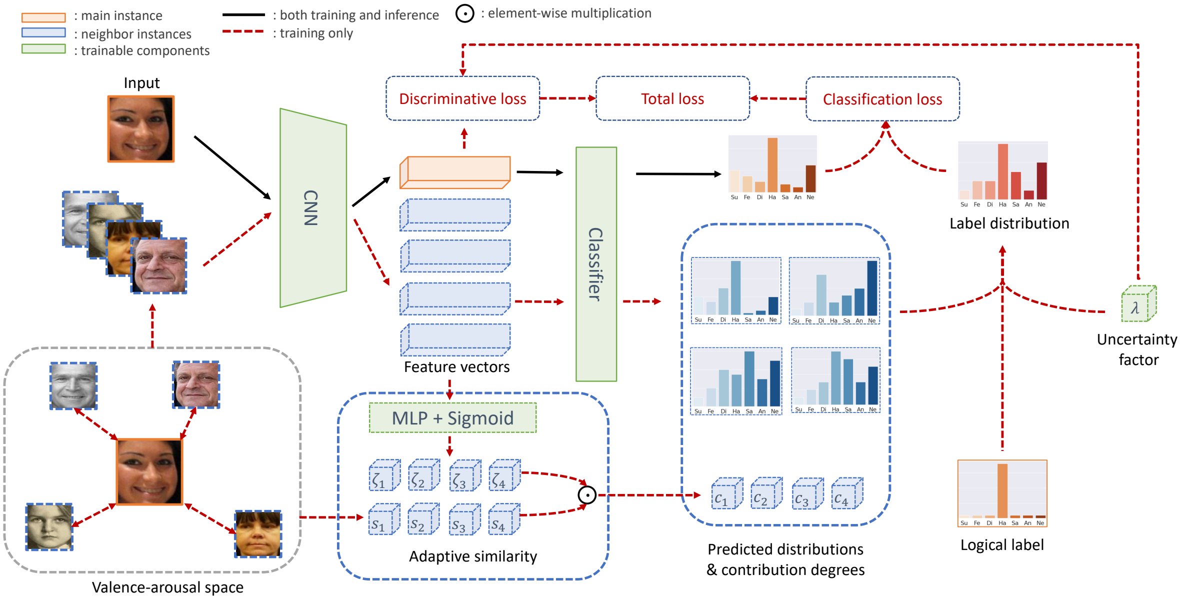

Figure 2: An overview of our Label Distribution Learning with Valence-Arousal (LDLVA) for facial expression recognition under ambiguity. Dotted lines denote components used in training only while solid lines denote components used in both training and testing.

However, since valence-arousal values are not always available in practice, we leverage an existing method [49] to generate pseudo-valence-arousal. Consequently, these values can be inaccurate and lead to incorrect calculation of s_k^i. Therefore, we proposed to correct these potential errors with our adaptive similarity mechanism. Specifically, we calculate a *calibration score* for each $(\boldsymbol{x}^i, \boldsymbol{x}^k)$ pair using the feature vectors $(\boldsymbol{v}^i, \boldsymbol{v}^k)$ extracted by the CNN backbone of $\boldsymbol{x}^i$ and its neighbor instance $\boldsymbol{x}^k \in N(i)$ as follows:

$$\zeta_k^i = \text{Sigmoid}\left(g([\boldsymbol{v}^i, \boldsymbol{v}^k]; \phi)\right), \tag{3}$$

where $[\cdot, \cdot]$ is the concatenation operator, g is a three-layer perceptron (MLP) with parameter ϕ. The dimensionality of each layer is 512, 256, and 1, respectively. We also apply layer normalization and ReLU non-linearity in the first two layers.

The final *contribution degrees* of neighbor instances are calculated as the product of the local similarity and the calibration score:

$$c_k^i = \begin{cases} \zeta_k^i s_k^i, & \text{for } \boldsymbol{x}^k \in N(i), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

3.3. Uncertainty-aware Label Distribution Construction

After obtaining the contribution degree of each neighbor $\boldsymbol{x}^k \in N(i)$, we can now generate the target label distribution $\boldsymbol{d}^i$ for the main instance $\boldsymbol{x}^i$. The target label distribution is calculated using the logical label $\boldsymbol{l}^i$ and the aggregated distribution $\tilde{\boldsymbol{d}}^i$ defined as follows:

$$\tilde{\boldsymbol{d}}^i = \frac{\sum_k c_k^i f(\boldsymbol{x}^k; \theta)}{\sum_k c_k^i}, \tag{5}$$

$$\boldsymbol{d}^i = (1 - \lambda^i)\boldsymbol{l}^i + \lambda^i \tilde{\boldsymbol{d}}^i, \tag{6}$$

where $\lambda^i \in [0, 1]$ is the *uncertainty factor* for the logical label. It controls the balance between the provided label $\boldsymbol{l}^i$ and the aggregated distribution $\tilde{\boldsymbol{d}}^i$ from the local neighborhood. Intuitively, a high value of λ^i indicates that the logical label is highly uncertain, which can be caused by ambiguous expression or low-quality input images as illustrated in Figure 6, thus we should put more weight towards neighborhood information $\tilde{\boldsymbol{d}}^i$. Conversely, when λ^i is small, the label distribution $\boldsymbol{d}^i$ should be close to $\boldsymbol{l}^i$ since we are certain about the provided manual label. In our implementation, λ^i is a trainable parameter for each instance and will be optimized jointly with the model's parameters using gradient descent.

Mathematically speaking, consider Equation 1 and 6, the derivative of $\mathcal{L}_{cls}$ with respect to λ^i can be computed as:

$$\frac{\partial \mathcal{L}_{cls}}{\partial \lambda^i} = \frac{\partial \text{CE}\left(\boldsymbol{d}^i, f(\boldsymbol{x}^i; \theta)\right)}{\partial \lambda^i} \tag{7}$$

$$= -\sum_j \tilde{\boldsymbol{d}}_j^i \log f_j(\boldsymbol{x}^i; \theta) + \sum_j \boldsymbol{l}_j^i \log f_j(\boldsymbol{x}^i; \theta) \tag{8}$$

$$= \text{CE}(\tilde{\boldsymbol{d}}^i, f(\boldsymbol{x}^i; \theta)) - \text{CE}(\boldsymbol{l}^i, f(\boldsymbol{x}^i; \theta)). \tag{9}$$

If $\text{CE}(\boldsymbol{l}^i, f(\boldsymbol{x}^i; \theta))$ is smaller than $\text{CE}(\tilde{\boldsymbol{d}}^i, f(\boldsymbol{x}^i; \theta))$, the derivative of $\mathcal{L}_{cls}$ with respect to λ^i is positive, which leads to a negative update for λ^i following gradient descent optimization scheme. This is desirable because in this case, the network output is in more agreement with the logical label than the aggregated neighborhood distribution. In other words, it is more confident about the provided label and thus, we should decrease the value of the uncertainty factor λ^i. The same reasoning can be applied in the opposite situation.

3.4. Loss Function

Recent literatures have shown the benefits of learning discriminative features in FER [4, 27, 14, 15]. Inspired by this, we believe it is beneficial to encourage the network to learn good facial descriptions because it can help improve the model's ability to discriminate between ambiguous emotions. We find that the center loss [55] is suitable for our purpose because of its simplicity and efficacy in reducing the intra-class variations of the learned representations. Nevertheless, in the traditional formulation of the center loss [55], the features of a sample are "blindly" pulled towards its corresponding class center given its label. This means when the provided label is incorrect, it can cause the network to learn imprecise features. We propose to overcome this problem by incorporating the label uncertainty factor λ^i to adaptively penalize the distance between the sample and its corresponding center. For instances with high uncertainty, the network can effectively tolerate their features in the optimization process. Furthermore, we also add pairwise distances between class centers to encourage large margins between different classes, thus enhancing the discriminative power. Our discriminative loss is calculated as follows:

$$\begin{aligned} \mathcal{L}_D = &\frac{1}{2}\sum_{i=1}^{n}(1-\lambda^i)\|\boldsymbol{v}^i - \boldsymbol{\mu}_{y^i}\|_2^2 \\ &+ \sum_{j=1}^{m}\sum_{\substack{k=1 \\ k\neq j}}^{m} \exp\left(-\frac{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|_2^2}{\sqrt{V}}\right), \end{aligned} \tag{10}$$

where y^i is the class index of the i-th sample while $\boldsymbol{\mu}_j$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\mu}_{y^i} \in \mathbb{R}^V$ are the center vectors of the j-th, k-th, and y^i-th classes, respectively. During the training phase, all center vectors are zero-initialized and optimized using Equation 10. Intuitively, the first term of $\mathcal{L}_D$ encourages the feature vectors of one class to be close to their corresponding center [55] while the second term improves the inter-class discrimination by pushing the cluster centers far away from each other.

Combining Equation 1 and Equation 10, we obtain the total loss for training:

$$\mathcal{L} = \mathcal{L}_{cls} + \gamma\mathcal{L}_D, \tag{11}$$

where γ is the hyperparameter balancing between the two losses.

4. Experiments

In this section, we first validate the effectiveness of our approach on synthetic ambiguity caused by noisy label data. Next, we evaluate the performance of our LDLVA in handling inconsistent labels caused by ambiguous facial expressions. We then compare LDLVA with state-of-the-art methods to demonstrate the robustness of our approach towards annotation ambiguity that inherently exists in real-world data. Finally, we conduct ablation studies and present qualitative results to investigate the effectiveness of each component as well as the advantages of our method.

4.1. Datasets

We perform experiments on three popular in-the-wild FER datasets: AffectNet [37], RAF-DB [33] and SFEW [10]. They are created by collecting data from the Internet and reflect real-life scenarios. AffectNet [37] has more than 400,000 facial images manually annotated with discrete emotions and valence-arousal. Following previous work [5, 58, 15], we select approximately 280,000 and 3,500 images for training and testing, all of which belong to six basic emotions (`surprise`, `fear`, `disgust`, `happy`, `sad`, and `angry`) and `neutral` expression. RAF-DB [33] is split into training and test sets with more than 12,000 and 3,000 images, respectively. SFEW [10] has 879 training images and 406 testing images, all of which are extracted from movie videos.

4.2. Experimental Settings

By default, we use the pretrained ResNet-50 [21] as the CNN backbone. We align the input image and perform on-the-fly augmentation during training by randomly flipping the image horizontally and taking a random crop of size 224×224 after padding 16 pixels on each side. At test time, we use the central crop of the image as input for the model. During training, for each instance, we consider 8 nearest neighbors and initialize its uncertainty factor λ^i to zero. To optimize the discriminative loss (Equation 10), we follow the same settings as in [55]. We train the network using Adam optimizer [26] with batch size 32 for 30 epochs with an initial learning rate of 0.001. The parameters δ in Equation 2 and γ in Equation 11 are set to 0.5 and 0.1 based on validation results. Similar to previous works [5, 60, 52], we use the overall accuracy as the metric to evaluate the models.

4.3. Experiments with Noisy Labels

The two main aspects of annotation ambiguity in FER are noisy labels and uncertain visual features [52]. In particular, it can be difficult for people to accurately recognize the emotions on ambiguous facial images, which can result in noisy and incorrect labels. Therefore, we conduct experiments to study the robustness of our LDLVA on mislabelled

Table 1: Accuracy with synthetic noise.

Noise ratio	Method	Accuracy (%)		
		AffectNet	RAF-DB	SFEW
10%	Baseline	60.14 ± 0.23	83.28 ± 0.45	45.98 ± 0.93
	SCN [52]	61.57 ± 0.15	84.65 ± 0.32	49.51 ± 0.76
	RUL [60]	62.89 ± 0.13	86.24 ± 0.22	47.82 ± 1.32
	LDLVA (ours)	**64.37 ± 0.11**	**87.98 ± 0.10**	**53.33 ± 0.57**
20%	Baseline	58.37 ± 0.35	81.89 ± 0.61	41.25 ± 1.12
	SCN [52]	60.83 ± 0.19	83.21 ± 0.49	46.26 ± 1.24
	RUL [60]	61.74 ± 0.18	84.49 ± 0.24	44.78 ± 1.04
	LDLVA (ours)	**63.89 ± 0.14**	**86.81 ± 0.12**	**51.53 ± 0.92**
30%	Baseline	56.94 ± 0.43	78.92 ± 0.59	38.51 ± 1.69
	SCN [52]	58.80 ± 0.32	80.61 ± 0.54	43.28 ± 2.06
	RUL [60]	60.77 ± 0.15	82.59 ± 0.42	41.79 ± 0.81
	LDLVA (ours)	**62.57 ± 0.15**	**85.85 ± 0.09**	**50.3 ± 0.88**

data by adding synthetic noise to AffectNet, RAF-DB, and SFEW datasets. More specifically, we randomly flip the manual labels to one of the other categories. Three levels of noise are studied in our experiment. We quantitatively evaluate our method and compare with the baseline ResNet-50 [21] and recent noise-tolerant FER methods including SCN [52] and RUL [60].

We perform each experiment three times and report the mean accuracy and standard error in Table 1. The results clearly show that our method consistently outperforms other approaches in all cases. Particularly, our model makes significant improvements over the baseline with an average accuracy margin of 5.13%, 5.52%, and 9.81% on the AffectNet, RAF-DB, and SFEW datasets, respectively. We also observe that the improvements are even more apparent when the noise ratio increases, for example, the accuracy improvement on RAF-DB is 4.7% with 10% noise and 6.93% with 30% noise. The consistent results under various settings demonstrate the ability of our method to effectively deal with noisy annotation, which is crucial in the robustness against label ambiguity.

4.4. Experiments with Inconsistent Labels

Table 2: Accuracy with inconsistent labels.

Method	Accuracy (%)			
	AffectNet	RAF-DB	SFEW	Average
AIR [2]	54.23	67.37	49.88	57.16
NAL [20]	55.97	84.22	58.13	66.11
IPA2LT [58]	57.85	83.80	53.15	64.93
LDL-ALSG [5]	58.29	85.33	55.87	66.50
LDLVA (ours)	**62.89**	**87.26**	**58.70**	**69.62**

Due to the ambiguous nature of facial expressions, different individuals can assign different labels for the same image as illustrated in Figure 1. Since the annotations for large-scale FER data are commonly obtained via crowdsourcing, this can create label inconsistency, especially between different datasets. Therefore, to examine the effectiveness of the proposed methods in dealing with this problem, we follow the cross-dataset protocol in previous state-of-the-art methods [5, 58] and adopt the experimental settings as proposed in [5] for a fair comparison. Specifically, the model is trained using the joint training dataset from RAF-DB and AffectNet. The resulting model is then tested on all three RAF-DB, AffectNet, and SFEW datasets.

Table 3: Accuracy of different methods on original datasets.

Method	Accuracy (%)		
	AffectNet	RAF-DB	SFEW
Island Loss [4]	-	-	52.52
IPFR [51]	57.40	-	55.10
EfficientFace [66]	63.70	88.36	-
DACL [15]	65.20	87.78	-
MViT [32]	64.57	88.62	-
RAN [53]	-	86.90	56.4
SCN [52]	-	87.03	-
DMUE [45]	-	88.76	57.12
RUL [60]	-	88.98	-
PSR [50]	63.37	88.98	-
LDLVA (ours)	**66.23**	**90.51**	**59.90**

Table 2 reports the results of our experiments. Our method achieves the best performance on all three datasets and the highest average accuracy. Notably, LDLVA surpasses the current state-of-the-art LDL-ALSG [5] with an improvement of 3.12% on average accuracy. Compared to our approach, LDL-ALSG only uses the neighbors' distributions to constrain the network prediction without constructing a label distribution for the center instance. It also lacks a mechanism to adaptively measure the contribution of each neighbor and the uncertainty of the provided annotation. The favorable performance confirms the advantages of our method over previous works and demonstrates the generalization ability to data with label inconsistency, which is essential for real-world FER applications.

4.5. Experiments on Original Datasets

We further perform experiments on the original AffectNet, RAF-DB, and SFEW to evaluate the robustness of our method to the uncertainty and ambiguity that unavoidably exists in real-world FER datasets. We compare the proposed LDLVA with several state-of-the-art methods in Table 3. By leveraging label distribution learning on valence-arousal space, our model outperforms other methods and achieves state-of-the-art performance on AffectNet, RAF-DB, and SFEW. Although these datasets are considered to be "clean", the results suggest that they indeed suffer from uncertainty and ambiguity.

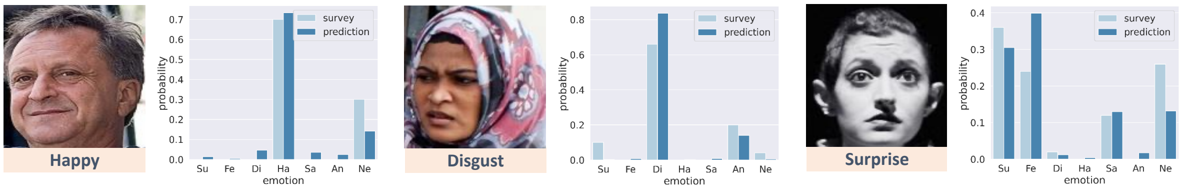

Figure 3: Comparison of the results from our survey and our model. More results can be found in the supplementary materials.

4.6. Qualitative Analysis

Real-world Ambiguity. To understand more about real-world ambiguous expressions, we conducted a user study in which we asked 50 participants to choose the most clearly expressed emotion on random test images from RAF-DB and AffectNet datasets. The numbers of votes per class are normalized to obtain the emotion distribution. We compare our model's predictions with the survey results in Figure 3. We can see that these images are ambiguous as they express a combination of different emotions, hence the participants do not fully agree and have different opinions about the most prominent emotion on the faces. It is further shown that LDLVA can give consistent results and agree with the perception of humans to some degree, which suggests that our model can effectively address the ambiguity problem in facial expressions.

Adaptive Similarity. Figure 4 presents the normalized calibration scores of different neighbors with respect to the center instance computed by our adaptive similarity mechanism. It can be seen that some neighbors may look visually similar to the central image but do not express the same emotion. By giving low calibration values, our method can effectively suppress the negative influence of these neighbors and lower their contribution, hence resulting in a more robust and accurate estimation of the emotion distribution.

Constructed Label Distribution. In Figure 5, we visualize the emotion distributions reconstructed by our method on mislabelled images. Despite the incorrect annotations, our approach is able to construct plausible distributions and discover the correct labels. It is also noteworthy that some expressions manifest multiple emotions rather than only the single provided category, which means the discovered distribution can provide more supervision for training.

Uncertainty Factor. Figure 6 shows the estimated uncertainty factors of some training images in the RAF-DB dataset and their original labels. The uncertainty values decrease from top to bottom. Highly uncertain labels can be caused by low-quality inputs (as shown in `Angry` and `Surprise` columns) or ambiguous facial expressions. In contrast, when the emotions can be easily recognized as those in the last row, the uncertainty factors are assigned low values. This characteristic can guide the model to decide whether to put more weights on the provided label or the neighborhood information. Therefore, the model can be more robust against uncertainty and ambiguity.

Table 4: Component analysis (LD: Label Distribution, AS: Adaptive Similarity, UF: Uncertainty Factor, DL: Discriminative Loss)

Setting	LD	AS	UF	DL	RAF-DB (original)	RAF-DB (30% noise)
`(i)`	-	-	-	-	87.06	78.92
`(ii)`	✓	✓	-	-	88.95	82.69
`(iii)`	✓	✓	✓	-	89.57	84.38
`(iv)`	✓	-	✓	✓	89.31	83.56
`(v)`	✓	✓	✓	✓	90.51	85.85

4.7. Ablation Study

Contribution of Each Component. In Table 4, we present the accuracy corresponding to different combinations of our components: label distribution (whether to construct $\boldsymbol{d}^i$ or not), adaptive similarity (whether to compute calibration scores or directly use local similarity scores as contribution degrees), uncertainty factor (whether to use separate λ^i for each instance or share a fixed value λ for all training samples), and discriminative loss (whether to incorporate $\mathcal{L}_D$ in Equation 11 or not). By employing label distribution with adaptive similarity `(ii)`, we can significantly improve the accuracy of the vanilla approach `(i)` by 1.89% on original RAF-DB and 3.77% on 30%-noise RAF-DB. Further integrating uncertainty factor and discriminative loss consistently boost the performance of the model, as shown in the results of `(iii)` and `(v)`, respectively. The results show the effectiveness of each component as well as the advantages of their combination in our LDLVA method.

Number of Nearest Neighbors. We present the effect of the number of nearest neighbors K on the model performance in Figure 7. For original RAF-DB data, higher values of K give better results but also require more training time. In particular, our training time with $K = 8$ and $K = 16$ on AffectNet is 12 hours and 20 hours, respectively. Under noisy conditions, the best result is obtained with $K = 8$ while larger or smaller K can lead to slightly worse performance. The reason is that using a large K might include more corrupted labels while using too few neighbors can limit the amount of exploitable information.

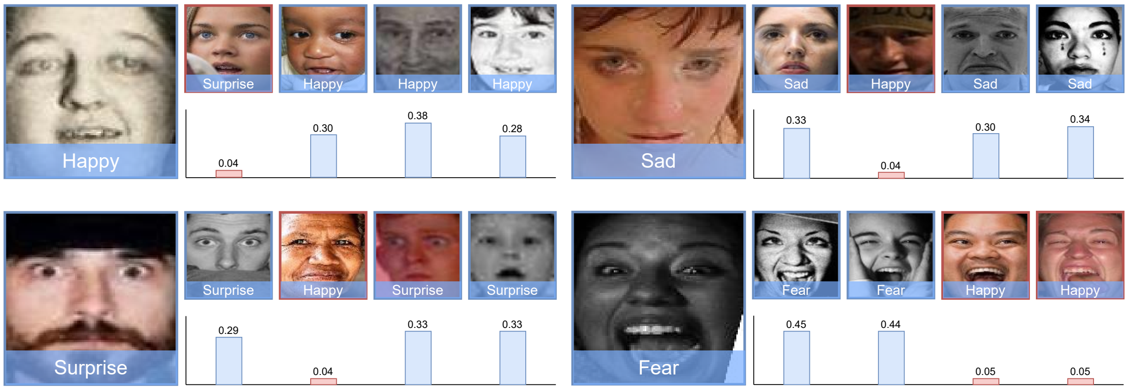

Figure 4: The calibration scores of neighbor images with respect to the main instance. The large image on the left is the main instance. The neighbor images are shown at the top, and their corresponding scores are shown at the bottom.

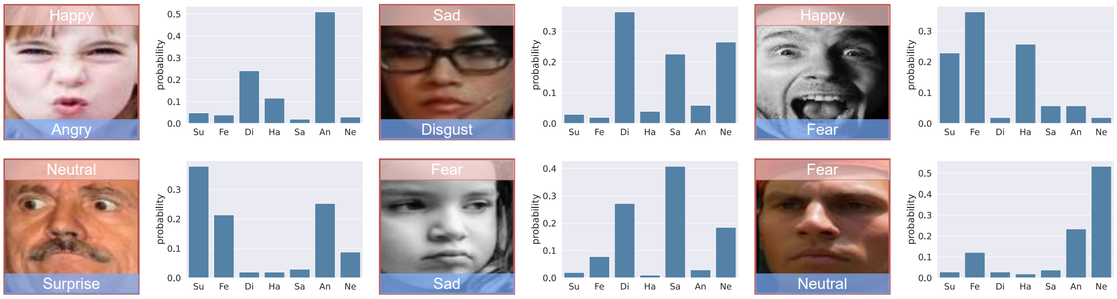

Figure 5: Examples of the emotion distribution recovered by our method when the dataset is contaminated with noisy labels. The label on top of each image is the synthetic noisy label and the bottom denotes the human annotation.

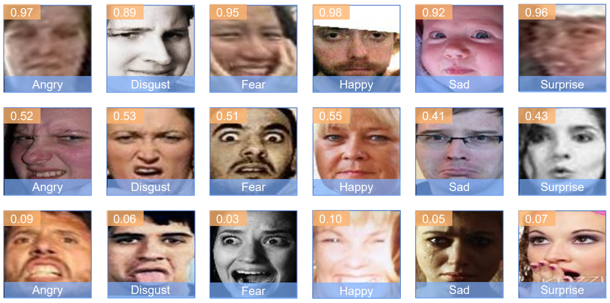

Figure 6: Visualization of uncertainty values of some examples from RAF-DB dataset.

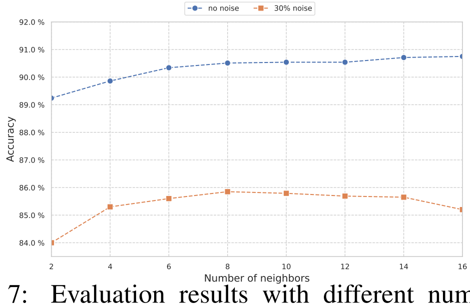

Figure 7: Evaluation results with different numbers of neighbors.

5. Conclusion

This paper introduces a new label distribution learning method for facial expression recognition by leveraging structure information in the valence-arousal space to recover the intensities distributed over emotion categories. We first employ the adaptive similarity to account for the errors caused by pseudo valence-arousal and robustly measure the contribution degree of each neighbor. Then, the target label distribution is constructed by incorporating both the provided single label and the combination of neighbor distribution guided by the uncertainty value. The constructed label distribution provides rich information about the emotions, thus can effectively describe the ambiguity degree of the facial image. Intensive experiments on popular datasets demonstrate the effectiveness of our method over previous approaches under inconsistency and uncertainty conditions in facial expression recognition.

Acknowledgement. The authors want to express our appreciation to Dr. Thieu Vo for his valuable feedback during the preparation of this paper. We also want to thank all volunteers that participated in our user study.

References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, 2017.

[2] Samaneh Azadi, Jiashi Feng, Stefanie Jegelka, and Trevor Darrell. Auxiliary image regularization for deep CNNs with noisy labels. In *ICLR*, 2016.

[3] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016.

[4] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018.

[5] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, 2020.

[6] Elizabeth A. Clark, J'Nai Kessinger, Susan E. Duncan, Martha Ann Bell, Jacob Lahne, Daniel L. Gallagher, and Sean F. O'Keefe. The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review. *Frontiers in Psychology*, 2020.

[7] Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 2008.

[8] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 2001.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[10] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *ICCVW*, 2011.

[11] Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Fine-grained visual classification using self assessment classifier. *arXiv preprint arXiv:2205.10529*, 2022.

[12] P Ekman and WV Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 1971.

[13] Neamat El Gayar, Friedhelm Schwenker, and Günther Palm. A study of the robustness of knn classifiers trained using soft labels. In *Artificial Neural Networks in Pattern Recognition*. Springer Berlin Heidelberg, 2006.

[14] Amir Hossein Farzaneh and Xiaojun Qi. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In *CVPRW*, 2020.

[15] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *WACV*, 2021.

[16] B. Gao, C. Xing, C. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 2017.

[17] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*.

[18] Xin Geng, Qin Wang, and Yu Xia. Facial age estimation by adaptive label distribution learning. In *ICPR*, 2014.

[19] Xin Geng and Yu Xia. Head pose estimation based on multivariate label distribution. In *CVPR*, 2014.

[20] J. Goldberger and E. Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[22] Peng Hou, Xin Geng, and Min-Ling Zhang. Multi-label manifold learning. In *AAAI*, 2016.

[23] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *CVPR*, 2019.

[24] Xiufeng Jiang, Zhang Yi, and Jian Lv. Fuzzy svm with a new fuzzy membership function. *Neural Computing and Applications*, 2006.

[25] Corentin Kervadec, Valentin Vielzeuf, Stéphane Pateux, Alexis Lechervy, and Frédéric Jurie. Cake: Compact and accurate k-dimensional representation of emotion. In *BMVC*, 2018.

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[27] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. In *BMVC*, 2019.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[29] Yi Kun and Wu Jianxin. Probabilistic End-to-end Noise Correction for Learning with Noisy Labels. In *CVPR*, 2019.

[30] Nhat Le, Khanh Nguyen, Anh Nguyen, and Bac Le. Global-local attention for emotion recognition. *Neural Computing and Applications*, pages 1–15, 2021.

[31] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, 2019.

[32] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvt: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021.

[33] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

[34] Yu-Kun Li, Min-Ling Zhang, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *IEEE International Conference on Data Mining*, 2015.

[35] Christine Lisetti, Fatma Nasoz, Cynthia Lerouge, Onur Ozyer, and Kaye Alvarez. Developing multimodal intelligent affective interfaces for tele-home health care. *International Journal of Human-Computer Studies*, 2003.

[36] D. Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 1992.

[37] 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*, 2019.

[38] Binh X Nguyen, Binh D Nguyen, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Graph-based person signature for person re-identifications. In *CVPRW*, 2021.

[39] Koichiro Niinuma, Laszlo A. Jeni, Itir Onal Ertugrul, and Jeffrey F. Cohn. Unmasking the devil in the details: What works for deep facial action coding? In *BMVC*, 2019.

[40] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: a loss correction approach. In *CVPR*, 2017.

[41] Silke Paulmann, Martin Bleichner, and Sonja A. Kotz. Valence, arousal, and task effects in emotional prosody processing. *Frontiers in Psychology*, 2013.

[42] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980.

[43] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[44] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 2009.

[45] 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 *CVPR*, 2021.

[46] Kai Su and Xin Geng. Soft facial landmark detection by label distribution learning. In *AAAI*, 2019.

[47] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir D. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv*, 2014.

[48] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, 2018.

[49] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021.

[50] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 2020.

[51] Can Wang, Shangfei Wang, and Guang Liang. Identity-and pose-robust facial expression recognition through adversarial feature learning. In *ACM International Conference on Multimedia*, 2019.

[52] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, 2020.

[53] 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*, 2020.

[54] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

[55] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[56] Ning Xu, An Tao, and Xin Geng. Label enhancement for label distribution learning. In *IJCAI*, 2018.

[57] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *CVPR*, 2018.

[58] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, 2018.

[59] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *CVPR*, 2019.

[60] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. In *NIPS*, 2021.

[61] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. *arXiv preprint arXiv:2207.10299*, 2022.

[62] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *CVPR*, 2019.

[63] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NIPS*, 2018.

[64] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, 2018.

[65] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *AAAI*, 2021.

[66] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *AAAI*, 2021.

[67] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *ACM International Conference on Multimedia*, 2015.

[68] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-Supervised Learning with Graphs*. PhD thesis, USA, 2005. AAI3179046.