Soft Label Mining and Average Expression Anchoring for Facial Expression Recognition

Haipeng Ming, Wenhuan Lu, and Wei Zhang*

Tianjin University, Tianjin, China
{minghaipeng,wenhuan,tjuzhangwei}@tju.edu.cn

Abstract. Facial expression recognition (FER) suffers from high interclass similarity and large intraclass variation, leading to ambiguity or uncertainty and further confusing annotators. They also hinder the network in learning the valuable features of facial expression. Recently, many studies have revealed that the uncertainty or ambiguity is one of the key challenges in FER. In this paper, we propose a new method to address this issue from two aspects: a soft label mining module to convert the original hard labels to soft labels dynamically during training, and an average facial expression anchoring module to separate unique expression features from similarity expression features. The soft label mining module breaks the limits of the categorical model and mitigates the uncertainty or ambiguity. And the average facial expression anchoring module suppresses the high interclass similarity of facial expressions. Our method can train any backbone network for facial expression recognition. The experiments on the popular datasets show that our method achieves state-of-the-art results by 92.82% on RAF-DB and 67.91% on SFEW, and achieves a comparable result of 62.26% on AffectNet. The code is available at *https://github.com/HaipengMing/SLM-AEA*.

1 Introduction

Facial expression is one of the most natural, powerful, and universal signals for human beings to convey their emotional states and intentions [6, 32]. In the past years, facial expression recognition (FER) has attracted much attention due to its important role in human-computer interaction, health care, and many other applications. Similar to other modalities in affective computing, a facial expression is commonly characterized as one of several discrete affective states(*e.g.*, basic emotions defined by Ekman and Friesen [10, 9]), which is also known as the categorical model. Generally, annotating facial expressions with the categorical model is much easier and cheaper than other models (FACS [11] and dimension models [26]). It's more consistent with people's intuition as well. Existing facial expression datasets are mostly annotated with categorical model, such as Oulu-CASIA [44], SFEW/AFEW [7], FERPlus [1], RAF-DB [16], AffectNet [20] (AffectNet also annotated Valence-Arousal dimensions), etc.

* Corresponding Author.

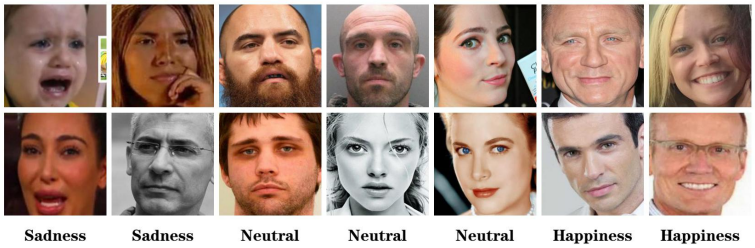

Fig. 1. An illustration of the uncertainty or ambiguity. These 14 images are from AffectNet and their labels are attached to the bottom of the images. From left to right, sadness, neutral, happiness. But there is no clear boundary between sadness and neutral as well as neutral and happiness.

However, the categorical model is limited in the ability to represent the complexity and subtlety of facial expressions, especially in the wild. As illustrated in Fig. 1, there is not usually a clear boundary between the different expression categories. Even worse, due to the high interclass similarity, annotators may even annotate incorrectly. This means that the quality and consistency of the datasets are difficult to guarantee because of the subjectivity of the annotators. Compared to discrete affective states, the soft label has a greater expressivity, which can describe ambiguity appropriately. High interclass similarity can be described as other expression components, while high intraclass variation can be described as different extensions of the real expression component. However, it is expensive and time-consuming to provide soft labels for large-scale FER datasets. A compromised way is soft label mining. Based on [41], we designed a simple yet efficient *soft label mining module* to mine soft labels from original annotations, *i.e.*, the hard labels. Specifically, we introduce the label smoothing method to initially transfer the hard labels to soft labels according to an artificially set value p, which reflects the confidence level of the original annotations. The initialized soft labels act as targets to train the network parameters. We update them by taking into account both the network predicted distribution and the original distribution(*i.e.*, the hard labels) during training. To make a trade-off between them, we design a ramp function to achieve a balance dynamically. Note that the soft label mining module is designed for training, imposes no additional burden on inference, and adds only a very small additional burden to training. Meanwhile, we also design a novel *average facial expression anchoring module* to suppress the high interclass similarity. Specifically, we take the average feature of a mini-batch as the anchor feature, *i.e.*, the average expression, and introduce a learnable vector with the same size as the attention weight of the anchor feature. The weighted anchor features are further summed with the expression features as the final extracted features. We suppress the uncertainty or ambiguity in FER through these two modules, especially the soft label mining module, which has an excellent and stable performance on different FER datasets. Our method can be used to train any backbone network for facial expression recognition.

Overall, the main contributions can be summarized as follows:

(1) We propose a novel method to address the uncertainty or ambiguity problem in FER by designing a soft label mining module and an average expression anchoring module. In comparison with existing related work, our approach is simpler and more efficient.

(2) Our approach is extensively evaluated on laboratory databases and real-world datasets. Experimental results show that our method achieves the state-of-the-art performance by 92.82% on RAF-DB and 67.91% on SFEW, and achieves a comparable result of 62.26% on AffectNet.

2 Related work

2.1 Facial Expression Recognition

Facial expression recognition has been an active topic for many years. Early work focused on using *handcraft features* (*e.g.*, SIFT [22], HOG [5], LBP [27], Gabor Wavelets [2].etc) to exact the emotion feature and recognize facial expressions. While with the development of deep learning, now the *learning-based* methods have become mainstream, which can be roughly classified into three groups: data-focused [21, 42], model-focused [3, 40, 34, 25] and label-focused [4, 35, 28]. The results observed by Ng *et al.* [21] show that pre-fine-tuning on an additional FER dataset can improve the performance. Zeng *et al.* [42] propose a new model termed IPA2LT to address the inconsistency issue in fusion of different FER datasets. Yang *et al.* [40] and Wang *et al.* [34] introduce adversarial mechanism into FER. Ruan *et al.* [25] propose a novel model to take into account subtle differences between different facial expressions. Recently, some researchers began to consider the uncertainty of annotations in FER datasets. Chen *et al.* [4] introduce label distribution learning and draw on an auxiliary space of FACS or landmarks. Wang *et al.* [35] propose a Self-Cure Network (SCN) to suppress the uncertainty by correcting possible mislabeling. She *et al.* [28] proposed a model named DMUE, which achieved previous leading performance, to mine latent distribution and estimation uncertainty. Zhao *et al.* [45] propose a lightweight FER network considering both model and label.

2.2 Methods for Uncertainty or Ambiguity

Low-quality annotations caused by uncertainty or ambiguity can be considered as label noise, which also appears in other computer vision tasks. Numerous methods have been proposed to resolve this issue. Some methods leverage a small set of clean data [33, 17]. Qu *et al.* [23] proposed a label-noise robust network by matching the feature distributions. In the field of FR (Face Recognition), a GCN-based model [43] is proposed to address the large-scale label noise. Recently, label distribution has been proposed to mitigate the adverse impact of label noise by converting logical labels to discretized bivariate Gaussian label distribution with the help of prior knowledge [12, 13, 31]. In the classification problem, label

distribution means soft label. Label enhancement (LE) [39, 38] is the universal way to find the latent ground truth. Uncertainty estimation methods can be used to [30, 37] address the inconsistent data quality. The uncertainty or ambiguity in FER is intrinsic. Compared to other classification tasks, facial expressions of different classes suffers from high interclass similarity but also high intraclass variation.

3 Method

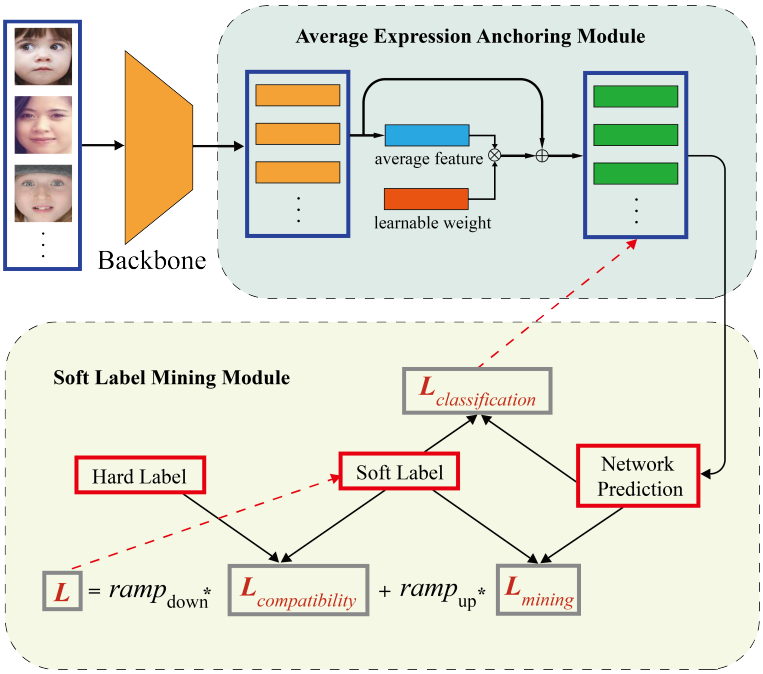

Fig. 2. The overall framework of our method. Face images are first fed into a backbone network for feature extraction. The extracted features are then averaged to obtain the anchor feature, further multiplied element-wise by a learnable weight vector. The obtained results and the extracted initial features are added element-wise as the final extracted features of the network. We view the soft label as the target to train the network. When training, the soft labels are updated dynamically with a newly designed loss function. The solid line in the figure represents forward propagation while the dashed line represents back-propagation.

Notation. Given a FER dataset $\mathcal{X}$, we donate its corresponding hard label space as $\mathcal{H} = \{\boldsymbol{y} : \boldsymbol{y} \in \{0,1\}^C, \|\boldsymbol{y}\|_1 = 1\}$, and the soft label space as $\mathcal{S} = \{\boldsymbol{y} : \boldsymbol{y} \in [0,1]^C, \|\boldsymbol{y}\|_1 = 1\}$, where C is the number of classes. For the i-th

facial image $\boldsymbol{x}_i$ belonging to $\mathcal{X}$, we denote $\hat{\boldsymbol{y}}_i = \{\hat{y}_{i1}, \hat{y}_{i2}, ..., \hat{y}_{iC}\} \in \mathcal{H}$ as its annotated deterministic hard label, which is a one-hot vector. The output vector of backbone network is donated as $\boldsymbol{z}_i = \{z_{i1}, z_{i2}, ..., z_{iC}\}$.

3.1 Soft Label Mining

We first transfer the hard labels to soft labels with the help of label smoothing [15], which is a trick to improve the performance in classification tasks. A hyperparameter p is introduced for initialization. The more uncertain or ambiguous the FER dataset is, the lower the p is set. As for hard labels, p is 1. We take the initialized soft labels as the new target to train the network. While training, we update the soft labels dynamically.

Specifically, in a C-classes classification task, the process of prediction with the model can be formulated as:

$$\boldsymbol{z}_i = f(\boldsymbol{x}_i; \boldsymbol{\theta}) \tag{1}$$

where f is a model and $\boldsymbol{\theta}$ is the set of network parameters. Then the predicted result of the model f will be normalized by softmax function.

$$\bar{\boldsymbol{z}}_i = softmax(\boldsymbol{z}_i) \tag{2}$$

$$\bar{z}_{ij} = \frac{exp(z_{ij})}{\sum_{j=1}^{C} exp(z_{ij})} \tag{3}$$

Obviously, $\bar{\boldsymbol{z}}_i$ belongs to $\mathcal{S}$. It can be viewed as the label distribution generated by model f. Generally, the purpose of training is to minimize the cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \hat{y}_{ij} log \bar{z}_{ij} \tag{4}$$

For the i-th image belongs to c-th class,

$$\mathcal{L}_i = -z_{ic} + log(\textstyle\sum_{j=1}^{C} exp(z_{ij})) \tag{5}$$

Its optimal solution is $z_{ic}^* = \boldsymbol{inf}$, while keeping others small enough. But many ambiguous facial expressions have similar intensity in different emotional components. It is not reasonable to force other components small enough. The idea of label smoothing is to transfer $\hat{\boldsymbol{y}}_i \in \mathcal{H}$ to $\bar{\boldsymbol{y}}_i \in \mathcal{S}$ by changing the construction as:

$$\bar{y}_{ij} = \begin{cases} 1 - \varepsilon & \text{if } j = c, \\ \varepsilon / (C - 1) & otherwise \end{cases} \tag{6}$$

where ε is a small constant. $\bar{\boldsymbol{y}}_i$ is a label distribution. We donate its predecessor before the softmax operator as $\tilde{\boldsymbol{y}}_i$, its components are computed as:

$$\tilde{y}_{ij} = \begin{cases} k & \text{if } j = c, \\ k - log\frac{(1-\varepsilon)(C-1)}{\varepsilon} & otherwise \end{cases} \tag{7}$$

where k is an arbitrary number. It determines the data scale before softmax. But when ε is fixed, k has no effect on $\bar{\boldsymbol{y}}$, that is, the distribution of $\bar{\boldsymbol{y}}$ in $\mathcal{S}$ is only related to ε. In our experiments, k is set to 5. The $\tilde{\boldsymbol{y}}_i$ is used as the score-form soft labels and updated with a learning rate during training.

Note that we use p ($p = 1-\varepsilon$) instead of ε as the hyperparameter. It is because p responds to the confidence level of the original annotations. Our experimental results show that in the synthetic noise datasets, p should be adjusted to lower as the noise ratio increases.

As mentioned earlier, $\tilde{\boldsymbol{y}}_i$ is used as the score-form soft label. The cross entropy loss function in Equation 4 now is:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\bar{y}_{ij}log\bar{z}_{ij} \tag{8}$$

$$\bar{\boldsymbol{y}}_i = softmax(\tilde{\boldsymbol{y}}_i) \tag{9}$$

$\mathcal{L}_{cls}$ is the loss function to update the network parameter $\boldsymbol{\theta}$. We update $\tilde{\boldsymbol{y}}_i$ and $\boldsymbol{\theta}$ in two continuous but different backpropagation stages. After $\boldsymbol{\theta}$ has been updated, $\tilde{\boldsymbol{y}}_i$ is regarded as the learnable parameters. Both [35] and [28] show that the network prediction $\bar{\boldsymbol{z}}_i$ of some ambiguous facial expressions is more credible than the annotations during training. So an intuitive approach is to take full advantage of $\bar{\boldsymbol{z}}_i$.

$$\mathcal{L}_{m} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\bar{z}_{ij}log\bar{y}_{ij} \tag{10}$$

We mine the soft label by Equation 10. But certain samples should benefit from the original annotations, so we also utilize another loss function (Equation 11) to explore the compatibility between the original label and the latent label distribution.

$$\mathcal{L}_{cpt} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\hat{y}_{ij}log\bar{y}_{ij} \tag{11}$$

We introduce a ramp function [28] to adjust the weight between different loss functions. The overall loss function guiding $\tilde{\boldsymbol{y}}_i$ to update is:

$$\mathcal{L} = ramp_{up}(e)\cdot\mathcal{L}_{m} + ramp_{down}(e)\cdot\mathcal{L}_{cpt} \tag{12}$$

$$ramp_{up}(e) = \begin{cases} exp(-(\alpha - \frac{e}{\beta})^2) & e \leq \beta, \\ 1 & e > \beta \end{cases} \tag{13}$$

$$ramp_{down}(e) = \begin{cases} 1 & e \leq \beta, \\ exp(-(\alpha - \frac{\beta}{e})^2) & e > \beta \end{cases} \tag{14}$$

where α and β are hyperparameters. α is used to adjust the slope of ramp function, which is fixed at 1 in [28]. β is the epoch threshold. In the training process, the network trusts the original annotations more before β-th epoch, and trusts itself more after β-th epoch.

$$\tilde{\boldsymbol{y}} \leftarrow \tilde{\boldsymbol{y}} - \lambda \frac{\partial \mathcal{L}}{\partial \tilde{\boldsymbol{y}}} \tag{15}$$

The backpropagation of $\mathcal{L}$ takes another learning rate λ to update $\tilde{\boldsymbol{y}}$ as shown in Equation 15. Note that λ has a strong relationship with k in Equation 7. k decides the scale of score-form soft label, while λ is the learning rate to update it. In our experiment, k is fixed to 5 and λ is fixed to 200. Different k has different optimal λ, but once a pair of k and λ is determined, there is no need to change it. For different datasets, only the hyperparameter p, α and β need to be adjusted.

3.2 Average Expression Anchoring

To handle the high interclass similarity in FER, we design an average expression anchoring module to separate the common feature and the unique feature. We share a similar mindset with a small part of [29].

In general, the features extracted from backbone networks are processed by a fully-connected layer to get $\boldsymbol{z}_i$:

$$\boldsymbol{z}_i = \boldsymbol{W}^T \boldsymbol{F}_i \tag{16}$$

where $\boldsymbol{F}_i \in \boldsymbol{R}^D$ and belongs to feature space.

The average expression anchoring module is comprised of a learnable attention weight $\tilde{\boldsymbol{W}} \in \boldsymbol{R}^D$, and an element-wise summation operation layer which can be donated as:

$$\tilde{\boldsymbol{F}}_i = \tilde{\boldsymbol{W}} * \bar{\boldsymbol{F}} + \boldsymbol{F}_i \tag{17}$$

$$\bar{\boldsymbol{F}} = \frac{1}{b} \sum_{i=1}^{b} \boldsymbol{F}_i \tag{18}$$

where $*$ means element-wise multiplication and $+$ means element-wise summation, b is the batch size of the train set.

The average expression anchoring module replaces $\boldsymbol{F}_i$ with $\tilde{\boldsymbol{F}}_i$, and we think it addresses the uncertainty or ambiguity problem from two aspects. First, because the expressions of the different categories have very similar information, which disturbs the network to distinguish, it is necessary to separate the unique features from the common features. $\bar{\boldsymbol{F}}$ represents the highly similar information while $\boldsymbol{F}_i$ represents the special information. Experimental results show that the network advantages of the separation. Second, for low-quality samples, anchoring the average expression allows them to take advantage of information from high-quality samples in the same batch, thus improving the overall quality of the dataset.

Note that the improvements of this module depends on “average expression", *i.e.*, $\bar{\boldsymbol{F}}$. That means it will not work if inferencing a single expression rather than a batch of expressions.

3.3 The overall framework

As illustrated in Fig. 2, given a FER dataset with hard labels, we first evaluate the confidence level of the dataset annotations by p. Then we fix p and soften the hard labels by label smoothing to get the initialized soft labels. The initialized soft labels are used as the target to train the backbone networks. An average expression anchoring module is introduced at the end of the backbone to mitigate the high interclass similarity in FER, by separating the unique features from common features. The soft labels are further updated dynamically during training by minimizing multiple loss functions. We force them to approximate both the network prediction and the original annotations by $\mathcal{L}_{cls}$ and $\mathcal{L}_m$. We designed different weights to make the network trust original annotations more in the beginning and trust itself more as the training progressed. Not that we use not only traditional CNNs but also Swin-transformers [18] as our backbones to extract facial expression features. The effects of different backbone networks will be compared in Section 4.

4 Experiments

4.1 Datasets

Oulu-CASIA [44] contains videos captured in controlled lab conditions. Subjects were asked to pose six basic expressions (happiness, surprise, sadness, anger, disgust, fear). We select the last three frames in each sequence in the condition of the visible light and strong illumination (consisting of 1,440 images in total), to construct the training set and the test set. Similar to [35], we employ the subject-independent tenfold cross-validation protocol for evaluation.

RAF-DB [16] is the real-world facial expression dataset, and contains 15,339 facial images annotated with six basic expressions and neutral expression. Among them, 12,271 images are used for training, and the other 3,068 images for testing. We report the overall sample accuracy of the testing set for measurement.

AffectNet [20] is by far the largest FER dataset collected in the wild. It was annotated with both categorical and Valence-Arousal dimension labels. It contains more than 100,000 images from the Internet by querying 1,250 expression-related keywords in three search engines, of which 450,000 images are manually annotated with seven expression classes the same as RAF-DB and the extra contempt expression. Among them, 280K images are used for training and the remaining 4K images for testing. The overall sample accuracy is used for measurement.

SFEW [7] is created by selecting static frames from Acted Facial Expressions in the Wild (AFEW). The images in SFEW are labeled with six basic expressions and neutral expression same as RAF-DB. We use 958 images for training and 436 images for testing. The overall sample accuracy is used for measurement.

4.2 Implementation Details

We take ResNet18 [14] pre-trained on MS-Celeb-1M as the default backbone network with the standard routine for a fair comparison. We also aligned the faces of the in-the-wild datasets for pose normalization. The facial expression images are resized to 256×256 pixels and further augmented by random cropping to 224×224 pixels, horizontally flipped, and added Gaussian noise with a probability of 0.5. As mentioned earlier, the parameter k and λ are set to 5 and 200, respectively. The hyperparameter α is set to 1.6. But the other parameters (β and p) need to be adjusted according to the datasets. p reflects the confidence level of the original annotations and β is the epoch that the backbone has learned enough knowledge so that it can trust its prediction. Considering carry out in practice, β is set to 7 and p is set to 0.9 can achieve a performance that greatly exceeds the baseline generally. We use Adam with a weight decay of 10^{-4}. The initial learning rate is 10^{-3}, which is further reduced to 10^{-6} as a cosine function. The training ends at epoch 40. We use the Pytorch toolbox to implement our model on a single Nvidia 2080Ti GPU and train it in an end-to-end manner.

4.3 Ablation Studies

We conduct ablation experiments to observe the effect of key parameters and different modules on the final performance. We choose RAF-DB and AffectNet as the benchmark since they are two of the popular largest real-world FER datasets.

Influence of different modules. In order to evaluate the contribution of different modules, we conduct an ablation study to investigate the soft label mining module and the average expression anchoring module on RAF-DB and AffectNet. Considering some related works did not pre-train the network on large scale face recognition datasets, we also investigate the effect of pertaining on MS-Celeb-1M. The experimental results are shown in Table 1. Some observations can be concluded. First, the soft label mining module improves the performance by 1.82% on RAF-DB and 3.86% on AffectNet without pertaining on MS-Celeb-1M, and improves the performance by 1.98% on RAF-DB and 4.31% on AffectNet with pre-training. It plays the most important role in our method due to its outstanding and stable improvements. Second, pre-trained on large scale face recognition datasets and the average expression anchoring module improves the accuracy on RAF-DB observably. But pre-training achieves very little improvement on AffectNet. The average expression anchoring module can even degrade the accuracy on AffectNet slightly. We think it can be explained by the consistency of the datasets. Due to the uncertainty or ambiguity problem in FER, larger scale FER datasets tend to be more inconsistent. At this point, the requirement for the soft mining module is even greater because it breaks the limitations of the categorical model. Third, pre-trained on large scale face recognition datasets and the union of two modules can improve the performance greater. Taking the result of pre-training on MS-Celeb-1M as a baseline, our

Table 1. Accuracy(%) comparison of different modules on RAF-DB and AffectNet. SLM and AEA are abbreviations for the soft label mining module and the average expression anchoring module, respectively. ✓ on "Pre-train" means ResNet18 is pre-trained on MS-Celeb-1M, while × means the initialized parameters of ResNet18 are provided by Pytorch, which is pre-trained on ImageNet.

Pre-train	SLM	AEA	RAF-DB	AffectNet
×	×	×	86.34	59.34
×	✓	×	87.91	61.63
×	×	✓	89.12	58.78
×	✓	✓	91.65	61.82
✓	×	×	87.59	59.44
✓	✓	×	89.32	62.00
✓	×	✓	91.99	59.29
✓	✓	✓	**92.82**	**62.26**

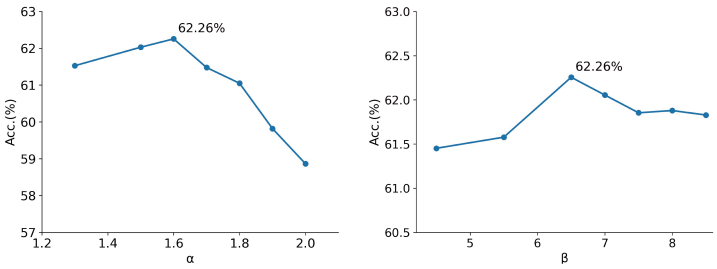

Fig. 3. The accuracy (%) with different α and β on AffectNet.

method improves the accuracy by 5.97% on RAF-DB and 4.74% on AffectNet finally.

Influence of α. α determines the slope and the ramp function's initial value. We introduced α in our model to improve the ramp function. This is because the original ramp function has a higher starting point, but we want to force the network to concentrate more on the original labels in the early phases of training. The experimental results show that the introduction of α helps to resolve this problem. Fig. 3 shows the influence of α. If α is set too small, the network's predictions at the beginning moments of training will have a greater impact on relabeling. If α is set too large, the slope of the ramp function will be too steep.

Influence of β. β is the epoch that the network starts to mine soft labels. Before β-th epoch, soft label mining depends more on the original hard label. But after β-th epoch, it depends more on the logits of the network. Theoretically, the lower the quality of the dataset, the lower the β. But too small β will harm learning enough useful features. Fig. 3 shows the impact of β on AffectNet, and proves this speculation well. Different datasets need different β. As shown in

Table 2, AffectNet has the minimum optimal β and RAF-DB has the maximum optimal β.

Table 2. Optimal β and p for different FER datasets.

	Oulu-CASIA	RAF-DB	AffectNet	SFEW
β	7	8	6.5	7
p	0.95	0.97	0.85	0.91

Optimal p of different FER datasets. We conduct plenty of experiments to test the optimal p of different datasets. As shown in Table 2, the result shows that the AffectNet has the minimum optimal p. This means that it may be the most uncertain or ambiguous of the four FER datasets. We think it can be explained by the scale of AffectNet. As early mentioned, larger-scale FER datasets tend to be more inconsistent.

Table 3. Accuracy(%) comparison of different backbone networks on RAF-DB and AffectNet. SLM-AEA is an abbreviation for our method. All the backbone networks are pre-trained on ImageNet for a fair comparison.

Backbone	SLM-AEA	RAF-DB	AffectNet
ResNet18 [14]	×	86.34	59.34
ResNet18 [14]	✓	91.65	61.82
ResNet50 [14]	×	85.39	59.04
ResNet50 [14]	✓	89.08	60.95
Swin Transformer [18]	×	87.61	59.74
Swin Transformer [18]	✓	90.63	62.17

Different backbone networks. Our framework can be used to train any backbone networks for facial expression recognition. We conduct experiments not only on CNNs but also on the recently popular transformer. We choose ResNet18, ResNet50, and Swin-transformer as the backbone networks for comparison. For a fair comparison, all the backbones are pre-trained on ImageNet. Pre-training on ImageNet makes the performance slightly inferior compared to pre-training on MS-Celeb-1M. As shown in Table 3, one interesting thing is that ResNet50 does not perform as well as ResNet18 on RAF-DB and AffectNet. In contrast, the performance of the Swin-transformer on AffectNet is a bit surprising. We believe the transformer has even more potential for facial expression recognition. However, on RAF-DB, the performance of the Swin-transformer decreases when our method is implemented.

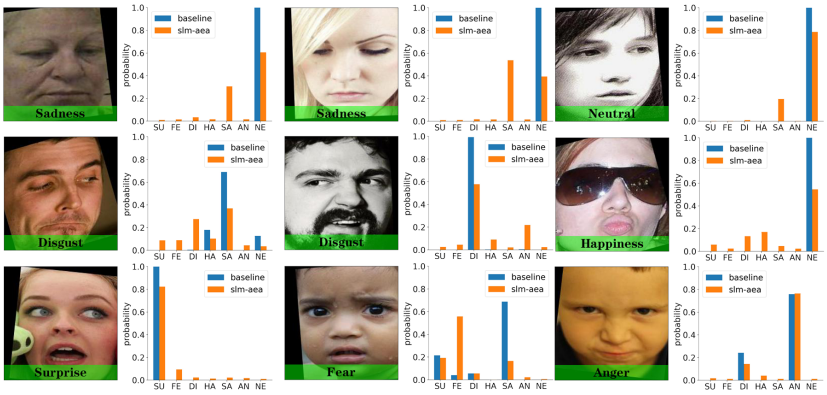

Fig. 4. Nine images and their predicted results from RAF-DB. There is some ambiguity in each image. The orange soft labels were mined by our model (Su: Surprise, Fe: Fear, Di: Disgust, Ha: Happy, Sa: Sad, An: Anger, Ne: Neutral). The blue logits were output from the baseline method.

4.4 Visualization Analysis

To demonstrate the superiority of our method, we analyze it visually in comparison with the baseline. We selected 9 images from RAF-DB with different degrees of ambiguity. As shown in Fig. 4, the label of each expression is attached to the green rectangle in the lower part of the image. Compared with the baseline, our method mined more latent ground truth. The rationality of the soft labels can be illustrated with the first two expressions labeled as sadness in Fig. 4. In the baseline method, both images are predicted to be “Neutral”, but our predictions have a larger weight on both “Neutral" and “Sadness”. Although our method predicts only one of them correctly, we think it is closer to the ground truth. The same conclusion can be observed in other images. Another interesting example is the fourth expression, it’s hard to determine its real label. And it seems that our model has some “confusion” with it, too. Despite the presence of many uncertain or ambiguous expressions in the FER dataset, we assume that the labels provided by the annotators are correct overall (at least for one component).

4.5 Evaluation on Synthetic datasets

One of the consequences of ambiguity is that annotators tend to mislabel expressions, which can be viewed as label noise. Label noise is usually classified into two types: symmetric label noise and asymmetric label noise. Asymmetric noise means that each expression can be incorrectly labeled as any other expression uniformly. We synthesize asymmetrical noise like [34], and quantitatively added 10%, 20%, and 30% noise to RAF-DB. We reproduced SCN [34] to compare with our method. To make a fair comparison, we use the same initialization parameters which pre-train on MS-Celeb-1M with the backbone of ResNet18. We

Table 4. Accuracy(%) comparison with the state-of-the-art method on synthetic noise RAF-DB datasets.

Method	n(%)	RAF-DB/best	RAF-DB/last
Baseline	10	82.88	81.64
SCN	10	86.77	86.40
SLM-AEA	10	**91.59**	**91.59**
Baseline	20	80.84	74.88
SCN	20	85.20	84.78
SLM-AEA	20	**89.86**	**89.57**
Baseline	30	79.19	63.88
SCN	30	82.69	82.34
SLM-AEA	30	**87.74**	**85.95**

Table 5. Comparison with the State-of-the-arts. $^{+}$ denotes both AffectNet and RAF-DB are used as the training set. * denotes oversampling is used since the train set of AffectNet is imbalanced.

Benchmark Dataset	Method	pre-trained Dataset	Acc.(%)
Oulu-CASIA	IPA2LT^{+} [42]	-	61.49
Oulu-CASIA	FN2EN [8]	2.6M face images	87.71
Oulu-CASIA	DeRL [40]	BU-4DFE & BP4D	88.00
Oulu-CASIA	DDL [24]	Multi-PIE	88.26
Oulu-CASIA	FDRL (ResNet18) [25]	MS-Celeb-1M	88.26
Oulu-CASIA	**SLM-AEA (ResNet18)**	**MS-Celeb-1M**	**88.61**
RAF-DB	LDL-ALSG^{+} [4]	-	85.53
RAF-DB	IPA2LT^{+} [42]	-	86.77
RAF-DB	SCN (ResNet18)$^{+}$ [35]	MS-Celeb-1M	88.14
RAF-DB	DMUE (ResNet18) [28]	MS-Celeb-1M	88.76
RAF-DB	FDRL (ResNet18) [25]	MS-Celeb-1M	89.47
RAF-DB	**SLM-AEA (ResNet18)**	**MS-Celeb-1M**	**92.82**
AffectNet	IPA2LT^{+} [42]	-	55.71
AffectNet	RAN* [36]	MS-Celeb-1M	59.50
AffectNet	SCN (ResNet18)* [35]	MS-Celeb-1M	60.23
AffectNet	CVT* [19]	MS-Celeb-1M	61.70
AffectNet	**SLM-AEA* (ResNet18)**	**MS-Celeb-1M**	**62.26**
AffectNet	DMUE (ResNet18)* [28]	MS-Celeb-1M	62.84
SFEW	RAN [36]	MS-Celeb-1M	56.40
SFEW	DMUE(ResNet18) [28]	MS-Celeb-1M	57.12
SFEW	IPA2LT^{+} [42]	-	58.29
SFEW	DDL [24]	Multi-PIE	59.86
SFEW	FDRL(ResNet18) [25]	MS-Celeb-1M	62.16
SFEW	**SLM-AEA (ResNet18)**	**MS-Celeb-1M**	**67.91**

report the mean accuracy of ten experiments. As shown in Table 4, our method outperforms both the baseline and the SCN for different ratios of label noise. That is, our method has better error correction capability. We believe that this error correction capability comes from representing expressions with soft labels rather than hard labels. Note that the value of p should be adjusted according to the synthesized noise ratio. With the noise ratio increasing, p should be tuned down slightly to achieve better performance.

4.6 Comparison with the State-of-the-arts

We compared our method with the state-of-the-arts as shown in Table 5. IPA2LT aims to address the inconsistency of different datasets. DDL is proposed to disentangle the disturbing factors in facial expression images. DeRL decomposes expressions into expressive components and unexpressed neutral components with the help of GAN. RAN is developed to deal with the occlusion and head pose in FER. LDL-ALSG introduces label distribution learning to FER and uses extra information from related tasks. SCN and DMUE are recently proposed and also committed to addressing the uncertainty or ambiguity problem of FER. [35] is the first paper to propose the uncertainty problem in FER. DMUE (ResNet50-IBN version) achieved previous leading performance. We choose version ResNet18 of the DMUE for a fair comparison. FDRL is also a recently proposed method and achieves leading performance on both the Oulu-CASIA and RAF-DB datasets. CVT introduced transformer into FER and also achieved state-of-the-art performance. Compared with these methods, we definitely achieve better performance and set new records on Oulu-CASIA, RAF-DB, and SFEW. SLM-AEA outperforms these state-of-the-art methods consistently with a huge margin on RAF-DB and SFEW. Although it did not outperform DMUE on AffectNet, a very competitive result was achieved and our method was simpler compared to DMUE.

5 Conclusion

In this paper, we proposed a novel method to address the uncertainty and ambiguity problem. Our method is composed of two main modules: the soft label mining module and the average expression anchoring module. The former is designed to break the limitations of the classification model by dynamically converting hard labels to soft labels, and the latter aims to alleviate the high interclass similarity. The soft label mining module plays the most important role in our method due to its outstanding and stable improvements. Compare with the state-of-the-art methods, our SLE-AEA is more simple yet effective. Experiments on popular benchmarks show that our method is extremely competitive.

Acknowledgements This work was supported by the Tianjin Municipal Science and Technology Program for New Generation of Artificial Intelligence (19ZXZNGX00030).

References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM International Conference on Multimodal Interaction. pp. 279–283 (2016)
2. Bazzo, J.J., Lamar, M.V.: Recognizing facial actions using gabor wavelets with neutral face average difference. In: International Conference on Automatic Face and Gesture Recognition. pp. 505–510 (2004)
3. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: International Conference on Automatic Face and Gesture Recognition. pp. 302–309 (2018)
4. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 13984–13993 (2020)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients forhuman detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
6. Darwin, C., Prodger, P.: The expression of the emotions in man and animals. Oxford University Press (1998)
7. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: International Conference on Computer Vision. pp. 2106–2112 (2011)
8. Ding, H., Zhou, S.K., Chellappa, L.: Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: International Conference on Automatic Face and Gesture Recognition (2017)
9. Ekman, P.: Strong evidence for universals in facial expressions: a replyto russell's mistaken critique. Psychological bulletin **115**(2), 268–287 (1994)
10. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality andSocial Psychology **17**(2), 124–129 (1971)
11. Ekman, P., Rosenberg, E.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press (1997)
12. Gao, B., Xing, C., Xie, C., Wu, J., Geng, X.: Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing **26**(6), 2825–2838 (2017)
13. Geng, X.: Deep label distribution learning with label ambiguity. IEEE Transactions on Knowledge and Data Engineering **28**(7), 1734–1748 (2016)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
15. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 558–567 (2019)
16. Li, S., Deng, W., Du, J., Zhang, Z.: Reliable crowd-sourcing and deep locality-preserving learning for expression recognition in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 285–2861 (2017)
17. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.: Learning from noisy labels with distillation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1910–1918 (2017)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision (2021)

19. Ma, F., Sun, B., Li, S.: Robust facial expression recognition with convolutional visual transformers. IEEE Transactions on Affective Computing pp. 1–1 (2021)
20. Mollahosseini, A., Hasani, B., Mahoor, M.H., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. TAC **10**(1), 18–31 (2017)
21. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: ACM International Conference on Multimodal Interaction. pp. 443–449 (2015)
22. Ng, P.C., Henikoff, S.: Sift: predicting aminoacid changes that affect protein function. Nucleic Acids Research **31**(13), 3812–3814 (2003)
23. Qu, Y., Mo, S., Niu, J.: Dat: Training deep networks robust to label-noise by matching the featuredistributions. In: IEEE Conference on Computer Vision and Pattern Recognition (2021)
24. Ruan, D., Yan, Y., Chen, S., Xue, J.H., HanziWang: Deep disturbance-disentangled learning for facial expression recognition. In: ACM International Conference on Multimedia (2020)
25. Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7660–7669 (2021)
26. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980)
27. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing **27**(6), 803–816 (2009)
28. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6248–6257 (2021)
29. Shi, J., Zhu, S., Liang, Z.: Learning to amend facial expression representation via de-albino and affinity (2021), arXiv:2103.10189
30. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learningan explicit mapping for sample weighting. In: Annual Conference on Neural Information Processing Systems (2019)
31. Su, K., Geng, X.: Soft facial landmark detection by label distribution learning. In: AAAI Conference on Artificial Intelligence. pp. 5008–5015 (2019)
32. Tian, Y.I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. T-PAMI **23**(2), 97–115 (2001)
33. Veit, A., Alldrin, N., Chechika, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 839–847 (2017)
34. Wang, C., Wang, S., Liang, G.: Identity- and pose-robust facial expression recognition through adversarial feature learning. In: ACM International Conference on Multimedia. pp. 238–246 (2019)
35. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6897–6906 (2020)
36. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020)

37. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robustlearning with noisy labels. In: International Conference on Computer Vision (2019)
38. Xu, N., Liu, Y., Geng, X.: Label enhancement for labeldistribution learning. IEEE Transactions on Knowledge and Data Engineering **33**(4), 1632–1643 (2021)
39. Xu, N., Shu, J., Liu, Y., Geng, X.: Variational label enhancement. In: Proceedings of the 37th International Conference on Machine Learning. pp. 10597–10606 (2020)
40. Yang, H., Ciftci, U.A., Yin, L.: Facial expression recognition by de-expression residue learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2168–2177 (2018)
41. Yi, K., Wu, J.: Probabilistic end-to-end noise correction for learning with noisy labels. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
42. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: European Conference on Computer Vision. pp. 22–237 (2018)
43. Zhang, Y., Deng, W., Wang, M., Hu, J., Li, X., Zhao, D., Wen, D.: Global-local gcn: Large-scale label noise cleansing for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
44. Zhao, G., Huang, X., Taini, M., Li, S.Z., äInen, M.P.: Facial expression recognition from near-infrared videos. Image and Vision Computing **29**(9), 607–619 (2011)
45. Zhao, Z., Liu, Q., Zhou, F.: Robust lightweight facial expression recognition network with label distribution training. In: AAAI Conference on Artificial Intelligence. pp. 3510–3519 (2021)