ReSup: Reliable Label Noise Suppression for Facial Expression Recognition

Xiang Zhang[1] *Member, IEEE,* Yan Lu[1], Huan Yan, Jinyang Huang* *Member, IEEE,* Yu Gu* *Senior Member, IEEE* Yusheng Ji *Fellow, IEEE* Zhi Liu *Senior Member, IEEE* and Bin Liu

Abstract—Because of the ambiguous and subjective property of the facial expression, the label noise is widely existing in the FER dataset. For this problem, in the training phase, current methods often directly predict whether the label is noised or not, aiming to reduce the contribution of the noised data. However, we argue that this kind of method suffers from the low reliability of such noise data decision operation. It makes that some mistakenly abounded clean data are not utilized sufficiently and some mistakenly kept noised data disturbing the model learning. In this paper, we propose a more reliable noise-label suppression method called ReSup. First, instead of directly predicting noised or not, ReSup makes the noise data decision by modeling the distribution of noise and clean labels simultaneously according to the disagreement between the prediction and the target. Specifically, to achieve optimal distribution modeling, ReSup models the similarity distribution of all samples. To further enhance the reliability of our noise decision results, ReSup uses two networks to jointly achieve noise suppression. Specifically, ReSup utilize the property that two networks are less likely to make the same mistakes, making two networks swap decisions and tending to trust decisions with high agreement. Extensive experiments on popular datasets shows the effectiveness of ReSup.

Index Terms—Facial Expression, Label Noise, Affective Computing, Label Noise Modeling, Label Noise Suppression.

1 Introduction

Facial expression recognition (FER) has become a crucial service in various real-world applications, such as healthcare [1], [2], surveillance [3], [4], and virtual reality [5]. It aims to recognize specific human emotions from the given facial images. However, obtaining high-quality annotations for large-scale FER datasets collected from the Internet poses challenges due to the subjectivity of annotators and the ambiguity of facial expressions. Consequently, these low-quality annotations introduce label noises. Therefore, how to suppress label noises in FER tasks has become a research hotspot and attracted more and more attention [6]–[11].

To address this challenge, existing FER methods commonly integrate an importance learning branch to estimate the importance weight of each image, determining whether the label of the input image is noisy or not [6], [7], [9]. However, we argue that these methods suffer from the low reliability of such noise decision operation. Such operation usually generates unreliable weights and makes that some mistakenly abounded clean data are not utilized sufficiently and some mistakenly kept noised data disturbing the model learning process. These unreliable weights originate from two perspectives: the noise decision process and the FER model itself. The former unreliable weights are due to overfitting of the importance learning branch as a result of the strong learning ability of deep neural networks (DNNs) [9]. Furthermore, such noise decision process only considers information from a single sample [6] or a batch [7], neglecting global information [9] and resulting in unreliable decision-making. In addition to the unreliable weights caused by

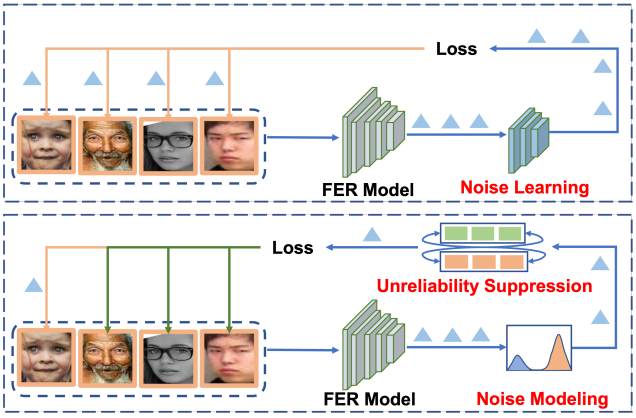


Fig. 1. Comparison of label noise suppression process of different noisy label FER methods. The top is the current schemes and the bottom is ReSuP. ReSuP generates more reliable weights by noise modeling and suppresses unreliable weights by unreliability suppression design in addition to suppressing noisy labels. Triangles represent unreliable weights.

Xiang Zhang, and Bin Liu, are with School of Cyber Science and Technology, University of Science and Technology of China, Hefei, 230026, China.
Xiang Zhang, Jinyang Huang, and Yu Gu, School of Computer and Information, Hefei University of Technology, Hefei, 230601, China.
Yan Lu, AI Lab, Shanghai, 200030, China
Huan Yan, School of Computer and Information, Guizhou Normal University, Guiyang, 230601, China.
Yusheng Ji is with the National Institute of Informatics, Tokyo 101-8430, Japan.
Zhi Liu, Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo, 1828585, Japan.
[1]*Equal contribution, * Corresponding authors.*
This work is supported by the Anhui Province Science Foundation for Youths (Grant No. 2308085QF230), Major scientific and technological project of Anhui Provincial Science and Technology Innovation Platform (Grant No. 202305a12020012), the National Natural Science Foundation of China (Grant No. 62302145, and No. 62462015), and the Guizhou Province Basic Research Plan Youth Guidance Project, China ([2024]345).

the noise decision process, the FER model itself inevitably produces some unreliable outputs (the inputs of the importance learning branch), resulting in further unreliable weights. These unreliable weights accumulate during the entire learning process and affect current and future learning stages. Unfortunately, existing methods do not address **how to mitigate the effects of these unreliable weights.** Novel methods are required to address these limitations and improve the reliability and accuracy of noisy label FER.

In this paper, we present a novel approach called ReSup. The main objective of ReSup is to suppress noisy labels and unreliable weights, as illustrated in Figure 1. Specifically, instead of directly predicting noised or not, ReSup makes the noise decision by modeling the joint distribution of noise and clean labels. This is motivated by the memorization effect of deep neural networks (DNNs), where the model tends to memorize correctly labeled samples first [12]–[14], leading to noisy samples having higher loss during early epochs of training [15], [16]. In order to achieve optimal distribution modeling, ReSup propose to model the similarity (cosine similarity of predictions and targets) distribution of all samples rather than the loss, which reduces the confusion between noisy and clean distributions. The fitted noise model is then used to provide importance weights for each sample based on its similarity, without using neural network branches to avoid overfitting. In addition, the proposed scheme can take into account the global distribution of all samples. Furthermore, ReSup mitigate the effect of the unreliable weights by leveraging the agreement maximization principles [17], [18], which suggest that two different networks would agree on most samples except for noisy samples [19] and thus can filter different types of errors. Inspired by the agreement maximization principles, ReSup employs two different networks to provide importance weights to each other, to prevent the accumulation of errors caused by unreliable weights. We also introduce a consistency loss that assigns large losses to samples with small agreement to prevent the model from fitting samples with unreliable weights, since the samples with small agreements usually are noisy samples. In summary, our contributions include:

- To avoid extra unreliable weights caused by the DNN-based importance learning branch, a novel label noise modeling method based on similarity distribution statistics is proposed to estimate the importance weights.
- We propose ReSup to suppress label noise in FER. ReSup satisfactorily mitigates the effect of the unreliable weights by leveraging a weight exchange strategy and a consistency loss.
- Extensive experiment results demonstrate that the proposed ReSup significantly outperforms state-of-the-art noisy label FER solutions on multiple FER benchmarks with different levels of label noise.

2 Related Work

2.1 Facial Expression Recognition

The categorization of FER methods can be broadly divided into two groups based on the features used: handcraft-based and learning-based approaches. Earlier research mainly relied on handcrafted features [20], [21], which capture the folds and geometry changes caused by facial expressions [22], [23]. Nevertheless, researchers have gradually uncovered the limitations of handcraft-based methods, particularly in-the-wild scenarios. Fortunately, with the advancement in computational ability and the rapid development of large-scale datasets, e.g., AffectNet [24], RAF-DB [25], and EmotioNet [26], recent studies mainly focus on deep learning [27], [28] as a better alternative. Ruan [29], [30] proposed the feature decomposition method and the reconstruction learning algorithm for effective FER. RAN [31] proposed a region attention network to overcome the pose and occlusion challenges in FER. Transformers [32] have also been introduced into the FER due to its powerful global information awareness. However, the presence of label noise in large-scale datasets remains a significant challenge for FER in in-the-wild scenarios.

In recent years, several algorithms are proposed to address the label noise challenge in FER. Among these, SCN [6], RUL [9], and DMUE [7] use neural network branches to estimate importance weight for each sample, determining whether the label of the input image is noisy or not. However, the strong learning capability of DNNs can lead to overfitting of the importance learning branch and the fact that these approaches do not exploit global information, both leading to unreliable weights beyond the FER model itself. To address this issue, PT [8] selects samples with small losses as clean samples for training, without relying on a neural network branch to learn importance weights. However, PT requires setting a threshold based on the exact noise level and dataset used, which is impractical in real-world scenarios. In contrast, our proposed scheme uses unsupervised noise modeling to mitigate these issues without requiring prior knowledge, and suppresses unreliable weights by unreliability suppression design in addition to suppressing noisy labels.

2.2 Noisy Label Learning

Learning with noisy labels has been extensively studied in the computer vision community, and current approaches can be broadly categorized into four following groups [33]: robust architecture, robust regularization, robust loss function, and robust data.

Robust architecture-based methods usually added a noise adaptation layer [34] at the top of the network [35] to learn label transition proces or designed a noise-tolerant architecture [36] to reliably support more diverse types of label noise. Robust regularization-based methods aimed to explicitly [37] or implicitly [38] induce DNNs to be less likely to overfit the noisy labels. By avoid overfitting in training, the robustness to label noise improves with regularization techniques such as data-augmentation [39] and weight decay [40].

Robust loss function-based methods seeked to design a loss function that is robust to noisy labels, to prevent DNNs from fitting to the noisy samples [41], [42]. Robust data-based methods aim to select clean samples from noisy data based on thresholds [14] or weights [15]. Threshold-based algorithms, which require an artificial threshold to select clean samples and discard noisy samples, have been

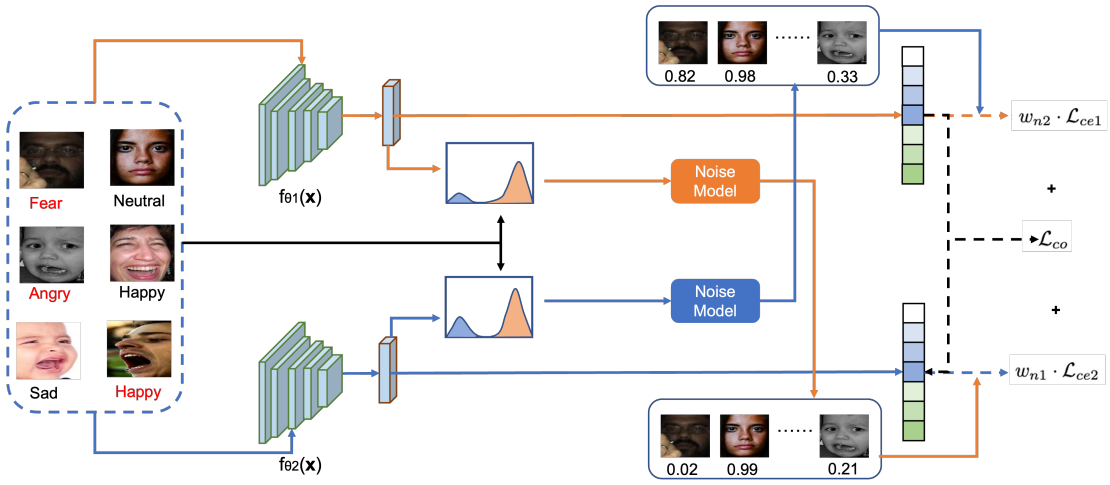


Fig. 2. The ReSup framework. During the training process, input images are simultaneously passed through two networks, $f_{\theta 1}(\boldsymbol{x})$ and $f_{\theta 2}(\boldsymbol{x})$. The outputs of the final layer of these networks are then used as inputs to the label noise modeling module for modeling label noise. Both networks simultaneously model the label noise, and the fitted noise model is used to generate importance weights for the corresponding inputs and to weight the loss of the other network. The weighted loss is then summarized and used in the backpropagation process to guide network learning.

explored in classical robust data-based methods [14], [43], [44]. Discarding noisy samples, however, potentially mistakenly discard clean samples and removes useful information about the data distribution [15], [45]. To avoid this problem, recent approaches propose the use of two-component beta mixture models (BMM) [15] or Gaussian mixture models (GMM) [46] that rely on losses to model the label noise and assign weights to the samples based on the fitted noise model to suppress the noisy labels. Based on the memory effect mentioned in section 1, these schemes believe that the loss distributions of clean and noisy samples are significantly different. However, these state-of-the-art schemes rely on loss distribution for noise modeling, which may not be effective in noisy label FER due to the inter-class similarity of facial expressions. We propose to model label noise by relying on the similarity (Section 3.2) and demonstrate its effectiveness in section 4.5.

3 Method

3.1 Overview of ReSup

Current approaches in FER that rely on DNNs [47] to learn how to identify label noise in facial expression datasets often result in the generation of unreliable weights. To mitigate this issue, we employ a statistical modeling approach to differentiate between noisy and clean labels. However, compared to related research in other domains, facial expressions pose unique challenges due to their inherent ambiguity. The ambiguity of facial expressions refers to the lack of clear distinction between different classes of expressions, leading to confusion and potential overlap. This can cause severe distributional confusion in the later stages of training in previous statistical modeling-based approaches, rendering the model unable to identify label noise effectively. The proposed solution addresses this challenge by employing similarity instead of loss for the statistical modeling of label noise, thus effectively resolving it. Furthermore, a weight exchange strategy and consistency loss are designed to further mitigate the impact of unreliable weights.

FER can be formulated as the problem of learning a model $f_\theta(\boldsymbol{x})$ from a set of facial images $T = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ with $\boldsymbol{y}_i \in \{0,1\}^C$ is the one-hot ground-truth label corresponding to $\boldsymbol{x}_i$.

In our case, f_θ is a CNN and θ is the model parameters. Besides, the label $\boldsymbol{y}_i$ could differ from the true label (noisy label). During the learning, the parameters of the model are fitted by optimizing the cross-entropy (CE) loss:

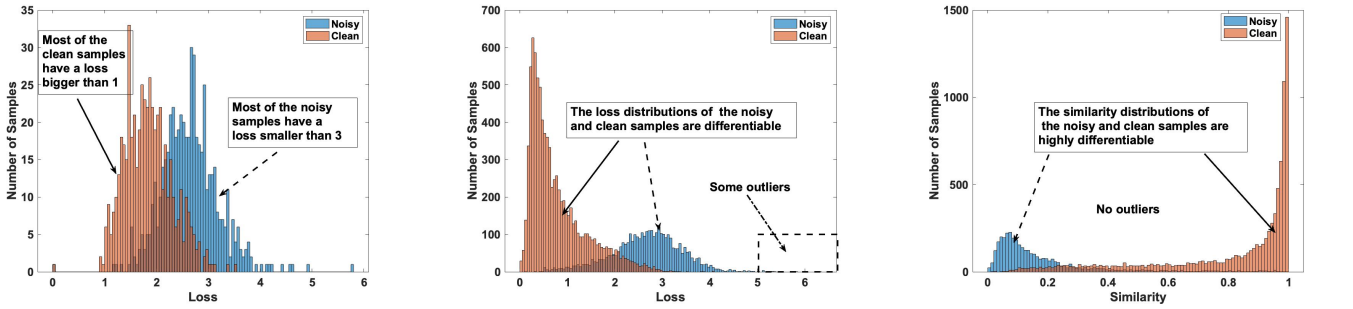
$$\mathcal{L}_{ce} = \sum_{i=1}^{N} \mathcal{L}_i = -\sum_{i=1}^{N} \boldsymbol{y}_i^T log(f_\theta(\boldsymbol{x}_i)) \tag{1}$$

where $f_\theta(\boldsymbol{x}_i)$ and $\mathcal{L}_i$ represent the softmax output and the loss produced by the model, respectively. For simplicity, we use $\boldsymbol{f}_i$ to represent $f_\theta(\boldsymbol{x}_i)$ in the remainder of the paper. The goal of FER model is to minimize the CE loss to obtain the optimal model parameters θ. However, noisy labels can negatively impact the training process and lead to poor performance. Therefore, our goal is to mitigate the impact of label noise during the training process.

The framework of ReSup is depicted in Figure 2. The proposed method comprises two components, namely label noise modeling and noise-robust learning. Label noise modeling aims to model the distribution of the label noise. The fitted noise model can provide importance weights for each sample according to its similarity. These assigned weights are then used by noise-robust learning to effectively learn informative representations in the presence of label noise. To better achieve this goal, noise-robust learning uses a weight exchange strategy and a consistency loss to mitigate the effect of unreliable weights.

3.2 Label Noise Modeling

To enable a noise-robust [48] learning approach (section 3.3), it is necessary to identify noisy samples in the training set T first. Given that deep neural networks (DNNs) often tend to memorize correctly labeled samples initially and then gradually fit to noisy labels [12]–[14], leading to larger losses for noisy samples during the early epochs of training, the



(a) The loss distribution of samples in the disgust class after the 1st epoch of training.

(b) The loss distribution of all samples after the 1st epoch of training.

(c) The similarity distribution of all samples after the 1st epoch of training.

(d) The loss distribution of samples in the surprise class after the 1st epoch of training.

(e) The loss distribution of all samples after the 7th epoch of training.

(f) The similarity distribution of all samples after the 7th epoch of training.

Fig. 3. Loss and similarity distributions of training samples in the RAF-DB under 30% noise.

memorization effect is thus observed. Recently, inspired by the memorization effect, some state-of-the-art label noise modeling methods [15], [46] are proposed to unsupervised fit the loss distribution of all samples by a two-component mixture model. By relying on the clear distinction between the loss distributions of clean and noisy samples, this mixture model can effectively predict the importance weights.

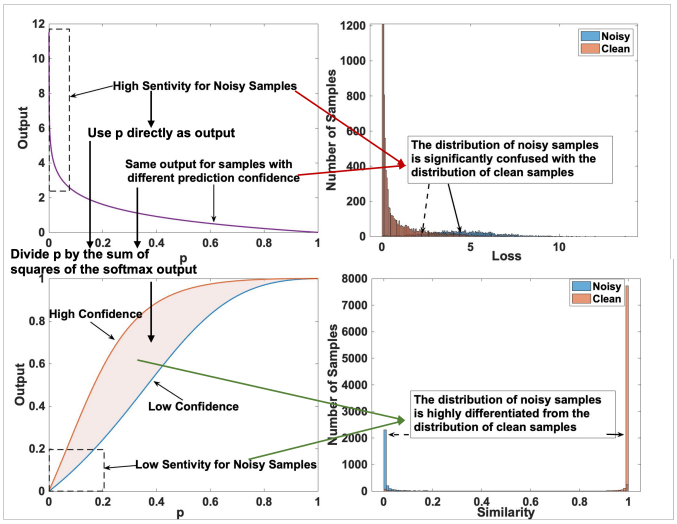


Fig. 4. Comparison of using loss and similarity to model label noise. The figures in the upper and lower left corners show the change in loss and similarity with respect to the prediction probability p, respectively. The figures in the upper and lower right corners show the corresponding loss and similarity distributions of samples, respectively.

Regrettably, since the loss distributions of clean and noisy samples are significantly confused, the performance of label noise modeling algorithms, such as [15], [46], may be inadequate in the context of noisy label FER tasks. This confusion arises from the ambiguity of facial expressions, which is exacerbated by the cross-entropy (CE) loss function. In particular, certain expressions, such as fear and surprise, are more ambiguous than others and result in larger losses for clean samples and smaller losses for noisy samples in these classes relative to others [29] (Because FER model has higher probability of incorrect predictions for these classes.), as shown in Figures 3(a) and 3(d). The prediction probability p_i for a sample $\boldsymbol{x}_i$ is defined as:

$$p_i = \sum_{k=1}^{c} y_{ik} \cdot f_{ik} \tag{2}$$

where c is the number of classes. y_{ik} and f_{ik} denote the one-hot label and the softmax output corresponding to $\boldsymbol{x}_i$ in class k. Hence, the CE loss can be represented as:

$$\mathcal{L}_{cei} = -log(p_i) \tag{3}$$

where L_{cei} is the CE loss corresponding to $\boldsymbol{x}_i$.

As depicted in Figure 4, the CE loss, which has been widely used in previous studies, exhibits high sensitivity to small prediction probabilities. As a result, even a slight change in the prediction probability can lead to a significant variation in the resulting loss. This sensitivity amplifies the difference in the loss distribution of noisy samples from different classes, resulting in an approximately uniform distribution of all noisy sample losses (Figure 3(e)). This

phenomenon ultimately leads to confusion between the distributions of noise and clean samples and hinders the ability to model them using unsupervised mixture models. Moreover, the CE loss only takes into account the prediction probability, which can be the same for both noisy and clean samples, thus further exacerbating the confusion between the two types of samples. To address this limitation, a new function should be employed the following requirements:

- Whether the model is learning well enough can be measured by the chosen function (similar to loss);
- The chosen function should exhibits low sensitivity to prediction probability change;
- The chosen function should be able to differentiate between clean and noisy samples even when their prediction probabilities are the same.

In order to model label noise, directly using prediction probability has the advantages of meeting the first and second requirements. However, it falls short in meeting the third requirement. To overcome this limitation, we propose using the cosine similarity (S) of the prediction to the given label as a suitable function for the task. This choice allows us to better distinguish between clean and noisy samples, even in cases where the prediction probabilities for these samples may be the same. The formula for calculating the similarity is as follows:

$$\mathcal{S}_i = \frac{p_i}{\sqrt{\sum_{k=1}^{c}(f_{ik})^2}} = cos(\boldsymbol{f}_i, \boldsymbol{y}_i) \tag{4}$$

where c is the number of classes

As illustrated in Figure 4, compared to the CE loss, the similarity-based approach is less sensitive to variations in the prediction probability. This implies that the same difference in probability yield a smaller difference in the cosine similarity distance compared to CE loss, leading to a more focused distribution of noise samples. In contrast to methods that solely rely on the prediction probability, the similarity-based approach considers the entire output of the network, providing more information to effectively discriminate between the distributions of clean and noisy samples. Specifically, for a given $\boldsymbol{x}_i$ and p_i, if $\exists f_{ij} = 1-p_i, j \in [1..c]$ & $y_{ij} \neq 1$, then S_i get the minimum. Besides, if $\forall f_{ij} = (1-p_i)/(c-1), j = 1..c$ & $y_{ij} \neq 1$, then S_i obtain the maximum. Apparently, according to the memorization effect, $\boldsymbol{y}_i$ is more likely to be a noisy label when S_i obtains the minimum. Intuitively, the memorization effect suggests that noisy labels are not the first to be memorized, implying a mismatch between the network's prediction and the label. And if S_i obtains the minimum, there is a greater probability that $f_{ij} > 1 - p_i$ exists.

Following previous studies [15], [46], we first model the similarity distribution using a mixture model, and then use the generated noise model to provide a probability for each sample that it belongs to clean samples. The probability density function (pdf) of a mixture model of K components on the similarity $\mathcal{S}$ can be defined as:

$$p(\mathcal{S}) = \sum_{m=1}^{K} \delta_m p(\mathcal{S}|m) \tag{5}$$

where δ_m are the mixing coefficients of each pdf $p(\mathcal{S}|m)$. In our case, we fit a two-components (i.e. $K = 2$) mixture model to model the distribution of clean and noisy samples.

We choose the more flexible BMM to model the similarity distribution. The pdf of the beta distribution is:

$$p(\mathcal{S}|\alpha_m, \beta_m) = \frac{\Gamma(\alpha_m + \beta_m)}{\Gamma(\alpha_m)\Gamma(\beta_m)} \mathcal{S}^{\alpha_m - 1}(1-\mathcal{S})^{\beta_m - 1} \tag{6}$$

where $\Gamma(\cdot)$ is the Gamma function, $\mathcal{S}$ is the similarity, $\alpha_m, \beta_m > 0$, and the mixture pdf is given by substituting the above into equation 5.

We use an Expectation Maximization (EM) procedure to fit the BMM to the similarity of all samples in this study. After the noise model is fitted, we can obtain the probability of a sample belongs to clean samples (importance weights) through the posterior probability:

$$w_n = p(m|\mathcal{S}) = \frac{p(m)p(\mathcal{S}|m)}{p(\mathcal{S})} \tag{7}$$

where $m = 1(0)$ denotes clean (noisy) classes. $p(\mathcal{S}|m)$ is defined to be the posterior probability of the similarity $\mathcal{S}$ having been generated by component m.

3.3 Noise-Robust Learning

Label noise modeling generates a noise model to assign importance weights to each training sample in T, and noise-robust learning aims to suppress the label noise in T and extract meaningful knowledge by leveraging these importance weights. To better achieve this goal, we propose a weight exchange strategy along with a consistency loss to eliminate the effect caused by unreliable weights. The proposed method consists of two DNNs denoted by $f_{\theta 1}(\boldsymbol{x})$ and $f_{\theta 2}(\boldsymbol{x})$.

Network. For ReSup, both $f_{\theta 1}(\boldsymbol{x})$ and $f_{\theta 2}(\boldsymbol{x})$ can be used to predict facial expression alone, but during the training stage, the parameters of the two networks are updated simultaneously by a joint loss. Specifically, the joint loss function $\mathcal{L}_{jo}$ is constructed as follows:

$$\mathcal{L}_{jo} = \mathcal{L}_{wc} + \lambda \mathcal{L}_{co} \tag{8}$$

In the joint loss function, the first part $\mathcal{L}_{wc}$ is the weighted CE loss of the two networks, which reliably uses the importance weights by relying on the weight exchange strategy to suppress label noise and learn useful knowledge. The second part $\mathcal{L}_{co}$ is the consistency loss, which is used to further attenuate the effect of unreliable weights. λ is the hyperparameter to control the influence of $\mathcal{L}_{co}$.

Weight Exchange Strategy Based Weighted CE Loss. The importance weights are utilized to weigh the CE loss for suppress label noise. To avoid errors caused by unreliable weights being accumulated, the weights of $f_{\theta 1}(\boldsymbol{x})$ and $f_{\theta 2}(\boldsymbol{x})$ are exchanged. Intuitively, due to the presence of memorization effects, DNNs tend to prioritize learning from clean samples, allowing us to select clean samples based on DNN predictions. However, neural networks inevitably produce unreliable predictions, such as misclassifying noise samples as clean ones. If this error flow is directly fed back into the network during the learning process, the error will gradually accumulate. This is because the errors generated during subsequent learning stages are likely to be of the same type, causing the DNN to treat this type of error as correct knowledge. However, different networks have varying learning capabilities and can therefore identify and

filter out different types of errors [14], [44]. Consequently, in the weight exchange procedure, error flows can be mitigated by peer networks mutually. When errors from noisy data flow into the peer network, they are attenuated due to its robustness. The weighted CE loss is formulated as:

$$\mathcal{L}_{wc} = w_{n2} \cdot \mathcal{L}_{ce1} + w_{n1} \cdot \mathcal{L}_{ce2} \tag{9}$$

where $\mathcal{L}_{ce1}$ and $\mathcal{L}_{ce2}$ represents the CE loss of $f_{\theta 1}(\boldsymbol{x})$ and $f_{\theta 2}(\boldsymbol{x})$, respectively. w_{n1} and w_{n2} denotes the importance weight from $f_{\theta 1}(\boldsymbol{x})$ and $f_{\theta 2}(\boldsymbol{x})$, respectively.

Consistency Loss. From the perspective of agreement maximization principles [17], [18], different networks are unlikely to agree on noisy labels, meaning that the probability of two networks simultaneously making the same unreliable decisions for the same noisy samples is small. Thus, we utilize a consistency loss to evaluate the agreement between the two networks. The consistency loss imposes high loss values on samples with low agreement to further discourage the model from fitting samples with unreliable weights. When there is a small agreement, the probability that the sample belongs to noise is high. A high consistency loss makes the model more likely to fit the prediction of another network instead of the intended target. The consistency loss is as follows:

$$\mathcal{L}_{co} = \frac{1}{c} \sum_{i=1}^{N} \sum_{k=1}^{c} (\boldsymbol{f}_{ik1} - \boldsymbol{f}_{ik2})^2 \tag{10}$$

$\boldsymbol{f}_{ik1}$ and $\boldsymbol{f}_{ik2}$ denotes the outputs of the "softmax" layer in $f_{\theta 1}(\boldsymbol{x})$ and $f_{\theta 2}(\boldsymbol{x})$ for sample x_i on class k, respectively.

The closest approach to our model is JoCoR [19], which also employs two networks and a contrastive loss. However, JoCoR uses the average loss of the two networks to select clean samples based on a threshold. In contrast to JoCoR, ReSup utilizes two networks to provide importance weights to each other and employs a consistency loss to mitigate the impact of unreliable weights. To evaluate ReSup's effectiveness, we have also implemented a JoCoR-based scheme in our experiments. (Section 4.5).

4 Experiments

In this section, we evaluate the effectiveness of ReSup on synthetic label noise datasets and in-the-wild benchmarks.

4.1 Datasets

RAF-DB [25] is the first in-the-wild dataset containing basic or compound expressions, including nearly 30k facial images annotated with basic or compound expressions by 40 trained human annotators. In our experiments, we use only images of six basic expressions (happy, surprise, sad, anger, disgust, fear) as well as neutral, which leads to 12,271 images for training and 3,068 images for testing.

FERPlus [34] is an extension of FER2013, which is used in the ICML 2013 Challenge. It is a large-scale dataset collected through the Google search engine. It contains 28,709 training images, 3,589 validation images, and 3,589 test images, which are resized to 48×48 pixels. FERPlus includes eight emotional states (six basic emotions, neutral and contempt), and each image is labeled by ten human annotators.

TABLE 1
Comparison with other state-of-the-art noisy label FER schemes. Results are computed as the mean of the last 5 epochs.

Method	Noise	RAF-DB	FERPlus	AffectNet
SCN [6]	10%	84.97%	85.08%	61.71%
DMUE [7]	10%	83.19%	-	-
RUL [9]	10%	86.05%	86.32%	62.80%
PT [8]	10%	87.28%	85.04%	-
EAC [49]	10%	88.10%	86.87%	63.37%
La-net [50]	10%	88.75%	**88.02%**	62.05%
DR-FER [51]	10%	**88.92%**	87.58%	63.12%
DivFER [52]	10%	88.17%	87.75%	-
Ours	10%	88.43%	87.82%	**64.29%**
SCN [6]	20%	83.67%	84.87%	60.80%
DMUE [7]	20%	81.02%	-	-
RUL [9]	20%	84.83%	84.65%	61.69%
PT [8]	20%	86.25%	84.27%	-
EAC [49]	20%	86.76%	85.98%	62.74%
La-net [50]	20%	87.12%	86.85%	61.72%
DR-FER [51]	20%	86.82%	87.07%	61.33%
DivFER [52]	20%	87.05%	86.97%	-
Ours	20%	**87.29%**	**87.08%**	**63.97%**
SCN [6]	30%	80.61%	83.32%	59.00%
DMUE [7]	30%	79.41%	-	-
RUL [9]	30%	81.16%	83.73%	60.71%
PT [8]	30%	84.32%	83.73%	-
EAC [49]	30%	85.07%	85.36%	62.60%
La-net [50]	30%	85.33%	86.01%	60.82%
DR-FER [51]	30%	84.31%	84.88%	59.40%
DivFER [52]	20%	85.63%	83.65%	-
Ours	30%	**86.86%**	**86.74%**	**62.89%**

AffectNet [24] is the largest in-the-wild facial expression dataset by far. It contains nearly 450K manually annotated facial images collected from the Internet by three major search engines with emotion-related keywords. This dataset has an imbalanced training set and a balanced validation set. Following previous work [10], [53], [54], we selected approximately 280,000 and 3,500 images for training and testing, and the classes are the same as RAF-DB.

4.2 Implementation Details

In our experiments, we adopt ResNet-18 [55] pretrained on the MS-Celeb-1M [56] as $f_{\theta 1}(\boldsymbol{x})$ and ResNet-18 pretrained on the ImageNet [57] as $f_{\theta 2}(\boldsymbol{x})$ for fair comparisons with previous works [6], [8], [9]. The images we used are aligned and cropped with three landmarks, then resized to 224×224 pixels, and augmented by random horizontal flipping, random erasing, and random cropping.During training, we use a batch size of 96 and employ Adam as the optimizer with an initial learning rate of 0.0002. We divide the learning rate by 10 at epoch 10 and 20 for RAF-DB and FERPlus, and at epoch 5 and 10 for AffectNet. Training concludes at epoch 30 for RAF-DB and FERPlus, and at epoch 20 for AffectNet. We set the hyperparameter λ to 5 by default based on our ablation studies. Our implementation is based on the Pytorch toolbox, and all experiments are conducted on a single NVIDIA RTX 3090. Upon completion of model training and during inference, we can discard $f_{\theta 2}(\boldsymbol{x})$ and the label noise modeling component, retaining only $f_{\theta 1}(\boldsymbol{x})$ and the final softmax layer to obtain the final results, significantly reducing the computational and storage requirements during inference. Code:https://github.com/purpleleaves007/FERDenoise.

Fig. 5. Visualization of the confidence score of SCN, RUL, and ReSup. All models are trained on RAF-DB with 30% noise. The importance weights/uncertainty values are marked at the top. The true labels and the labels for training are marked in the lower right and left corners, respectively. For noisy samples, the uncertainty values should be large but the importance weights should be small.

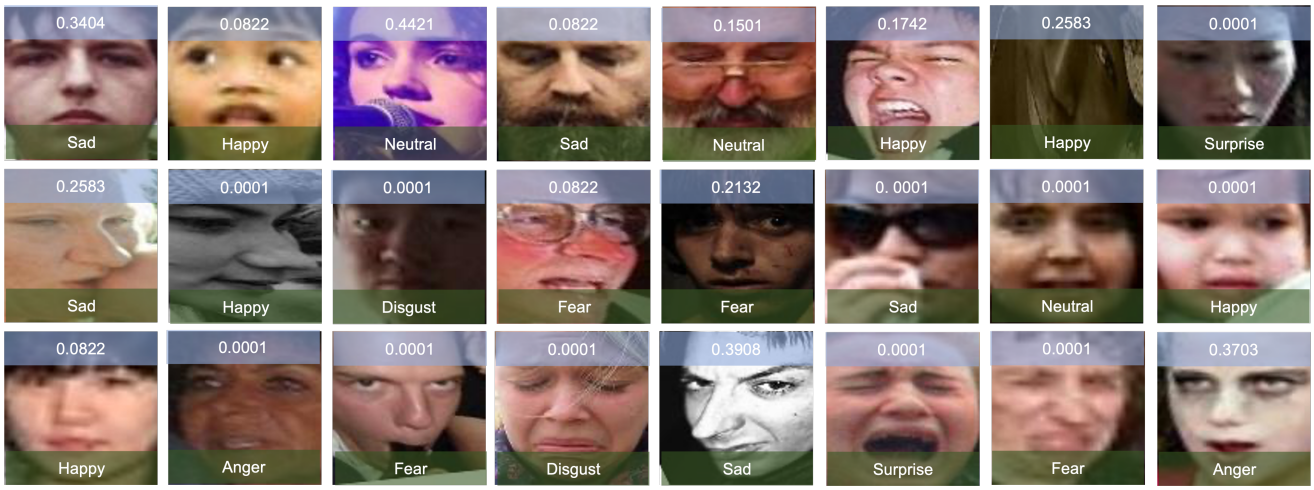


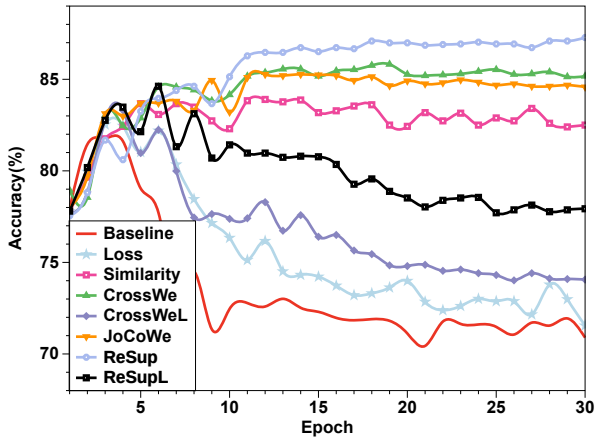
Fig. 6. Some examples of RAF-DB (w/o synthetic noisy labels) with low importance weights. The importance weights are marked at the top, and the true labels are marked at the bottom.

4.3 Evaluation on Noisy FER Datasets

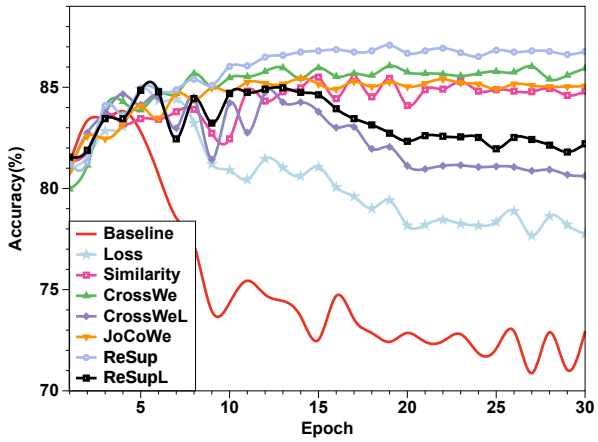
In this section, we present a quantitative evaluation of the proposed ReSup method compared to other state-of-the-art approaches for addressing noisy labels in FER on RAF-DB, FERPlus, and Affectnet datasets. Following prior studies [6]–[9], we randomly select a portion (10%, 20%, and 30%) of the training data and corrupt their labels by assigning them to other random facial expression categories to generate noisy labels.

The experimental results in Table 1 demonstrate that the proposed ReSup method achieves superior performance compared to other methods. For instance, under the noise rate of 30%, ReSup outperforms SCN by 6.25%, 3.42%, and 3.89% on RAF-DB, FERPlus, and AffectNet, respectively. Moreover, the performance degradation of ReSup is only 1.57% and 1.08% when adding 30% noise to 10% RAF-DB and FERPlus datasets, while SCN degrades by 4.36% and 1.76%, and PT drops by 2.96% and 1.31%.

It is noteworthy that ReSup uses a modeling-based approach for importance weight estimation, unlike SCN [6], DMUE [7], and RUL [9] that rely on a network branch for this task. As such, ReSup does not face the concern that the powerful learning ability of deep neural networks (DNNs) may degrade the importance weight decision process. In comparison to PT [8], ReSup does not require knowledge of the exact noise level to set a threshold for selecting clean samples. Moreover, PT [8] introduces a considerable amount of additional data ($280,000$ extra samples) for achieving better results through semi-supervised learning. The purpose of FENN [11], [58] is to suppress heteroscedastic uncertainty caused by label noise between classes, but its suppression method is also implemented through DNN, leading to its limited reliability and performance. Furthermore, SCN, RUL, and DMUE rely on a label correction module to im-



(a) Ablation studies on RAF-DB.



(b) Ablation studies on FERPlus.

Fig. 7. The test accuracy vs. epochs on RAF-DB and FERPlus with 30% noise level. Loss and Sim represent using a single network to model and suppress label noise by relying on the loss and the similarity distribution, respectively. ReSup is the proposed Cross-Denoising, while Crosswe is ReSup w/o the consistency loss. ReSupL and CrossWeL are similar to ReSup and CrossWe but use the loss to model noise. JoCoWe is the JoCoR-based method.

prove performance, whereas ReSup outperforms them with only a more reliable utilization of the importance weights. Compared to EAC, our approach exhibits superior generalization capability (Section4.12). Compared to La-net [50], DR-FER [51], and DivFER [52], ReSuP achieves superior noisy FER performance without relying on landmarks or additional representations, leveraging only the information disparity during the network learning process.

4.4 Visualization Analysis

In this section, we present a comparative analysis of the proposed ReSup method with other state-of-the-art FER schemes for noisy label data, namely RUL and SCN. We assess the effectiveness of these methods on the RAF-DB dataset with a 30% noise level by visualizing and comparing the estimated importance scores. The results are showcased in Figure 5, where the first three columns depict clearly mislabeled images, and the last two column features clean images. We observe that both the proposed ReSup method and RUL accurately distinguish between clean and noisy

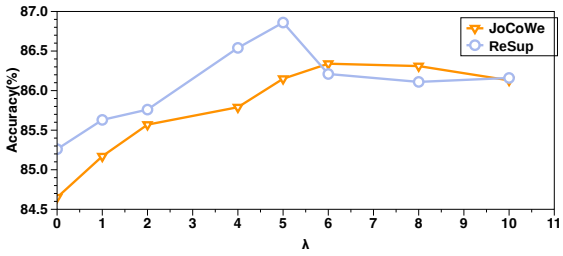


Fig. 8. The accuracy of CrossDe and JoCoWe with different λ on RAF-DB with 30% noise. (For JoCoWe, the hyperparameter for the contrastive loss is $0.1 * \lambda$)

samples. In contrast, SCN may confuse clean and noisy samples. However, we note that RUL does not provide significant differences in uncertainty values for clean and noisy samples. Additionally, the images in the fourth and fifth column is ambiguous, and the proposed method assigns an importance weight close to 0.5. However, SCN and RUL classify this image either as clean or noisy. Furthermore, the sixth column displays a meaningless image, and both RUL and ReSup assign small importance to it, while SCN does not.

Furthermore, we evaluated our ReSup on original FER datasets, which inevitably suffer from label noise, and the results are presented in section 4.6. We illustrated some samples with low importance weights from the original RAF-DB in Figure 6. We found that ReSup is more likely to assign low-importance weights to ambiguous, low-quality, and occluded images.

TABLE 2
Ablation studies on RAF-DB and FERPlus with 30% noise. Loss and Sim represents cross-entropy loss and similarity-based noise modeling, respectively. WeEx and CoL represents weight exchange and consistency loss, respectively.

Loss	Sim	WeEx	CoL	RAF-DB	FERPlus
				71.43%	72.14%
✓				72.67%	78.23%
	✓			83.05%	84.79%
✓		✓		74.14%	80.83%
	✓	✓		85.26%	85.74%
✓		✓	✓	77.91%	82.22%
	✓	✓	✓	86.86%	86.74%

4.5 Ablation Studies

To analyze the roles of different components in ReSup, we conduct ablation studies on RAF-DB and FERPlus datasets with a 30% noise level. The experimental results are depicted in Figure 7(a), 7(b), and Table 2, respectively. It is noteworthy that, to further illustrate the effectiveness of the weight exchange strategy, we also compare ReSup with JoCoR [19], which selects examples based on the average loss of two networks. The first row of Table 2 serves as a baseline without noise suppression measures.

The observations are as follows: **(1)** The use of the loss for label noise modeling is found to be less effective as the training progresses, while the proposed similarity-based approach is seen to remain effective. **(2)** The weight exchange

TABLE 3
Comparison with other state-of-the-art FER schemes.

Method	Publication	RAF-DB	FERPlus	AffectNet
LDL-ALSG [54]	CVPR 2020	85.53%	-	59.35%
EfficientFace [53]	AAAI 2021	88.36%	-	63.70%
SCN [6]	CVPR 2020	87.03%	88.01%	-
DMUE [7]	CVPR 2021	88.76%	88.64%	-
RUL [9]	NIPS 2021	88.98%	88.75%	-
PT [8]	TAFFC 2021	88.69%	86.60%	58.54%
EAC [49]	ECCV 2022	89.99%	89.64%	65.32%
POSTER [59]	ICCV 2023	92.05%	91.62%	67.31%
CA-FER [60]	TAFFC 2023	90.06%	-	64.31%
POSTER++ [61]	PR 2024	92.21%	-	67.49%
DR-FER [51]	TMM 2024	91.61%	91.91%	67.54%
Ours		89.70%	88.85%	65.46%

strategy is observed to improve performance by eliminating accumulated errors caused by unreliable weights. **(3)** The application of the consistency loss further enhances performance by preventing the networks from fitting unreliable weights when two networks' predictions significantly differ. **(4)** The proposed ReSup approach outperforms the JoCoR-based method due to the latter's clean sample selection method weak in eliminating accumulated errors.

We also evaluate the effect of the hyperparameter λ for both ReSup and JoCoR-based method as shown in Figure 8. ReSup achieved the best performance at $\lambda = 5$ and JoCoR-based method achieved the best performance at $\lambda = 0.6$, while ReSup performs better than JoCoR-based method with different λ. In additional analysis, we computed the cosine similarity between the predictions of $f\theta_1(x)$ and $f\theta_2(x)$ on RAF-DB. The results show that with consistency loss, the average similarity on the test set increased from 0.83 to 0.91, indicating that consistency loss enhances the alignment between the two networks' predictions, thereby improving overall performance.

4.6 Comparision with State-of-the-art Methods

We compare ReSup with several state-of-the-art FER methods on original RAF-DB, FERPlus, and AffectNet in Table 3. Besides the noisy label FER studies mentioned in previous sections, EfficientFace [53] incorporates local and global features to learn a FER model. gACNN [62] leverage attention mechanism to conduct an occlusion-aware FER model. POSTER [59] concentrates on inter-class similarity, intra-class discrepancy, and scale sensitivity issues in FER and proposes a two-stream Pyramid Cross-fusion Transformer network to tackle these challenges. DR-FER [63] designs a ResNet-50-based network to extract discriminative and robust representations from facial expressions. Additionally, CA-FER [60] devises an IR-50-based network that uses causal reasoning to simultaneously optimize feature discrimination and diversity to mitigate spurious correlations in expression datasets.

Although our performance on the original dataset doesn't match that of the SOTA approaches, our work demonstrates the effectiveness of the proposed solution in addressing the issue of noisy labels in facial expressions. For facial expressions, the problem of ambiguous labels is widespread due to the ambiguity and combinability between expressions, as well as the cost considerations in annotating large-scale datasets. Therefore, besides focusing on learning better facial expression features like SOTA approaches, it's crucial to avoid learning erroneous knowledge introduced by noisy labels during model training. This becomes particularly important in the context of semi-supervised learning-based FER models, which involve generating large-scale pseudo-labels to leverage abundant unlabeled facial expression data, and it is a potential next research direction for us. Moreover, the proposed similarity modeling approach is well-suited for facial expressions, as it can address the traditional modeling breakdown caused by the ambiguity of facial expressions.

We also implement two plain methods using threshold from Co-teaching [14] and PT [8] with performance of (86.53%,83.21%,82.37%) and (84.91%,77.93%,72.46%) on the RAF-DB (noise 10%-30%), respectively. We also meticulously tune the threshold-based method from [8], [14] on our scheme, with the best performance of 87.32%, 84.68%, and 83.51%. The performance of ReSup exceeds all these solutions. Our ReSup outperforms all of these solutions and offers an advantage over them by not requiring prior knowledge or estimation of the noise ratio to carefully tune the parameters for optimal performance.

4.7 Experiments on Real Noisy FER Dataset

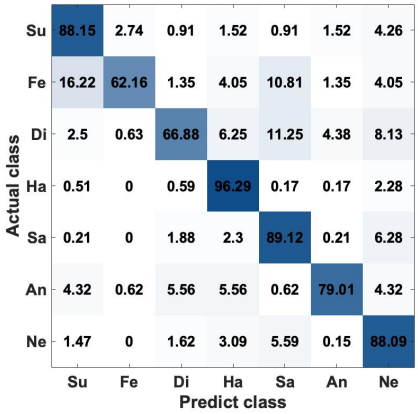
To validate the effectiveness of our proposed method, ReSup, we conducted experiments on a real-world noisy Facial Expression Recognition (FER) dataset, namely ExpW [64], which contains a significant number of low-quality annotations. We trained our models on this dataset and evaluated their performance on the test sets of ExpW and RAF-DB. The experimental results demonstrate the superior efficacy of our ReSup approach (73.12%/75.65%) compared to the ResNet-18 baseline (67.87%/71.77%), POSTER (70.96%/73.04%) and EAC (71.01%/73.96%).

4.8 Extended Experiments Based on the Similarity Distribution

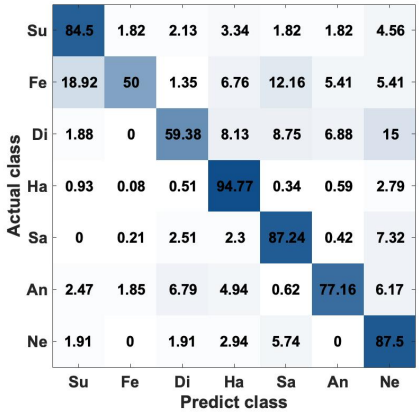
We observed that leveraging the statistical characteristics of the similarity distribution can contribute to achieving more equitable outcomes. Specifically, the similarity values of clean samples in easier classes tend to be closer to 1, while those of noisy samples are closer to 0, resulting in higher variance of similarity for samples in the easier classes. Conversely, harder classes exhibit lower variance due to their inherent learning difficulty. To address this, we propose dividing the estimated importance weight by the variance of the corresponding class, thus obtaining a balanced importance weight. The effectiveness of this approach is demonstrated in Figures 9(a), 9(c), 9(b), and 9(d), where a more balanced performance is achieved with only a slight degradation in overall accuracy, both for the original and synthetic noisy datasets.

4.9 Experiments on Asymmetric Label Noise

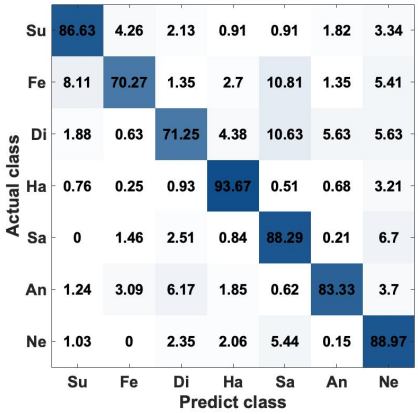
The majority of previous approaches have primarily focused on evaluating their performance under symmetric noise, where all samples are mislabeled with equal probability across all labels. However, to address the more realistic



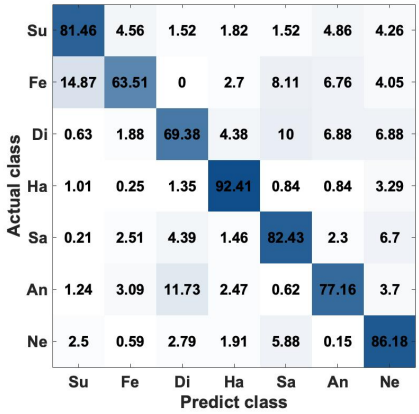
(a) Confusion matrix of the proposed ReSup.



(b) Confusion matrix of ReSup under 30% noise.



(c) Confusion matrix of ReSup using balanced importance weights.



(d) Confusion matrix of ReSup under 30% noise using balanced importance weights.

Fig. 9. Confusion matrices on RAF-DB, the overall accuracy of Figure 9(a), 9(c), 9(b), and 9(d) is 89.21%, 88.76%, 87.03% and 85.59%, respectively.

TABLE 4
Comparison with other state-of-the-art noisy label FER schemes on RAF-DB with the setting asymmetric noise.

Method	Accuracy
Baseline	82.72%
SCN [6]	83.10%
RUL [9]	84.24%
Ours	**86.94%**

scenario of asymmetric noise [33], where different classes have varying probabilities of being mislabeled, we conducted additional experiments. For instance, emotions such as happy and neutral are typically less prone to mislabeling, whereas surprise and fear are more likely to be confused due to their similarity in terms of mouth opening. In our experiments, we intentionally did not introduce noise to the happy and neutral classes. However, the surprise, anger, disgust, sadness, and fear classes were relabeled with probabilities of 10%, 20%, 30%, 40%, and 50%, respectively, using a uniform distribution across all expressions. In this case, by employing the balanced importance weights, we compared our proposed ReSup approach with two existing methods that have released code. The results, as presented in Table 4, clearly indicate the superiority of our solution over both SCN [6] and DUL [9].

4.10 Experiments on Different Network Structures

TABLE 5
The influence of different backbones on ReSup. We carry out experiments on RAF-DB and * means baseline.

Noise	MobileNet	ResNet18	ResNet50	VGG16
20%*	74.54%	74.25%	76.01%	72.49%
20%	84.49%	84.55%	86.86%	83.28%
30%*	68.52%	68.19%	68.87%	66.33%
30%	83.57%	82.76%	85.24%	82.50%

We conducted experiments using different network structures (MobileNet, ResNet18/50 and Vgg16), as presented in Table 5, where all networks were pre-trained solely on ImageNet. Our proposed scheme demonstrated effectiveness across diverse network architectures.

4.11 Experiments on DivideMix

To evaluate the proposed similarity-based label noise modeling approach, we performed experiments on the CIFAR10

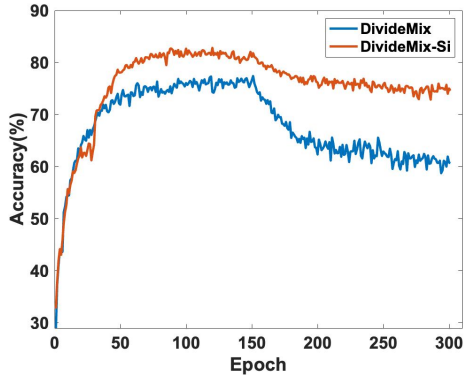


Fig. 10. Comparison of using loss and similarity to model label noise in DivideMix on CIFAR-10 with 80% noise. DivideMix and DivideMix-Si means using loss and similarity to model label noise, respectively.

dataset using DivideMix [46] as the baseline, which is the current state-of-the-art loss-based label noise modeling and correction method. We replaced the noise modeling method in DivideMix with the proposed similarity-based approach and conducted experiments for 300 epochs. The results presented in Table 6, demonstrate that our similarity-based approach outperforms the loss-based approach.

DivideMix determines the cleanliness of a sample by utilizing a fitted noise model and employs mixup data augmentation to improve its noise suppression capabilities. Moreover, it uses a semi-supervised technique to assign pseudo-labels to noisy samples to address the challenge of data size reduction caused by noisy labels. To further test the efficacy of the proposed similarity-based label noise modeling method, we removed the semi-supervised module of DivideMix and performed experiments on the CIFAR-10 dataset with 80% noisy labels. The performance metrics, as depicted in Figure 10 also demonstrate that the proposed similarity-based noise modeling approach outperforms the previous loss-based approach.

TABLE 6
Comparison of using loss and similarity to model label noise in DivideMix on CIFAR-10 with symmetric noise. DivideMix and DivideMix-Si means using loss and similarity to model label noise, respectively.

Method/Noise	20%	50%	80%	90%
DivideMix	95%	93.7%	92.4%	74.2%
DivideMix-Si	**95.7%**	**94.5%**	**93.3%**	**76.9%**

4.12 Experiments on CIFAR10 and CIFAR 100

The generalization performance of ReSup was evaluated on two widely used datasets, namely CIFAR10 and CIFAR100, using the ResNet18 and the 7-layer CNN utilized in JoCoR [19] for a fair comparison. Specifically, we employed the Adam optimizer with an initial learning rate of 0.001 and decayed the learning rate by a factor of 0.1 at the 50th, 100th, and 150th epochs for the 7-layer CNNs. For the ResNet18 network, we used the SGD optimizer with an initial learning rate of 0.1, and the learning rate was decayed to 0 from the 100th to the 200th epoch. The training process was run for a total of 200 epochs with a batch size of 128. ReSup was applied to suppress noise starting from the 3rd epoch for the 7-layer CNN and from the 80th epoch for ResNet18.

Table 7 and 8 present the experimental findings. The results indicate that ReSup has an edge over the existing method even when used with a shallower neural network, demonstrating its effectiveness for non-FER tasks. However, it should be noted that ReSup performs poorly on CIFAR100 with high noise ratios due to the limited number of accurate samples per category. In such scenarios, Co-learning [70] provides an advantage by enhancing information acquisition through a self-supervised approach.

5 Conclusion

In this paper, we have proposed ReSup, a novel approach for tackling the problem of noisy label FER. Our method consists of two key components: label noise modeling and noise-robust learning. The former enables us to reliably model the label noise in FER, while the latter allows us to mitigate the negative impact caused by unreliable weights and learn from noisy datasets. Extensive experiments on three public datasets have demonstrated the effectiveness of our approach, which outperforms several state-of-the-art FER methods. Our approach is also shown to be effective on various network structures and able to generalize well to other datasets. Overall, our work provides a promising solution for improving the performance of FER models in the presence of noisy labels.

References

[1] C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, and K. Alvarez, "Developing multimodal intelligent affective interfaces for tele-home health care," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 245–255, 2003.

[2] Y. Gu, X. Zhang, H. Yan, Z. Liu, and Y. Ji, "Real-time vital signs monitoring based on cots wifi devices," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1320–1324.

[3] A. A. M. Al-Modwahi, O. Sebetela, L. N. Batleng, B. Parhizkar, and A. H. Lashkari, "Facial expression recognition intelligent security system for real time surveillance," in *Proc. of World Congress in Computer Science*, 2012.

[4] X. Zhang, Y. Gu, H. Yan, Y. Wang, M. Dong, K. Ota, F. Ren, and Y. Ji, "Wital: A cots wifi devices based vital signs monitoring system using nlos sensing model," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, pp. 629–641, 2023.

[5] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, and H. Yu, "Realistic facial expression reconstruction for vr hmd users," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 730–743, 2019.

[6] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6897–6906.

[7] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.

[8] J. Jiang and W. Deng, "Boosting facial expression recognition by a semi-supervised progressive teacher," *IEEE Transactions on Affective Computing*, 2021.

[9] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 616–17 627, 2021.

[10] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

TABLE 7
Test accuracy on CIFAR-10, and ‡ means asym noise.

Method	Publication	Architecture	20%	40%	50%	60%	80%	40%‡
Standard	-	7-L-CNN	69.18%	57.34%	42.71%	34.09%	16.24%	69.43%
Co-teaching [14]	NIPS2018	7-L-CNN	78.23%	-	71.30%	-	26.58%	69.43%
Co-teaching+ [44]	ICML2019	7-L-CNN	78.71%	-	71.30%	-	26.58%	69.43%
JoCoR [19]	CVPR2020	7-L-CNN	85.73%	-	57.05%	-	24.19%	68.84%
Jo-SRC [65]	CVPR2021	7-L-CNN	87.80%	-	73.67%	-	34.30%	79.96%
Li [66]	MM2022	8-L-CNN	87.94%	84.90%	-	80.08%	-	-
TS^3-Net [67]	TNNLS2022	8-L-CNN	88.52%	86.54%	-	-	-	-
ReSup	-	7-L-CNN	**90.08%**	**87.73%**	**86.06%**	**84.68%**	**70.78%**	**85.31%**
Standard	-	ResNet18	84.81%	66.58%	61.49%	46.55%	28.98%	76.30%
Class2Simi [68]	ICML2021	ResNet26	91.38%	88.22%	-	79.45%	-	87.79%
HOC [69]	ICML2021	ResNet34	90.03%	85.49%	-	77.40%	-	-
Co-learning [70]	MM2021	ResNet18	92.21%	-	84.49%	-	61.20%	81.42%
ReSup	-	ResNet18	**94.13%**	**92.65%**	**89.81%**	**86.49%**	**72.43%**	**90.28%**

TABLE 8
Test accuracy on CIFAR-100, and ‡ means asym noise.

Method	Publication	Architecture	20%	40%	50%	60%	80%	40%‡
Standard	-	7-L-CNN	35.14%	23.05%	16.97%	10.52%	4.41%	27.29%
Co-teaching [14]	NIPS2018	7-L-CNN	43.73%	-	34.96%	-	15.15%	28.35%
Co-teaching+ [44]	ICML209	7-L-CNN	49.27%	-	40.04%	-	13.44%	33.62%
JoCoR [19]	CVPR2020	7-L-CNN	53.01%	-	43.49%	-	15.49%	32.70%
Jo-SRC [65]	CVPR2021	7-L-CNN	58.15%	-	51.26%	-	23.80%	38.52%
Li [66]	MM2022	8-L-CNN	58.84%	52.54%	-	42.62%	-	-
ReSup	-	7-L-CNN	**62.08%**	**57.76%**	**56.08%**	**52.26%**	**28.78%**	**47.66%**
Standard	-	ResNet18	57.79%	45.77%	33.75%	24.30%	8.64%	42.49%
Class2Simi [68]	ICML2021	ResNet56	60.26%	54.85%	-	40.38%	-	52.99%
HOC [69]	ICML2021	ResNet34	68.82%	62.29%	-	**52.96%**	-	-
Co-learning [70]	MM2021	ResNet18	66.58%	-	54.54%	-	**35.45%**	47.62%
TS^3-Net [67]	TNNLS2022	VGG-16	66.87%	60.99%	-	-	-	-
EAC [49]	ECCV2022	ResNet18	66.73%	60.59%	-	-	-	-
ReSup	-	ResNet18	**72.62%**	**64.35%**	**58.83%**	51.96%	29.19%	**53.37%**

[11] H. Yan, Y. Gu, X. Zhang, Y. Wang, Y. Ji, and F. Ren, "Mitigating label-noise for facial expression recognition in the wild," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[13] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[14] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[15] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *International conference on machine learning*. PMLR, 2019, pp. 312–321.

[16] C. Zhang, S. Bengio, M. Hardt, M. C. Mozer, and Y. Singer, "Identity crisis: Memorization and generalization under extreme overparameterization," in *International Conference on Learning Representations*, 2019.

[17] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[18] V. Sindhwani, P. Niyogi, and M. Belkin, "A co-regularization approach to semi-supervised learning with multiple views," in *Proceedings of ICML workshop on learning with multiple views*, vol. 2005. Citeseer, 2005, pp. 74–79.

[19] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 726–13 735.

[20] J. J. Bazzo and M. V. Lamar, "Recognizing facial actions using gabor wavelets with neutral face average difference," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* IEEE, 2004, pp. 505–510.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[22] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2562–2569.

[23] S. Khan, L. Chen, and H. Yan, "Co-clustering to reveal salient facial features for expression recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 348–360, 2017.

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

[25] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[26] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.

[27] 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.

[28] R. Wang, J. Huang, J. Zhang, X. Liu, X. Zhang, Z. Liu, P. Zhao, S. Chen, and X. Sun, "Facialpulse: An efficient rnn-based depression detection via temporal facial landmarks," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 311–320.

[29] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7660–7669.

[30] D. Ruan, R. Mo, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Adaptive deep disturbance-disentangled learning for facial expression recognition," *International Journal of Computer Vision*, pp. 1–23, 2022.

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

[32] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3601–3610.

[33] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

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

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2691–2699.

[37] H. Wei, L. Tao, R. Xie, and B. An, "Open-set label noise can improve robustness against inherent label noise," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7978–7992, 2021.

[38] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar, "Does label smoothing mitigate label noise?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 6448–6458.

[39] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[40] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems*, vol. 4, 1991.

[41] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[42] Y. Liu and H. Guo, "Peer loss functions: Learning from noisy labels without knowing noise rates," in *International conference on machine learning*. PMLR, 2020, pp. 6226–6236.

[43] E. Malach and S. Shalev-Shwartz, "Decoupling" when to update" from" how to update"," *Advances in neural information processing systems*, vol. 30, 2017.

[44] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 7164–7173.

[45] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia, "Iterative learning with open-set noisy labels," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8688–8696.

[46] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.

[47] X. Zhang, J. Huang, H. Yan, P. Zhao, G. Zhuang, Z. Liu, and B. Liu, "Wiopen: A robust wi-fi-based open-set gesture recognition framework," *IEEE Transactions on Human-Machine Systems*, 2025.

[48] J. Huang, B. Liu, C. Miao, X. Zhang, J. Liu, L. Su, Z. Liu, and Y. Gu, "Phyfinatt: An undetectable attack framework against phy layer fingerprint-based wifi authentication," *IEEE Transactions on Mobile Computing*, vol. 23, no. 7, pp. 7753–7770, 2023.

[49] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 418–434.

[50] Z. Wu and J. Cui, "La-net: Landmark-aware learning for reliable facial expression recognition under label noise," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 698–20 707.

[51] M. Li, H. Fu, S. He, H. Fan, J. Liu, J. Keppo, and M. Z. Shou, "Dr-fer: Discriminative and robust representation learning for facial expression recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 6297–6309, 2024.

[52] W. Nie, Z. Wang, X. Wang, B. Chen, H. Zhang, and H. Liu, "Diving into sample selection for facial expression recognition with noisy annotations," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 7, no. 1, pp. 95–107, 2025.

[53] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 3510–3519.

[54] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 984–13 993.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[58] Y. Gu, H. Yan, X. Zhang, Y. Wang, Y. Ji, and F. Ren, "Towards facial expression recognition in the wild via noise-tolerant network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[59] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3146–3155.

[60] P.-J. Huang, H. Xie, H.-C. Huang, H.-H. Shuai, and W.-H. Cheng, "Ca-fer: Mitigating spurious correlation with counterfactual attention in facial expression recognition," *IEEE Transactions on Affective Computing*, 2023.

[61] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang, "Poster++: A simpler and stronger facial expression recognition network," *Pattern Recognition*, p. 110951, 2024.

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

[63] M. Li, H. Fu, S. He, H. Fan, J. Liu, J. Keppo, and M. Z. Shou, "Dr-fer: Discriminative and robust representation learning for facial expression recognition," *IEEE Transactions on Multimedia*, 2023.

[64] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, pp. 550–569, 2018.

[65] Y. Yao, Z. Sun, C. Zhang, F. Shen, Q. Wu, J. Zhang, and Z. Tang, "Jo-src: A contrastive approach for combating noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5192–5201.

[66] J. Li and H. Sun, "Correct twice at once: Learning to correct noisy labels for robust deep learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5142–5151.

[67] R. Jiang, Y. Yan, J.-H. Xue, B. Wang, and H. Wang, "When sparse neural network meets label noise learning: A multistage learning framework," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[68] S. Wu, X. Xia, T. Liu, B. Han, M. Gong, N. Wang, H. Liu, and G. Niu, "Class2simi: A noise reduction perspective on learning with noisy labels," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 285–11 295.

[69] Z. Zhu, Y. Song, and Y. Liu, "Clusterability as an alternative to anchor points when learning with noisy labels," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 912–12 923.

[70] C. Tan, J. Xia, L. Wu, and S. Z. Li, "Co-learning: Learning from noisy labels with self-supervision," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1405–1413.

Xiang Zhang received the B.E. degree from Hefei University of Technology, China, in 2017, and his D.E. degree from the same university in 2023. Currently, he is a postdoc with the School of Cyber Science and Technology, University of Science and Technology of China. His research interests include wireless sensing and affective computing. He is a TPC Member of ACM MM, IEEE ICME and Globecom. He has served as a reviewer for ACM IMWUT, CHI, TECS, CogSci, IEEE TNNLS, TMM, THMS, TCSVT and TNSM. He is the recipient of the IEEE SMC Society Andrew P. Sage Best Transactions Paper Award and the IEEE HITC Distinguished Phd Dissertation Award.

Yan Lu received a BS degree in Telecommunication Engineering from the Anhui Polytechnic University, Wuhu, Anhui, China, in 2013 and 2017 and an MS degree in electrical engineering from the University of Science and Technology of China, Hefei, Anhui, China in 2020. His research interests include multimedia processing and computer vision.

Huan Yan received the B.E. degree from Hefei University of Technology, China, in 2017, and his D.E. degree from the same university in 2023. He is currently a lecturer in the School of Big Data and Computer Science at Guizhou Normal University. His research interests include Wireless Security and Wireless Sensing.

Jinyang Huang is a lecturer at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT) and the Secretary-General of the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligence Machine (led by Prof. Meng Wang (IEEE Fellow)). He obtained his Ph.D. in Computer Science and Technology from the School of Cyberspace Security, University of Science and Technology of China (USTC) in 2022. He was sponsored by China Scholarship Council (CSC) (from 2020.12.1 to 2021.11.31) for a joint Ph.D. study supervised by Assoc. Prof. Lu Su from Purdue University and Chang Wen Chen from University at Buffalo, USA. His research interests include Multimodal Perception, Human-computer Interaction, Wireless Security, and Signal Processing. In this area, he has published 35 papers in international peer-reviewed journals and conferences, including ToN, TMC, TIFS, TAFFC, THMS, TVT, IOTJ, MobiCom, Infocom, ACM MM, and ECCV. He has served as a TPC member for conferences, including ACM MM, IEEE ICME, and Globecom, and has the honor of becoming ACM MM 2024 Outstanding Reviewers. He is a Guest Editor for Applied science. He is the recipient of the Young Scientist of Anhui Computer Federation and IEEE HITC Distinguished PhD Dissertation Award.

Yu Gu (M'10-SM'12) received his B.E. degree from the Special Classes for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2004, and his D.E. degree from the same university in 2010. In 2006, he was an Intern with Microsoft Research Asia, Beijing, China, for seven months. From 2007 to 2008, he was a Visiting Scholar at the University of Tsukuba, Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow with the National Institute of Informatics, Tokyo, Japan. Since 2012, he has been a Professor and Dean Assistant at the School of Computer and Information, Hefei University of Technology. His current research interests include pervasive computing and affective computing. He was the recipient of the IEEE Scalcom 2009 Excellent Paper Award, NLP-KE2017 Best Paper Award, and IEEE CCIS 2018 Best Student Paper Award. He is a member of ACM and a senior member of IEEE.

Yusheng Ji received the B.E., M.E., and D.E. degrees in electrical engineering from the University of Tokyo. She joined the National Center for Science Information Systems (NACSIS), Japan, in 1990. She is currently a Professor at the National Institute of Informatics (NII), and the Graduate University for Advanced Studies (SOKENDAI), Japan. Her research interests include resource management in wireless networks and mobile computing. She has served as Editor of IEEE Transactions of Vehicular Technology, Symposium Co-chair of IEEE GLOBECOM 2012 and 2014, Track Co-chair of IEEE VTC 2016 Fall and 2017 Fall, General Co-chair of ICT-DM 2018, MSN2020, and Symposium Co-chair of IEEE ICC 2020. She is a Distinguished Lecturer of IEEE Vehicular Technology Society, Associate Editor of IEEE Vehicular Technology Magazine, TPC Co-chair of IEEE INFOCOM 2023, and TPC member of IEEE ICC, GLOBECOM, WCNC, etc.

Zhi Liu (Senior Member, IEEE) received the Ph.D. degree in informatics from the National Institute of Informatics. He is currently an Associate Professor with The University of Electro Communications. His research interests include video network transmission and mobile edge computing. He is an Editorial Board Member of Wireless Networks (Springer) and IEEE OPEN JOURNAL OF THE COMPUTER SOCIETY.

Bin Liu Bin Liu received the BS and MS degrees in electrical engineering from the University of Science and Technology of China, Hefei, Anhui, China, in 1998 and 2001, respectively, and the PhD degree in electrical engineering from Syracuse University, Syracuse, New York, in 2006. Currently, he is an associate professor with the School of Cyber Science and Technology, University of Science and Technology of China. His research interests include multimedia processing, computer vision and AI security.