Suppressing Uncertainties for Large-Scale Facial Expression Recognition

Kai Wang*[1,2], Xiaojiang Peng*[1], Jianfei Yang[3], Shijian Lu[3], and Yu Qiao†[1]

[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[2]University of Chinese Academy of Sciences, China
[3] Nanyang Technological University Singapore

Abstract

*Annotating a qualitative large-scale facial expression dataset is extremely difficult due to the uncertainties caused by ambiguous facial expressions, low-quality facial images, and the subjectiveness of annotators. These uncertainties lead to a key challenge of large-scale Facial Expression Recognition (FER) in deep learning era. To address this problem, this paper proposes a simple yet efficient Self-Cure Network (SCN) which suppresses the uncertainties efficiently and prevents deep networks from over-fitting uncertain facial images. Specifically, SCN suppresses the uncertainty from two different aspects: 1) a self-attention mechanism over mini-batch to weight each training sample with a ranking regularization, and 2) a careful relabeling mechanism to modify the labels of these samples in the lowest-ranked group. Experiments on synthetic FER datasets and our collected WebEmotion dataset validate the effectiveness of our method. Results on public benchmarks demonstrate that our SCN outperforms current state-of-the-art methods with **88.14**% on RAF-DB, **60.23**% on AffectNet, and **89.35**% on FERPlus. The code will be available at https://github.com/kaiwang960112/Self-Cure-Network.*

1. Introduction

Facial expression is one of the most natural, powerful and universal signals for human beings to convey their emotional states and intentions [7, 41]. Automatically recognizing facial expression is also important to help the computer understand human behavior and to interact with them. In the past decades, researchers have made significant progress on facial expression recognition (FER) with algorithms[17, 47] and large-scale datasets, where datasets can be collected in laboratory or in the wild, such as CK+ [31], MMI [42], Oulu-CASIA [54], SFEW/AFEW [10], FERPlus [4], AffectNet [35], EmotioNet [11], RAF-DB [24], etc.

Figure 1: Illustration of uncertainties on real-world facial images from RAF-DB. The right samples are extremely difficult for machines and even human which are better to be suppressed in training.

However, for the large-scale FER datasets collected from the Internet, it is extremely difficult to annotate with high quality due to the uncertainties caused by the subjectiveness of annotators as well as ambiguous in-the-wild facial images. As illustrated in Figure 1, the uncertainties increase from high-quality and evident facial expressions to low-quality and micro expressions. These uncertainties usually lead to inconsistent labels and incorrect labels, which are suspending the progress of large-scale Facial Expression Recognition (FER), especially for the one of data-driven deep learning based FER. Generally, training with uncertainties of FER may lead to the following problems. First, it may result in over-fitting on the uncertain samples which may be mislabeled. Second, it is harmful for a model to learn useful facial expression features. Third, a high ratio of incorrect labels even makes the model disconvergence in the early stage of optimization.

To address these issues, we propose a simple yet efficient method, termed as Self-Cure Network (SCN), to suppress the uncertainties for large-scale facial expression recognition. The SCN consists of three crucial modules: self-

*Equally-contributed first authors (kai.wang, xj.peng@siat.ac.cn)
†Corresponding author (yu.qiao@siat.ac.cn)

attention importance weighting, ranking regularization, and noise relabeling. Given a batch of images, a backbone CNN is first used to extract facial features. Then the self-attention importance weighting module learns a weight for each image to capture the sample importance for loss weighting. It is expected that uncertain facial images are assigned low importance weights. Further, the ranking regularization module ranks these weights in descending order, splits them into two groups (i.e. high importance weights and low importance weights), and regularizes the two groups by enforcing a margin between the average weights of the two groups. This regularization is implemented with a loss function, termed as Rank Regularization loss (RR-Loss). The ranking regularization module ensures that the first module learns meaningful weights to highlight certain samples (*e.g.* reliable annotations) and to suppress uncertain samples (*e.g.* ambiguous annotations). The last module is a careful relabeling module that attempts to relabel these samples from the bottom group by comparing the maximum predicted probabilities to the probabilities of given labels. A sample is assigned to a pseudo label if the maximum prediction probability is higher than the one of given label with a margin threshold. In addition, since the main evidence of uncertainties is the incorrect/noisy annotation problem, we collect an extreme noisy FER dataset from the Internet, termed as WebEmotion, to investigate the effect of SCN with extreme uncertainties.

Overall, our contributions can be summarized as follows,

- We innovatively pose the uncertainty problem in facial expression recognition, and propose a Self-Cure Network to reduce the impact of uncertainties.
- We elaborately design a rank regularization to supervise the SCN to learn meaningful importance weights, which also provides a reference for the relabeling module.
- We extensively validate our SCN on synthetic FER data and a new real-world uncertain emotion dataset (WebEmotion) collected from the Internet. Our SCN also achieves performance **88.14**% on RAF-DB, **60.23**% on AffectNet, and **89.35**% on FERPlus, which set new records on them.

2. Related Work

2.1. Facial Expression Recognition

Generally, a FER system mainly consists of three stages, namely face detection, feature extraction, and expression recognition. In face detection stage, several face detectors like MTCNN [51] and Dlib [2]) are used to locate faces in complex scenes. The detected faces can be further aligned alternatively. For feature extraction, various methods are designed to capture facial geometry and appearance features caused by facial expressions. According to the feature type, they can be grouped into engineered features and learning-based features. For the engineered features, they can be further divided into texture-based local features[48], geometry-based global features, and hybrid features. The texture-based features mainly include SIFT [37], HOG [6], Histograms of LBP [38], Gabor wavelet coefficients [28], etc. The geometry-based global features are mainly based on the landmark points around noses, eyes, and mouths. Combining two or more of the engineered features refers to the hybrid feature extraction, which can further enrich the representation. For the learned features, Fasel [12] finds that a shallow CNN is robust to face poses and scales. Tang [40] and Kahou *et al*. [23] utilize deep CNNs for feature extraction, and win the FER2013 and Emotiw2013 challenge, respectively. Liu *et al*. [29] propose a Facial Action Units based CNN architecture for expression recognition. Recently, both Li *et al.* [27] and Wang *et al.* [45] have designed region-based attention networks for pose and occlusion aware FER, where the regions are either cropped from landmark points or fixed positions.

2.2. Learning with Uncertainties

Uncertainties in the FER task mainly come from ambiguous facial expressions, low-quality facial images, inconsistent annotations, and incorrect annotations (*i.e*. noisy labels). Particularly, learning with noisy labels is extensively studied in the computer vision community while the other two aspects are rarely explored. In order to handle noisy labels, one intuitive idea is to leverage a small set of clean data that can be used to assess the quality of the labels during the training process [43, 25, 8], or to estimate the noise distribution [39], or to train the feature extractors [3]. Li *et al*. [25] propose a unified distillation framework using 'side' information from a small clean dataset and label relations in knowledge graph, to 'hedge the risk' of learning from noisy labels. Veit *et al*.[44] use a multi-task network that jointly learns to clean noisy annotations and to classify images. Azadi *et al*.[3] select reliable images by an auxiliary image regularization for deep CNNs with noisy labels. Other methods do not need a small clean dataset but they may assume extra constrains or distributions on the noisy samples [34], such as a specific loss for randomly flipped labels [36], regularizing the deep networks on corrupted labels by a MentorNet [22], and other approaches that model the noise with a softmax layer by connecting the latent correct labels to the noisy ones [13, 50]. For the FER task, Zeng *et al*. [50] first consider the inconsistent annotation problem among different FER datasets, and propose to leverage these uncertainties to improve FER. *In contrast, our work focuses on suppressing these uncertainties to learn better facial expression features.*

3. Self-Cure Network

To learn robust facial expression features with uncertainties, we propose a simple yet efficient Self-Cure Network (SCN). In this section, we first provide an overview of the SCN, and then present its three modules. We finally present the detailed implementation of SCN.

3.1. Overview of Self-Cure Network

Our SCN is built upon traditional CNNs and consists of three crucial modules: i) self-attention importance weighting, ii) ranking regularization, and iii) relabeling, as shown in Figure 2.

Given a batch of face images with some uncertain samples, we first extract the deep features by a backbone network. The self-attention importance weighting module assigns an importance weight for each image using a fully-connected (FC) layer and the sigmoid function. These weights are multiplied by the logits for a sample reweighting scheme. To explicitly reduce the importance of uncertain samples, a rank regularization module is further introduced to regularize the attention weights. In the rank regularization module, we first rank the learned attention weights and then split them into two groups, i.e. high and low importance groups. We then add a constraint between the mean weights of these groups by a margin-based loss, which is called rank regularization loss (RR-Loss). To further improve our SCN, the relabeling module is added to modify some of the uncertain samples in the low importance group. This relabeling operation aims to hunt more clean samples and then to enhance the final model. The whole SCN can be trained in an end-to-end manner and easily added into any CNN backbones.

3.2. Self-Attention Importance Weighting

We introduce the self-attention importance weighting module to capture the contributions of samples for training. It is expected that certain samples may have high importance weights while uncertain ones have low importance. Let $\mathbf{F} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in R^{D \times N}$ denotes the facial features of N images, the self-attention importance weighting module takes $\mathbf{F}$ as input, and output an importance weight for each feature. Specifically, the self-attention importance weighting module is comprised of a linear fully-connected (FC) layer and a sigmoid activation function, which can be formulated as ,

$$\alpha_i = \sigma(\mathbf{W}_a^\top \mathbf{x}_i), \tag{1}$$

where α_i is the importance weight of the *i*-th sample, $\mathbf{W}_a$ is the parameters of the FC layer used for attention, and σ is the sigmoid function. This module also provides reference for the other two modules.

Logit-Weighted Cross-Entropy Loss. With the attention weights, we have two simple choices to perform loss weighting inspired by [19]. The first choice is to multiply the weight of each sample by the sample loss. In our case, since the weights are optimized in an end-to-end manner and are learned from the CNN features, they are doomed to be zeros as this trival solution makes zero loss. MentorNet [22] and other self-paced learning methods [21, 32] solve this problem by alternating minimization, i.e. optimize one at a time while the other is held fixed. In this paper, we choose the logit-weighted one of [19] which is shown to be more efficient. For a multi-class Cross-Entropy loss, we call our weighted loss as Logit-Weighted Cross-Entropy loss (WCE-Loss), which is formulated as,

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\alpha_i \mathbf{W}_{y_i}^\top \mathbf{x}_i}}{\sum_{j=1}^{C} e^{\alpha_i \mathbf{W}_j^\top \mathbf{x}_i}}, \tag{2}$$

where $\mathbf{W}_j$ is the *j*-th classifier. As suggested in [30], the $\mathcal{L}_{WCE}$ has a positive correlation with the α.

3.3. Rank Regularization

The self-attention weights in the above module can be arbitrary in (0, 1). To explicitly constrain the importance of uncertain samples, we elaborately design a rank regularization module to regularize the attention weights. In the rank regularization module, we first rank the learned attention weights in descending order and then split them into two groups with a ratio β. The rank regularization ensures that the mean attention weight of high-importance group is higher than the one of low-importance group with a margin. Formally, we define a rank regularization loss (RR-Loss) for this purpose as follows,

$$\mathcal{L}_{RR} = \max\{0, \delta_1 - (\alpha_H - \alpha_L)\}, \tag{3}$$

with

$$\alpha_H = \frac{1}{M}\sum_{i=0}^{M} \alpha_i, \alpha_L = \frac{1}{N-M}\sum_{i=M}^{N} \alpha_i, \tag{4}$$

where δ_1 is a margin which can be a fixed hyper parameter or a learnable parameter, α_H and α_L are the mean values of the high importance group with $\beta * N = M$ samples and the low importance group with $N - M$ samples, respectively. In training, the total loss function is $\mathcal{L}_{all} = \gamma\mathcal{L}_{RR} + (1 - \gamma)\mathcal{L}_{WCE}$ where γ is a trade-off ratio.

3.4. Relabeling

In the rank regularization module, each mini-batch is divided into two groups, i.e. the high-importance and the low-importance groups. We experimentally find that the uncertain samples usually have low importance weights, thus an intuitive idea is to design a strategy to relabel these samples.

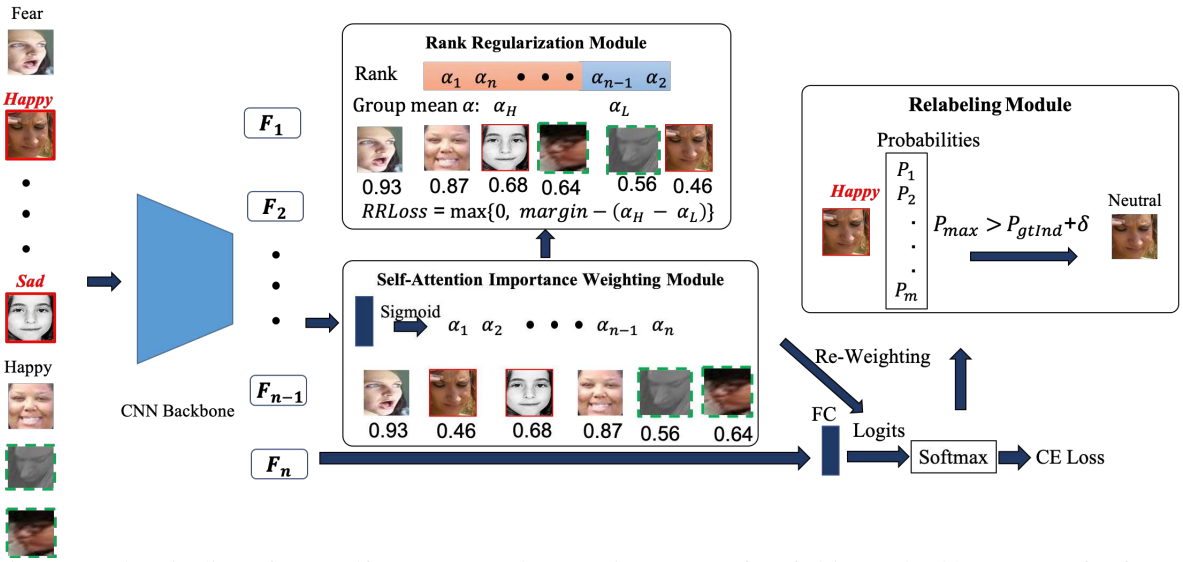

Figure 2: The pipeline of our Self-Cure Network. Face images are first fed into a backbone CNN for feature extraction. The self-attention importance weighting module learns sample weights from facial features for loss weighting. The rank regularization module takes as input the sample weights and constrain them with a ranking operation and a margin-based loss function. The relabeling module hunts reliable samples by comparing maximum predicted probabilities to the probabilities of given labels. *Mislabeled samples are marked in red solid rectangles and ambiguous samples in green dash ones. It is worth noting that SCN mainly resorts to the re-weighting operation to suppress these uncertainties and only modifies some of the uncertain samples.*

Table 1: The statistics of our WebEmotion.

Category	Happy	Sad	Surprise	Fear	Angry	Disgust	Contempt	Neutral	**Total**
# Videos	4,231	5,670	4,573	5,328	5,668	5,197	5,266	5,406	**41,339**
# Clips	27,854	29,667	27,418	29,822	31,483	20,764	6,454	26,687	**200,149**

The main challenge to modify these annotations is to know which annotation is incorrect.

Specifically, our relabeling module only considers the samples in the low-importance group and is performed on the Softmax probabilities. For each sample, we compare the maximum predicted probability to the probability of given label. A sample is assigned to a new pseudo label if the maximum prediction probability is higher than the one of given label with a threshold. Formally, the relabeling module can be defined as,

$$y' = \begin{cases} l_{max} & \text{if } P_{max} - P_{gtInd} > \delta_2, \\ l_{org} & \text{otherwise,} \end{cases} \tag{5}$$

where y' denotes the new label, δ_2 is a threshold, P_{max} is the maximum predicted probability, and P_{gtInd} is the predicted probability of the given label. l_{org} and l_{max} are the original given label and the index of the maximum prediction, respectively.

In our system, uncertain samples are expected to obtain low importance weights thus to degrade their negative impacts with re-weighting, and then fall into the low-importance group, and finally may be corrected as certain samples by relabeling. Those corrected samples may obtain high important weights in the next epoch. *We expect the network can be cured by itself with either re-weighting or relabeling, which is the reason why we call our method as self-cured network.*

3.5. Implementation

Pre-processing and facial features. In our SCN, face images are detected and aligned by MTCNN [52] and further resized to 224 × 224 pixels. The SCN is implemented with Pytorch toolbox and the backbone network is ResNet-18 [16]. By default, the ResNet-18 is pre-trained on the MS-Celeb-1M face recognition dataset and the facial features are extracted from its last pooling layer.

Training. We train our SCN in an end-to-end manner with 8 Nvidia Titan 2080ti GPU, and set the batch size as 1024. In each iteration, the training images are divided into two groups including 70% high importance samples and 30% low importance samples by default. The margin δ_1 between the mean value of high and low importance groups

can be either set at 0.15 by default or designed as a learnable parameter. Both strategies will be evaluated in the ensuing Experiments. The whole network is jointly optimized with RR-Loss and WCE-Loss. The ratio of the two losses is empirically set at 1:1, and its influence will be studied in the ensuing ablation study of Experiments. The leaning rate is initialized as 0.1 which is further divided by 10 after 15 epochs and 30 epochs, respectively. The training stops at 40 epochs. The relabeling module is included for optimization from the 10th epoch, where the relabeling margin δ_2 is set at 0.2 by default.

4. Experiments

In this section, we first describe three public datasets and our WebEmotion dataset. We then demonstrate the robustness of our SCN under uncertainties of both synthetic and real-world noisy facial expression annotations. Further, we conduct ablation studies with qualitative and quantitative results to show the effectiveness of each module in SCN. Finally, we compare our SCN to the state-of-the-art methods on public datasets.

4.1. Datasets

RAF-DB [24] contains 30,000 facial images annotated with basic or compound expressions by 40 trained human coders. In our experiment, only images with six basic expressions (neutral, happiness, surprise, sadness, anger, disgust, fear) and neutral expression are used which leads to 12,271 images for training and 3,068 images for testing. The overall sample accuracy is used for measurement.

FERPlus [4] is extended from FER2013 as used in the *ICML 2013 Challenges*. It is a large-scale dataset collected by the Google search engine. It consists of 28,709 training images, 3,589 validation images and 3,589 test images, all of which are resized to 48×48 pixels. *Contempt* is included which leads to 8 classes in this dataset. The overall sample accuracy is used for measurement

AffectNet [35] is by far the largest dataset that provides both categorical and Valence-Arousal annotations. It contains more than one million images from the Internet by querying expression-related keywords in three search engines, of which 450,000 images are manually annotated with eight expression labels as in FERPlus. It has imbalanced training and test sets as well as a balanced validation set. The *mean class accuracy* on the validation set is used for measurement.

The collected WebEmotion. Since the main evidence of uncertainties is the incorrect/noisy annotation problem, we collect an extreme noisy FER dataset from the Internet, termed as WebEmotion, to investigate the effect of SCN with extreme uncertainties. The WebEmotion is a video dataset (though we use it as image data by assigning labels to frames) downloaded from YouTube with a set of keywords including 40 emotion-related words, 45 countries from *Asia, Europe, Africa, America*, and 6 age-related words (i.e. *baby, lady, woman, man, old man, old woman*). It consists of the same 8 classes with FERPlus, where each class is connected to several emotion-related keywords, *e.g*. *Happy* is connected to the keywords *happy, funny, ecstatic, smug, and kawaii*. To obtain meaningful correlation between the keywords and the searched videos, only the top 20 crawled videos with less then 4 minutes are selected. This leads to around 41,000 videos which are further segmented into 200,000 video clips with a constraint that a face (detected by MTCNN) appears at least 5 seconds. For evaluation, we only use WebEmotion for pretraining since annotating is extremely difficult. Table 1 shows the statistics of WebEmotion. The meta videos and video clips will be public to the research community.

4.2. Evaluation of SCN on Synthetic Uncertainties

The uncertainties of FER mainly come from ambiguous facial expressions, low-quality facial images, inconsistent annotations, and incorrect annotations (*i.e*. noisy labels). Considering that only noisy labels can be analyzed quantitatively, we explore the robustness of SCN with three levels of label noises including the ratio of 10%, 20%, and 30% to RAF-DB, FERPLus, and AffectNet datasets. Specifically, we randomly choose 10%, 20%, and 30% of training data for each category and randomly change their labels to others. In Table 2, we use ResNet-18 as CNN backbone and compare our SCN to the baseline (traditional CNN training without considering label noises) with two training schemes: i) training from scratch and ii) fine-tuning with a pretrained model on Ms-Celeb-1M [15]. We also compare our SCN to two state-of-the-art noise-tolerant methods on RAF-DB, namely CurriculumNet [14] and MetaCleaner [53].

As shown in Table 2, our SCN consistently improves the baseline by a large margin. For scheme i) with noise ratio 30%, our SCN outperforms the baseline by 13.80% , 1.07%, and 1.91% on RAF-DB, FERPLus, and AffectNet, respectively. For scheme ii) with noise ratio 30%, our SCN still gain improvements of 2.20%, 2.47%, and 3.12% on these datasets though the performance of them are relatively high. For both schemes, the benefit from SCN becomes more obvious as the noise ratio increases up. CurriculumNet designs training curriculum by measuring data complexity using cluster density which can avoid training noisy-labeled data in early stages. MetaCleaner aggregates the features of several samples in each class into a weighted mean feature for classification which can also weaken the noisy-labeled samples. Both CurriculumNet and MetaCleaner improve the baseline largely but are still inferior to the SCN which is simpler. Another interesting find is that the improvement of SCN on RAF-DB is much higher than on other

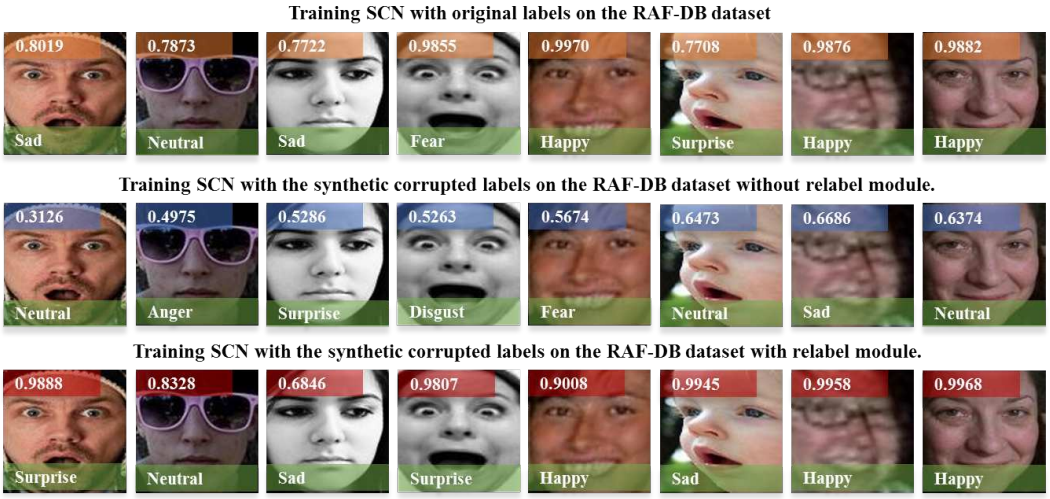

Figure 3: Visualization of the learned importance weights in our SCN, we show these weights on randomly selected images with original labels (1st row) and synthetic noisy labels before and after relabeling (2nd row and 3rd row).

Table 2: The evaluation of SCN on synthetic noisy FER datasets. 'Pretrain' means we use a pretrained model from face recognition, otherwise we train from scratch.

Pretrain	SCN	Noise(%)	RAF-DB	AffectNet	FERPlus
×	CurriculumNet [14]	10	68.5	-	-
×	MetaCleaner [53]	10	68.45	-	-
×	×	10	61.43	44.68	77.15
×	✓	10	**70.26**	**45.23**	**78.53**
×	CurriculumNet [14]	20	61.23	-	-
×	MetaCleaner [53]	20	61.35	-	-
×	×	20	55.5	41.00	71.88
×	✓	20	**63.50**	**41.63**	**72.46**
×	CurriculumNet [14]	30	57.52	-	-
×	MetaCleaner [53]	30	58,89	-	-
×	×	30	46.81	38.35	68.54
×	✓	30	**60.61**	**39.42**	**70.45**
✓	×	10	80.81	57.18	83.39
✓	✓	10	**82.18**	**58.58**	**84.28**
✓	×	20	78.18	56.15	82.24
✓	✓	20	**80.10**	**57.25**	**83.17**
✓	×	30	75.26	52.58	79.34
✓	✓	30	**77.46**	**55.05**	**82.47**

datasets. It may be explained by the following reasons. On the one hand, RAF-DB consists of compound facial expressions and is annotated by 40 people with crowdsourcing, which make the data annotations more inconsistent. Thus, our SCN may also gain improvement on the original RAF-DB without synthetic label noises. On the other hand, AffectNet and FERPlus are annotated by experts, thus less inconsistent labels are involved, leading to less improvement on RAF-DB.

Visualization of α in SCN. To further investigate the effectiveness of our SCN under noisy annotations, we visualize the importance weight α during the training phase of SCN on RAF-DB with noise ratio 10% . In Figure 3, the first row indicates the importance weights when SCN is trained with original labels. The images of the second row are annotated with synthetic corrupted labels, and we use SCN (without Relabel module) to train the synthetic noisy dataset. Indeed, the SCN regards those label-corrupted images as noises and automatically suppresses the weights of them. After sufficient training epochs, the relabeling module are added into SCN, and these noisy-labeled images are relabeled (of course many others may be not relabeled since we have relabeling constraint). After several other epochs, the importance weights of them become high (the 3rd row), which demonstrates that our SCN can 'self-cure' the corrupted labels. It is worth noting that the new labels from relabeling module may be inconsistent with "ground-truth" labels (see the 1st, 4th, and 6th columns) but they are also reasonable in visualization.

Table 3: The effect of SCN on WebEmotion for pretraining. The 2nd column indicates finetuning with or without SCN.

WebEmoition	SCN	RAF-DB	AffectNet	FERPlus
×	×	72.00	46.58	82.4
w/o SCN	×	78.97	56.43	84.20
w/o SCN	✓	80.42	57.23	85.13
SCN	✓	82.45	58.45	85.97

4.3. Exploring SCN on Real-World Uncertainties

Synthetic noisy data proves the effectiveness of the 'self-curing' ability of SCN. In this section, we apply our SCN to real-world FER datasets which can include all types of uncertainties.

SCN on WebEmotion for pretraining. Our collected WebEmotion dataset consists of massive noises since the

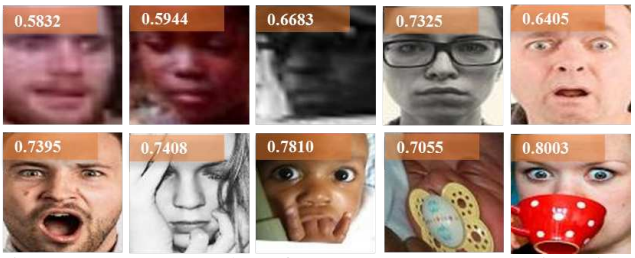

Figure 4: Ten examples of RAF-DB (w/o synthetic noisy labels) with low importance weights. Each column corresponds to a basic emotion. One can guess their labels and the ground-truth labels of RAD-DB are included in the text.

Table 4: SCN on real-world FER datasets. The improvements of SCN suggests that these public datasets more or less suffer from uncertainties.

Pretrain	SCN	RAF-DB	AffectNet	FERPlus
×	×	72.00	46.58	82.4
×	✓	**78.31**	**47.28**	**83.42**
×	CurriculumNet [14]	74.67	-	-
×	MetaCleaner [53]	77.18	-	-
✓	×	84.20	58.5	86.80
✓	✓	**87.03**	**60.23**	**88.01**

searching keywords are regarded as labels. To better validate the effect of SCN on real-world noisy data, we apply our SCN to WebEmotion for pretraining and then finetune the model on target datasets. We show the comparison experiments in Table 3. From the 1st and the 2nd rows, we can see that pretraining on WebEmotion without SCN improves the baseline by 6.97%, 9.85%, and 1.80% on RAF-DB, FERPlus and AffectNet, respectively. Fine-tuning with SCN on target datasets obtains gains ranged from 1% to 2%. Pretraining on WebEmotion with SCN further boosts the performance from 80.42% to 82.45% on RAF-DB. This suggests that SCN learns robust features on WebEmotion which is better for further fine-tuning.

SCN on Original FER datasets. We further conduct experiments on original FER datasets to evaluate our SCN since these datasets inevitably suffer from uncertainties such as ambiguous facial expressions, low-quality facial images, etc. Results are shown in Table 4. When training from scratch, our proposed SCN improves the baseline consistently with gains of 6.31%, 0.7%, and 1.02% on RAD-DB, AffectNet, and FERPlus, respectively. MetaCleaner also boosts the baseline on RAF-DB but slightly worse than our SCN. With pretraining, we still obtain improvements of 2.83%, 1.73%, and 1.21% on these datasets. The improvement of SCN and MetaCleaner suggests that there indeed exists uncertainties in those datasets. To validate our speculation, we rank the importance weights of RAF-DB, and show some examples with low importance weights in Figure 4. The ground-truth labels from top-left to bottom-right are *surprise, neutral, neutral, sad, surprise, surprise, neutral, surprise, neutral, surprise*. We find that images with low quality and occlusion are difficult to annotate and are more likely to have low-importance weights in SCN.

Table 5: Evaluation of the three modules in SCN.

Weight	Rank	Relabel	RAF-DB	RAF-DB (pretrain)
×	×	×	72.00	84.20
×	×	✓	71.25	83.78
×	✓	×	74.15	85.14
✓	×	×	76.26	86.09
✓	✓	×	76.57	86.63
✓	✓	✓	78.31	87.03

Table 6: Evaluation of the ratio γ between RR-Loss and WCE-Loss.

0.2	0.3	0.5	0.6	0.8
76.12%	76.35%	**78.31%**	76.57%	71.75%

4.4. Ablation Studies

Evaluation of the three modules in SCN. To evaluate the effect of each module of SCN, we design an ablation study to investigate WCE-Loss, RR-Loss and Relabel modules on RAF-DB. We show the experimental results in Table 5. Several observations can be concluded in the following. First, for both training schemes, a naive relabeling module (2nd row) added into the baseline (1st row) can degrade performance slightly. This may be explained by that many relabeling operations are wrong from the baseline model. It indirectly indicates that our elaborately-designed relabeling in the low-importance group with rank regularization is more effective. Second, when adding one module, we obtain the highest improvement by WCE-Loss which improves the baseline from 72% to 76.26% on RAF-DB. This suggests that the re-weighting is the most contributed module for our SCN. Third, the RR-Loss and the relabeling module can further boost WCE-Loss by 2.15%.

Evaluation of the ratio γ. In Table 6, we evaluate the effect of different ratios between the RR-Loss and WCE-Loss. We find that setting equal weight for each loss achieves the best results. Increasing the weight of RR-Loss from 0.5 to 0.8 dramatically degrades performance which suggests that WCE-Loss is more important.

Evaluation of δ_1 and δ_2. δ_1 is a margin parameter to control the mean margin between the high- and low-importance groups. For fixed setting, we evaluate it from 0 to 0.30. Figure 5 (left) shows the results for both fixed and learned δ_1. The default $\delta_1 = 0.15$ obtains the best performance, which shows that the margin should be an appropriate value. We also design a learnable paradigm of

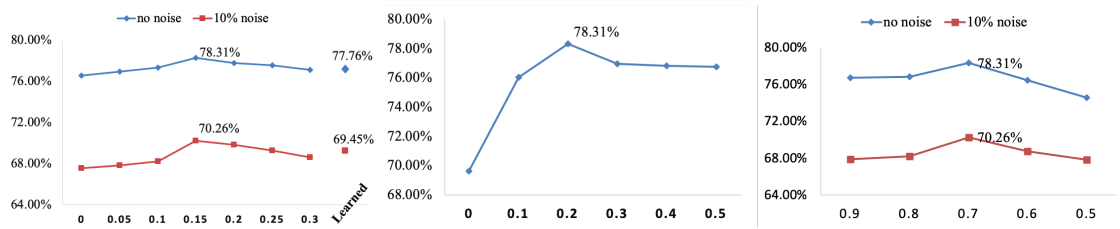

Figure 5: Evaluation of the margin δ_1 and δ_2, and the ratio β on the RAF-DB dataset.

Table 7: Comparison to the state-of-the-art results.*These results are trained using label distributions. $^+$Oversampling is used since AffectNet is imbalanced. ‡RAF-DB and AffectNet are jointly used for training. Note that IPA2LT tests with 7 classes on AffectNet.

(a) Comparison on RAF-DB.

Method	Acc.
DLP-CNN [24]	84.22
IPA2LT [50]	86.77
gaCNN [26]	85.07
RAN [45]	86.90
Our SCN (ResNet18)	**87.03**
Our SCN (ResNet18) ‡	**88.14**

(b) Comparison on AffectNet.

Method	mean Acc.
Upsample [35]	47.00
Weighted loss [35]	58.00
IPA2LT‡ [50] (7 cls)	55.71
RAN [45]	52.97
RAN$^+$ [45]	59.5
Our SCN$^+$(ResNet18)	**60.23**

(c) Comparison on FERPlus

Method	Acc.
PLD* [5]	85.1
ResNet+VGG [20]	87.4
SeNet50* [1]	88.8
RAN [45]	88.55
RAN-VGG16* [45]	89.16
Our SCN (ResNet18/IR50)	88.01/**89.35**

δ_1, and initialize it to 0.15. The learnable δ_1 converges to 0.142 ± 0.05 and the performances are 77.76% and 69.45% in original and noise RAF-DB datasets, respectively.

δ_2 is a margin to determine when to relabel a sample. The default δ_2 is 0.2. We evaluate δ_2 from 0 to 0.5 on original RAF-DB, and show the results in Figure 5 (middle). $\delta_2 = 0$ means we relabel a sample if the max prediction probability is larger than the probability of the given label. Small δ_2 leads to a lot of incorrect relabeling operations which may hurt performance significantly. Large δ_2 leads to few relabeling operations which converges to no relabeling. We get the best performance in 0.2.

Evaluation of the β. β is the ratio of high importance samples in a minibatch. We study different ratios from 0.9 to 0.5 in both synthetic noisy and original RAF-DB dataset. The results are shown in Figure 5 (right). Our default ratio is 0.7 that achieves the best performance. Large β degrades the ability of SCN since it considers few of the data is uncertain. Small β leads to over-consideration of uncertainties which decreases the training loss unreasonably.

4.5. Comparison to the State of the Art

Table 7 compares our method to several state-of-the-art methods on RAF-DB, AffectNet, and FERPlus. IPA2LT [50] introduces the latent ground-truth idea for training with inconsistent annotations across different FER datasets. gaCNN [26] leverages a patch-based attention network and a global network. RAN[45] utilizes face regions and original face with a cascade attention network. gaCNN and RAN are time-consuming due to the cropped patches and regions. Our proposed SCN does not increase any cost in inference. Our SCN outperforms these recent state-of-the-art methods with **88.14%**, **60.23%**, and **89.35%** (with IR50 [9]) on RAF-DB, AffectNet, and FERPlus, respectively.

5. Conclusion

This paper presents a self-cure network (SCN) to suppress the uncertainties of facial expression data thus to learn robust feature for FER. The SCN consists of three novel modules including self-attention importance weighting, ranking regularization, and relabeling. The first module learns a weight for each facial image with self-attention to capture the sample importance for training and is used for loss weighting. The ranking regularization ensures that the first module learns meaningful weights to highlight certain samples and suppress uncertain samples. The relabeling module attempts to identify mislabeled samples and modify their labels. Extensive experiments on three public datasets and our collected WebEmotion show that our SCN achieves state-of-the-art results and can handle both synthetic and real-world uncertainties effectively.

6. Acknowledge

This work is partially supported byScience and Technology Service Network Initiative of Chinese Academy of Sciences (KFJ-STS-QYZX-092), Guangdong Special Support Program (2016TX03X276), and National Natural Science Foundation of China (U1813218, U1713208), Shenzhen Basic Research Program (JCYJ20170818164704758, CXB201104220032A), the Joint Lab of CAS-HK.

References

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. *arXiv preprint arXiv:1808.05561*, 2018. 8

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. 2

[3] Samaneh Azadi, Jiashi Feng, Stefanie Jegelka, and Trevor Darrell. Auxiliary image regularization for deep cnns with noisy labels. *arXiv preprint:1511.07069*, 2015. 2

[4] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM ICMI*, 2016. 1, 5

[5] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM ICMI*, pages 279–283, 2016. 8

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[7] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 1

[8] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision. *arXiv preprint 1711.00313*, 2017. 2

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 8

[10] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *ICCV*, pages 2106–2112, 2011. 1

[11] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016. 1

[12] B. Fasel. Robust face analysis using convolutional neural networks. In *ICPR*, pages 40–43, 2002. 2

[13] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016. 2

[14] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *ECCV*, September 2018. 5, 6, 7

[15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *CoRR*, abs/1607.08221, 2016. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[17] Guosheng Hu, Li Liu, Yang Yuan, Zehao Yu, Yang Hua, Zhihong Zhang, Fumin Shen, Ling Shao, Timothy Hospedales, Neil Robertson, et al. Deep multi-task learning to recognise subtle facial expressions of mental states. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018. 1

[18] Guosheng Hu, Yongxin Yang, Dong Yi, Josef Kittler, William Christmas, Stan Z Li, and Timothy Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 142–150, 2015.

[19] Wei Hu, Yangyu Huang, Fan Zhang, and Ruirui Li. Noise-tolerant paradigm for training face recognition cnns. In *CVPR*, pages 11887–11896, 2019. 3

[20] Christina Huang. Combining convolutional neural networks for emotion recognition. In *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–4, 2017. 8

[21] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. In *NIPS*, pages 2078–2086, 2014. 3

[22] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint:1712.05055*, 2017. 2, 3

[23] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, and Raul Chandias Ferrari. Combining modality specific deep neural networks for emotion recognition in video. In *International Conference on Multimodal Interaction*, pages 543–550, 2013. 2

[24] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2852–2861, 2017. 1, 5, 8

[25] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, pages 1910–1918, 2017. 2

[26] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *TIP*, 28(5):2439–2450, 2018. 8

[27] Y. Li, J. Zeng, S. Shan, and X. Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, May 2019. 2

[28] Chengjun Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, April 2002. 2

[29] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159(C):126–136, 2015. 2

[30] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017. 3

[31] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and

emotion-specified expression. In *CVPRW*, pages 94–101, 2010. 1

[32] Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. Self-paced co-training. In *ICML*, pages 2275–2284, 2017. 3

[33] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3866–3870. IEEE, 2019.

[34] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *ICML*, pages 567–574, 2012. 2

[35] Ali Mollahosseini, Behzad Hasani, Mohammad H Mahoor, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *TAC*, 10(1):18–31, 2017. 1, 5, 8

[36] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, pages 1196–1204, 2013. 2

[37] Pauline C. Ng and Steven Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003. 2

[38] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803 – 816, 2009. 2

[39] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint:1406.2080*, 2(3):4, 2014. 2

[40] Yichuan Tang. Deep learning using linear support vector machines. *Computer Science*, 2013. 2

[41] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *T-PAMI*, 23(2):97–115, 2001. 1

[42] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France, 2010. 1

[43] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, pages 839–847, 2017. 2

[44] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, July 2017. 2

[45] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *arXiv preprint:1905.04075*, 2019. 2, 8

[46] Kai Wang, Jianfei Yang, Da Guo, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. Bootstrap model ensemble and rank loss for engagement intensity regression. In *2019 International Conference on Multimodal Interaction*, pages 551–556, 2019.

[47] Kai Wang, Xiaoxing Zeng, Jianfei Yang, Debin Meng, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. Cascade attention networks for group emotion recognition with face, body and image cues. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 640–645, 2018. 1

[48] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[49] Jianfei Yang, Kai Wang, Xiaojiang Peng, and Yu Qiao. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 594–598, 2018.

[50] Jiabei Zeng, Shiguang Shan, Xilin Chen, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, pages 222–237, 2018. 2, 8

[51] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 2

[52] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letter*, 23(10):1499–1503, 2016. 4

[53] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *CVPR*, June 2019. 5, 6, 7

[54] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 1