

SCN-SAM: A Modified Self-Cure Network for Facial Expression Recognition Under Face Masks

Qiang Wu, Kouichi Hamada, Masayuki Arai

Graduate School of Science and Engineering, Teikyo University, 1-1 Toyosatodai, Utsunomiya, Tochigi, Japan

21m107fl@stu.teikyo-u.ac.jp, hamada.kouichi.bq@teikyo-u.ac.jp, arai@ics.teikyo-u.ac.jp

Abstract—Due to COVID-19, wearing masks has become more common. However, it is challenging to recognize expressions in the images of people wearing masks. In general facial recognition problems, blurred images and incorrect annotations of images in large-scale image datasets can make the model's training difficult, which can lead to degraded recognition performance. To address this problem, the Self-Cure Network (SCN) effectively suppresses the over-fitting of the network to images with uncertain labeling in large-scale facial expression datasets. However, it is not clear how well the SCN suppresses the uncertainty of facial expression images with masks. This paper verifies the recognition ability of SCN on images of people wearing masks and proposes a self-adjustment module to further improve SCN (called SCN-SAM). First, we experimentally demonstrate the effectiveness of SCN on the masked facial expression dataset. We then add a self-adjustment module without extensive modifications to SCN and demonstrate that SCN-SAM outperforms state-of-the-art methods in synthetic noise-added FER datasets.

Index Terms—computer vision, facial expression recognition, occlusion, face mask, SCN-SAM

I. INTRODUCTION

Facial expression is one of the most natural, powerful, and universal signals that humans use to communicate their intentions and emotional state [1] [2]. The technology of facial expression recognition (FER) is becoming increasingly mature. This technology can be used in human-computer interaction [3], fatigue detection [4] [5], and many other fields. As FER algorithms have continuously improved over the past decades, the algorithms now perform well on some large-scale datasets collected in the laboratory or the wild, such as RAF-DB [6], FERPlus [7], AffectNet [8], CK+ [9], and EmotioNet [10].

Deep learning algorithms are also becoming popular. Most algorithms that achieve high accuracy on large-scale datasets incorporate deep learning. The development of FER research based on deep learning methods is mainly focused on model optimization and dataset expansion. However, for large-scale FER datasets collected from the Internet, high-quality annotation is difficult to achieve due to the subjectivity of the annotators and the uncertainty associated with the ambiguity in the images. Therefore, there are often noisy and incorrectly annotated images in these large-scale datasets. These images

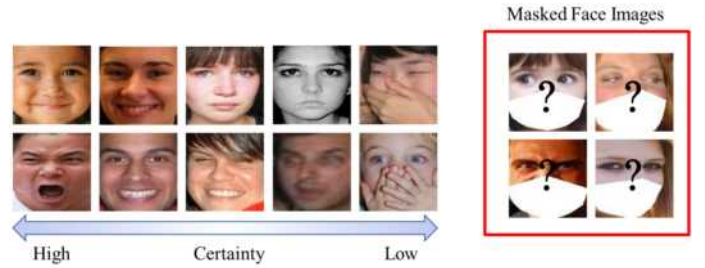


Fig. 1. A set of real-world facial expression images from RAF-DB. The images in the left group are facial expressions without masks, and those in the right group are facial expressions with masks. Some facial expressions without masks were difficult to distinguish. Images of facial expressions with masks are harder for both machines and humans to distinguish.

often interfere with the model's training process and make the model's training more difficult. Furthermore, due to the spread of COVID-19, it is becoming more common for people to wear masks daily. As shown in Fig. 1, the quality of facial expression images of people without masks affects the recognition of the images. However, the facial expression images of people with masks causes some features of the face to be lost due to occlusion, which makes expression recognition more difficult. And the labeling of sample data also becomes more uncertain in the deep learning approach. Therefore, from the perspective of sample training of models, this paper proposes the problem of ambiguous annotation for facial images of people wearing masks.

To address these problems, this paper proposes a Self-Cure Network (SCN) with a self-adjustment module (hereafter referred to as SCN-SAM) to suppress the uncertainties for large-scale facial expression recognition caused by the presence of masks in the images. This method is an improvement on the research of Wang [11]. We mainly introduce a self-adjustment module to the SCN to further improve the model.

The main contributions of this work are summarized as follows:

- For facial expression recognition under masks, we innovatively propose the uncertainty problem of masks on image annotation and propose SCN-SAM to reduce the

impact of uncertainty caused by masks.

- We improve the SCN and make it capable of classifying more expressions in the recognition of facial images with masks.
- SCN-SAM not only suppresses uncertainties on large-scale facial image datasets with masks but also performs very well on large-scale facial image datasets without masks.

II. RELATED WORK

A. Facial Expression Recognition Under Face Masks

Although most of the research on facial expression recognition is based on facial images without masks, with the impact of COVID-19 in recent years, research on facial expression recognition for images of people wearing masks has also begun to appear. There have been two primary types of research on facial expression recognition through facial images with masks. One is a method for direct recognition based on facial images with masks. The other is a method for recognition by restoring the features of the mask parts with a generative adversarial network (GAN). Yang et al. [12]’s research, which was based on the first method, aimed to solve the problem of classifying facial expressions into three categories (Positive, Neutral, Negative). They proposed a two-stage attention model to improve the accuracy of face-mask-aware facial expression recognition. They first trained a binary classifier to distinguish between facial images with or without masks. The attention heatmaps generated by this binary classifier are then used to guide the attention of the FER classifier to the critical regions of facial expression classification. In the second approach, Dong et al. [13] realized facial expression recognition based on the reconstruction of the parts occluded by masks. They proposed the two-stage Occlusion-Aware GAN (OA-GAN), which used one GAN to separate the occluded parts, and combined it with a second GAN to reconstruct the face. Nevertheless, using the GAN will increase the complexity and computational cost of the algorithm. For the labeling uncertainty of the sample caused by the mask, the above two methods do not solve the problem well.

B. Uncertainty Learning for Facial Expression Recognition

With the popularity of deep learning, FER research has begun to develop in the direction of large-scale datasets. However, high-quality annotation is challenging for large-scale FER datasets collected from the Internet. This is because ambiguous facial expressions, low-quality facial images, and inconsistent annotations increase the uncertainty in large-scale FER datasets. Zhang et al. [14] regarded uncertainty as a relative concept and proposed an innovative uncertainty learning method called Relative Uncertainty Learning (RUL). They used uncertainties as weights to mix facial features and designed an add-up loss to encourage uncertainty learning. She et al. [15] proposed a method to solve this problem from two perspectives: latent Distribution Mining and pairwise Uncertainty Estimation (DMUE). They introduced an auxiliary multi-branch learning framework to better mine and described

the latent distribution in the label space. They fully exploited the pairwise relationship of semantic features between instances to estimate the ambiguity extent in the instance space. Wang et al. [11] proposed a simple and effective Self-Cure Network (SCN) which suppresses the uncertainties efficiently and prevents deep networks from over-fitting to uncertain facial images. The above approaches address uncertainties in large-scale facial datasets in recent years. However, these methods do not introduce the uncertainty in recognizing facial expressions caused by the person wearing a mask. Our research is to improve SCN to reduce the influence of mask-wearing on facial expression uncertainty.

III. METHOD

To suppress the effect of masks on facial expressions, we improve the SCN and propose the SCN-SAM. This section provides an overview of the SCN-SAM and describes the role of each module within the model.

A. Overview of SCN-SAM

The framework proposed in this research is based on research by Wang [11]. The details of the framework are shown in Fig. 2.

Since we are trying to solve the problem of uncertainty in facial expression recognition caused by the person in the image wearing a mask, the first thing we need to do is data augmentation. A batch of images we obtained from the facial expression image dataset is processed by adding face masks to the images. We then use the backbone CNN network to extract the deep features of the masked facial images. The attention importance weighting module (AIWM) uses the sigmoid function to calculate the importance weight of each extracted deep feature. In order to reduce the importance of uncertain samples, we design a self-adjustment module (SAM) to divide the samples into high- and low-importance-weight categories. In the importance weighting regularization module (IWRM), we sort the obtained samples with high and low importance weights in descending order and divide them into two groups. In this module, we are constrained by a loss function called Importance Weighting Regularization loss (IWR-Loss). Its role is to suppress uncertain samples and strengthen the model’s learning of high-importance-weight samples. In order to repair the labels of those wrong samples, in the last relabeling module (RM), after training a certain number of epochs, we compare the labels with maximum predicted probabilities of the samples of low importance weights with the original labels when the difference exceeds a certain threshold. Afterward, the sample label is changed to the label with maximum predicted probabilities, thereby repairing the model’s training process.

B. Attention Importance Weighting Module (AIWM)

This module can calculate the importance of samples in model training. The higher the uncertainty of the sample, the lower the importance weight will be. Let $M = [X_1, X_2, X_3, \dots, X_N] \in R^{D \times N}$ denotes the facial features of N images. D represents the number of channels. Take M as

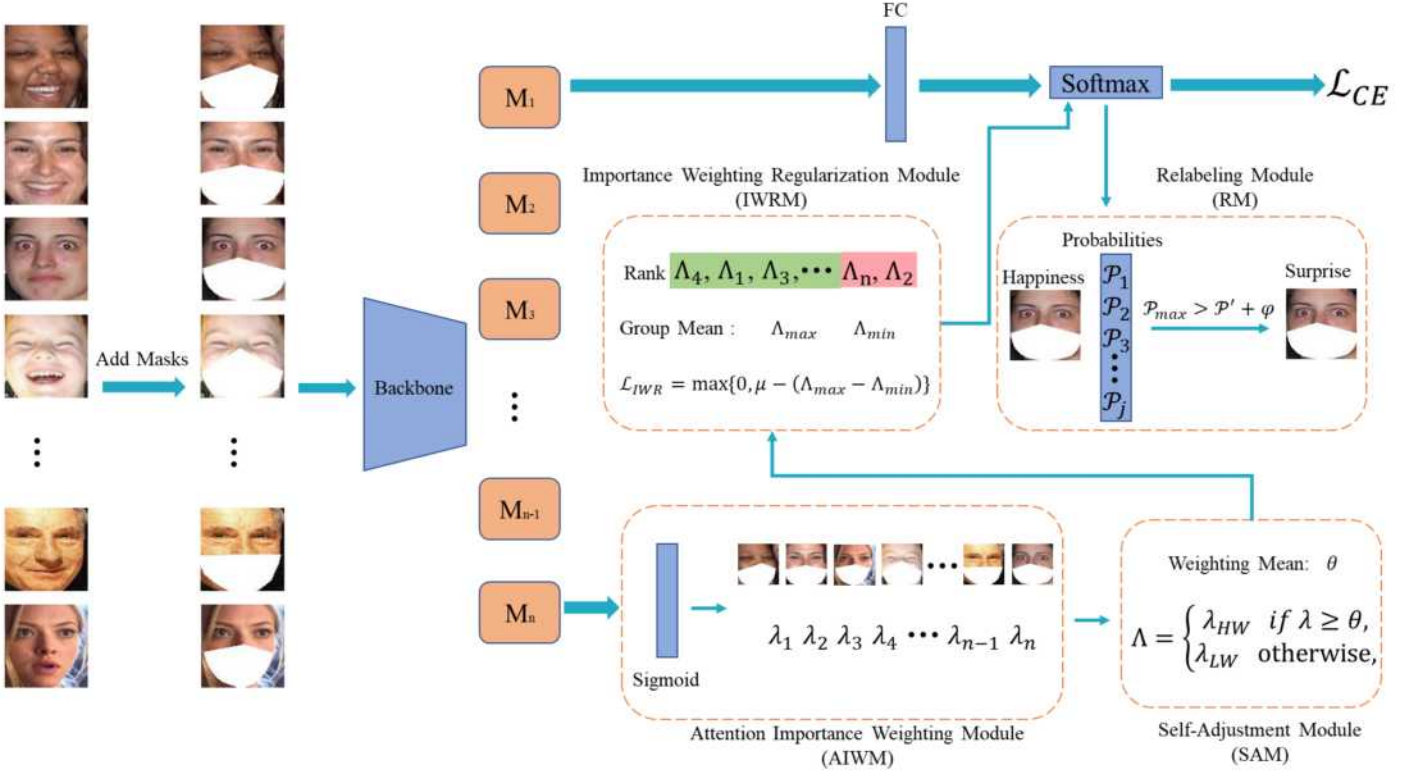


Fig. 2. The structure of our SCN-SAM, which includes an Attention Importance Weighting Module (AIWM), a Self-Adjustment Module (SAM), an Importance Weighting Regularization Module (IWRM), and a Relabeling Module (RM).

the input of this module, and the output is importance weight of each feature. The module consists of a fully-connected (FC) layer and a sigmoid activation function, which can be formulated as

$$\lambda_i = \text{Sigmoid}(W_\lambda^T X_i), \quad (1)$$

where λ_i is the importance weight of the i -th sample, and W_λ are the parameters of the FC layer.

C. Self-Adjustment Module (SAM)

Because the importance weight is determined according to the image's uncertainty, the weight of each image is not the same. The self-adjustment module classifies high and low weights according to the importance weights λ_i inputted to the module. We then average the obtained batch of importance weights. We called the weighting mean θ . This mean is used as a discriminant threshold, and when the weight is more significant than the threshold, it is classified as a high importance weight, λ_{HW} . Otherwise, the weight is classified as a low importance weight, λ_{LW} . Formally, it can be defined as

$$\theta = \frac{1}{N} \sum_{i=1}^N \lambda_i \quad (2)$$

with

$$\Lambda = \begin{cases} \lambda_{HW} & \text{if } \lambda \geq \theta, \\ \lambda_{LW} & \text{otherwise,} \end{cases} \quad (3)$$

where Λ is importance weight outputted by the module. λ_{HW} and λ_{LW} represent the high and low importance weights, respectively. θ is the mean value of importance weights with N samples.

D. Importance Weighting Regularization Module (IWRM)

In the importance weight regularization module, each batch of samples is automatically divided into two groups, the high-importance group and the low-importance group. Then, we count the data volume of incoming high-weight and low-weight samples separately. We calculate the mean of these high- and low-weight samples according to the data volume of the high- and low-weight samples. The definition of the IWR-loss function is

$$\mathcal{L}_{IWR} = \max\{0, \mu - (\Lambda_{max} - \Lambda_{min})\} \quad (4)$$

with

$$\Lambda_{max} = \frac{1}{\delta} \sum_{i=0}^{\delta} \Lambda_i, \quad (5)$$

$$\Lambda_{min} = \frac{1}{N - \delta} \sum_{i=\delta}^N \Lambda_i, \quad (6)$$

where μ is a margin that can be a learnable parameter, Λ_{max} is the mean value of the high-importance group, and Λ_{min} is the mean value of the low-importance group. Additionally, δ is the number of samples representing the high-importance

group. Therefore, the sample size of the low-importance group is $N - \delta$. Due to the effect of the self-adjustment module, δ changes with the training process. Therefore, the information of high-importance samples can be better captured. The total loss function for model training can be defined as

$$\mathcal{L}_{total} = (1 - \rho)\mathcal{L}_0 + \rho\mathcal{L}_{IWR} \quad (7)$$

ρ in the formula is the trade-off ratio of the two losses. In this study, it is set to 0.5. $\mathcal{L}_0$ is the cross-entropy loss.

E. Relabeling Module (RM)

High importance weights can make the model pay more attention to the training of these samples. However, to effectively utilize the samples of the low-importance group, this module compares the maximum predicted probability to the probability of a given label if the maximum predicted probability is higher than the given label with a threshold value. The sample is updated to a new label. It can be defined as

$$\ell = \begin{cases} \ell_{max} & \text{if } \mathcal{P}_{max} > \mathcal{P}' + \varphi, \\ \ell' & \text{otherwise,} \end{cases} \quad (8)$$

where ℓ denotes the new label, $\mathcal{P}_{max}$ is the maximum predicted probability, $\mathcal{P}'$ is the probability of the original label, and φ is a threshold. Here, ℓ_{max} and ℓ' are the label of the maximum predicted probability and the original label, respectively.

IV. EXPERIMENTS

A. Datasets

In FER, there are already collected datasets for researchers to conduct experiments. In this study, experiments were conducted on the following two datasets. Seven primary expression images (Neutral, Happiness, Surprise, Sadness, Anger, Disgust, Fear) from these two datasets are used in the preprocessing step where masks are added to the facial expression images.

RAF-DB [6]: It contains 30,000 facial images tagged with basic or compound expressions by 40 independent taggers. Among them, 12,271 images were used for the model's training, and 3,068 images were used in the testing phase of the model.

FERPlus [7]: It is an extension based on the FER2013 [16]. These images are from the Google search engine. We used 24,941 training images and 3,186 testing images for experiments. These images are all 48×48 pixels in size, and we also used seven basic facial expression images.

B. Preprocessing

Fig. 3 shows the flow of the mask-applying process used in this study. First, an unmasked facial image is inputted to the Face Alignment Network (FAN) [17]. FAN is a network based on the Hourglass network [18] and can detect 68 landmarks from a facial image, as shown in Fig. 3. In our study, 14 (red) of the 68 landmarks are used to draw a mask.

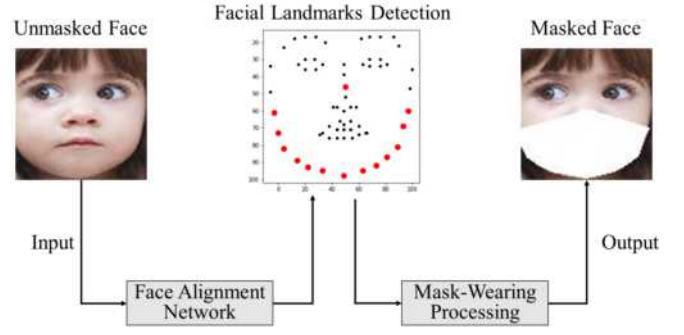


Fig. 3. A flowchart for adding a mask on an unmasked facial image. Left: Original image. Middle: Output of the Face Alignment Network. Right: Final result of preprocessing.

The results of the preprocessing are summarized in Table I. FER+ has a total of 28,127 images, of which 795 (2.83%), and RAF-DB has a total of 15,339 images, of which 114 (0.74%) could not be masked. For some face expression images, masks could not be added in the preprocessing stage because the facial landmarks could not be detected due to the low pixel quality of the images. Although there are images that could not be processed, the proportion of these images was not high, so this was not considered to affect the comparison experiments.

C. Evaluation of SCN-SAM on masked datasets

In order to show that our proposed model SCN-SAM can effectively suppress the uncertainty of facial expressions caused by masks, we choose ResNet-18 [19] as the baseline and the backbone of SCN-SAM and compare it to SCN [11]. Since SCN-SAM improves SCN, the parameters of the same module are set unchanged during the experiment to facilitate the comparison of experimental results and all models are pretrained on ImageNet with the backbone of ResNet-18. Finally, after completing 70 epochs of training, We select the optimal accuracy for comparison. The results are shown in Fig. 4.

As shown in Fig. 4, the SCN-SAM outperforms each baseline and SCN consistently. In the evaluation performed on the masked RAF-DB dataset, SCN does not suppress the uncertainty caused by the presence of masks well, and the performance is almost the same as the baseline model. However, our improved model, SCN-SAM, can suppress the influence of uncertainty caused by masks. The accuracy of the masked RAF-DB is 0.53% higher than that of SCN. The experimental results on the masked FERPlus are even more apparent. In the masked FERPlus, the SCN shows a negative performance, slightly worse than the accuracy of the baseline model. With the help of our proposed self-adjustment module, the accuracy of SCN-SAM is improved by 2.28%. This is because the data in FERPlus is more complex and contains a higher number of ambiguous and low-quality images. The increase in blurred images and the influence of ambiguous annotations due to the presence of masks causes the number δ of highly-importance samples to become very unstable during training instead remaining at a constant value. In model

TABLE I
TOTAL NUMBER OF IMAGES BY CLASS AND NUMBER OF IMAGES THAT COULD NOT HAVE MASKS APPLIED TO THEM.

	FERPlus						
	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
Total	3,564	594	143	8,152	3,365	2,387	9,922
Failures	113	19	0	134	238	61	230
Failure Ratio (%)	3.2	3.2	0	1.6	7.1	2.6	2.3

	RAF-DB						
	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
Total	1,619	355	877	5,957	2,460	867	3,204
Failures	11	4	2	21	29	13	34
Failure Ratio (%)	0.7	1.1	0.2	0.4	1.2	1.5	1.1

training, the self-adjustment module of SCN-SAM mines high-importance samples and classifies high and low-importance samples so that the δ value is no longer constant, thus solving this problem well. It can also be seen that our proposed SCN-SAM has more robust performance in more complex datasets and better ability to repair data labels.

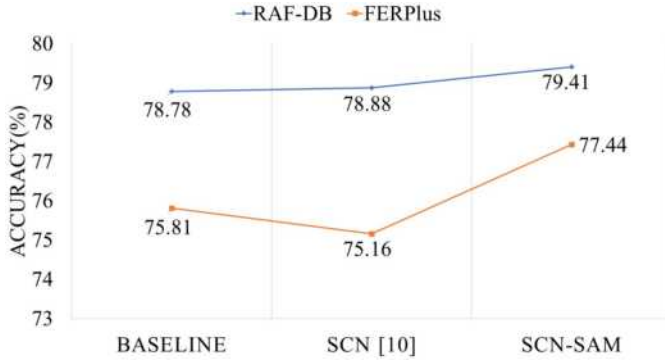


Fig. 4. Comparison to SCN on masked RAF-DB and FERPlus.

D. Evaluation on Synthetic Uncertainties

We have demonstrated that the proposed SCN-SAM can effectively suppress the uncertain effect of masks on facial expressions. In order to verify whether the model also works in the annotation ambiguity of unmasked facial expression datasets, we conduct the following experiments. The annotation ambiguity in the dataset is mainly manifested in the uncertain annotation of the annotator caused by the annotation error and the uncertainty of the image. We quantitatively evaluate the improvement of SCN-SAM against the mislabeled annotations on RAF-DB. In detail, we randomly select a part of the training samples and modify the labels to other categories. Then we also choose ResNet-18 [19] as the backbone of SCN [11], DMUE [15], and SCN-SAM, and finally compare these three models. SCN and DMUE are state-of-the-art noise-tolerant FER methods. In this study, we use ImageNet for the pre-training parameters of the three models and randomly select 10%, 20%, and 30% of the training samples as noise samples, respectively.

As summarized in Table II, our proposed SCN-SAM achieves the best results after adding 10%, 20%, or 30%

TABLE II
EVALUATION OF SCN-SAM ON RAF-DB WITH SYNTHETIC NOISE-ADDED ANNOTATIONS.

Method	Noise (%)	Accuracy (%)
SCN [11]	10	82.18
DMUE [15]	10	83.19
SCN-SAM (Ours)	10	85.37
SCN [11]	20	80.10
DMUE [15]	20	81.02
SCN-SAM (Ours)	20	83.05
SCN [11]	30	77.46
DMUE [15]	30	79.41
SCN-SAM (Ours)	30	80.22

synthetic noise samples to the training samples. With a 10%, 20%, and 30% noise ratio, SCN-SAM improves the accuracy by 3.19%, 2.95%, and 2.76%, respectively. This is attributed to the fact that SCN-SAM can mine high-importance-weight samples to change the proportion of high- and low-importance-weight samples, thereby improving the model's learning of the training samples. Additionally, the accuracy results obtained by SCN-SAM under different noise ratios are 2.18%, 2.03%, and 0.81% higher than the latest study DMUE [15], respectively. We cannot analyze the excellence of the two models in detail cause of the different methods and purposes. Nevertheless, in terms of accuracy, SCN-SAM outperforms the latest DMUE [15]. Thus, SCN-SAM can also play a role in suppressing label ambiguity on FER datasets without masks.

In order to further verify the effectiveness of the self-adjustment module (SAM), we also performed ablation experiments on the FERPlus dataset. The experimental results are summarized in Table III. Under different ratios of noise data, the experimental group with the SAM module can obtain the best performance. Therefore, focusing on the high importance samples during the training phase is essential. SAM's mining of high-importance samples can more accurately guide the direction of model learning and prevent the model from learning low-importance samples by mistake.

V. CONCLUSIONS AND FUTURE WORKS

In order to effectively suppress the problem of annotation ambiguity of images of people wearing masks, we propose

TABLE III
EVALUATION OF SELF-ADJUSTMENT MODULE (SAM) OF SCN-SAM ON FERPLUS WITH SYNTHETIC NOISE-ADDED ANNOTATIONS.

SCN	SAM	Noise (%)	Accuracy (%)
✓		10	84.12
✓	✓	10	84.28
✓	✓	10	86.63
✓		20	83.18
✓	✓	20	83.17
✓	✓	20	85.12
✓		30	79.88
✓	✓	30	82.47
✓	✓	30	83.27

SCN-SAM and design a self-adjustment module. This module can mine high-importance-weight samples in training samples and change the proportion of high- and low-importance-weight samples so that the model can pay more attention to learning the sample features of high importance weight. For making full use of the samples with low importance weights, after completing a certain number of training epochs, SCN-SAM will modify the labels of the samples with low importance weights to repair the dataset. Experiments on masked datasets and unmasked synthetic ambiguous datasets show that SCN-SAM can suppress the uncertainty in recognizing masked facial expressions and suppress annotation ambiguity in the facial expression datasets without masks.

In the follow-up work, we will use SCN-SAM to investigate further different computer vision fields to improve the model's generality. In addition to facial expression recognition, there are many other fields in computer vision where ambiguous annotation is a problem, such as object detection or face recognition. We believe that the proposed method will also be effective in these fields. After completing the improvement of the model's generality, we will extend the model to real life to develop the application.

REFERENCES

- [1] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. "Recognizing action units for facial expression analysis". In: *IEEE TPAMI* 23.2 (2001), pp. 97–115.
- [2] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.
- [3] Huan-Huan Wang and Jing-Wei Gu. "The applications of facial expression recognition in human-computer interaction". In: *ICAM*. 2018, pp. 288–291.
- [4] Mohammad Amin Assari and Mohammad Rahmati. "Driver drowsiness detection using face expression recognition". In: *ICSIPA*. 2011, pp. 337–341.
- [5] Sunsern Cheamanunkul and Sanchit Chawla. "Drowsiness detection using facial emotions and eye aspect ratios". In: *ICSEC*. IEEE. 2020, pp. 1–4.
- [6] Shan Li, Weihong Deng, and JunPing Du. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild". In: *CVPR*. 2017, pp. 2584–2593.
- [7] Emad Barsoum et al. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution". In: *ACM ICMI*. 2016.
- [8] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Transactions on Affective Computing* 10.1 (2019), pp. 18–31.
- [9] Patrick Lucey et al. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In: *CVPRW*. 2010, pp. 94–101.
- [10] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. "EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild". In: *CVPR*. 2016, pp. 5562–5570.
- [11] Kai Wang et al. "Suppressing uncertainties for large-scale facial expression recognition". In: *CVPR*. 2020, pp. 6897–6906.
- [12] Bo Yang, Wu Jianming, and Gen Hattori. "Face mask aware robust facial expression recognition during the COVID-19 pandemic". In: *ICIP*. 2021, pp. 240–244.
- [13] Jiayuan Dong et al. "Occlusion-aware gan for face de-occlusion in the wild". In: *ICME*. 2020, pp. 1–6.
- [14] Yuhang Zhang, Chengrui Wang, and Weihong Deng. "Relative Uncertainty Learning for Facial Expression Recognition". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17616–17627.
- [15] Jiahui She et al. "Dive Into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition". In: *CVPR*. June 2021, pp. 6248–6257.
- [16] Ian J Goodfellow et al. "Challenges in representation learning: A report on three machine learning contests". In: *International conference on neural information processing*. Springer. 2013, pp. 117–124.
- [17] Adrian Bulat and Georgios Tzimiropoulos. "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)". In: *ICCV*. 2017, pp. 1021–1030.
- [18] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation". In: *European conference on computer vision*. Springer. 2016, pp. 483–499.
- [19] Kaiming He et al. "Deep residual learning for image recognition". In: *CVPR*. 2016, pp. 770–778.