IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 May 2024, accepted 13 June 2024, date of publication 17 June 2024, date of current version 24 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3415154

RESEARCH ARTICLE

Enhancing Facial Expression Recognition Under Data Uncertainty Based on Embedding Proximity

NING CHEN[1], VEN JYN KOK[1], AND CHEE SENG CHAN[2], (Senior Member, IEEE)
[1]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

Corresponding author: Ven Jyn Kok (vj.kok@ukm.edu.my)

ABSTRACT Facial Expression Recognition (FER) on unconstrained datasets poses a significant challenge, primarily due to data uncertainty stemming from human subjectivity and ambiguous facial expressions. Previous methods attempt to address this issue through relabeling strategies. However, this work reveals a relabel inconsistency problem. Specifically, the model weights are not updated simultaneously with the relabeling process. Consequently, the feature representations of the noisy samples remain associated with the previous label despite being relabeled. As a result, the relabeling mechanism reverts the new label back to the previous one, initiating a cycle between the two classes during the subsequent training. The failure to ''shift'' the feature representations closer to the new label centers hinders the model from learning discriminative features capable of handling data uncertainty, leading to degraded performance. In this work, a new framework based on embedding proximity is proposed to ensure consistent updating of feature representations with rectifications made during relabeling to overcome this limitation. This is achieved by pushing relabeled images closer to their newly assigned class centers and farther away from their previous class (wrong) centers in the feature embedding space. Through comprehensive experiments, this work utilizes existing models–SCN, RUL, and DMUE–to map the original feature space and then applies the proposed embedding proximity technique to update the feature representations. The updated models, denoted as SCN-C, RUL-C, and DMUE-C, demonstrate significant improvements in addressing inconsistency issues and enhancing overall performance. The proposed models outperform state-of-the-art methods, achieving accuracies of 65.73% on AffectNet, 89.51% on RAF-DB, and 71.83% on FER2013.

INDEX TERMS Data uncertainty, facial expression recognition, feature embedding space, relabeling.

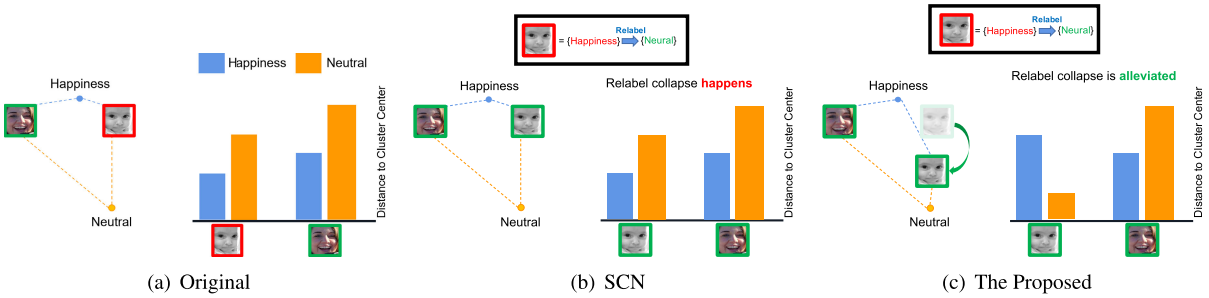
I. INTRODUCTION

Facial expressions play a pivotal role in the daily interactions of humans, serving not only as a medium for conveying emotions but also as a nonverbal way of interpersonal communication [1]. Automatic facial expression recognition (FER) holds tremendous practical importance across various fields, including psychology [2], human-computer interaction (HCI) [3], healthcare [4], service robots [5], and security systems [6]. In recent years, many deep learning-based strategies have been proposed and achieved promising performance with the emergence of unconstrained large-scale datasets, such as AffectNet [7] and FERPLUS [8].

Nevertheless, unconstrained large-scale datasets often suffer from inconsistent and noisy annotations due to the inherent uncertainty of human expressions and the subjective nature of annotators [9]. Consequently, the training of FER models with label noise has consistently remained a focal point in ongoing research efforts [9], [10], [11], [12], [13].

Various methods have been proposed to tackle the challenge of uncertainty in emotion recognition, demonstrating significant progress. Some approaches involve using a small, clean dataset to assess annotations [14], [15], while others focus on learning label distributions [12], [16]. Another

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.



(a) Original (b) SCN (c) The Proposed

FIGURE 1. (a) Original distance to cluster centers [Blue: *Happiness* and Orange: *Neutral*] represented in a bar chart (before relabeling). (b) After SCN (Self-Cure Network) relabeling module [9] (*Happiness* to *Neutral*), relabel inconsistency happens. Despite rectifying the label of the mislabeled image, the distance of the relabeled image to its new corresponding label centers (i.e., *Neutral*) remains unchanged (the blue and orange bars of the baby image are the same as in (a)). (c) The proposed framework encourages the model to update, aligning with the changes made during relabeling. The relabeled image is closer to its new corresponding class centers (i.e., *Neutral*) in the embedding space. A lower bar indicates a closer distance to its corresponding cluster center, and vice-versa. Best viewed in color.

set of methods addresses uncertainty by rectifying noisy annotations through relabeling noisy images [9], [17], [18]. Despite the observed performance gains, the first two strategies often demand substantial labour and time, rendering their application in real-world scenarios impractical [19], [20]. Consequently, there has been significant attention on suppressing uncertainties by directly correcting noisy annotations. However, empirical findings indicate that existing relabeling-based methods, even after introducing new (corrected) labels, result in the model producing feature representations that remain closer to the centers of the previous labels in the feature embedding space.

This inconsistency between the new labels and feature representations adversely impacts model training, preventing the model weights from being appropriately updated in the next epoch to "shift" the feature representations closer to the new respective label centers. As a consequence, in the following epoch, the relabeling mechanism consistently reverts the new label to the previous label due to the lack of proximity to the new class center. This cycle repeats, creating a persistent thread in the feature embedding space between the two class centers.

For example, as illustrated in Fig. 1(b), when the (baby) image is relabeled from its original expression of *Happiness* to *Neutral*, it is observed that its embedding space remains unchanged compared to the original setting in Fig. 1(a). In other words, the (baby) image is still associated with the *Happiness* class. This work terms this phenomenon the ***relabel inconsistency problem***. In principle, after an image is relabeled, the model should be updated to reflect the rectification made. In the aforementioned example, the relabeled (baby) image should be closer to the newly relabeled class center (*Neutral*) instead of the *Happiness* class center, as shown in Fig. 1(c).

To alleviate this *relabel inconsistency problem*, this work proposed a new framework based on embedding proximity to ensure that the FER model is updated when relabeling ambiguous images. Specifically, the proposed framework learns the center for each facial expression class. During training, the proposed framework concurrently updates these centers while minimizing the distances between image samples and their respective class centers. Consequently, following the relabeling of an image, the feature embedding space accurately reflects the rectification made. In other words, the relabeled image is now closer to the new corresponding class center, rather than persisting with the old class center in the embedding space during subsequent training epochs, as depicted in Fig. 1(c). As such, the proposed framework empowers the network to learn a discriminative embedding space characterized by compactness within classes and separation between classes. This capability ensures that the network can effectively manage data uncertainty.

Empirically, this paper shows that the proposed framework can improve the performance of existing relabeling-based methods (i.e. SCN [9]) by 0.63%, 0.88%, 2.07%, and 1.34% on AffectNet, RAF-DB, FERPLUS, and FER2013 datasets, respectively. Extensive experiments conducted on these four benchmarks also demonstrate that the proposed solution can significantly enhance the performance of state-of-the-art (SOTA) methods that do not utilize relabeling techniques.

The main contributions of this work can be summarized as follows:

- This work identifies and highlights a fundamental issue, namely the *relabel inconsistency problem*, inherent in the existing relabeling strategy. Specifically, after relabeling ambiguous images, the feature representations fail to update to accurately reflect the rectifications made in the embedding space. Consequently, the training process continues to grapple with the uncertainty problem stemming from the data.
- To tackle the *relabel inconsistency problem*, this work proposes a new framework that penalises the distance between image samples and the learned centers of facial expression classes during training. The proposed solution is crucial for bridging the gap in the existing relabeling mechanism used to suppress data uncertainty in emotion recognition, especially in the wild. The aim is to ensure that the feature representations are appropriately updated after relabeling noisy images.

- Extensive experiments on four in-the-wild FER datasets validate the superiority of the proposed solution over the existing relabeling methods and SOTA alternatives in suppressing data uncertainty for accurate emotion recognition.

II. RELATED WORK

A. FACIAL EXPRESSION RECOGNITION

Generally, a FER task includes three steps: face detection, expression feature extraction and emotion recognition. These three steps are indispensable to accomplish a precise FER, in which feature extraction determines the performance [21]. In the detection stage, common tools like Multi-task CNN (MTCNN) [22] and Dlib [23] are utilized to locate faces, which can be further aligned alternatively. In terms of feature extraction, emotion recognition strategies can be categorized into traditional and deep learning methods, depending on the techniques utilized. Local binary pattern (LBP) [24], [25], Gabor wavelet transform [26], k-nearest neighbour (KNN) [27], and support vector machine (SVM) [28], to cite a few, are examples of traditional approaches. Instead, [10], [13], [29], [30], [31], [32], [33], [34], [35], [36] concern deep learning based methods. For example, Zeng et al. [30] were the first to consider uncertainties and the challenge of inconsistent annotations. Kim et al. [35] proposed a novel scheme for expression recognition systems based on hierarchical deep learning. Zhu et al. [13] developed a convolutional relation network (CRN) for emotion recognition in the wild, leveraging feature similarity comparisons among sufficient expression samples to identify new classes using limited training images.

B. UNCERTAINTY LEARNING IN THE FACE DOMAIN

Noisy samples are often outliers with high variability in the embedding space, which can hamper or even harm performance [37]. A high proportion of noisy labels can also prevent the model from converging in the early optimization stage [9]. Among the recent contributions, the general methods of dealing with data uncertainty are to use a small set of clean data to evaluate the annotation quality during training [14], [15], [38], to estimate the noise distribution [8], [39], [40], [41], or to train a feature extractor [42], [43]. For example, Li et al. [14] utilized a small clean dataset and knowledge graph to guide the unified distillation framework to disambiguate mislabeled labels. Veit et al. [15] proposed a multi-task framework to clean incorrect labels and classify images. Sukhbaatar and Fergus [39] proposed a noise layer on top of softmax to "absorb" label noise. In [44], a noise-tolerant paradigm was introduced for learning facial features. This method employed Θ values from samples in a Θ-distribution to determine the likelihood of their cleanliness. Moreover, Zhao et al. [43] trained the EfficientFace network based on the local and global-salient features, which benefited from the feature extractor. However, annotating a clean dataset is usually expensive and time-consuming, sometimes even impossible [19].

Other methods adopt an alternative strategy by not relying on clean datasets but imposing additional constraints or distributions on noisy labels. One common strategy is to design a specific loss function, such as triplet loss [45], [46], [47]. Xie et al. [45] proposed a novel triplet loss that relied on class-pair margins and multistage outlier suppression, aimed to achieve inter-class separability and intra-class compactness of feature embedding spaces. However, mining hard triplets is time-consuming, and the criterion for defining "good" hard triplets remains unclear [48]. These methods also suffer the limitation of random sampling of triplets, which leads to slow convergence in the training process [49]. Another exploration focuses on introducing label distribution [12], [16], [50], [51]. The latent Distribution Mining and the pairwise Uncertainty Estimation (DMUE) [12] used an auxiliary multi-branch to model the latent label distribution of emotion images and used cosine similarity to capture the uncertainty. Like [45], this work also focused on minimizing the intra-class distance to facilitate the network to learn discriminative features as a better recognition system is built on an efficiently discriminated space [52]. However, these methods still suffer from the uncertainty problem inherent in datasets that cannot be directly addressed from the single instance perspective [12]. Moreover, it is costly to provide label distributions when dealing with large-scale datasets [20]. Some researchers attempt to deal with the noise label problem in FER from the perspective of regularization [53], [54], [55]. Zhang et al. [53] introduced a new Erasing Attention Consistency (EAC) approach to suppress noisy facial images during training. Gao et al. [54] proposed a SNEFER method to stop the negative effect of noisy annotations adaptively by a contrastive regularization term. However, regularization-based methods typically aim to smooth the learning process, which might lead to the loss of subtle but essential features in the sample that could be critical for distinguishing similar facial expressions.

Recently, there has been a growing interest among researchers in directly rectifying noisy labels through a relabeling mechanism during the training process [9], [17], [18]. For example, the Self-Cure Network (SCN) method proposed by Wang et al. [9] relied on a relabeling module to correct potentially inaccurate emotion labels based on the maximum predicted probabilities of labels. In contrast to a fixed threshold in SCN, Li et al. [17] suggested a dynamic relabeling module where the threshold is adjusted according to the probabilities of the given labels. While these methods have proven useful in suppressing uncertainties, they often overlook the importance of maintaining consistency between labels and feature representations, causing the noisy samples to be relabeled back and forth during training. This limitation hinders the effectiveness of FER models. This observation prompted this work to investigate more deeply into the inconsistency problem associated with relabeling techniques and explore potential solutions.

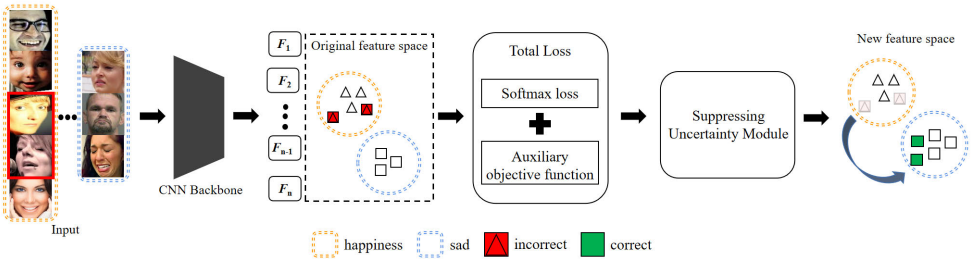


FIGURE 2. The pipeline of the proposed framework. Images in the orange and blue dotted bounding box correspond to the *happiness* and *sad* expression classes, respectively. The images in the red bounding box correspond to noisy samples belonging to the *sad* expression class but are incorrectly labeled as *happiness*. Visualization of the image features in the original feature space is shown in the black dotted bounding box. Noisy samples are represented as red rectangles with embedded triangles in the feature embedding space. The green rectangles correspond to samples accurately relabeled and "shifted" closer to the sad expression class center. Best viewed in color.

Given the challenges posed by noisy labels and relabeling inconsistencies in existing FER methods, this study introduces a novel framework that leverages embedding proximity to update FER models when relabeling ambiguous images. By simultaneously updating the center for each facial expression class and minimizing the distances between image samples and their respective class centers, the proposed approach significantly enhances the discriminative capability of the learned features. This effectively reduces data uncertainty, leading to more robust and accurate FER models.

III. METHODOLOGY

This work opts to elucidate the relabeling technique using SCN [9] as a baseline study. Hence, this section begins by revisiting SCN and is followed by the motivation behind this work. Finally, the proposed framework is detailed. The pipeline of the proposed framework is illustrated in Fig. 2, and its elaboration is in Sec. III-C.

A. SCN REVISIT

SCN [9] comprises three main modules: i) self-attention importance weighting module that consists of a fully connected layer and a sigmoid function. It is responsible for assigning an importance weight to each sample where noisy samples are expected to be given a low importance weight. ii) ranking regulation module first ranks the important weights in descending order. Then, it divides the samples into high and low importance groups according to the weight values. In this module, a rank regularization loss (RR Loss) is proposed to ensure that the mean importance weight of the high-importance group is higher than that of the low-importance group with a margin. iii) relabeling module that attempts to relabel samples in the low importance group to suppress uncertainty.

The relabeling module of SCN is performed after the softmax probabilities. Specifically, if the maximum predicted probability for a sample is greater than the probability of the given label with a threshold, a pseudo-label will be assigned to the sample. Otherwise, the original label will be retained.

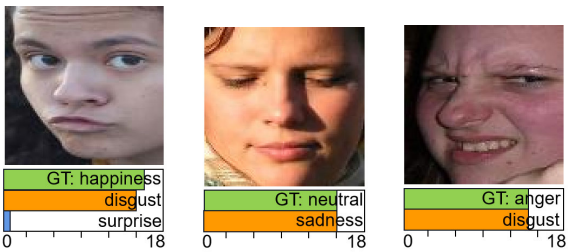


FIGURE 3. Sample images that are being alternately relabeled between the ground truth (GT) label and other facial expression labels. The different color bars (i.e. green, orange and blue) represent the number of times (x-axis) the image is relabeled with the corresponding expressions. Best viewed in color.

The relabeling module is defined as follows:

$$y' = \begin{cases} y_{max} & \text{if } P_{max} - P_{gt} > \delta, \\ y_{org} & \text{otherwise,} \end{cases} \tag{1}$$

where y' represents the new label of the sample, P_{max} and P_{gt} are the maximum predicted probability and the probability of the given label from training, respectively. δ is a threshold hyper parameter. y_{max} is the label corresponding to the maximum predicted probability, and y_{org} is the original label.

B. MOTIVATION

Practically, Equation (1) in SCN [9] has shown to be useful in rectifying noisy labels, where samples are relabeled based on the maximum predicted probability of the label. However, this work empirically observed a peculiar phenomenon where the label of a random set of noisy images tends to be relabeled back and forth in each epoch during the training stage (sample images as shown in Fig. 3). As an example, in the second image in Fig. 3, the facial expression was relabeled between *Neutral* and *Sadness* 30 times.

As the relabeling module is implemented at the end of the pipeline, following the computation of softmax probabilities in SCN, the conducted experimental investigations in this work revealed a phenomenon where the model weights cannot be updated simultaneously. As a result of this, the

Algorithm 1 The Algorithm of the Proposed Framework

Input: Training dataset, $(x_i, y_i)_{i=1}^{M}$,
Initialize CNN parameters;
Output: Updated FER model weights and feature embedding space;
while *not converged* **do**
  Sample a mini-batch of size N from the training dataset, $(x_i, y_i)_{i=1}^{N}$;
  Compute the image features using the CNN;
  Compute the auxiliary objective function L_C;
  Compute the softmax loss L_S by (2);
  Compute the total loss L_{Total} by (3);
  Update the model parameters by loss backpropagation;
  Relabel ambiguous sample images by (1);
end

image feature embedding space remains unchanged, causing the feature representations to persistently stay closer to the previous labels in the feature embedding space despite being relabeled. Consequently, in the subsequent training epoch, the model weights cannot be adequately updated to ''shift'' feature representations closer to the new label centers, resulting in the noisy samples still being associated with the previous label. This lack of update causes the model to repeatedly relabel the noisy samples back and forth between the two labels. This paper terms this phenomenon the *relabel inconsistency problem*.

Ideally, after a sample is relabeled, the feature representation should be updated simultaneously to reflect the rectification made. In other words, the relabeled image should be closer to the new corresponding class center rather than the old class center, as depicted in Fig. 2. This is essential to ensure efficient FER model training to suppress data uncertainty.

In summary, existing relabeling techniques suffer from the *relabel inconsistency problem*, where model weights are not updated simultaneously with the relabeling process. This limitation hampers the model's potential ability to cope with data uncertainty. Therefore, solving this problem is critical for enhancing emotion recognition, especially in wild conditions.

C. PROPOSED FRAMEWORK

The core idea of the proposed solution to mitigate the *relabel inconsistency problem* is to ensure the feature representations are updated concurrently with the relabeling of noisy samples. This work accomplishes this by ensuring the model will undergo a weights update stage at each epoch during the training phase. Fig. 2 illustrates the key steps of the proposed framework.

Given a mini-batch of training images with feature vectors $F = [x_1, x_2, \cdots, x_N] \in R^{D\times N}$, the deep features of the original feature space are first extracted using a convolutional neural network (CNN) backbone network. Here, x_i is the feature vector of the i-th image in the feature embedding space, y_i as the expression label, where $y_i \in \{1, \cdots, K\}$, N denotes the number of images in the mini-batch, and D is the feature dimension. Here, K is the number of facial expression classes. The total number of training samples in the dataset is M.

To circumvent the *relabel inconsistency problem*, ensuring that the feature representations in the embedding space are updated simultaneously to reflect the rectification made during each training epoch is critical. Additionally, encouraging samples to be closer to the corresponding class center in the embedding space is equally important. Hence, the proposed framework adopts the center loss [56] as the auxiliary objective function, denoted as L_C. This is because center loss has proven effective in minimizing intra-class variations while maintaining features of different classes separable.

Technically, L_C separates the embedding space into K clusters (i.e. in this work, K different facial expression clusters) and minimizes the sum of the squared distance of the feature vectors in a batch from the cluster centers, $L_C = \sum_{i=1}^{N} \|x_i - c_{y_i}\|_2^2$, where c_{y_i} represents the center of the i-th cluster, which is updated as the feature space changes. Specifically, the class center c_{y_i} is updated by averaging all deep features of the same class in each iteration. Concurrently, the softmax loss (L_S) is employed to estimate the probability distribution over K facial expression classes and measure the prediction error. The discrepancy between predicted label and true label, y_i, is computed using the cross-entropy loss, as follows:

$$L_S = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^{\top} x_i + b_{y_i}}}{\sum_{j=1}^{K} e^{W_j^{\top} x_i + b_j}}, \tag{2}$$

where W is the parameter of the fully connected layer used to weight the feature vector x_i and $b_j \in R^K$ is the bias. For simplicity, this work sets $b_j = 0$ as in [56].

Hence, the total loss function L_{Total} in this work can be formulated as:

$$L_{Total} = L_S + \gamma L_C, \tag{3}$$

where γ is the parameter of the balance ratio. The calculated total loss will be backpropagated to update the model weights utilizing the Adam optimization [57].

Finally, based on P_{max} and P_{gt} from the softmax function, the suppressing uncertainty module (i.e. the relabeling module) relabel ambiguous training images with a new expression label, y_i', based on the definition in (1). Note that the feature representations of relabeled samples will be shifted closer to the new (relabeled) cluster center when the total loss in the next epoch is backpropagated to update the model weights. The algorithm of the proposed framework is summarised in Algorithm 1.

IV. EXPERIMENTS

A. DATASETS

To verify the effectiveness of the proposed framework, this paper conducts experiments using four popular in-the-wild FER datasets.

1) RAF-DB

RAF-DB [58] contains 29672 100 × 100 pixels real-world images downloaded from the Internet that are crowdsourced to 40 annotators to annotate basic or composite expressions. In the experiments of this work, only aligned images labeled as six basic expressions and neutral expressions are used, consisting of a training set of 12271 images and a test set of 3068 images.

2) FER2013

FER2013 [59] contains 28709 training sets, 3589 public test sets and 3589 private test sets. All images are normalized to grayscale 48 × 48 pixels and annotated with one of seven expressions, including six basic expressions and a neutral expression. Since this database is collected using the Google image search engine, the samples contain more variation, including facial occlusions, head pose changes, low resolution, etc.

3) FERPLUS

FERPLUS [8] is an extended version of FER2013, which includes the same image and image number allocation as FER2013. The difference is that the images are re-annotated by ten crowdsourced annotators, and contempt is added to form 8 expression categories.

4) AFFECTNET

AffectNet [7] is currently the largest FER dataset, which contains about 1M facial images collected from the Internet by querying three major search engines using 1250 emotion-related keywords in six different languages. A mini version that contains approximately 280K imbalanced training images and 4K balanced test images manually annotated into eight discrete expressions and cropped to 224 × 224 pixels is adopted in this work. For a fair comparison with SOTA FER methods, this paper conducted experiments with seven emotion classes excluding contempt expression.

B. IMPLEMENTATION DETAILS

For image preprocessing, this work uses MTCNN [22] to align the images of FER2013, FERPLUS, and AffectNet. RAF-DB uses the open-source aligned samples. The images in each batch are first resized to 224 × 224 pixels and then fed to the ResNet18 [60] backbone network to extract deep features, which is pretrained on MS-Celeb-1M [61]. ResNet18 is a widely used, efficient standard CNN with 18 layers, known for its simplicity and effectiveness in image classification tasks due to its residual learning framework. The balance ratio γ of AffectNet, RAF-DB, FERPLUS, and FER2013 is set to 0.02, 0.01, 0.1 and 0.002, respectively. These ratio values are determined experimentally to optimize performance for each dataset. These above experimental setups are identical on SCN [9], RUL (Relative Uncertainty Learning) [10], and DMUE [12] with the following exception: (i) For batch size, SCN is set to 1024, whereas RUL is 64 and DMUE is 72, and (ii) For initial learning rate, SCN is set to 0.1, and the MultiStepLR optimization scheduler is adopted. Meanwhile, the learning rate for RUL is initialized to 0.0002 and uses the ExponentialLR optimization scheduler. For DMUE, the initial learning rate is 0.01, which is further divided by 10 at epochs 10 and 20. Note that, this work re-implemented SCN,[1] RUL,[2] and DMUE[3] based on the original release code. This allows this work to evaluate all the models fairly under a single setting, sharing the same datasets and testing protocols.

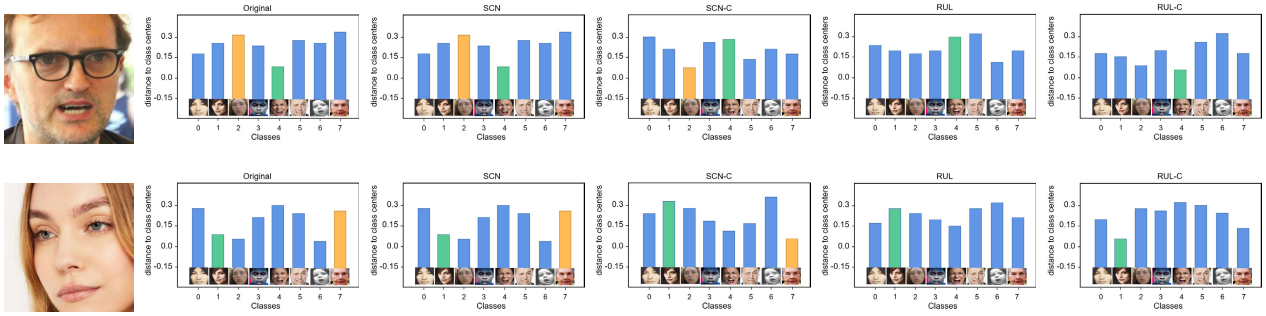
Unlike SCN, RUL and DMUE methods do not rely on relabeling techniques to suppress uncertainty in FER. On the contrary, RUL [10] builds an extra branch for learning uncertainty as a weight to mix the expression features of one easy sample with another ambiguous sample and design an add-up loss to encourage uncertainty learning. Similarly, DMUE [12] utilizes an auxiliary multi-branch to model the latent label distribution of emotion samples and uses cosine similarity to capture the uncertainty.

To ensure a consistent and fair comparison, SCN-C utilizes the existing SCN method to map the original feature space, followed by the application of the proposed embedding proximity to update feature representations. Specifically, this work reproduces the SCN using its publicly released code to map the feature space of the trained SCN model. The three modules described in Sec. III-A–self-attention

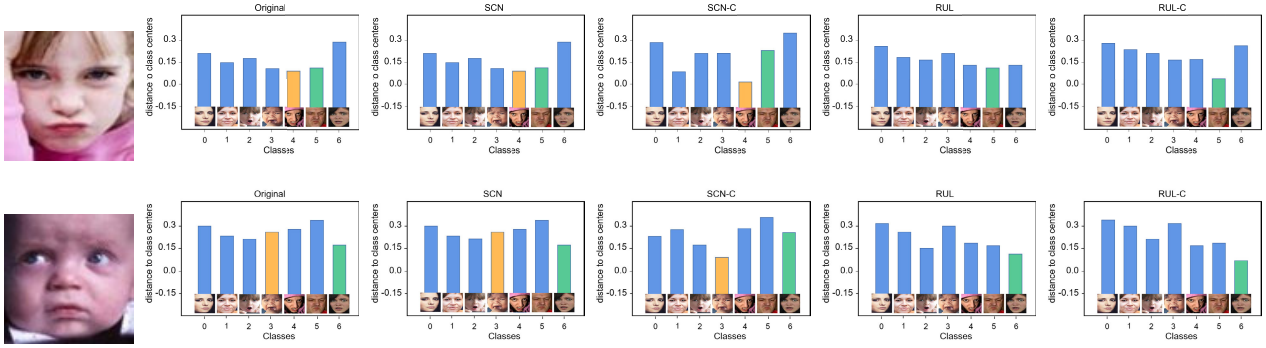
TABLE 1. Comparison of FER accuracy (%) with SOTA methods on AffectNet, RAF-DB, FERPLUS, and FER2013 datasets. † denotes training with both AffectNet and RAF-DB datasets on AffectNet accuracy. * denotes the test with seven classes on AffectNet. + indicates the accuracy improvement (%) of SCN-C, DMUE-C, and RUL-C over SCN, DMUE, and RUL, respectively. The best and second-best results are highlighted in bold and underlined, respectively.

Method	AffectNet	RAF-DB	FERPLUS	FER2013
DLP-CNN [59]	-	84.22	-	-
IPA2LT† [30]	57.31	86.77	-	-
gaCNN [62]	58.78	85.07	-	-
DACL * [52]	<u>65.20</u>	87.78	-	-
PLD [8]	-	-	85.10	-
ResNet+VGG [63]	-	-	85.07	-
RAN [64]	52.97	86.90	88.55	-
Conv+Inception [65]	-	-	-	66.40
Deep-Emotion [66]	-	-	-	70.02
KTN * [19]	63.97	88.07	**90.49**	-
MVT * [17]	64.57	88.62	<u>89.22</u>	-
SCN * [9]	63.50	87.03	85.24	68.82
DMUE * [12]	64.18	86.85	84.57	69.27
RUL * [10]	63.58	<u>88.98</u>	86.89	<u>71.19</u>
SCN-C (ours) *	64.13 $_{+0.63}$	87.91 $_{+0.88}$	87.31 $_{+2.07}$	70.16 $_{+1.34}$
DMUE-C (ours) *	**65.73** $_{+1.55}$	88.04 $_{+1.19}$	86.03 $_{+1.46}$	70.60 $_{+1.33}$
RUL-C (ours) *	65.00 $_{+1.42}$	**89.51** $_{+0.53}$	87.37 $_{+0.48}$	**71.83** $_{+0.64}$

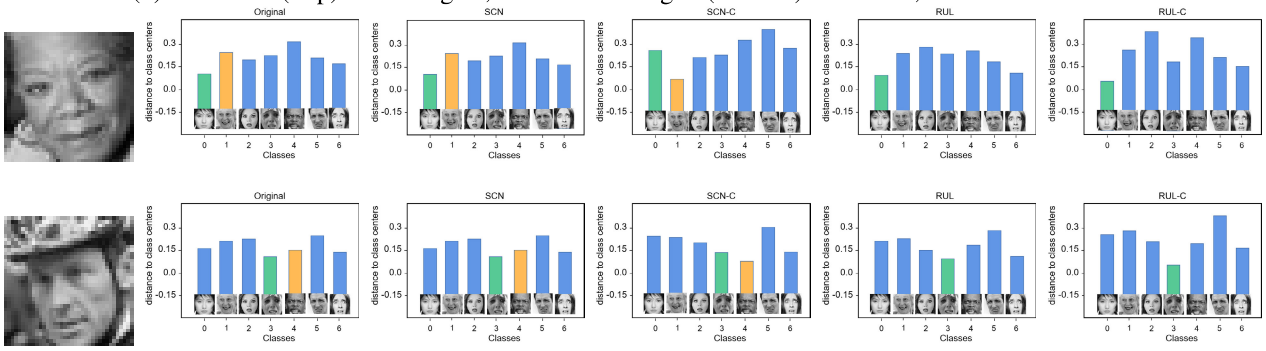
[1] https://github.com/kaiwang960112/Self-Cure-Network
[2] https://github.com/zyh-uaiaaaa/Relative-Uncertainty-Learning
[3] https://github.com/JDAI-CV/FaceX-Zoo/tree/main/addition_module/DMUE



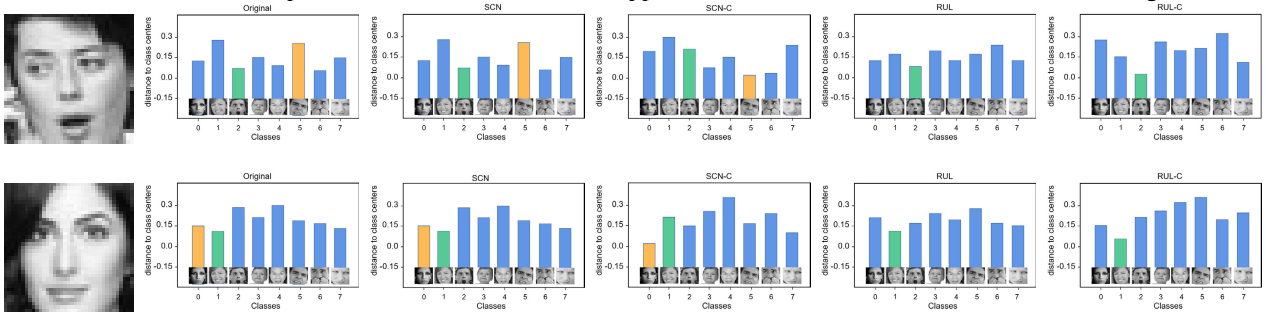
(a) AffectNet: (Top) GT = anger, relabeled → surprise. (Bottom) GT = happiness, relabeled → contempt.



(b) RAF-DB: (Top) GT = disgust, relabeled → anger. (Bottom) GT = fear, relabeled → sadness.



(c) FER2013: (Top) GT = neutral, relabeled → happiness. (Bottom) GT = sadness, relabeled → anger.



(d) FERPLUS: (Top) GT = surprise, relabeled → disgust. (Bottom) GT = happiness, relabeled → neutral.

FIGURE 4. **Visualization of the image feature of sample ambiguous images distances to their respective class centers on the (a) AffectNet, (b) RAF-DB, (c) FER2013, and (d) FERPLUS datasets. The x-axis label (0-7) corresponds to the different facial expressions (0: *Neutral*, 1: *Happiness*, 2: *Surprise*, 3: *Sadness*, 4: *Anger*, 5: *Disgust*, 6: *Fear*, and 7: *Contempt*). [Green: GT label, Yellow: relabeled label, and Blue: other labels]. A lower yellow bar indicates closer proximity to its corresponding relabeled cluster center in the embedding space and vice-versa. Note that the RUL model does not perform relabeling; instead, an ambiguous image is used to enable the model to learn uncertainty through the relativity of the two samples (feature mixup). Best viewed in color.**

importance weighting, ranking regulation, and relabeling–are integrated into the proposed framework, replacing the module originally designed to suppress uncertainty. Following this framework, SCN-C applies embedding proximity, supervised by L_{Total}, to refine the training of the extracted expression features. Consequently, the relabeled samples are drawn

TABLE 2. Compare the cluster compactness of different methods. The mean and standard deviation are computed by calculating the sample points belonging to the same facial expression to their class centers. Lower numbers indicate better compactness.

(a) AffectNet

Method	neutral	happiness	surprise	sadness	anger	disgust	fear	contempt
original	21.19 ± 8.64	21.4 ± 7.91	26.78 ± 9.10	26.50 ± 11.24	26.03 ± 9.91	23.68 ± 11.83	26.76 ± 10.85	15.82 ± 7.54
SCN	21.19 ± 8.64	21.4 ± 7.91	26.78 ± 9.10	26.50 ± 11.24	26.03 ± 9.91	23.68 ± 11.83	26.76 ± 10.85	15.82 ± 7.54
SCN-C	**18.06 ± 7.64**	**17.01 ± 7.01**	**25.13 ± 6.94**	**21.02 ± 8.30**	**21.06 ± 8.24**	**19.22 ± 7.56**	**26.11 ± 7.01**	**14.96 ± 5.71**
RUL	20.53 ± 8.32	24.17 ± 9.71	22.68 ± 8.89	23.20 ± 9.73	24.86 ± 8.22	25.7 ± 9.53	25.27 ± 8.10	20.89 ± 9.69
RUL-C	**19.63 ± 7.96**	**22.91 ± 9.63**	**21.54 ± 6.97**	**22.35 ± 8.46**	**23.24 ± 7.54**	**20.00 ± 7.76**	**24.73 ± 6.79**	**20.11 ± 8.77**

(b) RAF-DB

Method	neutral	happiness	surprise	sadness	anger	disgust	fear
original	23.51 ± 7.96	30.32 ± 10.25	14.67 ± 10.17	16.69 ± 6.88	10.46 ± 7.51	12.34 ± 7.39	7.41 ± 4.03
SCN	23.51 ± 7.96	30.32 ± 10.25	14.67 ± 10.17	16.69 ± 6.88	10.46 ± 7.51	12.34 ± 7.39	7.41 ± 4.03
SCN-C	**19.96 ± 6.62**	**27.52 ± 9.77**	**13.35 ± 8.76**	**15.49 ± 5.64**	**9.47 ± 5.98**	**9.98 ± 4.85**	**6.52 ± 3.81**
RUL	23.54 ± 10.62	41.98 ± 21.66	49.74 ± 13.22	53.43 ± 14.09	34.70 ± 12.97	52.46 ± 16.32	16.36 ± 16.41
RUL-C	**21.89 ± 10.23**	**41.32 ± 19.40**	**46.49 ± 9.77**	**47.76 ± 13.39**	**26.74 ± 12.90**	**41.85 ± 15.23**	**8.69 ± 6.07**

(c) FERPLUS

Method	neutral	happiness	surprise	sadness	anger	disgust	fear	contempt
original	24.55 ± 12.10	23.15 ± 12.87	16.66 ± 10.26	22.71 ± 14.72	15.09 ± 15.63	41.23 ± 26.12	18.19 ± 17.22	24.46 ± 17.59
SCN	24.55 ± 12.10	23.15 ± 12.87	16.66 ± 10.26	22.71 ± 14.72	15.09 ± 15.63	41.23 ± 26.12	18.19 ± 17.22	24.46 ± 17.59
SCN-C	**19.60 ± 11.08**	**20.80 ± 11.97**	**14.74 ± 8.70**	**16.98 ± 12.00**	**13.76 ± 13.09**	**21.16 ± 13.93**	**9.75 ± 13.40**	**17.51 ± 15.84**
RUL	40.79 ± 16.12	45.63 ± 17.96	48.80 ± 23.37	53.31 ± 19.00	58.60 ± 15.85	44.81 ± 14.73	46.38 ± 22.09	38.81 ± 16.73
RUL-C	**39.23 ± 15.43**	**42.31 ± 17.70**	**34.95 ± 18.47**	**39.51 ± 14.01**	**53.02 ± 13.22**	**41.56 ± 14.03**	**42.14 ± 16.19**	**34.88 ± 9.65**

(d) FER2013

Method	neutral	happiness	surprise	sadness	anger	disgust	fear
original	18.44 ± 9.75	20.12 ± 8.81	15.05 ± 13.93	18.05 ± 9.86	16.63 ± 10.69	14.30 ± 19.73	17.88 ± 13.36
SCN	18.44 ± 9.75	20.12 ± 8.81	15.05 ± 13.93	18.05 ± 9.86	16.63 ± 10.69	14.30 ± 19.73	17.88 ± 13.36
SCN-C	**17.40 ± 8.37**	**18.11 ± 8.67**	**14.26 ± 10.69**	**17.12 ± 8.01**	**13.49 ± 9.32**	**7.89 ± 13.84**	**16.76 ± 11.29**
RUL	46.26 ± 18.10	41.42 ± 15.16	35.48 ± 16.50	36.53 ± 16.02	47.26 ± 19.45	56.17 ± 28.03	53.70 ± 14.78
RUL-C	**40.63 ± 14.84**	**40.26 ± 14.39**	**29.03 ± 16.13**	**27.52 ± 14.17**	**42.14 ± 16.47**	**40.91 ± 21.00**	**40.21 ± 13.95**

closer to their new class centers within the updated feature space.

Similarly, RUL and DMUE are reproduced to map the original feature space, with the updated models designated as RUL-C and DMUE-C, respectively. For these models, which do not incorporate relabeling techniques, the suppressing uncertainty module in the proposed framework is replaced by their respective dual-branch and multi-branch networks. Additionally, the auxiliary branches in RUL-C and DMUE-C are enhanced with L_{Total} to ensure that the feature representations cluster effectively around their class centers.

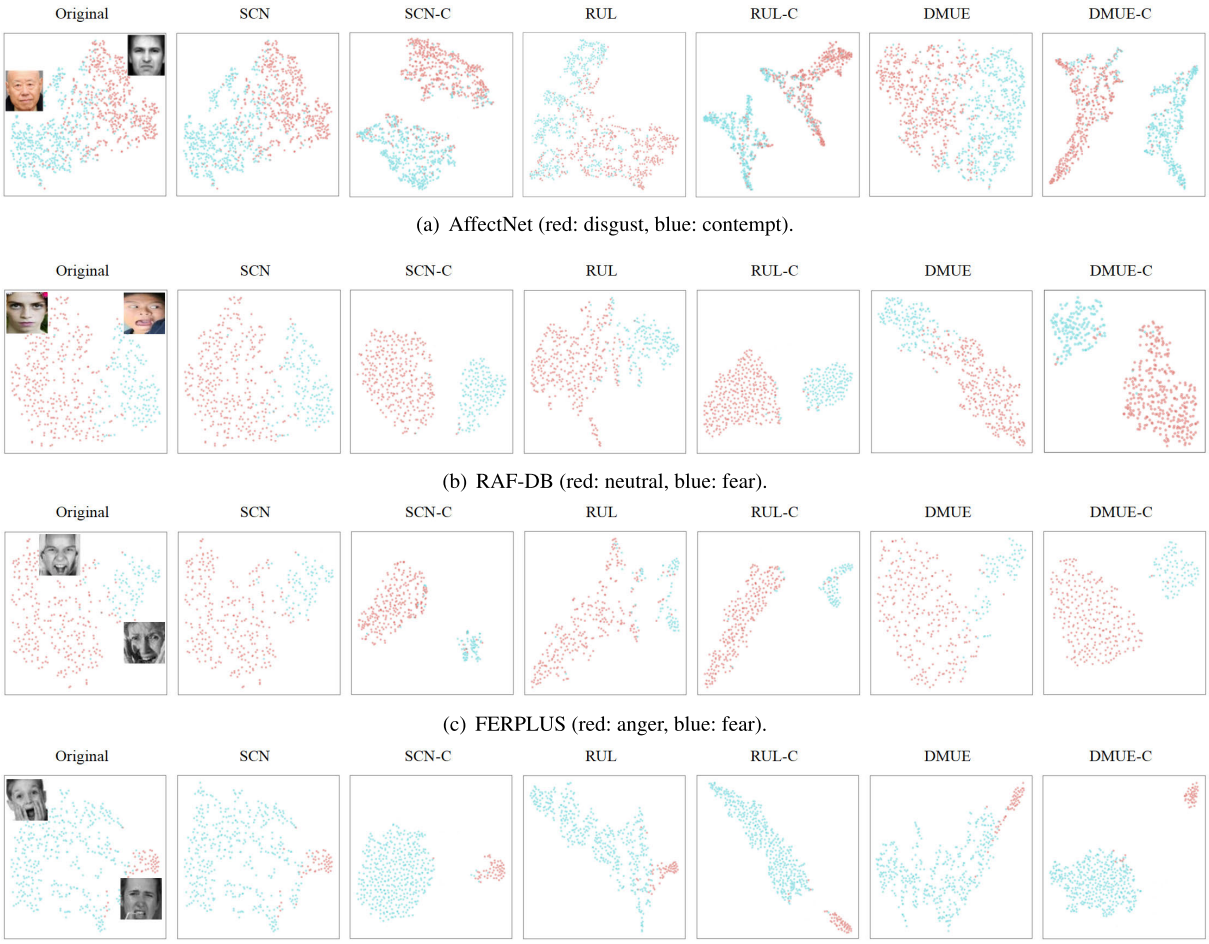
For consistency, the experimental setups for SCN-C, RUL-C, and DMUE-C mirror those used for SCN, RUL, and DMUE, respectively.

C. COMPARISON TO THE STATE OF THE ART

The quantitative and qualitative comparison with SOTA methods on AffectNet, RAF-DB, FERPLUS, and FER2013 datasets are shown in Table 1 and Fig. 4, respectively. It is worth noting that the bulk of this work's effort and primary contribution is to demonstrate the inherent *relabel inconsistency problem* within the existing relabeling-based method, and the proposed framework aimed to circumvent the problem. Nevertheless, this paper demonstrates in Table 1 that the proposed framework outperforms the current SOTA approaches on AffectNet, RAF-DB, and FER2013 datasets. Fig. 4 shows that the proposed framework is able to alleviate the *relabel inconsistency problem*. This is examined by measuring the distances from the image feature of sample ambiguous images (from the AffectNet, RAF-DB, FERPLUS, and FER2013 datasets) to their respective class centers.

1) COMPARISON ON AFFECTNET

Herein, experiments are conducted on the seven expression classes (excluding contempt) and compared with the SOTA methods. Table 1 (the second column) shows that the accuracy of SCN-C, RUL-C, and DMUE-C on the AffectNet dataset is 64.13%, 65.00%, and 65.73%, respectively, and the performance gains based on SCN, RUL, and DMUE are 0.63%, 1.42%, and 1.55%. The proposed DMUE-C achieves the best result at 65.73%. For qualitative analysis, it can be observed in the first row, the third column of Fig. 4 (SCN) that after the sample image has been relabeled, the height of the *surprise* bar (yellow) and the *anger* bar (green) remains the same. This indicates that the feature



(a) AffectNet (red: disgust, blue: contempt).

(b) RAF-DB (red: neutral, blue: fear).

(c) FERPLUS (red: anger, blue: fear).

(d) FER2013 (red: disgust, blue: surprise).

FIGURE 5. T-SNE visualization of cluster compactness of different methods. More compact and distinct clusters mean the model has learned a more discriminative embedding space. Best viewed in color.

representations are not updated to reflect the rectification made after relabeling. However, for SCN-C (fourth column), the *surprise* bar (yellow) is now shorter, and the *anger* bar (green) is taller compared to SCN. This suggests that the relabeled sample is being pushed closer to its corresponding surprise center and away from the anger center in the embedding space. Although the RUL [10] method does not use the relabeling technique for suppressing uncertainty, the proposed framework significantly reduces the distance of the image features to the ground truth class center (green bar). This is an additional advantage of the proposed framework, as it further enhances the discriminative power of deeply learned features by penalizing the distances between the deep features and their corresponding class centers. These observations are consistent across all four datasets, demonstrating the improvement of inter-class separation and intra-class compactness.

2) COMPARISON ON RAF-DB

From Table 1 (the third column), it can be seen that RUL-C outperforms all SOTA methods in terms of accuracy. RUL-C improves by 0.88% compared to RUL, DMUE-C improves by 1.19% compared to DMUE, and SCN-C improves by 0.53% compared to SCN.

3) COMPARISON ON FERPLUS

It can be noticed from Table 1 (the fourth column) that the accuracy of the proposed (SCN-C) has increased to 87.31%, which exceeds SCN by 2.07%. The proposed DMUE-C achieves 86.03%, which is 1.46% better than DMUE. Meanwhile, a similar trend can also be observed for RUL and RUL-C, where RUL-C obtained an accuracy of 86.89% and 87.37%, respectively. The proposed RUL-C exceeds the baseline RUL result by 0.48%. Although KTN (Knowledgeable Teacher Network) [19], MVT (Mask Vision Transformer) [17], and RAN (Region Attention Network) [64] achieve 3.09%, 1.85%, and 1.18% superiority over the proposed models on FERPLUS, these methods require an additional data augmentation technique [64], use a different backbone network [17], or employ a class imbalance strategy [19] compared to the proposed models.

TABLE 3. FER accuracy (%) on AffectNet, RAF-DB, FERPLUS, and FER2013 with synthetic noisy labels. The baseline method refers to ResNet18 pretrained on MS-Celeb-1M, the same as the implementation in RUL [10]. The best results are highlighted in bold.

Method	Noisy(%)	AffectNet	RAF-DB	FERPLUS	FER2013
Baseline	10	56.85 ± 0.14	80.43 ± 0.72	81.73 ± 0.19	66.36 ± 0.22
DUL [11]	10	58.26 ± 0.10	85.08 ± 0.21	-	-
DMUE [12]	10	58.15 ± 0.41	82.05 ± 0.73	82.16 ± 0.75	67.72 ± 0.71
DMUE-C	10	59.37 ± 0.30	85.98 ± 0.53	84.22 ± 0.43	68.84 ± 0.42
SCN [9]	10	58.72 ± 0.20	81.92 ± 0.69	83.29 ± 0.15	67.09 ± 0.18
SCN-C	10	$\mathbf{60.76 \pm 0.18}$	85.40 ± 0.17	85.01 ± 0.11	68.76 ± 0.12
RUL [10]	10	58.51 ± 0.21	86.22 ± 0.29	84.76 ± 0.31	68.15 ± 0.29
RUL-C	10	59.43 ± 0.17	$\mathbf{86.92 \pm 0.16}$	$\mathbf{85.57 \pm 0.23}$	$\mathbf{69.24 \pm 0.21}$
Baseline	20	54.74 ± 0.62	78.01 ± 0.29	78.28 ± 0.27	62.05 ± 0.35
DUL [11]	20	56.25 ± 0.09	81.95 ± 0.32	-	-
DMUE [12]	20	56.46 ± 0.44	79.89 ± 0.58	79.78 ± 0.84	65.74 ± 0.49
DMUE-C	20	57.79 ± 0.32	83.86 ± 0.31	82.39 ± 0.52	67.19 ± 0.34
SCN [9]	20	56.35 ± 0.61	80.02 ± 0.32	81.23 ± 0.38	65.21 ± 0.31
SCN-C	20	58.47 ± 0.29	83.63 ± 0.09	83.07 ± 0.31	66.94 ± 0.27
RUL [10]	20	57.42 ± 0.34	84.34 ± 0.29	82.75 ± 0.26	66.14 ± 0.28
RUL-C	20	$\mathbf{58.52 \pm 0.26}$	$\mathbf{85.23 \pm 0.10}$	$\mathbf{83.88 \pm 0.21}$	$\mathbf{67.33 \pm 0.23}$
Baseline	30	51.46 ± 0.52	75.12 ± 0.78	76.87 ± 0.42	61.26 ± 0.52
DUL [11]	30	55.09 ± 0.32	78.90 ± 0.80	-	-
DMUE [12]	30	52.87 ± 0.58	77.59 ± 0.57	78.32 ± 0.61	62.38 ± 0.48
DMUE-C	30	54.47 ± 0.43	81.67 ± 0.35	81.12 ± 0.34	64.11 ± 0.35
SCN [9]	30	52.60 ± 0.86	77.46 ± 0.64	80.19 ± 0.37	62.02 ± 0.48
SCN-C	30	55.05 ± 0.41	81.21 ± 0.11	82.33 ± 0.29	63.85 ± 0.43
RUL [10]	30	56.16 ± 0.27	82.06 ± 0.44	81.53 ± 0.32	63.10 ± 0.37
RUL-C	30	$\mathbf{57.48 \pm 0.22}$	$\mathbf{83.17 \pm 0.09}$	$\mathbf{82.83 \pm 0.23}$	$\mathbf{64.45 \pm 0.31}$

4) COMPARISON ON FER2013

As shown in Table 1 (the fifth column), the proposed SCN-C achieves 70.16%, which is a 1.34% improvement compared to the original SCN. The proposed DMUE-C achieves 70.60%, which is 1.33% better than DMUE. Meanwhile, the result of RUL-C is 71.83%, which improves the RUL accuracy by 0.64%. Like SCN, the proposed framework cannot correct the noisy labels of FER2013 to reach FERPLUS performance. This is reasonable mainly because of the considerable within-class sample variance in the FERPLUS seven classes (excluding contempt) after relabeling compared to FER2013.

In summary, by leveraging the proposed framework, there are marginal improvements in the FER models. This is because the relabeled images are now being updated accordingly (closer to their new corresponding class centers) in the embedding space. Specifically, from the bar charts in column 2 to column 3 from Fig. 4, it can be observed that the distance of the image features to the ground truth label (green bar) and the relabeled label (yellow bar) remains unchanged after relabeling, i.e., the *relabel inconsistency problem*. However, as shown in column 4, it is evident that the proposed models significantly improve the distance of image features to the ground truth label (green bar) and the relabeled label (yellow bar) to reflect the relabeling made. Similar results are also observed in the RUL model without the relabeling module. It shows that the updates of the feature embedding space are essential to suppress data uncertainty for accurate emotion recognition.

D. INTRA-CLASS COMPACTNESS OF FACIAL EXPRESSION CLASSES

The proposed framework imposes constraints on the distribution of expression representations in high-dimensional space. Hence, this subsection provides quantitative and qualitative demonstrations of the intra-class compactness for each expression cluster in the embedding space to verify its effect. The evaluations are conducted on the AffectNet, RAF-DB, FERPLUS, and FER2013 datasets, and the results are presented in Table 2 and Fig. 5, respectively.

In Table 2, within-cluster-mean and -standard deviation are used to measure within-cluster compactness. The mean and standard deviation are computed by calculating the sample points belonging to the same facial expression to their class centers. A large standard deviation indicates a large intra-class distance between samples in a class. In the third and fifth rows of Table 2, it can be seen that the improvement of intra-class compactness on SCN-C and RUL-C, compared to SCN and RUL, for all facial expression classes and consistent for all four datasets. In comparison to SCN, the average mean and standard deviation reductions achieved by SCN-C on AffectNet, RAF-DB, FERPLUS, and FER2013 datasets are 3.20 ± 2.33, 1.87 ± 1.25, 6.47 ± 3.31, and 2.21 ± 2.28, respectively. Similarly, in comparison to RUL, the average mean and standard deviation reductions obtained by RUL-C on AffectNet, RAF-DB, FERPLUS, and FER2013 datasets are 1.60 ± 1.04, 5.35 ± 2.62, 6.19 ± 3.39, and 8.02 ± 2.44, respectively.

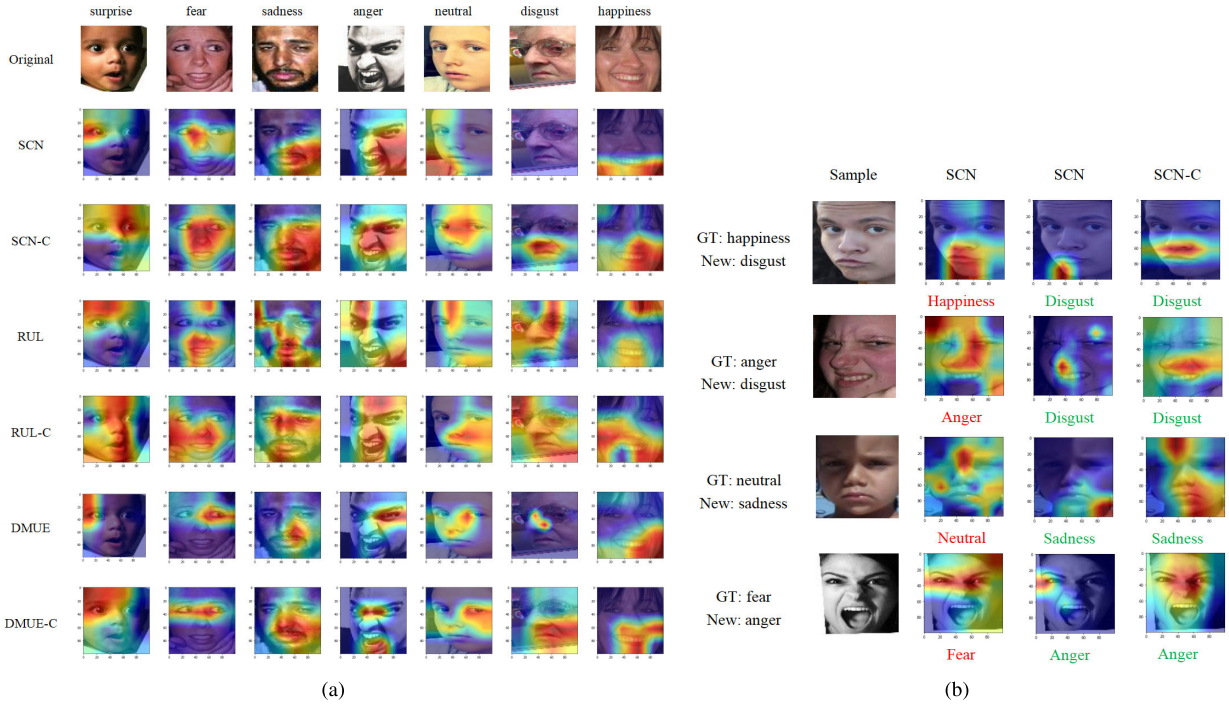


FIGURE 6. (a) Grad-CAM visualization on the RAF-DB dataset. (b) Samples of Grad-CAM visualization of SCN and SCN-C on noisy images. The activation maps in the third column of (b) are SCN under GT labels (red), and the activation maps in the fourth column of (b) are SCN under new labels (green). The activation maps are calculated for the last convolutional outputs. Best viewed in color.

For qualitative evaluation, this paper adopts the T-SNE visualization technique [67] to examine whether the embedding space has been updated from the visualization results. In principle, after ambiguous images have been relabeled, the feature embedding space should also be updated to reflect the rectification made. This ensures that relabeled images are closer to their new corresponding class centers in the embedding space. In Fig. 5, the T-SNE visualization is plotted on the proposed methods (SCN-C, RUL-C, DMUE-C) and three other methods (SCN, RUL, DMUE) to demonstrate the effectiveness of the proposed framework. Two random classes are selected, respectively, from (a) AffectNet, (b) RAF-DB, (c) FERPLUS, and (d) FER2013 to plot T-SNE visualization. It can be observed that the figures generated by SCN-C, RUL-C, and DMUE-C are more compact, well separated, and have a more discriminant embedding space compared to SCN, RUL, and DMUE, respectively. For instance, the T-SNE visualization of the AffectNet dataset on SCN shows that the *Disgust* expression denoted by red dots is cluttered together with *Contempt* (blue). However, for SCN-C, the two different expression clusters are better separated (as shown in the third column of (a)). Similar observations can also be seen for the RUL and DMUE models without the relabeling module.

In summary, as shown in Table 2 and Fig. 5, the proposed models (i.e., SCN-C, RUL-C, and DMUE-C) achieves intra-class compactness and inter-class separation in the embedding space across four datasets. This indicates that the proposed framework effectively updates the feature representations, ensuring that ambiguous images are positioned closer to their respective class centers in the embedding space. Consequently, the improved compactness within clusters contributes to enhanced accuracy and overall performance of the emotion recognition models.

E. EVALUATION ON NOISY LABELS

Following SCN, RUL, and DMUE, this work conducts extensive experiments to quantitatively analyze the robustness of the proposed framework on AffectNet, RAF-DB, FERPLUS, and FER2013 containing 10%, 20%, and 30% noisy labels. For a fair comparison with SOTA methods, this work follows the protocols in RUL [10] where the mean and standard deviation are derived from the accuracy of the last epoch.

Table 3 shows that SCN-C, RUL-C, and DMUE-C outperform SCN, RUL, and DMUE on all three noisy label ratios and four datasets. For example, SCN improves accuracy by 2.04%, 2.12%, and 2.45% with the noise ratios of 10%, 20%, and 30% on AffectNet, 3.48%, 3.61%, and 3.75% on RAF-DB, 1.72%, 1.84%, and 2.14% on FERPLUS, and 1.67%, 1.73%, and 1.83% on FER2013 compared with SCN. RUL-C and DMUE-C have similar performance gains over RUL and DMUE.

For SCN-C, this improvement is attributed to the proposed framework updating the embedding space, effectively reflecting the relabeling made on ambiguous image samples. In the case of RUL-C and DMUE-C, the proposed framework reduces the distance of the image features to the ground

TABLE 4. Comparison of FER accuracy (%) with different loss functions.

Method	AffectNet	RAF-DB	FERPLUS	FER2013
SCN (baseline) [9]	60.23	87.03	85.24	68.82
SCN w/o RR loss	60.14 (−0.09)	86.97 (−0.06)	85.17 (−0.07)	68.71 (−0.11)
The proposed framework	**61.37 (+1.14)**	**87.91 (+0.88)**	**87.31 (+2.07)**	**70.16 (+1.34)**

truth class center (as shown in Fig. 4). This ensures that image features are closer to the cluster center, effectively suppressing data uncertainty. The results indicate that the proposed models are more robust against noisy labels. Moreover, the findings demonstrate that the proposed models achieve SOTA results even with increased noisy labels.

F. GRAD-CAM VISUALIZATION

Gradient-weighted Class Activation Mapping (Grad-CAM) [68] is a novel visualization technique that utilizes gradients to compute the significance of spatial positions in convolutional layers. By computing gradients for a specific class, Grad-CAM generates a heatmap highlighting the attention regions in an image, which provides valuable insights into the network's understanding of relevant features and information. Hence, the Grad-CAM visualization facilitates a comprehensive analysis of the trained model's performance and capacity to capture essential visual cues.

Fig. 6(a) shows the visualization results. The attention regions are characterized by high intensity in the Grad-CAM heatmaps. The Grad-CAM masks of the SCN-C, RUL-C, and DMUE-C models show better coverage of expression-relevant facial components than their counterparts. Hence, the proposed models learned from more explanatory regions that improved their FER accuracies.

This study also included Grad-CAM visualization experiments on noisy samples to evaluate the effectiveness of the proposed framework in addressing the relabeling inconsistency problem. The results, presented in Fig. 6(b), offer insights into the behaviour of the SCN model when using both the ground truth (GT) labels (noisy labels) and the revised labels, as well as the SCN-C model under the revised labels.

From the third and fourth columns of Fig. 6(b), it is apparent that the SCN model, even after relabeling, still associates some noisy samples with their GT labels, as evidenced by the SCN heatmaps under the GT labels demonstrating better coverage of facial regions. In contrast, the SCN-C heatmaps under the revised labels (shown in the fifth column) specifically activate the facial components relevant to expressions. This observation confirms that the proposed framework ensures consistent model updates in line with the corrections made during the relabeling process.

G. ABLATION ON THE LOSS FUNCTION OF SCN

The SCN model reproduced in this work does not include RR loss, as empirical observations indicate that the RR loss does not contribute to improving the model during training. As illustrated in Table 4, the drops in accuracy are negligible across datasets when excluding the RR loss from the model. In contrast, the accuracy of the proposed framework outperforms the baseline model by 1.14%, 0.88%, 2.07%, and 1.34% on AffectNet, RAF-DB, FERPLUS, and FER2013 datasets, respectively. These improvements stem from the proposed framework ensuring that the feature embedding space is updated, particularly reflecting the relabeling of ambiguous image samples.

V. CONCLUSION

This paper uncovers and addresses the *relabel inconsistency problem* inherent in existing relabeling techniques. This work introduces a new framework based on embedding proximity to circumvent this inconsistency problem, ensuring that the relabeling of noisy samples is updated accordingly during training. Moreover, this work highlights the importance of embedding space proximity, which is critical in maintaining the accuracy and reliability of FER systems. The proposed approach with embedding proximity ensures that feature representations are updated consistently with rectifications made during relabeling. Extensive experiments on four widely used benchmark datasets demonstrate that the proposed framework achieves remarkable quality gains over SOTA methods, with or without relabeling techniques. However, this work fundamentally relies on the labeling and relabeling of noisy samples, and thus, its effectiveness is constrained by the accuracy of these labeling and relabeling algorithms. Future work is planned to integrate advanced techniques such as feature pyramid networks into the proposed framework, creating a more carefully designed network ensemble method and extending its application to other computer vision domains.

REFERENCES

[1] B. Balasubramanian, P. Diwan, R. Nadar, and A. Bhatia, "Analysis of facial emotion recognition," in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 945–949.

[2] G. Simcock, L. T. McLoughlin, T. De Regt, K. M. Broadhouse, D. Beaudequin, J. Lagopoulos, and D. F. Hermens, "Associations between facial emotion recognition and mental health in early adolescence," *Int. J. Environ. Res. Public Health*, vol. 17, no. 1, p. 330, Jan. 2020.

[3] J. Pu and X. Nie, "Convolutional channel attentional facial expression recognition network and its application in human–computer interaction," *IEEE Access*, vol. 11, pp. 129412–129424, 2023.

[4] K. S. Zaman and M. M. B. I. Reaz, "Secure and efficient implementation of facial emotion detection for smart patient monitoring system," *Quant. Biol.*, vol. 11, no. 2, pp. 175–182, Jun. 2023.

[5] T. M. W. Vithanawasam and B. G. D. A. Madhusanka, "Dynamic face and upper-body emotion recognition for service robots," in *Proc. IEEE/ACIS 17th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2018, pp. 428–432.

[6] Y. An, "Deep facial emotion recognition using local features based on facial landmarks for security system," *Comput., Mater. Continua*, vol. 76, no. 2, pp. 1817–1832, 2023.

[7] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[8] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.

[9] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6896–6905.

[10] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17616–17627.

[11] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5709–5718.

[12] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6244–6253.

[13] Q. Zhu, Q. Mao, H. Jia, O. E. N. Noi, and J. Tu, "Convolutional relation network for facial expression recognition in the wild with few-shot learning," *Expert Syst. Appl.*, vol. 189, Mar. 2022, Art. no. 116046.

[14] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1928–1936.

[15] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6575–6583.

[16] N. Le, K. Nguyen, Q. Tran, E. Tjiputra, B. Le, and A. Nguyen, "Uncertainty-aware label distribution learning for facial expression recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6077–6086.

[17] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, "MVT: Mask vision transformer for facial expression recognition in the wild," 2021, *arXiv:2106.04520*.

[18] S. Li, W. Li, S. Wen, K. Shi, Y. Yang, P. Zhou, and T. Huang, "Auto-FERNet: A facial expression recognition network with architecture search," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 3, pp. 2213–2222, Jul. 2021.

[19] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via C–F labels and distillation," *IEEE Trans. Image Process.*, vol. 30, pp. 2016–2028, 2021.

[20] J. Shao, Z. Wu, Y. Luo, S. Huang, X. Pu, and Y. Ren, "Self-paced label distribution learning for in-the-wild facial expression recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 161–169.

[21] N. Samadiani, G. Huang, B. Cai, W. Luo, C. H. Chi, Y. Xiang, and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, Apr. 2019.

[22] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[23] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," *CMU School Comput. Sci.*, vol. 6, no. 2, p. 20, 2016.

[24] S. Yasmin, R. K. Pathan, M. Biswas, M. U. Khandaker, and M. R. I. Faruque, "Development of a robust multi-scale featured local binary pattern for improved facial expression recognition," *Sensors*, vol. 20, no. 18, p. 5391, Sep. 2020.

[25] A. Elmadhoun and M. J. Nordin, "Facial expression recognition using uniform local binary pattern with improved firefly feature selection," *ARO-Sci. J. Koya Univ.*, vol. 6, no. 1, pp. 23–32, Apr. 2018.

[26] S. Qin, Z. Zhu, Y. Zou, and X. Wang, "Facial expression recognition based on Gabor wavelet transform and 2-channel CNN," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 18, no. 2, Mar. 2020, Art. no. 2050003.

[27] H. I. Dino and M. B. Abdulrazzaq, "Facial expression classification based on SVM, KNN and MLP classifiers," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, Apr. 2019, pp. 70–75.

[28] A. Ali Alhussan, F. M. Talaat, E.-S. M. El-Kenawy, A. A. Abdelhamid, A. Ibrahim, D. Sami Khafaga, and M. Alnaggar, "Facial expression recognition model depending on optimized support vector machine," *Comput., Mater. Continua*, vol. 76, no. 1, pp. 499–515, 2023.

[29] R. Zhi, C. Zhou, T. Li, S. Liu, and Y. Jin, "Action unit analysis enhanced facial expression recognition by deep neural network evolution," *Neurocomputing*, vol. 425, pp. 135–148, Feb. 2021.

[30] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 222–237.

[31] Z. M. Zohreh Madhoushi, A. R. Hamdan, and S. Zainudin, "Aspect-based sentiment analysis methods in recent years," *Asia–Pacific J. Inf. Technol. Multimedia*, vol. 8, no. 1, pp. 79–96, Jun. 2019.

[32] M. M. Stofa, M. A. Zulkifley, and M. A. A. M. Zainuri, "Micro-expression-based emotion recognition using waterfall atrous spatial pyramid pooling networks," *Sensors*, vol. 22, no. 12, p. 4634, Jun. 2022.

[33] Y. Fan, X. Jiang, S. Lan, and J. Lan, "Facial expression transfer based on conditional generative adversarial networks," *IEEE Access*, vol. 11, pp. 82276–82283, 2023, doi: 10.1109/ACCESS.2023.3294697.

[34] G. Perveen, S. F. Ali, J. Ahmad, S. Shahab, M. Adnan, M. Anjum, and I. Khosa, "Multi-stream deep convolution neural network with ensemble learning for facial micro-expression recognition," *IEEE Access*, vol. 11, pp. 118474–118489, 2023.

[35] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE Access*, vol. 7, pp. 41273–41285, 2019.

[36] W. Gong, Z. La, Y. Qian, and W. Zhou, "Hybrid attention-aware learning network for facial expression recognition in the wild," *Arabian J. Sci. Eng.*, pp. 1–15, Jan. 2024, doi: 10.1007/s13369-023-08538-6.

[37] Y. Shi and A. Jain, "Probabilistic face embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6901–6910.

[38] M. Dehghani, A. Severyn, S. Rothe, and J. Kamps, "Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision," 2017, *arXiv:1711.00313*.

[39] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," 2014, *arXiv:1406.2080*.

[40] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7010–7018.

[41] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11236–11245.

[42] R. K. Pandey, S. Karmakar, A. Ramakrishnan, and N. Saha, "Improving facial emotion recognition systems with crucial feature extractors," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2019, pp. 268–279.

[43] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3510–3519.

[44] W. Hu, Y. Huang, F. Zhang, and R. Li, "Noise-tolerant paradigm for training face recognition CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11879–11888.

[45] W. Xie, H. Wu, Y. Tian, M. Bai, and L. Shen, "Triplet loss with multistage outlier suppression and class-pair margins for facial expression recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 690–703, Feb. 2022.

[46] M.-I. Georgescu and R. T. Ionescu, "Teacher–student training and triplet loss for facial expression recognition under occlusion," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2288–2295.

[47] Y. Tian, Z. Wen, W. Xie, X. Zhang, L. Shen, and J. Duan, "Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 46–50.

[48] Z. Li, H. Shao, L. Niu, and N. Xue, "PLA: Progressive learning algorithm for efficient person re-identification," *Multimedia Tools Appl.*, vol. 81, no. 17, pp. 24493–24513, Jul. 2022.

[49] D. Shi, M. Orouskhani, and Y. Orouskhani, "A conditional triplet loss for few-shot learning and its application to image co-segmentation," *Neural Netw.*, vol. 137, pp. 54–62, May 2021.

[50] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13981–13990.

[51] H. Shin, B. Lee, B. Ku, and H. Ko, "Noisy label facial expression recognition via face-specific label distribution learning," *Image Vis. Comput.*, vol. 143, Mar. 2024, Art. no. 104901.

[52] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2401–2410.

[53] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2022, pp. 418–434.

[54] Y. Gao, W. Ren, Q. Wang, X. Chen, Z. Wang, and H. Liu, "SNEFER: Stopping the negative effect of noisy labels adaptively in facial expression recognition," *IEEE Sensors J.*, vol. 24, no. 11, pp. 18622–18632, Jun. 2024.
[55] R. Dong and K.-M. Lam, "Bi-center loss for compound facial expression recognition," *IEEE Signal Process. Lett.*, vol. 31, pp. 641–645, 2024.
[56] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vision. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 499–515.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[58] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.
[59] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2013, pp. 117–124.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[61] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–102.
[62] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[63] C. Huang, "Combining convolutional neural networks for emotion recognition," in *Proc. IEEE MIT Undergraduate Res. Technol. Conf. (URTC)*, Nov. 2017, pp. 1–4.
[64] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[65] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[66] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, Apr. 2021.
[67] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html
[68] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

NING CHEN received the B.S. degree in engineering from Southwest University, Chongqing, China, in 2020. She is currently pursuing the master's degree in artificial intelligence with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. Her current research interests include artificial intelligence, computer vision, and machine learning.

VEN JYN KOK received the bachelor's degree in electrical and electronics engineering from Universiti Tenaga Nasional, Malaysia, the master's degree in data communications from the University of Sheffield, U.K., and the Ph.D. degree from Universiti Malaya, Kuala Lumpur, in 2016. She is currently a Senior Lecturer with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor. Her research interests include image processing, pattern recognition, and machine learning.

CHEE SENG CHAN (Senior Member, IEEE) was with the Ministry of Science, Technology and Innovation (MOSTI), as the Undersecretary of the Division of Data Strategic and Foresight, from 2020 to 2022. He is currently a Full Professor with the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia. He is also leading a Research Team that specializes in computer vision and machine learning, where his team has published more than 100 papers in top peer-reviewed conferences and journals (IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON MULTIMEDIA).

Dr. Chan received the Top Research Scientists Malaysia (TRSM) in 2022, the Young Scientists Network Academy of Sciences Malaysia (YSN-ASM) in 2015, and the Hitachi Research Fellowship in 2013. Besides that, he is also a Professional Engineer (BEM) and a Chartered Engineer (IET).