IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 November 2023, accepted 25 November 2023, date of publication 5 December 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3339763

RESEARCH ARTICLE

Enhancing Noisy Label Facial Expression Recognition With Split and Merge Consistency Regularization

JIHYUN KIM[1], (Graduate Student Member, IEEE), JUNEHYOUNG KWON[1], MIHYEON KIM[1], EUNJU LEE[2], AND YOUNGBIN KIM[1,2], (Member, IEEE)

[1]Department of Artificial Intelligence, Chung-Ang University, Dongjak, Seoul 06974, South Korea
[2]Department of Imaging Science, Multimedia and Film, Chung-Ang University, Dongjak, Seoul 06974, South Korea

Corresponding author: Youngbin Kim (ybkim85@cau.ac.kr)

This work was supported in part by the Chung-Ang University Graduate Research Scholarship in 2022; in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2022R1C1C1008534; and in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) through the Korea Government (MSIT) under Artificial Intelligence Graduate School Program, Chung-Ang University, under Grant 2021-0-01341.

ABSTRACT Facial expression recognition (FER) has been extensively studied in various applications over the past few years. However, in real facial expression datasets, labels can become noisy due to the ambiguity of expressions, the similarity between classes, and the subjectivity of annotators. These noisy labels negatively affect FER and significantly reduce classification performance. In previous methods, overfitting can occur as the noise ratio increases. To solve this problem, we propose the split and merge consistency regularization (SMEC) method that is robust to noisy labels by examining various image regions rather than just one part of facial expression images without negatively affecting the meaning. We split facial expression images into two images and input them into the backbone network to extract class activation maps (CAMs). This approach merges two CAMs and improves robustness to noisy labels by normalizing the consistency between the CAM of the original image and the merged CAM. The proposed SMEC method aims to improve FER performance and robustness against highly noisy labels by preventing the model from focusing on only a single part without losing the semantics of the facial expression images. The SMEC method demonstrates robust performance over state-of-the-art noisy label FER models on an unbalanced facial expression dataset called the real-world affective faces database (RAF-DB) regarding class-wise accuracy for clean and noisy labels, even at severe noise rates of 40% to 60%.

INDEX TERMS Consistency regularization, deep learning, facial expression recognition, image classification, noisy label learning.

I. INTRODUCTION

Facial expression is an essential communication method that cannot be avoided in our daily lives. Ekman [1] categorized the basic human emotions into six categories: joy, sadness, anger, fear, disgust, and surprise. Based on this categorization, a facial emotion coding system [2] is developed that anatomically analyzes and explains the facial muscles used to make facial expressions. The exploration of facial expression recognition (FER) serves multiple purposes within academic research and has practical applications closely related to everyday experiences. Facial expression representation has been studied in various areas, such as medicine [3], [4], autonomous driving [5], [6], and human-machine interaction (HMI) [7], and has emerged as an essential task in computer vision.

Recently, numerous studies on deep learning-based FER have required considerable data [40], [41], [42], [43]. However, accurate annotation of facial expression images is a complex and time-consuming problem because it is

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

expensive and requires professional knowledge. In addition, label noise is generated by including annotator subjectivity in annotation operations. Furthermore, label noise occurs due to ambiguity in facial expressions (e.g., fear, disgust, and anger) or inter-class similarity and low-quality images captured from the internet.

Several studies have been conducted to solve this noisy label problem in FER [8], [9], [10], [11]. Existing studies have revealed that only certain areas of the face are used for classification; therefore, noisier labels are more vulnerable regarding performance. The random erasing [31] method has been used to observe various parts [8], but more substantial label noise increases result in lower performance. However, to our knowledge, research and experiments specifically addressing severe noisy labels in this context have not yet been conducted. In real-world situations that do not involve such processed datasets, the noise rate in facial expression images is higher than that previously studied.

To address this problem, we propose the split and merge consistency regularization (SMEC) method, which can enhance the robustness of FER models under noisy label conditions. To improve the focus of traditional FER models on one part without examining the whole, we split the facial expression image into two, allowing the extraction of class activation maps (CAMs) [20] from various areas of the facial expression image. The extracted CAMs are merged to match the size of the original image. To learn facial expression labels, we adopt the loss of attention consistency between the two CAMs extracted from the split and original images [12]. We apply early learning regularization (ELR) loss [13] to prevent remembering noisy labels in the early stages of learning.

The proposed SMEC method allows the model to split and merge input images, enabling it to observe various facial components while preserving the facial expression semantics. Thus, the model displays robust performance even if the noisy label increases. Compared to previous studies [8], [9], [15], this method proves effective in improving FER performance when using balanced accuracy [30] and best accuracy metrics. In summary, the main contributions of this study are as follows:

- We propose SMEC, which performs consistency regularization by splitting and merging while preserving facial expression image semantics.
- The proposed approach achieves up to a 40.37%p improvement in balanced accuracy per class compared to existing methods by allowing the model to view facial images from diverse perspectives.
- The proposed SMEC method demonstrates robust performance even under severe noise rates, attributed to the model's ability to observe various aspects of the face without losing semantic information.

The rest of the paper is organized as follows. Section II presents related work on FER with noisy labels and consistency regularization methods. Next, Section III describes the proposed approach. Then, Section IV reports the experimental results and ablation studies. The last section presents the conclusion and future research directions.

II. RELATED WORK

This section reviews the existing noisy label FER approaches and consistency regularization methods. Then, we emphasize the distinctiveness and novelty of this work.

A. NOISY LABEL FACIAL EXPRESSION RECOGNITION

Noisy labels heavily influence real-time FER tasks due to the ambiguity in facial expressions and inter-class similarity. Accordingly, many recent studies have proposed improving the performance of the noisy label FER task. Thus, this section reviews the existing methods that have improved the performance and robustness of the noisy label FER task.

Recently, noisy label FER methods have been categorized into two main approaches: sample selection and label ensembling. Sample selection methods involve relabeling noisy samples and retraining them as clean samples [9], [15], [37]. Relative uncertainty learning (RUL) [9] used uncertainty as a weight to mix features of different labels with a new branch that learns uncertainty while minimizing the total loss. Self-cure network (SCN) [15] recommended re-evaluating samples in subgroups based on comparing the maximum predicted probability with the probability of a given label. Reliable label noise suppression [37] simultaneously models noise distributions and clean labels to make noise data decisions based on inconsistent predictions and goals. The models used the similarity distribution of all samples to achieve optimal distribution modeling.

Label ensembling methods that assign multiple labels to a single sample facilitate creating a more robust latent truth [10], [14], [16], [38]. The distribution mining and the pairwise uncertainty estimation [10] model used latent distribution mining and pairwise uncertainty estimation to solve the ambiguity problem. It provided model-informative semantic features to manage ambiguous images flexibly and an orthogonal uncertainty estimation module based on pairwise relationships between samples. Inconsistent pseudo annotations to latent truth [14] employed a three-step framework for datasets with mismatched annotations. It matches the latent truth from the non-transferable pseudo-labels, ignores mapping between latent truth and input data, and puts little effort into predicting predictors for unseen samples. The label distribution learning with valence-arousal [16] model is an uncertainty-aware label distribution learning method that trains end-to-end to solve annotation ambiguity and configures the expression distribution for training samples. Dual-domain affect fusion method [38] introduced an approach exploring the relationship between discrete emotion classes and continuous representations through dual-domain influence fusion and addressed the fundamental label uncertainty by formalizing a mixed label set using a dual-domain label fusion module to leverage unique relationships. However, previous studies have focused on the

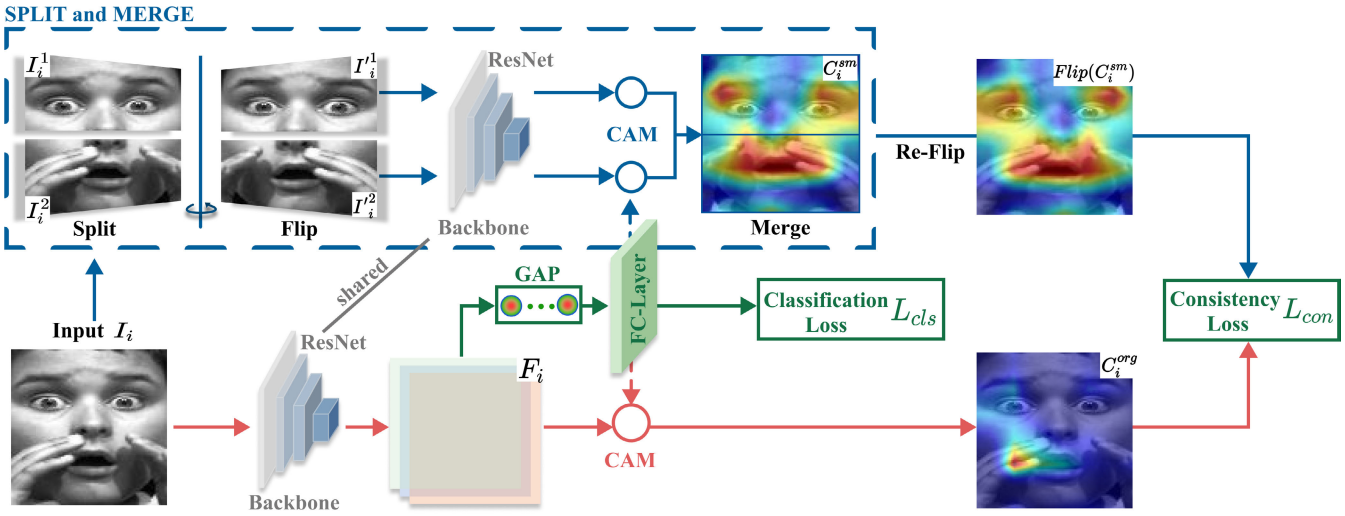


FIGURE 1. Overall procedures for split and merge consistency regularization (SMEC). To improve noisy label facial expression recognition (FER), facial expression images were divided into two parts, and class activation maps (CAMs) were extracted from various facial regions. The extracted CAMs were merged to match the original image size. The model employs the attention consistency loss [12] between these two CAMs to facilitate learning facial expression labels. Early learning regularization (ELR) loss [13] was applied to prevent the model from memorizing noisy labels during the early training stages. The proposed approach allows the model to perceive various parts of the face through splitting and merging.

latent truth region of the label by remembering the noise label of the dataset during training.

In the context of FER with noisy labels, erasing attention consistency (EAC) [8] addresses the problem of noisy labels through novel feature learning techniques. This method focuses explicitly on regions of attention related to noisy samples, and the researchers hypothesized that the FER model only observes a subset of features associated with noisy labels to memorize these challenging instances. The EAC mitigated label noise by training the model with attention consistency [12] and incorporating random erasing [31]. Despite these meticulous efforts, the EAC exhibits limitations in generalization performance during testing when trained on biased data due to a substantial noise level.

The proposed SMEC method splits facial images and obtains CAMs, ensuring that the semantic information of facial expression images is preserved. By merging these two images, the model can observe various parts of the facial image. Thus, SMEC outperformed previous methods in FER tasks. Furthermore, its ability to preserve the semantic information of facial expression images allowed it to exhibit robust performance even under severe noise rates.

B. CONSISTENCY REGULARIZATION

Consistency regularization [17], [32], [33] starts from the assumption that prediction results are consistent even when unlabeled data are perturbed [18]. The prediction result is unknown in unlabeled data; thus, if the class changes slightly enough to not alter through data augmentation techniques [19], it learns by assigning unsupervised loss so that the original data and prediction results are the same.

Additionally, studies [34] have leveraged the knowledge distillation framework [36], with some incorporating different input images through various techniques, such as CutMix [35]. Puzzle-CAM [22] reconstructed the regularization loss using an attention-based feature learning method between the CAM of a tile image consisting of tile images and the CAM of the original image to detect the integrated area of the object. SEAM [39] uses a self-attention mechanism because conventional CAM methods are inconsistent. This method refined CAM through unsupervised post-processing, employed a siamese network structure [44], and designed additional loss functions to ensure consistent performance across various transformations.

Within this transformation, visual attention consistency [12] was proposed for perceptual coherence of visual attention regions for multi-label image classification. The researchers built a two-branch network that inputs the original and converted images and introduces a consistent attention heatmap between the two branches. They assumed that when the input image is spatially transformed, the attention region for classification follows the same transformation. They found that the attention region of a convolutional neural network classifier can be derived from an attention heatmap in the middle network layer.

III. METHOD

To enhance the robustness of the FER model against noisy labels, we introduced the SMEC method. The proposed approach involves splitting the original image into images and inputting them into the backbone network to produce two CAMs. Subsequently, we merged these CAMs with those generated from the complete original image, incorporating attention consistency regularization [12]. To prevent the model from memorizing noisy labels, we adopted the ELR loss [13] as part of the classification loss. The approach demonstrates superior performance compared to existing methods, particularly when the proportion of noisy labels

FIGURE 2. Example images for the facial expression class in RAF-DB [28]

increases. The overall architecture is depicted in Fig. 1. This section provides a detailed explanation of the proposed SMEC method.

A. SPLIT AND MERGE MODULE

Existing methods for generating CAMs from a single image [20], [21] are vulnerable to label noise because they focus on specific parts of the object to identify the primary features of the class. Therefore, relying solely on capturing partial object regions is inadequate for extracting features and classifying facial expressions effectively.

Inspired by previous research [22], [23], [24] that extracted high-level CAMs by manipulating and transforming images, we propose an approach that splits a facial image into two and merges the images. Existing methods focus on only one part of the facial image. By dividing the image into two pieces without losing the semantics of the facial expression image, the model can observe various parts for each split image. The consistency regularization between the CAM of the merged image and the CAM of the original image enhances robust performance even if the label noise increases.

Given N images, $\{I\}_{i=1}^{N}$, $I_i \in \mathbb{R}^{C\times W\times H}$ denotes a single image in which C, W, and H represent the channel size, width, and height, respectively. The split and merge module splits the input image horizontally into non-overlapping images $\{I_i^1, I_i^2\}$ of size $\mathbb{R}^{C\times \frac{W}{2}\times H}$.

B. CLASS ACTIVATION MAP

Next, we generate the original CAM $C_i^{org} \in \mathbb{R}^{c\times w\times h}$ for original image I_i, where c, w, and h denote the number of classes, width, and height of the CAM, respectively. Then, we generate the split and horizontally flipped images I'^1_i and I'^2_i and process them through the neural network to generate the CAMs $C_i^1, C_i^2 \in \mathbb{R}^{c\times \frac{w}{2}\times h}$. The expression for creating the CAMs is provided below:

$$C_j(m,n) = \sum_{k=1}^{c} W(j,k)F_k(m,n), \tag{1}$$

where $C_j(m,n)$ represents the heatmap of the spatial position (m,n) of the j-th label, and $W(j,k)$ denotes the weight corresponding to the j-th label in the feature map channel k. Finally, $F_k(m,n)$ represents the feature map of the final convolutional layer. After creating a CAM of the split images, we combined the two CAMs of the same size as the original CAM through the split and merge module. The merged CAM represents $C_i^{sm} \in \mathbb{R}^{c\times w\times h}$.

C. LOSS

1) CLASSIFICATION LOSS TERM

The original image I_i undergoes feature extraction in the final convolutional network layer, generating feature maps $F_i \in \mathbb{R}^{c\times w\times h}$. We obtain $f_i \in \mathbb{R}^{1\times c}$ by applying global average pooling on the feature map F_i. Then, f_i is input into the softmax layer, producing the corresponding class scores p_i. In the generated p_i, W_{y_i} is the yth weight of the fully connected layer with the given label of the ith images as y_i, which matches the expression in (2). We computed the cross-entropy (CE) loss using p_i and the label y_i. Equation (3) represents the conventional CE loss.

$$p_i = \frac{e^{W_{y_i}f_i}}{\sum_j^c e^{W_j f_i}} \tag{2}$$

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y_i log p_i. \tag{3}$$

The CE loss is typically employed as the classification loss in existing FER or image classification models. However, when training with noisy labels, a model optimized with the CE loss may exhibit non-smooth learning because it tends to retain incorrect early learning outcomes if incorrect labels are introduced. Due to this memorization effect, FER models also face challenges in smoothly learning from noisy labels during the early training stages. To address this problem, we used the ELR loss [13], implicitly discouraging memorizing incorrect labels.

A regularization term is introduced to the current CE loss in (4). This additional term leverages the progress made during the initial learning phase to prevent retaining incorrect labels, providing a robust effect on noisy labels:

$$L_{cls} = L_{ce} + \frac{\lambda}{N}\sum_{i=1}^{n} log(1- < p_i, t_i >). \tag{4}$$

In (4), N refers to the number of data in a mini-batch, and λ denotes a weighted value to balance the existing CE loss. Moreover, p_i represents the conditional probability estimate for each given class generated through the softmax function, based on the value obtained through a c-dimensional encoding mapping of each input I_i using a neural network. Further, t_i is calculated by averaging the p_i

values that changed during the training process. Finally, $<,>$ indicates the inner product of the vector.

2) CONSISTENCY REGULARIZATION LOSS TERM

We introduced consistency regularization loss into the model to enhance robustness against label noise and improve performance. The $C_i^{sm} \in \mathbb{R}^{c\times w\times h}$ generated through the split and merge module to calculate the consistency regularization loss undergoes the process of reflipping to match the spatial with the original CAM C_i^{org}. Consistency regularization loss follows the equation below to reduce the gap between the CAM $C_i^{org} \in \mathbb{R}^{c\times w\times h}$ and the re-flipped CAM $Flip(C_i^{sm}) \in \mathbb{R}^{c\times w\times h}$:

$$L_{con} = \frac{1}{NcHW}\sum_{i=1}^{N}\sum_{j=1}^{c}||(C_{i,j}^{org}) - Flip(C_{i,j}^{sm})||_2, \quad (5)$$

where N is the number of images, and c is the number of labels.

The total loss consists of classification and consistency loss, which applies hyperparameter λ to balance the two losses:

$$L_{total} = L_{cls} + \lambda L_{con}. \quad (6)$$

IV. EXPERIMENT

This section compares the proposed SMEC method and various existing studies [8], [9], [15] using the real-world affective faces database (RAF-DB) [28], explaining the experimental setup and dataset in detail. Subsequently, this section presents the quantitative assessment of the performance of the proposed model by incorporating various comparative models, conducting experiments under different noise levels, and performing ablation studies.

A. DATASET

We leveraged RAF-DB [28], a large-scale facial expression dataset containing approximately 30,000 facial images, each labeled by about 40 annotators. This paper focuses on seven facial expressions (i.e., neutral, surprise, fear, disgust, happiness, sadness, and anger), as illustrated in Fig. 2. We used 12,271 images as training data and 3,068 as testing data.

B. IMPLEMENTATION DETAILS

We adopted ResNet-18 [27] as a backbone architecture pre-trained with MS-Celeb-1M [25]. The input image was resized to 224 × 224 pixels and aligned using three landmarks [29]. As a data augmentation technique, only random horizontal flips were used. The batch size was 64. The initial learning rate was 0.0001, and the weight decay was 1e-4. The network employed the Adam optimizer [26], and the entire training ended at 60 epochs. All experiments were performed on a single GeForce RTX 3090.

In experiments related to consistency loss, the hyperparameter that balances the two losses according to the ratio of label noise was fixed to 2 under clean labels, 3 under 10% to 40% noise, and 5 under 50% to 60% noise. Modifying the hyperparameter values according to the noise ratio is a pragmatic approach to fine-tuning the model training process. These hyperparameters ensure resilience against label noise while effectively harnessing severely noisy labels.

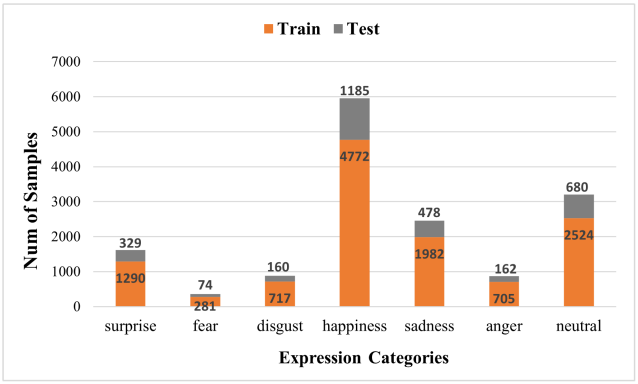


FIGURE 3. Class distribution of training and testing samples in RAF-DB [28].

C. METRICS

We used two performance metrics to evaluate the effectiveness of the proposed approach in improving resilience to label noise: the best accuracy and balanced accuracy [30], the arithmetic mean of the accuracy per class. The highest accuracy for the testing data at the end of training is the best accuracy. We estimated the accuracy as a recall for each class and averaged the values to obtain the balanced accuracy value. By adopting balanced accuracy, we can assess the overall performance of all large and small classes with the same importance as dealing with an imbalanced data distribution. The RAF-DB [28] exhibits an imbalanced class distribution within the training samples, as depicted in Fig. 3.

D. EVALUATION OF THE PROPOSED APPROACH ON THE NOISY RAF-DB

We evaluated the effectiveness of the SMEC method by employing state-of-the-art methods [8], [9], [15] for handling noisy label FER. The experimental results cover a range of noise rates, from 10% to 60% (including clean labels). We followed previous work [10], [20] to generate the noisy labels.

The proposed SMEC method performs better than all comparative models in the experimental results measured under clean labels and the 10% label noise in Table 1. The model outperforms state-of-the-art approaches by achieving balanced accuracy values of 82.58%, 81.09%, 81.06%, 81.06%, 75.22%, and 75.22% for each noise ratio. When the noise ratio falls within the range of 20% to 60%, the model exhibits slightly lower performance regarding the best accuracy compared to EAC [8]. However, this approach excels in balanced accuracy [30] over all existing models, suggesting that, as the noise increases, the model demonstrates greater robustness against the influence of class imbalance than the

TABLE 1. Evaluate split and merge consistency regularization (SMEC) with balanced accuracy [30], best accuracy and per-class accuracy on RAF-DB [28] with noise (including clean labels) ranging from 10% to 60%.

Methods	Noise (%)	Accuracy		Per-class accuracy						
		Balanced accuracy	Best accuracy	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
EAC [8]	0	81.32	89.54	**89.05**	54.05	**69.38**	**96.46**	87.87	**87.86**	84.57
SCN [15]		76.32	86.57	86.32	45.95	56.25	93.67	85.77	77.78	88.53
RUL [9]		81.04	88.89	88.75	62.16	61.88	95.95	88.49	82.1	87.94
SMEC (Ours)		**82.58**	**89.57**	87.87	**64.86**	**69.38**	95.7	**88.28**	84.57	**90.74**
EAC [8]	10	70.43	**88.3**	84.4	59.46	**68.75**	93.76	85.36	**83.95**	86.76
SCN [15]		72.84	85.01	86.02	43.24	42.5	93.92	81.59	74.07	88.53
RUL [9]		77.95	86.05	83.89	59.46	58.75	94.18	82.85	79.63	86.91
SMEC (Ours)		**81.09**	87.55	**86.93**	**60.81**	67.5	**95.19**	**88.28**	77.16	**91.76**
EAC [8]	20	67.99	**87.13**	84.8	**58.11**	57.5	92.24	79.29	**83.95**	86.03
SCN [15]		69.28	83.7	84.5	40.54	28.13	94.26	83.47	67.28	86.76
RUL [9]		73.23	83.67	81.16	51.35	51.88	94.35	72.18	72.84	88.82
SMEC (Ours)		**81.06**	87.06	**86.32**	**58.11**	**66.25**	**95.78**	**87.66**	82.72	**90.59**
EAC [8]	30	67.06	**85.33**	83.59	52.7	61.25	92.66	78.03	**80.25**	**85**
SCN [15]		63.12	78.36	78.12	40.54	13.75	93.16	79.08	61.73	75.44
RUL [9]		70.62	81.13	80.85	48.65	50.63	93.16	83.05	64.2	73.82
SMEC (Ours)		**81.09**	84.65	**86.93**	**60.81**	**67.5**	**95.19**	**88.28**	77.16	84.65
EAC [8]	40	63.91	**83.93**	79.94	56.76	58.13	91.39	69.87	72.22	78.97
SCN [15]		46.53	73.6	72.95	0	0	90.3	65.69	60.49	78.82
RUL [9]		59.19	79.95	83.28	31.08	37.5	90.13	79.5	67.28	80.74
SMEC (Ours)		**81.06**	82.92	**86.32**	**58.11**	**66.25**	**95.78**	**87.66**	**82.72**	**90.59**
EAC [8]	50	61.94	**80.31**	70.82	55.41	**82.19**	69.67	64.2	67.94	80.31
SCN [15]		34.56	63.7	57.75	0	0	90.3	22.8	22.84	77.8
RUL [9]		63.69	73.01	83.59	43.24	33.13	83.88	71.13	61.73	69.12
SMEC (Ours)		**75.22**	79.76	**85.41**	**56.76**	45	**92.74**	**75.31**	**80.86**	**90.44**
EAC [8]	60	47.97	**76.73**	58.36	44.59	41.25	71.73	45.4	56.79	59.71
SCN [15]		27.85	52.74	28.88	0	0	**96.2**	1.26	17.28	51.32
RUL [9]		50.98	67.14	69.3	4.05	14.38	89.54	29.92	74.69	75
SMEC (Ours)		**75.22**	75.39	**85.41**	**56.76**	**45**	92.74	**75.31**	**80.86**	**90.44**

standard models, improving generalization. In particular, the model performs well in classes with a small of the samples in the dataset (e.g., fear, disgust, and anger). Compared to comparable noisy label FER models, the proposed SMEC method demonstrates enhanced FER performance and robustness to noisy labels, primarily when applied across imbalanced classes with varying noise levels. Thus, the proposed approach demonstrates robustness to noisy labels in the FER task while preserving the meaningfulness of facial expressions. Furthermore, SMEC exhibits robust performance even in situations with a high proportion of noisy labels.

E. ABLATION STUDY

1) WHY DOES SPLITTING AN IMAGE IN TWO EFFECTIVELY CLASSIFY FACIAL EXPRESSIONS?

To demonstrate the effectiveness of SMEC, we conducted a qualitative analysis by dividing facial images into two, four, and 16 images. As presented in Table 2, we observed better classification performance when we split facial expression images into two because it preserves the meaning of the facial images more effectively. As the proportion of noisy labels increases, we divided them into two, displaying superior classification performance and resistance compared to dividing them into four or 16. In a previous study [8], random erasing [31] of facial expression images proved helpful in providing diverse perspectives on different components. However, this approach tended to overfit the data as the proportion of noisy labels increased. Additionally, we hypothesized that excessively partitioning the images may lead to a loss of facial expression semantics. The experimental results found that, although there were higher-performing classes when dividing facial images into more segments, this can be attributed to overfitting when considering the overall results.

2) WIDTH SPLIT IS EFFECTIVE FOR FER

This paper proposes a method of splitting facial expression images by width. By dividing the facial images into widths, we assumed that the model would better classify facial expressions without losing the semantics of the face, and we verified this assumption through the experiments. Thus, we divided the facial images into two categories: width and height. In this experiment, we applied the same noise rate from 10% to 60% (including clean labels) to check the FER ability. As listed in Table 3, when split by width, the classification accuracy performance remains robust despite an increase in the ratio of noisy labels.

In contrast, overfitting is evident when dividing images by height, even with a lower noise rate. As the noise ratio increases, FER performance significantly deteriorates.

TABLE 2. For comparison of facial expression semantic preservation ability, balanced accuracy [30] and best accuracy of split and merge consistency regularization (SMEC) on noisy RAF-DB [28], with noise ranging from 10% to 60%(including clean labels) with different splits(2, 4, 16).

Methods	Noise (%)	Accuracy		Per-class accuracy						
		Balanced accuracy	Best accuracy	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
Split 2	0	**82.58**	**89.57**	84.5	**64.86**	**69.38**	**95.7**	**88.28**	**84.57**	**90.74**
Split 4		77.95	87.94	**86.01**	44.59	65	94.94	85.36	79.01	90.73
Split 16		71.67	84.09	80.24	47.3	38.75	94.94	80.75	72.84	86.86
Split 2	10	**81.09**	**87.55**	86.93	**60.81**	**67.5**	95.19	**88.28**	**77.16**	**91.76**
Split 4		76.63	87.09	**88.15**	43.24	63.13	95.19	87.03	73.46	86.18
Split 16		70.84	83.8	83.28	43.24	38.13	**95.7**	79.5	71.6	84.41
Split 2	20	**81.06**	**87.06**	**86.32**	**58.11**	**66.25**	95.78	**87.66**	**82.72**	**90.59**
Split 4		74.15	85.33	85.11	41.89	55	**96.37**	82.43	75.93	82.35
Split 16		56.92	77.84	74.77	10.81	0.63	95.19	71.76	62.96	82.35
Split 2	30	**81.09**	**84.65**	**86.93**	**60.81**	**67.5**	**95.19**	**88.28**	**77.16**	**91.76**
Split 4		67.53	82.63	79.33	32.43	25.63	93.33	80.13	73.46	88.38
Split 16		54.09	76.6	69	1.35	0	94.35	69.67	59.88	84.41
Split 2	40	**81.06**	**82.92**	**86.32**	**58.11**	**66.25**	95.78	**87.66**	**82.72**	**90.59**
Split 4		63.61	81.16	70.21	17.57	29.38	95.44	82.01	67.28	83.38
Split 16		49.34	72.46	69	0	0	**96.29**	60.16	48.15	71.76
Split 2	50	**75.22**	**79.76**	**85.41**	**56.76**	**45**	92.74	75.31	**80.86**	**90.44**
Split 4		56.60	77.74	72.95	0	1.88	92.74	**77.62**	68.52	82.5
Split 16		32.86	60.63	45.59	0	0	**96.12**	14.64	0	73.68
Split 2	60	**75.22**	**75.39**	**85.41**	**56.76**	**45**	92.74	**75.31**	**80.86**	**90.44**
Split 4		53.23	74.54	72.04	0	0	92.24	59.83	65.46	83.09
Split 16		18.97	45.47	3.34	0	0	**99.92**	0.42	0	29.12

TABLE 3. Comparison of width and height splitting strategies evaluated on RAF-DB [28] with noise ranging from 10% to 60%, including clean labels.

Methods	Noise (%)	Accuracy		Per-class accuracy						
		Balanced accuracy	Best accuracy	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
width	0	**82.58**	**89.57**	84.5	**64.86**	**69.38**	95.7	**88.28**	**84.57**	**90.74**
height		78.87	88.07	**86.63**	48.65	66.87	**95.78**	85.98	80.25	87.94
width	10	**81.09**	**87.55**	**86.93**	**60.81**	**67.5**	**95.19**	**88.28**	77.16	**91.76**
height		61.44	86.44	84.19	4.05	8.13	92.49	75.1	**79.63**	86.47
width	20	**81.06**	**87.06**	**86.32**	**58.11**	**66.25**	**95.78**	**87.66**	**82.72**	**90.59**
height		73.22	85.04	85.41	40.54	48.13	94.43	81.17	75.93	86.91
width	30	**81.09**	**84.65**	**86.93**	**60.81**	**67.5**	**95.19**	**88.28**	**77.16**	**91.76**
height		68.15	82.82	**86.93**	31.08	33.75	93.33	82.43	65.43	84.12
width	40	**81.06**	**82.92**	**86.32**	**58.11**	**66.25**	**95.78**	**87.66**	**82.72**	**90.59**
height		62.61	81.09	82.67	8.11	30.63	94.35	74.48	61.73	86.32
width	50	**75.22**	**79.76**	**85.41**	**56.76**	**45**	**92.74**	**75.31**	**80.86**	**90.44**
height		61.44	80.35	84.19	4.05	8.13	92.49	75.1	79.63	86.47
width	60	**75.22**	**75.39**	**85.41**	**56.76**	**45**	**92.74**	**75.31**	**80.86**	**90.44**
height		53.97	74.02	81.16	0	0	91.14	61.72	67.28	76.47

We interpret these experimental results as follows. When dividing the face (the eyes, nose, and mouth) into two parts by width, we can provide visual diversity while preserving the meaning of the facial image. However, when divided by height, all components are split in half, resulting in a loss of meaning in facial expression and leading to overfitting in the FER performance. This observation confirms the hypothesis that using the facial image split at the width level results in less loss of facial semantics.

3) QUALITATIVE COMPARISON OF THE CAM

Fig. 4 presents examples of naïve CAMs [20] and CAMs that SMEC created for each facial expression. Although naïve CAMs activate only a tiny portion of the face, the CAMs generated using the proposed method capture a wider facial area, demonstrating better reasoning ability. In particular, for a neutral expression, the naive CAM activated only a minimal region around the nose, whereas SMEC captured a broader area spanning the eyes, nose, and mouth. We conjecture that SMEC's ability to capture a wider region is due to the split and merge module, which enables it to identify individual regions related to facial expressions from the split images. The qualitative results indicate that SMEC observes the entire facial region when inferring facial expressions, which is effective for FER and benefits from robustness even on noisy labels.

4) EFFECTS OF USING LOSS ROBUST TO LABEL NOISE

We derived an ablation study on each loss term in RAF-DB [28], assessing the FER performance under various noise

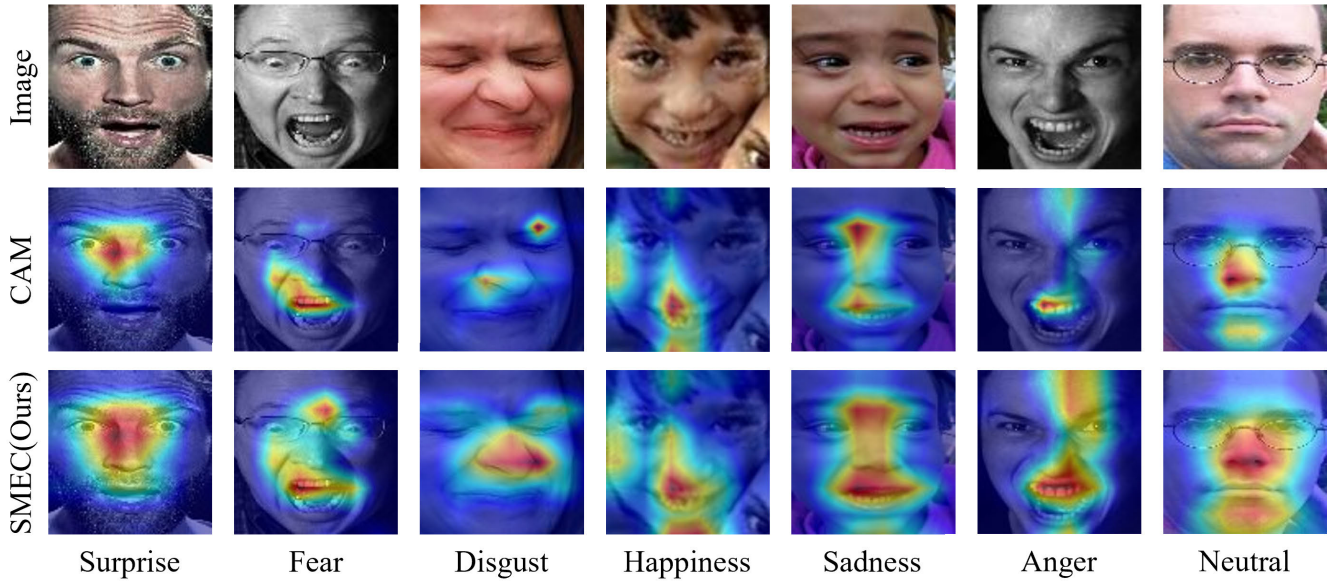


FIGURE 4. Qualitative comparison between split and merge consistency regularization (SMEC) and the naïve class activation map (CAM) [20] on the original images.

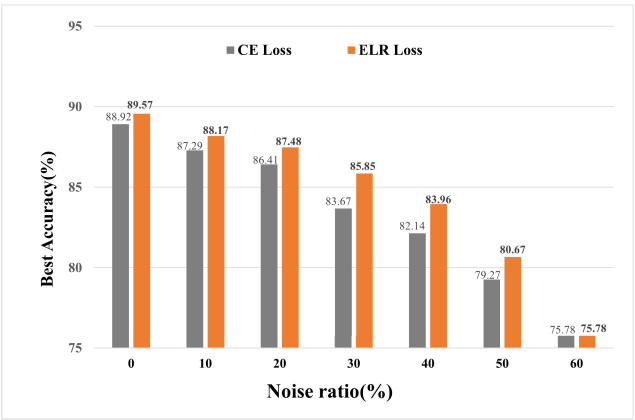


FIGURE 5. Effect of using cross-entropy (CE) loss and early learning regularization (ELR) loss [13] that is robust to label noise.

ratios (from 10% to 60%, including clean labels). As illustrated in Fig. 5, when the noise level is low, the effect of the ELR loss [13] is comparable to the performance of the conventional loss. However, a high noise ratio has a more pronounced effect on preventing the neural network from inverting the original predictions and memorizing incorrect outcomes. This classification loss demonstrates robust performance when noise is significant. At a noise level of 30%, this approach exhibits a notable performance improvement of approximately 2%p, indicating its effectiveness in regularization, as observed in the experimental results.

V. CONCLUSION AND FUTURE WORK

This paper focuses on various aspects of facial expression images and extracts and merges CAMs while dividing the face into two parts to maintain facial expression semantics. We proposed the SMEC method, which calculates the consistency loss between the merged and original CAMs and applies a loss to prevent memorization, enhancing the robustness of label noise. To the best of our knowledge, no studies or experiments have been conducted at severe noise rates (40% to 60%). Most existing methods fail to recognize facial expressions of suffer from overfitting at such high noise rates. Compared to these existing methods, SMEC demonstrates greater resilience to label noise across various noise rates, particularly concerning inter-class similarity.

The proposed SMEC method achieves a remarkable improvement of 47.37%p over the existing state-of-the-art noisy label FER models when the noisy label ratio is 60%, ensuring effective FER performance. Additionally, it exhibits robustness in the presence of noisy labels. These results emphasize the proficiency of this method in accurately recognizing facial emotions, irrespective of the diverse noises encountered in real-world scenarios.

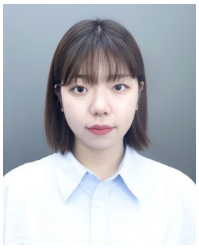
The research demonstrates superior performance even when increasing the ratio of noisy labels compared to existing studies. However, as a limitation, within the experimental results for SMEC, the performance underperforms when image data are scarce for certain classes. In future research, we will continue to enhance the robustness of low-data classes and explore methods to reduce computational costs, enabling the application of FER tasks in diverse fields.

REFERENCES

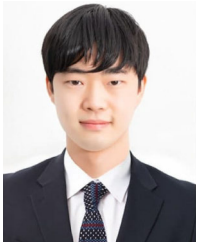
[1] P. Ekman, "The argument and evidence about universals in facial expressions of emotion," in *Handbook of Social Psychophysiology*, vol. 58, H. Wagner and A. Manstead, Eds. 1989, pp. 342–353.

[2] P. Ekman and W. V. Friesen, "Facial action coding system (FACS): A technique for the measurement of facial actions," *Rivista Di Psichiatria*, vol. 47, no. 2, pp. 38–126, 1978.

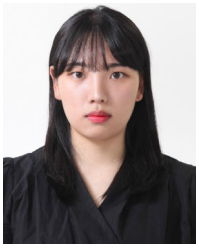
[3] E. M. Onyema, P. K. Shukla, S. Dalal, M. N. Mathur, M. Zakariah, and B. Tiwari, "Enhancement of patient facial recognition through deep learning algorithm: ConvNet," *J. Healthcare Eng.*, vol. 2021, pp. 1–8, Dec. 2021.
[4] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion recognition for healthcare surveillance systems using neural networks: A survey," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2021, pp. 681–687.
[5] Z. Chen, X. Feng, and S. Zhang, "Emotion detection and face recognition of drivers in autonomous vehicles in IoT platform," *Image Vis. Comput.*, vol. 128, Dec. 2022, Art. no. 104569.
[6] L. Yang, H. Yang, B.-B. Hu, Y. Wang, and C. Lv, "A robust driver emotion recognition method based on high-purity feature separation," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15092–15104, Dec. 2023.
[7] C.-S. Jiang, Z.-T. Liu, M. Wu, J. She, and W.-H. Cao, "Efficient facial expression recognition with representation reinforcement network and transfer self-training for human–machine interaction," *IEEE Trans. Ind. Informat.*, vol. 19, pp. 9943–9952, 2023.
[8] 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)*, 2022, pp. 418–434.
[9] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 1–12.
[10] 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.
[11] D. Gera, N. S. K. Badveeti, B. V. R. Kumar, and S. Balasubramanian, "Dynamic adaptive threshold based learning for noisy annotations robust facial expression recognition," 2022, *arXiv:2208.10221*.
[12] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 729–739.
[13] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 33, 2020, pp. 20331–20342.
[14] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 222–237.
[15] 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.
[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] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 29, 2016, pp. 1171–1179.
[18] L. Zhang and G.-J. Qi, "WCP: Worst-case perturbations for semi-supervised deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3911–3920.
[19] Z. Ke et al., "Guided collaborative training for pixel-wise semi-supervised learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 429–445.
[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[21] 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.
[22] S. Jo and I.-J. Yu, "Puzzle-CAM: Improved localization via matching partial and full features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 639–643.
[23] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.
[24] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12272–12281.
[25] 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. (ECCV)*, 2016, pp. 87–102.
[26] D. Kinga and J. B. Adam, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, vol. 5, 2015, p. 6.
[27] 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.
[28] 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.
[29] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6970–6980.
[30] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3121–3124.
[31] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.
[32] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.
[33] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of semi-supervised learning algorithms," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–15.
[34] G. French, S. Laine, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong varied perturbations," in *Proc. 31th Brit. Mach. Vis. Conf. (BMVC)*, 2020, pp. 1–21.
[35] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.
[36] G. Hinton, O. Vinyals, and J. F. Dean, "Distilling the knowledge in a neural network," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS) Workshop*, 2014, pp. 1–9.
[37] X. Zhang, Y. Lu, H. Yan, J. Huang, Y. Ji, and Y. Gu, "ReSup: Reliable label noise suppression for facial expression recognition," 2023, *arXiv:2305.17895*.
[38] D. Neo, T. Chen and S. Winkler, "Large-scale facial expression recognition using dual-domain affect fusion for noisy labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5691–5699.
[39] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12272–12281.
[40] Y. Chen and J. Joo, "Understanding and mitigating annotation bias in facial expression recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14960–14971.
[41] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 524–543, Apr. 2021.
[42] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022.
[43] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, May 2023.
[44] G. Kosh, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 2, 2015, pp. 1–8.



JIHYUN KIM (Graduate Student Member, IEEE) received the B.S. degree in information security from Baewha Women's University, Seoul, South Korea, in 2021. She is currently pursuing the M.S. degree with the Graduate School of Artificial Intelligence, Chung-Ang University. Her research interests include deep learning and computer vision.

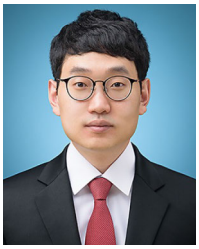


JUNEHYOUNG KWON received the B.S. degree in philosophy from Kyunghee University, Seoul, South Korea, in 2020, and the M.S. degree in digital imaging from Chung-Ang University, in 2022, where he is currently pursuing the Ph.D. degree with the Graduate School of Artificial Intelligence. His research interests include deep learning and computer vision.



MIHYEON KIM received the B.S. degree in mathematics from Chung-Ang University, Seoul, South Korea, in 2022, where she is currently pursuing the M.S. degree with the Graduate School of Artificial Intelligence. Her research interests include deep learning and natural language processing.

EUNJU LEE received the B.S. and M.S. degrees in imaging engineering from Chung-Ang University, Seoul, South Korea, in 2020 and 2022, respectively, where she is currently pursuing the Ph.D. degree in imaging engineering with the Graduate School of Advanced Imaging Science, Multimedia and Film. Her current research focuses on deep learning and computer vision.



YOUNGBIN KIM (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in visual information processing from Korea University, in 2010, 2012, and 2017, respectively. From August 2017 to February 2018, he was a Principal Research Engineer with Linewalks. He is currently an Assistant Professor with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University. His current research interests include data mining and deep learning.