



# Effective attention feature reconstruction loss for facial expression recognition in the wild

Weijun Gong<sup>1</sup> · Yingying Fan<sup>1</sup> · Yurong Qian<sup>1,2,3</sup>

Received: 7 October 2021 / Accepted: 30 January 2022 / Published online: 4 March 2022  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

## Abstract

Facial expression recognition (FER) in the wild is very challenging due to occlusion, posture, illumination, and other uncontrolled factors. Learning discriminant features for FER using Convolutional Neural Networks is a momentous task for the significant class imbalance, wrong labels, inter-class similarities, and intra-class variations. The traditional method utilizes the Cross entropy loss function to optimize the convolutional network to obtain discriminative features for classification. However, this loss function cannot effectively solve the above problems in practice and cannot contribute to obtaining highly discriminant facial features for further analysis. Center loss improves the learning efficiency by reducing the intra-class distance of similar expressions, while the improvement of inter-class similarity, class imbalance, and generalization is insufficient. In this paper, we propose a lightweight Effective Attention Feature Reconstruction loss (EAFR loss), which can further optimize the feature space and enhance the discriminability of expression. The loss model is composed of the Focal Smoothing loss (FS loss) and the Aggregation-Separation loss (AS loss). Firstly, the FS loss can improve the poor recognition performance caused by imbalanced classes and prevent paranoid knowledge learning behaviors. Meanwhile, AS loss further accurately condenses the intra-class expression features and expands the inter-class distance, which is achieved by using progressive stage max-pooling channel and position attention mechanism and lightweight asymmetric autoencoder model for feature reconstruction. Finally, the EAFR loss joins the above two loss functions to more comprehensively solve the above typical problems for FER in the wild. We validate the proposed loss function on three most commonly used large-scale wild expression datasets (RAF-DB, FERPlus, and AffectNet), and the results show that our model achieves superior performance to several state-of-the-art methods.

**Keywords** Facial expression recognition · Feature reconstruction · The joint loss · Deep learning

## 1 Introduction

As the most direct and effective way of expressing human emotions, facial expressions can be used to convey and understand emotions more realistically in interpersonal communication. With the continuous development of computer vision, the automatic recognition and analysis of

facial expressions have more application value in health assistance, automatic driving, mental analysis, public safety, etc. It has become an important computer vision research task. We call the research on the emotional state of facial expressions (neutral, happy, angry, sad, fear, disgust, surprise, contempt, etc.) as facial expression recognition (FER).

In recent years, with the continuous research and in-depth development of FER, the recognition of facial expressions taken in the laboratory under controlled conditions (correct posture, no occlusion, simple background, etc.) has achieved high recognition accuracy. Such expression datasets include JAFFE [23], CK + [22], Oulu CASIA [43], etc. However, with the continuous expansion of application scenarios, FER in real scenes including different occlusions, postures, lighting, and other factors is

✉ Yurong Qian  
qyr@xju.edu.cn

<sup>1</sup> College of Information Science and Engineering, Xinjiang University, Urumqi 830000, China

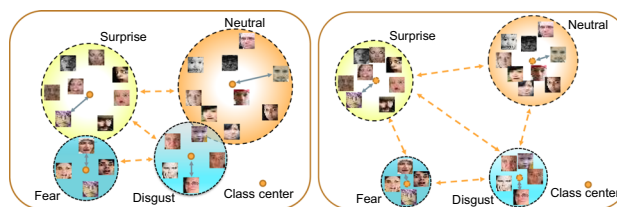
<sup>2</sup> College of Software, Xinjiang University, Urumqi 830000, China

<sup>3</sup> Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Urumqi 830000, China

facing greater challenges, and the controlled expression recognition in the laboratory environment has not been well accommodate the application research of FER in the wild. Along with the large-scale wild facial expression datasets are constantly being presented, such as RAF-DB [17], FERPlus [3], and AffectNet [25], more and more scholars pay attention to this field and carry out continuous research.

The early research on FER is based on the traditional machine learning method, which aims to find effective classification features in the feature space. This kind of method has achieved good results on the small-scale expression datasets in the lab [27, 29, 41]. With the introduction of the deep CNN method and its powerful feature extraction capability [14], many scholars have carried out researches on large-scale uncontrolled facial expression datasets and achieved certain results [1, 12, 30, 38, 40]. However, there are still some problems: (1) Due to the influence of posture, occlusion, illumination, scale, and other external factors, some of the expressions show significant intra-class variation and obvious inter-class similarity, which only have subtle differences, such as surprise and fear, contempt and neutral expressions, the recognition of these expression samples faces greater challenges. Therefore, it is crucial to pay more attention to the regions of interest and extract more discriminative expression features. In the meanwhile, we need to design a lightweight network to improve the overall efficiency of the model and further reduce feature redundancy and additional computational burden; (2) Due to the imbalance between classes in most of the existing wild expression datasets, the learning ability of minority classes is insufficient with few samples. On the other hand, the paranoid learning without any suspicion on the sample label of the dataset further reduces the generalization performance of the model; (3) As shown in Fig. 1 (left), the commonly used Cross entropy loss function cannot effectively deal with the spatial distribution of features of intra-class and inter-class, and the discrimination of learned facial features is insufficient [21], which is prone to misjudgment. Therefore, it cannot establish an efficient punishment mechanism in the classifier to improve the classification performance.

In response to the above problems, we propose a novel lightweight end-to-end Effective Attention Feature Reconstruction loss (EAFR loss) model for FER. In this model, the key features are further focused on by the progressive stage Max-pooling Coordinate Attention (MCA). Then the features are reconstructed by lightweight two-stage Asymmetric Autoencoder (AAE) to extract more discriminative expression space features, so it will pay more attention to the area of interest and obtain more adaptive expression feature weights. The self-adaptive Aggregation-Separation loss (AS loss) function constructed on the above new feature space weights can further condense intra-class features, maximize the



**Fig. 1** Cross entropy loss (left), EAFR loss (right). EAFR loss can make the features of intra-class more condensed toward the center of the class, while the distance of inter-class is more alienated

distance of inter-class, and reduce mutual interference. Finally, combined with the designed Focal Smoothing loss (FS loss) function with label smoothing, the imbalance and paranoid learning problems caused by unbalanced category and label errors will be further improved. As shown in Fig. 1 (right), the final joint loss model achieves better recognition performance.

Our main contributions are summarized as follows:

- (1) We present a novel loss function called Effective Attention Feature Reconstruction loss (EAFR loss) to effectively improve the poor recognition performance caused by unbalanced categories, paranoid learning for labels, and unreasonable intra-class and intra-class distances. This loss function combines the FS loss and the adaptive AS loss under the expression area of interest, which can efficiently solve the problems mentioned above of FER.
- (2) We propose an AS loss that combines the progressive stage channel and position MCA and AAE mechanisms to extract more discriminative facial features from area of interest and better reconstruct adaptive feature weights in expression space.
- (3) We design an FS loss function with a label smoothing mechanism to improve the low recognition rate caused by class imbalance effectively. At the same time, the smoothing mechanism can further improve the paranoid learning behavior and make the model have better generalization performance.
- (4) Our EAFR loss model is evaluated and visualized on three of the most popular large-scale wild datasets. Experiments show that our method outperforms several current state-of-the-art methods and achieves 89.80%, 89.57%, and 61.05% excellent performance on RAF-DB, FERPlus, and AffectNet, respectively.

## 2 Related work

In this section, we will focus on two aspects of related work done by predecessors: 1) FER in the wild. 2) Enhancing the performance of FER through loss function.

## 2.1 FER in the wild

The early FER mainly focus on the facial expression in the lab. However, as people attach great importance to the application research of expression in the wild, the research based on large-scale expression datasets in the real scene gradually gets more and more attention, such as RAF-DB, AffectNet, and FERPlus.

For focusing on regions of interest and extracting more effective latent features, Li et al. [18] and Wang et al. [35] respectively proposed global–local-based Attention CNN (gACNN) and Region Attention Networks (RAN) attention mechanisms to enhance robust recognition of wild expressions. Albanie et al. [2] used Squeeze-and-Excitation Network (SENet) to reconstruct the attention network to improve the learning ability of network expressions. Li et al. [15] developed a more robust Patch-Gated Convolution Neural Network (PG-CNN) for partial occlusion of real-world datasets and further improving the overall recognition rate. Zeng et al. [39] proposed an end-to-end training model called Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) for inconsistent labels in real datasets, which learns the potential association from inconsistent labels and outputs the hidden real labels for recognition. Barsoum et al. [3] further enhanced the network’s performance by constructing a more realistic label distribution. Georgescu et al. [7] improved recognition performance by combining multiple CNN networks and manual bag-of-visual-words (BOVW) features. Wang et al. [34] proposed the Self-Cure Network (SCN) to dynamically adjust labels during the training process for the current large-scale facial expression datasets in the wild with uncertain factors such as low image quality and labeling errors. Siqueira et al. [31] developed the ESR model, which integrates different network structures according to the type of datasets (in the lab or in the wild) to improve the recognition performance. For better feature extraction, Vo et al. [33] proposed a pyramid super-resolution (PSR) network structure to solve the FER task in the wild and achieve certain results. Karnati et al. [13] designed FER-net to extract relevant features accurately that can capture changes in facial expressions with high fidelity. Xia et al. [37] proposed an ADC-Net that combines attention and redesigns the reconstruction module to learn more distinctive facial expression features from the critical local subregions of scrambled facial expressions. Zhao et al. [42] designed the EfficientFace method from the feature extraction and training strategy for more effective and robust FER.

## 2.2 Enhancing the performance of FER through loss functions

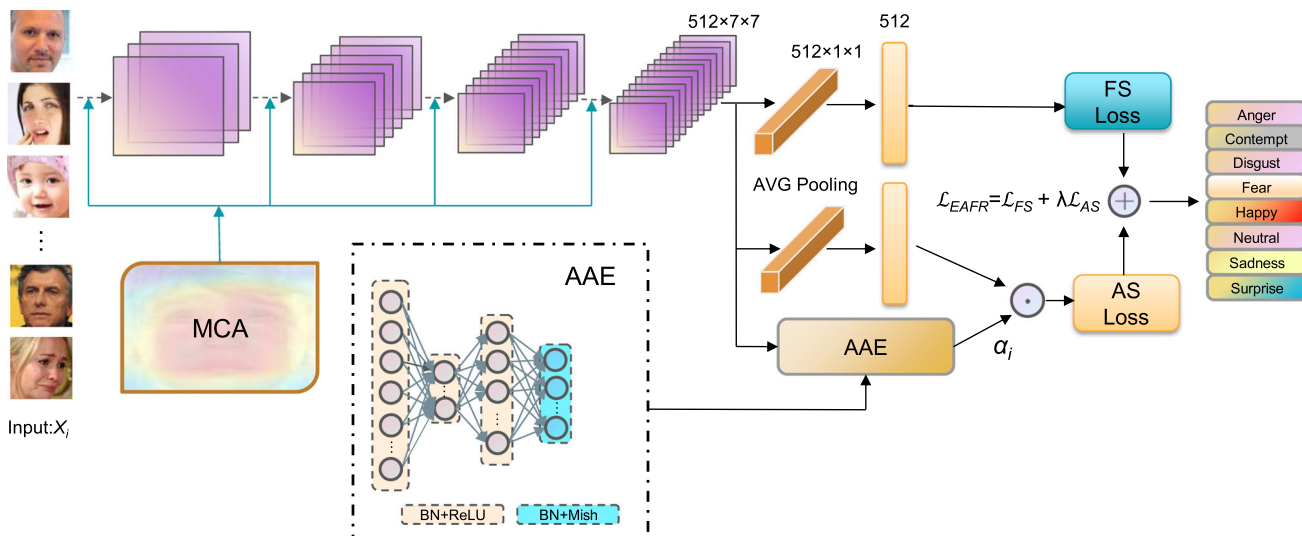
The loss function is usually an important approach for enhancing the discriminant ability of FER. To optimize the problems existing in FER in essence, some scholars make further improvements to the traditional cross entropy loss function. Wen proposed Center loss [36] to show good performance in face recognition by minimizing the distance between facial features and their corresponding class centers. Inspired by Center loss, Li et al. [17] introduced Locality-Preserving loss (LP loss) by condensing class features through the K-nearest neighbor algorithm for further reducing the distance within the class. Cai et al. [4] proposed isL and loss to further improve the inter-class distance. The Separate loss [16] uses normalized cosine similarity to adjust intra-class and inter-class distances to increase feature differentiation. Farzaneh et al. [6] further increased the penalty for the imbalance between classes to enhance the separability of each class. Fan et al. [5] proposed RW loss to learn discriminative features by enhancing the rationality of distance for intra-class and inter-class samples and suppressing uncertainties through a sample weighting scheme. Li et al. [19] built a knowledgeable teacher network (KTN) model based on adaptive regular loss (AdaReg loss), which can effectively enhance the discriminant ability by maximizing the distance between different facial expressions.

The above method enhances the discriminative ability of facial features by designing different loss functions. However, these methods do not pay special attention to the area of interest in expressions and the construction of adaptive feature weights. In contrast, our method combines the loss of feature weight reconstruction through the features of the area of interest extracted by attention with the loss function that improves the class imbalance problem and generalization performance, and the obtained features are more adaptable. It is conducive to learning the subtle differences between facial expressions, better handling the relationships of intra-classes and inter-classes, enhancing generalization performance, and ultimately improving the overall performance of facial expression recognition.

## 3 Proposed method

**Overview:** our proposed EAFR loss method includes a backbone network with the progressive stage Maxpooling Coordinate Attention (MCA), an Asymmetric Autoencoder (AAE), and a joint loss function  $\mathcal{L}$ . The overview of the proposed model is shown in Fig. 2.

Firstly, we use the lightweight ResNet18 as the backbone network to extract features. This network has a small



**Fig. 2** Overview of proposed EAFR loss model. The input expression  $X_i$  is sent to the backbone network (ResNet18) that joins the lightweight progressive stage attention MCA, and the attention features of the area of interest are sent to the two-stage lightweight

AAE for feature encoding and reconstruction, reconstructed feature weights and original features are reweighted to construct AS Loss. Finally, The EAFR loss is constructed as a joint loss through the combination of AS loss and FS loss

number of layers and parameters, and the residual network structure can effectively solve the gradient dispersion and explosion and perform better function fitting and feature extraction. To reduce the influence of non-expression regions and extract more expressive features in the area of interest without increasing excessive network burden, we gradually introduce a lightweight max-pooling space and location attention MCA in the backbone network in stages. Then the high-dimensional features of the expressions that pass through the last layer of the backbone network are sent to the lightweight two-stage AAE for feature weight reconstruction. The obtained latent feature weights with more adaptive capabilities participate in constructing the AS loss function proposed in this paper, which further promotes the cohesion for intra-class features and increases the distance on inter-class features. In the meantime, we send the final output of the backbone network to the designed FS loss function to solve the problems of class imbalance and overconfidence caused by mislabeling for improving the overall performance and generalization of the network. Finally, our proposed EAFR loss model combines the above two types of losses for learning and verifies the performance on three expression datasets in the wild.

### 3.1 Max-pooling coordinate attention (MCA)

Inspired by Coordinate Attention [10] and to capture features more sensitive for subtle expression changes, we propose MCA to embed the position information into the channel information for enhancing the orientation and

position perception capabilities of the feature map, while the general attention only focuses on the channel relationship to measure the importance of each channel. First, global max-pooling is used to decompose the channel attention into two spatial directions of aggregation features. The attention module can capture long-term dependencies along one spatial direction and retain accurate position information along the other spatial direction, which helps the network locate the target of interest more accurately. Then the generated feature maps of the two directions are respectively encoded into a pair of direction-aware and position-sensitive maps and finally merged with the input feature maps to further enhance the focus on the area of interest. The detailed structure of the MCA is shown in Fig. 3.

MCA encodes channel relationships and long-term dependencies through precise location information, and it is divided into two steps: information embedding and attention generation.

(1) To enable the attention module to capture long-term spatial interaction with accurate location information, we use adaptive max-pooling in two directions and convert it into a pair of one-dimensional feature encoding operations:

The output of channel  $c$  with a height of  $h$  can be represented as:

$$z_c^h(h) = \max_{i \in W} x_c(h, i) \tag{1}$$

The output of channel  $c$  with a width of  $w$  can be shown as:

$$z_c^w(w) = \max_{j \in H} x_c(j, w) \tag{2}$$

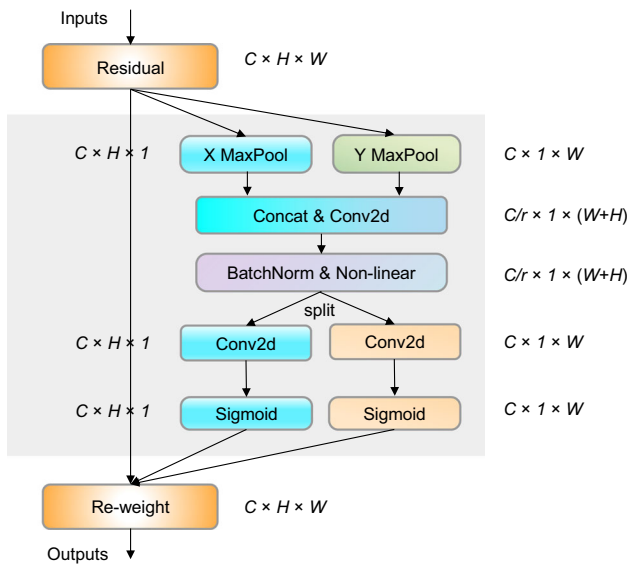


Fig. 3 The structure of MCA

The above transformations perform feature aggregation separately along two spatial directions and obtain corresponding feature maps that can perceive direction. So it can make the attention module capture long-term dependencies along one spatial direction and save accurate position information along another spatial direction, which will favor the model to more exactly locate the expression area of interest.

(2) To effectively capture the relationship between channels and make full use of the location information of the area of interest, we perform concatenate operation on the above transformations after embedding the above information, and then take transformation by using the  $1 \times 1$  convolution function  $F$ :

$$f = \delta(F([z^h, z^w])) \tag{3}$$

where  $[z^h, z^w]$  represents the concatenate operation along the spatial dimension,  $\delta$  denotes the non-linear activation function,  $f \in \mathbb{R}^{C/r \times (W+H)}$  donates feature mapping for encoding spatial information in two directions,  $r$  is used to control the reduction rate of the block size, here is 32.

Next,  $f$  is decomposed into two independent tensors of  $f^h \in \mathbb{R}^{C/r \times H}$  and  $f^w \in \mathbb{R}^{C/r \times W}$  along the spatial direction. Then using two  $1 \times 1$  convolutional transforms  $F_h$  and  $F_w$  to perform channel transformations with  $f^h$  and  $f^w$  respectively to obtain a tensor with the same number of channels, the formula is shown as follows:

$$a^h = \sigma(F_h(f^h)) \tag{4}$$

$$a^w = \sigma(F_w(f^w)) \tag{5}$$

where  $\sigma$  is the sigmoid activation function, and  $a^h$  and  $a^w$  are attention weights respectively.

The output feature of the final MCA attention is as follows:

$$y_c(i, j) = x_c(i, j) \times a_c^h(i) \times a_c^w(j) \tag{6}$$

### 3.2 Asymmetric autoencoder (AAE)

To extract the most representative features of facial expressions and reduce redundant information, we design a concise and effective AAE network that can greatly reduce redundant information while constructing optimal feature weights to enhance the discriminant ability of the network. Different from previous autoencoders, AAE uses a two-stage asymmetric encoding and decoding method to reconstruct key features. As shown in Fig. 2, First, the original high-dimensional information output from the backbone network is embedded into a 128-dimensional feature space. Then these low-dimensional features are reconstructed into a 1024-dimensional of sub-high dimensional feature space. Next, the 1024-dimensional feature space is reduced to a 512-dimensional feature space again, and the feature reconstruction is finally completed. After this two-stage asymmetric codec transformation, reconstructed features can fully focus on the latent expression feature space while avoiding excessive parameters and training costs, making the model easier to train. The finally obtained reconstructed latent features are activated using the Mish [24] activation function, which has good smoothness and allows better information to go deep into the neural network for obtaining better accuracy and generalization. Finally, the softmax function is utilized to obtain the final reconstruction feature weight. The specific process is as follows:

$$f_1 = W_2^T ReLU(BN(W_1^T x_i + b_1)) + b_2 \tag{7}$$

$$f_2 = W_3^T ReLU(BN(f_1)) + b_3 \tag{8}$$

$$f_3 = Mish(BN(f_2)) \tag{9}$$

$$\alpha_i = Softmax(f_3) \tag{10}$$

where  $x_i$  denotes the network features of the last layer after the backbone network,  $W_i$  and  $b_i$  represent the weight and bias of the  $i$ th linear network, where the bias is initialized to 0. The feature weight  $\alpha_i$  obtained through the network gains more critical and adaptable depth latent feature information with a lower dimension, which lays a good foundation for further analysis of facial expressions.

### 3.3 Effective attention feature reconstruction loss (EAFR Loss)

#### 3.3.1 Focal smoothing loss

Since the Cross entropy loss cannot solve the problem of class imbalance and the optimization ability is insufficient, the model’s performance cannot be effectively improved. Inspired by the idea of FocalLoss [20], our proposed Focal Smoothing loss (FS loss) model first effectively improves the problem of low accuracy of minority classification in multi-classification. By increasing the weight of minority samples, the loss function treats each category more fairly, prevents overlearning of the majority samples and ignores the minority samples, and lastly improves the overall performance of the model; At the same time, in order to reduce overfitting and enhance generalization, we embed a label smoothing regularization technique [26] for FS loss to soften the label and improve the classification performance, which will prevent paranoid learning and prediction of facial expressions due to overconfidence. Because many large-scale datasets, including the expression datasets we use, usually contain incorrectly labeled data, which means that our network model should be skeptical of the “correct answer” in essence. Adding label smoothing can reduce the modeling in extreme cases around wrong answers to a certain extent, perform network correction and enhance robustness, and produce more accurate predictions on invisible production data. Therefore, FS loss not only effectively improves the class imbalance problem but also enhances the generalization, and the overall recognition performance is further improved. The designed FS loss can be defined as follows:

$$y_s = p_{logits} * [(1 - s) * y + s * U] \tag{11}$$

$$\mathcal{L}_{FS} = -(1 - y_s)^\gamma \log y_s \tag{12}$$

where  $\gamma$  is a factor that adjusts the weight according to the probability of different samples and is set to 2.  $y_s$  is the sample probability output after label smoothing, where  $p_{logits}$  represents the output probability of the last layer after softmax,  $s$  is the smoothing factor with a value of 0.1,  $U$  is the introduced fixed distribution matrix, which is a matrix of all 1 s.

#### 3.3.2 Aggregation—separation loss

Inspired by the Center loss function [36], to make full use of potential deep information, enhance the expressive ability of features, and further strengthen the aggregation of intra-class expressions and the separation of inter-class expressions, we designed the Aggregation-Separation loss (AS loss) model. Firstly, the adaptive reconstructed feature

weights  $\alpha_i$  obtained after MCA and AAE are combined with the feature  $x_{ij}$  got after pooling of the backbone network and to be re-weighted as the input,  $x_i = [x_{i1}, x_{i2}, \dots, x_{ij}]^T \in R^D$ . Then using this input to combine with the class center for evaluation. In this way, we can use the features with stronger correlation and discrimination to construct our loss function to make the intra-class features more cohesive and is shown as follows:

$$\mathcal{L}_A = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^D \|\alpha_{ij} \circ x_{ij} - c_j\|_2^2 \tag{13}$$

where  $\alpha_{ij}$  denotes the weight after adaptive feature reconstruction,  $\circ$  represents the element dot product,  $c_j$  represents the feature center of the  $j$ th feature space,  $j \in R^D$ ,  $N$  is the number of images in each mini-batch, and  $\|\cdot\|_2$  represents  $L2$  norm.

To further increase the distance of inter-class and promote the separation of different types of expressions, we introduce the cosine distance for loss fusion, and the final Aggregation-Separation loss function  $\mathcal{L}_{AS}$  is as follows:

$$\mathcal{L}_{AS} = \mathcal{L}_A + \lambda_1 \sum_{c_j \in M} \sum_{\substack{c_k \in M \\ c_j \neq c_k}} \left( 1 + \frac{c_j \cdot c_k}{\|c_j\|_2 \|c_k\|_2} \right) \tag{14}$$

where  $\lambda_1$  represents the control factor of the contribution of the cosine distance loss to the  $\mathcal{L}_{AS}$  loss function,  $M$  represents the set of expression tags, and  $c_j$  and  $c_k$  denote the  $k$ th and  $j$ th expression category centers with  $L2$  norm, respectively. The former loss function can more accurately constrain intra-class differences, while the latter penalizes the expression similarity between inter-classes by increasing the distance between classes.

Our  $\mathcal{L}_{AS}$  also performs parameter optimization updates based on the SGD optimizer, and the class center of the  $j$ th class is updated according to the following formula:

$$\begin{aligned} \Delta c_j = & \frac{\sum_{i=1}^N (c_j - \alpha_i \circ x_i) \delta(y_i, j)}{\sum_{i=1}^N \delta(y_i, j) + 1} \\ & + \frac{\lambda_1}{|M| - 1} \sum_{\substack{c_k \in M \\ c_j \neq c_k}} \frac{c_k}{\|c_k\|_2 \|c_j\|_2} - c_j \left( \frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2^3} \right) \end{aligned} \tag{15}$$

where  $\delta(y_i, j) = \begin{cases} 1, & y_i = j \\ 0, & y_i \neq j \end{cases}$ ,  $|M|$  denotes the number of all expressions. For each mini-batch,  $c_j$  is updated with the learning rate  $\rho$ , which is set to 0.6, then:

$$c_j^t = c_j^{t-1} - \rho \Delta c_j^{t-1} \tag{16}$$

### 3.3.3 The joint loss

Our EAFR loss combines AS loss and FS loss as a total joint loss model and is trained in an end-to-end manner for model optimization and final evaluation as follows:

$$\mathcal{L} = \mathcal{L}_{FS} + \lambda \mathcal{L}_{AS} \quad (17)$$

where  $\mathcal{L}_{FS}$  and  $\mathcal{L}_{AS}$  denote Focal Smoothing loss and Aggregation and Separation loss respectively, and the hyperparameter  $\lambda$  represents the contribution of  $\mathcal{L}_{AS}$  loss, which is used to adjust the contribution degree of Aggregation-Separation loss in the total loss function.

## 4 Experiments

### 4.1 Datasets

We conduct experimental evaluations on the following three wild FER datasets, including RAF-DB, FERPlus, and AffectNet. These datasets are currently widely used wild expression datasets that fully reflect real scene emotions.

RAF-DB [17] is a dataset containing 29,672 facial expression images in the real world from the Internet annotated with basic or compound emotion datasets. We perform experiments on the basic dataset containing seven basic expressions (anger, disgust, fear, happiness, neutral, sadness, surprise), including 12,271 training samples and 3,068 test samples.

FERPlus [3] is an expression dataset in real scenes obtained by relabeling on the FER2013 [8] dataset. It contains 28,709 training images, 3,589 verification images, and 3,589 test images with a size of  $48 \times 48$  pixels, which have been relabeled as ten classes of extremely unbalanced expressions. For a more comprehensive assessment, we add contempt to the basic seven expressions (neutral, happiness, surprise, sadness, anger, disgust, fear) and conduct accuracy measurement on the test set containing eight basic expressions.

AffectNet [25] is the largest facial expression dataset to date, including 450 K labeled facial expression images with manual annotations. The dataset is very challenging that includes people of different races, background changes, lighting, posture, occlusion, etc., and the class data is very uneven. For a more comprehensive model evaluation, we choose eight basic facial expressions as FERPlus to evaluate the accuracy, including 287,651 images as training sets with imbalanced classes and 500 images as verification sets in each class with 4 K expressions in total.

### 4.2 Implementation details

Our EAFR loss model uses ResNet18 as a lightweight backbone network, which is pre-trained based on MS-Celeb-1 M [9] face dataset. To prevent overfitting and enhance generalization, we augment the input images on the fly before training the model, including resizing the original image to  $236 \times 236$ , crop to  $224 \times 224$  images by random five crops, random horizontal flip, random erasing, and final normalization. In addition, we crop  $236 \times 236$  images to  $224 \times 224$  by center crop and normalize it as input before the test. In the model training process, the SGD with momentum optimizer is used for optimization, where the momentum parameter is 0.9, the batch size is 64, and the weight decay is set to  $5e-4$ . We take model training on the RAF-DB and FERPlus datasets for 80 epochs with the initial learning rate of 0.04 decayed ten times every 20 epochs. For the AffectNet, we set the model training for 30 epochs with the initial learning rate of 0.035 decayed five times every 5 epochs. The hyperparameters  $\lambda$  and  $\lambda_1$  are both set to 0.01, and  $\rho$  is 0.6. The total parameter of our EAFR loss model is 15.09 M, and GFLOPs is 1.33. The training time of the model on the RAF-DB, FERPlus, and AffectNet datasets are 30.83, 52.11, and 191.47 min, respectively, and the test time is 2.94 ms per image. All the experiments in this article are programmed and implemented by using Python under the PyTorch framework on the NVIDIA RTX 2080Ti GPU with 11 GB RAM.

### 4.3 Ablation studies

To demonstrate the effectiveness of our proposed method in real scenes, we conducted ablation studies on three wild datasets to evaluate the influence of critical modules and parameters for our model.

1) *The influence of MCA and  $\mathcal{L}_{AS}$  on the model:* To verify the effect of each module, we perform ablation studies on the MCA and  $\mathcal{L}_{AS}$  modules in the EAFR loss model under the premise that the FS loss is used as the basic loss. As we can see from Table 1, when only using  $\mathcal{L}_{AS}$  with the combination of the basic loss function, the module improves by 0.59%, 0.67%, and 0.55% in the three datasets. We think that the  $\mathcal{L}_{AS}$  module plays a crucial role in promoting intra-class cohesion and increasing inter-class distance. When only the MAC attention module is added, the accuracy rates increase by 0.46%, 0.54%, and 0.32%, respectively, which indicates that the progressive stage lightweight attention can further focus the expression region and pay more attention to the area of interest. When we add both modules to form a complete algorithm model, we can see that the model achieves better performance and increased by 0.95%, 1.02%, and 0.97%, respectively.

Therefore, we believe that after MCA attention and two-stage feature reconstruction of lightweight AAE, our model can further reconstruct the more discriminative feature distribution in the area of interest and finally achieve the best results.

2) *The influence of hyperparameters  $\lambda$ ,  $\lambda_1$  on the network:* As shown in Table 2, we evaluate the performance of the hyperparameter in the final recognition. We first fix  $\lambda_1 = 0.01$  and set  $\lambda$  from 0.005 to 0.03. The results show that when  $\lambda = 0.01$ , our method achieves top performance in all three datasets and shows a downward trend as the parameter value increases. Then we fix  $\lambda = 0.01$ , the value of  $\lambda_1$  is from 0.001 to 1, the model obtains the best performance in the datasets when  $\lambda_1 = 0.01$ , and the accuracy decreases further as the value increases. Therefore, we finally set the value of  $\lambda = 0.01$ , and  $\lambda_1$  is also 0.01.

## 4.4 Visualization

### 4.4.1 Visualization of attention

To demonstrate the superiority of this MCA attention method that focuses on both channel direction and location information, we use GradCAM on some wild expressions to visualize the features after the fourth layer of ResNet18 residual network, as shown in Fig. 4. It is easy to see that the feature maps without MCA attention only focus on few areas of interest, and the feature energy distribution is not concentrated. The feature maps with progressive stage MCA attention can more accurately focus on areas of interest, such as eyes, nose, and mouth. Therefore, the features of the area of interest will be more discriminative when attention is added.

### 4.4.2 Visualization of loss function

To verify the effectiveness of our proposed EAFR loss method, we use t-SNE [32] to present the two-dimensional feature distribution on baseline (Cross entropy loss), Center loss, FS loss, and EAFR loss, respectively.

**Table 1** Ablation studies for MCA and  $\mathcal{L}_{AS}$  modules of our EAFR loss on RAF-DB, FERPlus, and AffectNet datasets. The overall accuracy (%) is used for experimental evaluation

MCA	$\mathcal{L}_{AS}$	RAF-DB	FERPlus	AffectNet
×	×	88.85	88.55	60.08
×	✓	89.44	89.22	60.63
✓	×	89.31	89.09	60.40
✓	✓	89.80	89.57	61.05

**Table 2** Ablation studies for the different values of  $\lambda$  and  $\lambda_1$  ( $\lambda$  represents the impact factor of the  $\mathcal{L}_{AS}$  loss module, and  $\lambda_1$  represents the impact factor of class separation cosine loss), the overall accuracy (%) is used for experimental evaluation

$\lambda$	RAF-DB	FERPlus	AffectNet
0.005	88.85	89.06	60.38
0.01	89.80	89.57	61.05
0.015	89.34	89.12	60.65
0.03	88.75	88.74	60.50
$\lambda_1$	RAF-DB	FERPlus	AffectNet
0.001	89.34	89.18	60.00
0.01	89.80	89.57	61.05
0.1	89.60	88.96	60.65
1	89.44	88.77	60.45

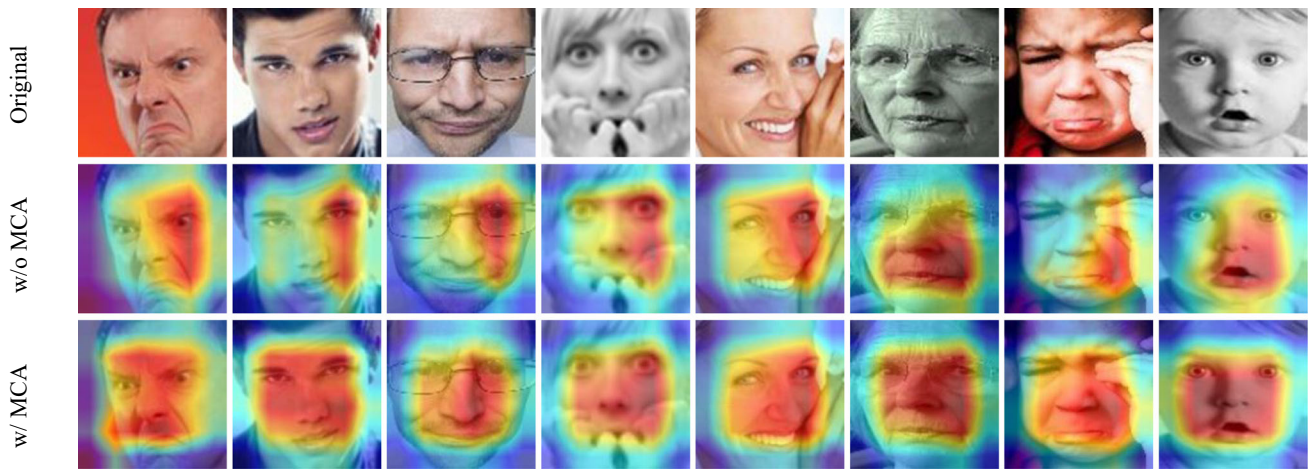
As shown in Fig. 5, we can clearly observe that the model based on Cross entropy loss cannot distinguish expressions accurately. The model based on Center loss has been improved, especially for the further condensation of intra-class expressions, but the adjustment of inter-class distance is not obvious. The FS loss further improves the classification performance for the problem of class imbalance and some wrong labels. Based on FS loss, our EAFR loss not only improves the imbalance between classes and increases generalization but also has a better constraint ability in high-dimensional space to construct more reasonable intra-class and inter-class distribution. The most impressive performance is that our EAFR loss model is clearer for different class boundaries and includes many blanks inside.

## 4.5 Comparison with state-of-the-art methods

Table 3 shows the comparison results of our method with several SOTA methods on facial expression datasets in the wild. We use ResNet18 with a pre-trained model as the backbone, and the Cross entropy loss (CE loss) is used as the baseline and Center loss as the basic comparison model. Our FS loss and EAFR loss function models are further compared with the above models to illustrate the effectiveness of the model we finally proposed.

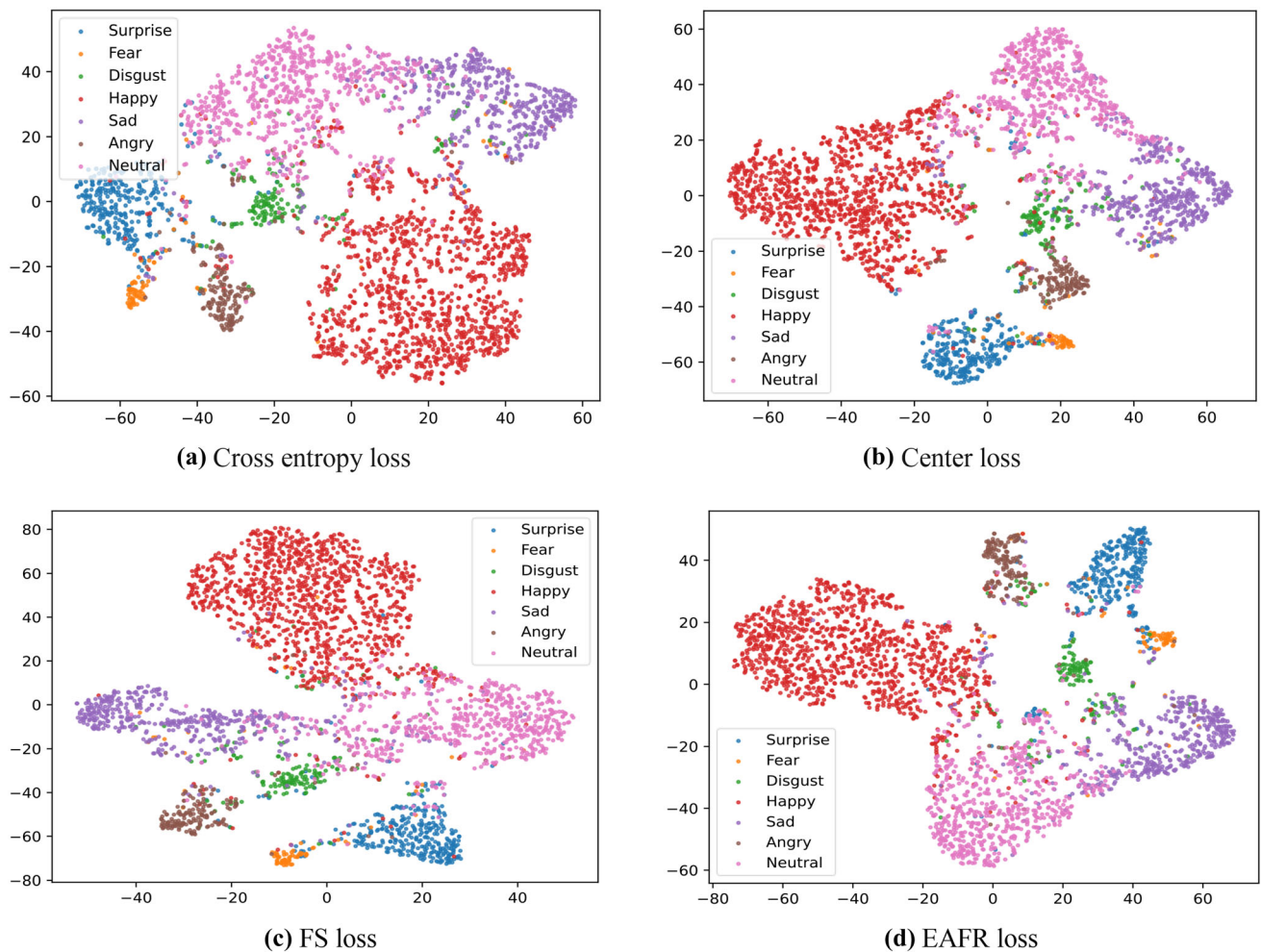
Among the contrast methods, RAN and gACNN respectively propose an attention network that enhances the robustness of regional occlusion. SCN and IPA2LT are aiming to solve the problem of label noise. PLD is to construct a more realistic label distribution. PSR handles the input problem of images with different resolutions through a pyramid structure. KTN enhances learning ability by building a progressive teacher knowledge





**Fig. 4** Visualization of feature maps produced by models with MCA methods after the last layer of backbone by using GradCAM [28] tool. Both feature maps without MCA and with MCA attention block are

visualized. It is obvious that our MCA can more precisely locate the expression area of interest



**Fig. 5** Visualization of features distribution based on t-SNE from RAF-DB dataset using four different losses

network. DDA further solves the problem of class learning accuracy caused by the imbalance of inter-class from the

perspective of the loss function. EfficientFace uses a local-feature extractor and a channel-spatial modulator for

feature extraction and uses label distribution learning as a training strategy for more effective and robust recognition.

**Table 3** Accuracy comparison with state-of-the-art methods on different wild FER datasets

Method	RAF-DB (%)
gACNN [17]	85.07
IPA2LT [21]	86.77
Separate loss [31]	86.38
RAN [18]	86.90
SCN [23]	87.03
DDA [32]	86.90
PSR [25]	88.98
KTN [34]	88.07
EfficientFace [28]	88.36
Baseline	88.17
Center loss	88.59
FS loss (ours)	<b>89.31</b>
EAFR loss (ours)	<b>89.80</b>
Method	FERPlus (%)
PLD [5]	85.35
ResNet + VGG [43]	87.40
SENet [19]	88.80
CNN + BOVW [22]	87.76
RAN [18]	88.55
ESR-9 [24]	87.25
SCN [23]	88.01
RW loss [33]	87.60
ADC-Net [27]	88.90
Baseline	87.95
Center loss	88.39
FS loss (ours)	<b>89.09</b>
EAFR loss (ours)	<b>89.57</b>
Method	AffectNet (%)
DownSampling [6]	50.00
IPA2LT [21]	57.31
PG-CNN [20]	55.33
gACNN [17]	58.78
RAN [18]	59.50
ESR-9 [24]	59.30
SCN [23]	60.23
PSR [25]	60.68
EfficientFace [28]	59.89
Baseline	59.63
Center loss	59.95
FS loss (ours)	<b>60.40</b>
EAFR loss (ours)	<b>61.05</b>

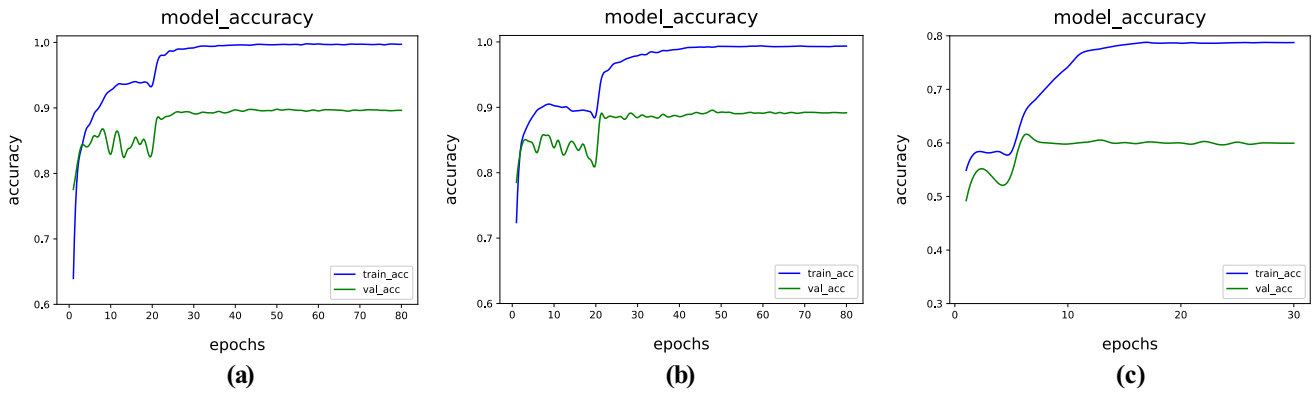
ADC-Net is constructed by reconstructing the semantic association between the subregions in the module. ESR-9 uses feature integration methods to reduce redundancy and computational load and enhance generalization capabilities. Separate loss and RW loss establish more reasonable intra-class and inter-class distances for identification. PG-CNN and SENet mainly increase the focus on the area of interest by re-weighting to improve network capabilities. ResNet + VGG and CNN + BOVW mainly increase the feature extraction performance of the network through the design of the network structure. The above methods have improved the performance of facial expression recognition from different perspectives to a certain extent. However, they do not pay too much attention to the variation and similarity of real scene expressions, and the effect of improving expression recognition rate from a certain aspect is limited. Our model builds a lightweight and effective network model from the comprehensive perspective of focusing on the area of interest, improving the problem of class imbalance, enhancing the generalization performance of tags, and aggregating the intra-class distance and separating the inter-class distance more reasonably, and finally reaches the level of SOTA.

The performance of our EAFR loss method in the RAF-DB, FERPlus, and AffectNet expression datasets outperforms other current SOTA methods and achieves excellent levels of 89.80%, 89.57%, and 61.05%, respectively. The above experimental results finally prove the effectiveness of our proposed method.

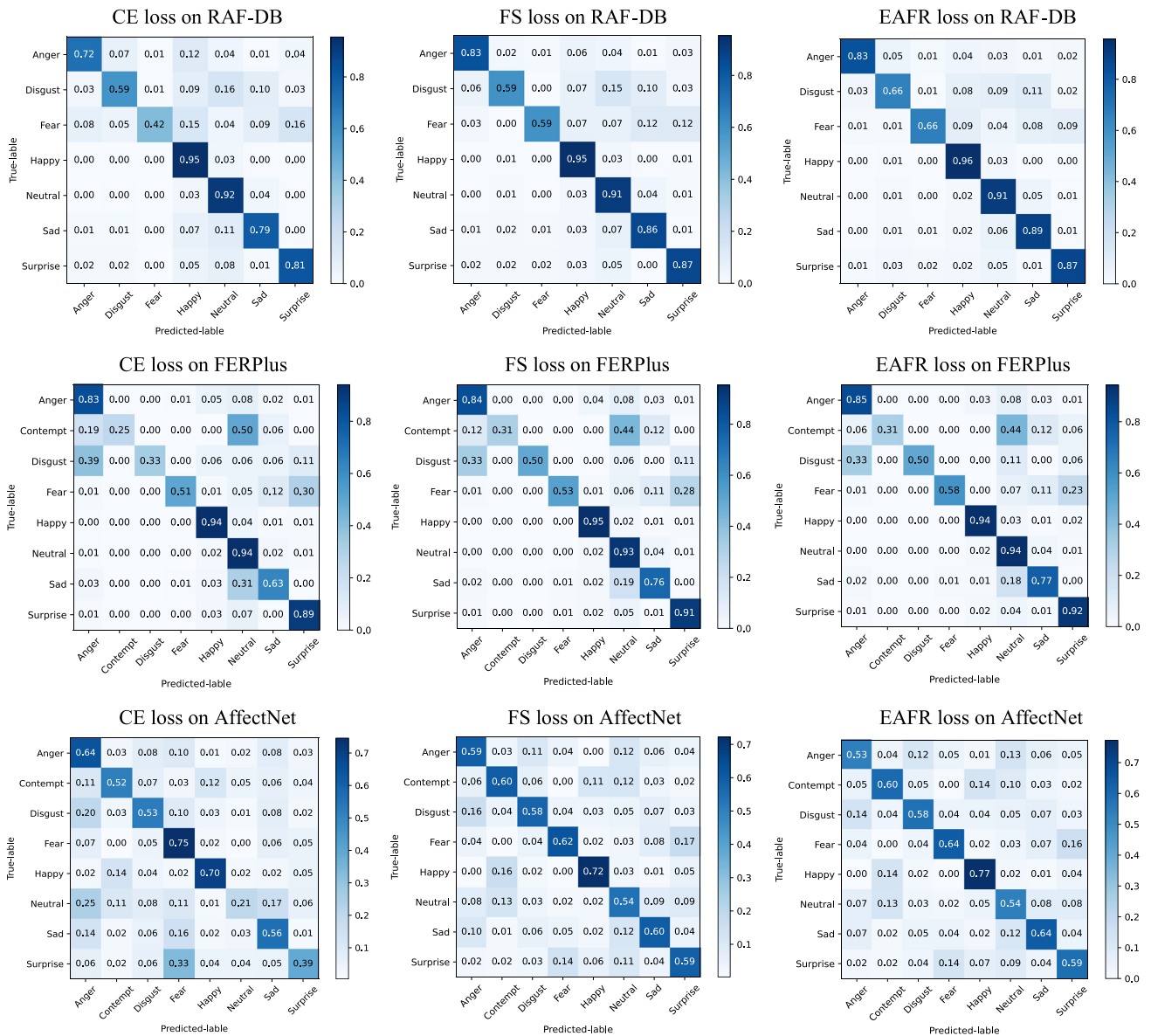
## 4.6 Discussion

We first show the training and validation accuracy curves of the three datasets of RAF-DB, FERPlus, and AffectNet, as shown in Fig. 6. It can be seen that our model finally reaches a stable state and achieves a competitive accuracy after relatively short learning epochs. Then we evaluate the confusion matrix of the CE loss, FS loss, and EAFR loss models in the above three different expression datasets. It can be seen from Fig. 7 that the model based on CE loss has a great deviation in class learning and recognition performance is not ideal. After adopting our optimization loss of FS loss, the comprehensive performance of the model has been improved to a certain extent. In particular, the learning ability of unbalanced category recognition and generalization is improved obviously. And finally, our EAFR loss model further enhances the intra-class expression cohesion and reasonably expands the inter-class distance, making the model have a better performance in the uncontrolled expressions.

In RAF-DB, since the distinction between fear and surprise, neutral and disgust expressions is not obvious, and fear, surprise, and disgust belong to the minority classes,



**Fig. 6** Training and validation accuracy curves of RAF-DB (a), FERPlus (b), AffectNet (c) datasets



**Fig. 7** Confusion matrix of CE loss (left), FS loss (middle), and EA FR loss models (right) on RAF-DB, FERPlus, and AffectNet expression datasets

we can clearly see the large difference in class recognition performance and poor recognition rate from the confusion matrix of the original CE loss. After applying the proposed FS loss, there have been significant improvements in class imbalance and paranoid learning. When the final EAFR loss is used, the model improves the recognition performance by further promoting intra-class condensation and expanding inter-class distance based on FS loss, which can better deal with subtle expression changes and show better comprehensive recognition performance.

In the confusion matrix of FERPlus, we can see that CE loss is still not very good at handling unbalanced classes, such as minority classes like contempt and disgust, and the performance of class recognition varies greatly. FS loss improves the problems of CE loss and enhances generalization. EAFR loss is more reasonable based on FS loss because it can better handle relationships of intra-class and inter-class, so the overall performance is significantly improved compared with the previous two methods.

For AffectNet, in addition to the class imbalance, there are also different classes of expressions with small amplitude and high similarity, such as neutral and anger, surprise and fear; Simultaneously, the dataset has the problem of low annotation accuracy. We observe that CE loss has insufficient processing capacity to deal with the above problems, FS loss improves the problem of category imbalance and paranoid learning caused by label errors, EAFR loss further optimizes the spatial distribution of features by adjusting the distance of intra-class and inter-class to maintain or improve the recognition rate of classes.

## 5 Conclusion

This paper proposes an Effective Attention Feature Reconstruction loss (EAFR loss) model for FER in real scene. First, the progressive stage attention MCA fully focuses on the area of interest. Then the FS loss effectively improves the problem of class imbalance, while the embedded smoothing mechanism further prevents paranoid learning and enhances generalization. At the same time, the AS loss model designed by lightweight AAE is used to reconstruct the more discriminant feature distribution space, strengthening the aggregate of intra-class distances and the separation of inter-class distances. Finally, we build the joint EAFR loss based on the above loss models, which effectively improves the ultimate expression recognition effect. The experiments are validated on three large-scale wild expression datasets and results indicate that our proposed model outperforms the baseline and Center loss models and shows superior performance on several state-of-the-art methods. However, our current research mainly focuses on recognizing discrete expressions. To better

reflect people's real feelings, we will pay more attention to recognizing continuous expressions in the following research.

**Acknowledgments** This research is supported by the National Science Foundation of China under Grant 61966035 and U1803261, by the Autonomous Region Science and Technology Department International Cooperation Project under Grant 2020E01023, by Tianshan Innovation Team Plan Project of Xinjiang Uygur Autonomous Region under Grant 202101642, and by the Funds for Creative Research Groups of Higher Education of Xinjiang Uygur Autonomous Region under Grant XJEDU 2017T002.

## Declarations

**Conflict of interest** We wish to submit a new manuscript entitled "Effective Attention Feature Reconstruction Loss for Facial Expression Recognition in the Wild" for consideration in the Neural Computing and Applications Journal. We declare that this work is original and there is no conflict of interest.

## References

- Acharya D, Huang Z, Pani Paudel D, Van Gool L (2018) Covariance pooling for facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 367–374. IEEE
- Albanie S, Nagrani A, Vedaldi A, Zisserman A (2018) Emotion recognition in speech using cross-modal transfer in the wild. In: Proceedings of the 26th ACM international conference on Multimedia, pp 292–301
- Barsoum E, Zhang C, Ferrer CC, Zhang Z (2016) Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI), pp 279–283
- Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y (2018) Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp 302–309. IEEE
- Fan X, Deng Z, Wang K, Peng X, Qiao Y (2020) Learning discriminative representation for facial expression recognition from uncertainties. In: 2020 IEEE International Conference on Image Processing (ICIP), pp 903–907. IEEE
- Farzaneh AH, Qi X (2020) Discriminant Distribution-Agnostic Loss for Facial Expression Recognition in the Wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 406–407. IEEE
- Georgescu MI, Ionescu RT, Popescu M (2019) Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access* 7:64827–64836
- Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Bengio Y (2013) Challenges in representation learning: A report on three machine learning contests. In: International Conference on Neural Information Processing, pp 117–124. Springer
- Guo Y, Zhang L, Hu Y, He X, Gao J (2016) Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Proceedings of the European conference on computer vision (ECCV), pp 87–102. Springer
- Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13713–13722. IEEE

11. Huang C (2017) Combining convolutional neural networks for emotion recognition. In: 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), pp 1–4. IEEE
12. Karnati M, Seal A, Krejcar O, Yazidi A (2020) Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. *IEEE Trans Instrum Meas* 70:1–12
13. Karnati M, Seal A, Krejcar O, Yazidi A (2021) FER-net: facial expression recognition using deep neural net. *Neural Comput Appl* 33:9125–9136
14. Karnati M, Seal A, Yazidi A, Krejcar O (2021) LieNet: a deep convolution neural networks framework for detecting deception. *IEEE Trans Cogn Develop Syst.* 126(5): 550–569
15. Li Y, Zeng J, Shan S, Chen X (2018) Patch-gated CNN for occlusion-aware facial expression recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp 2209–2214. IEEE
16. Li Y, Lu Y, Li J, Lu G (2019) Separate loss for basic and compound facial expression recognition in the wild. In: Asian Conference on Machine Learning (ACML), pp 897–911
17. Li S, Deng W, Du JP (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition (CVPR), pp 2852–2861. IEEE
18. Li Y, Zeng J, Shan S, Chen X (2018) Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans Image Process* 28(5):2439–2450
19. Li H, Wang N, Ding X, Yang X, Gao X (2021) Adaptively learning facial expression representation via CF labels and distillation. *IEEE Trans Image Process* 30:2016–2028
20. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 2980–2988. IEEE
21. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) SpheroFace: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 212–220. IEEE
22. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference On Computer Vision And Pattern Recognition-Workshops (CVPRW), pp 94–101. IEEE
23. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Third IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp 200–205. IEEE
24. Misra D (2019) Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*
25. Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans Affect Comput* 10(1):18–31
26. Müller R, Kornblith S, Hinton G (2019) When does label smoothing help? In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS), pp 4694–4703
27. Ou J, Bai XB, Pei Y, Ma L, Liu W (2010) Automatic facial expression recognition using Gabor filter and expression analysis. In: 2010 Second International Conference on Computer Modeling and Simulation (ICCMS), pp 215–218. IEEE
28. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 618–626. IEEE
29. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis Comput* 27(6):803–816
30. Shao J, Qian Y (2019) Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing* 355:82–92
31. Siqueira H, Magg S, Wermter S (2020) Efficient facial feature learning with wide ensemble-based convolutional neural networks. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 5800–5809
32. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11):2579–2605
33. Vo TH, Lee GS, Yang HJ, Kim SH (2020) Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access* 8:131988–132001
34. Wang K, Peng X, Yang J, Lu S, Qiao Y (2020) Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6897–6906. IEEE
35. Wang K, Peng X, Yang J, Meng D, Qiao Y (2020) Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans Image Process* 29:4057–4069
36. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision (ECCV), pp 499–515. Springer
37. Xia HY, Li C, Tan Y, Li L, Song S (2021) Destruction and reconstruction learning for facial expression recognition. *IEEE Multimedia* 28(2):20–28
38. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction (ICMI), pp 435–442
39. Zeng J, Shan S, Chen X (2018) Facial expression recognition with inconsistently annotated datasets. In: Proceedings of the European conference on computer vision (ECCV), pp 222–237
40. Zhang Z, Luo P, Loy CC, Tang X (2018) From facial expression recognition to interpersonal relation prediction. *Int J Comput Vis* 126(5):550–569
41. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans Pattern Anal Mach Intell* 29(6):915–928
42. Zhao Z, Liu Q, Zhou F (2021) Robust lightweight facial expression recognition network with label distribution training. In: Proceedings of the AAAI Conference on Artificial Intelligence. 35(4), pp 3510–3519
43. Zhao G, Huang X, Taini M, Li SZ, Pietikäinen M (2011) Facial expression recognition from near-infrared videos. *Image Vis Comput* 29(9):607–619

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.