Neural Processing Letters (2024) 56:96
https://doi.org/10.1007/s11063-024-11566-4

Harmonious Mutual Learning for Facial Emotion Recognition

Yanling Gan[1] · Luhui Xu[1] · Haiying Xia[2] · Gan Liu[2]

Accepted: 11 February 2024 / Published online: 7 March 2024

Abstract
Facial emotion recognition in the wild is an important task in computer vision, but it still remains challenging since the influence of backgrounds, occlusions and illumination variations in facial images, as well as the ambiguity of expressions. This paper proposes a harmonious mutual learning framework for emotion recognition, mainly through utilizing attention mechanisms and probability distributions without utilizing additional information. Specifically, this paper builds an architecture with two emotion recognition networks and makes progressive cooperation and interaction between them. We first integrate self-mutual attention module into the backbone to learn discriminative features against the influence from emotion-irrelevant facial information. In this process, we deploy spatial attention module and convolutional block attention module for the two networks respectively, guiding to enhanced and supplementary learning of attention. Further, in the classification head, we propose to learn the latent ground-truth emotion probability distributions using softmax function with temperature to characterize the expression ambiguity. On this basis, a probability distribution distillation learning module is constructed to perform class semantic interaction using bi-directional KL loss, allowing mutual calibration for the two networks. Experimental results on three public datasets show the superiority of the proposed method compared to state-of-the-art ones.

Keywords Facial emotion recognition · Mutual learning · Attention mechanisms · Probability distributions

✉ Luhui Xu
xlh@gxnu.edu.cn

Yanling Gan
gyl@gxnu.edu.cn

Haiying Xia
xhy22@mailbox.gxnu.edu.cn

Gan Liu
3037453964@qq.com

[1] School of Computer Science and Engineering, Guangxi Normal University, 15 Yucai Road, Qixing District, Guilin 541004, Guangxi, China

[2] School of Electronic and Information Engineering, Guangxi Normal University, 15 Yucai Road, Qixing District, Guilin 541004, Guangxi, China

1 Introduction

Facial emotion is an important clue of human behavior and intention. Its recognition can be applied in many scenarios, such as security monitoring [1], behavior understanding [2, 3], medical diagnosis [4], etc. However, facial emotion recognition in the wild through facial images is still a challenge, due to the presence of irrelevant information such as background, occlusion, illumination variation, and the expression ambiguity stemming from intra-class diversity and inter-class similarity. To address these challenges, many works have attempted to utilize attention mechanisms in deep models as an effective way to emphasize interesting features while suppressing irrelevant features [5–8]. At the same time, many studies have explored probability distributions in emotion recognition tasks [9, 10]. Probability distribution is the soft label that describes the confidence levels for each category, which can better characterize intra-class diversities and inter-class similarities for different facial expression images. Hence, it allows deep models to learn the ambiguity of expressions, which can better guide them to learn the diversities of the same expressions and the similarities between the different expressions.

Recently, mutual learning has been introduced in deep models. The core idea of mutual learning is the imitation learning between deep models at the levels of features or outputs, and which can usually be achieved by using attention mechanisms and probability distributions. However, there are few methods in expression recognition field that focus on exploring mutual learning. Besides, most of the existing mutual learning methods in other fields usually constrain the models for consistent learning, while generally neglecting to maintain the self-learning ability of the models, which may ruin the dynamics of mutual learning [11, 12]. Therefore, we propose harmonious mutual learning for emotion recognition. We construct a framework consisting of two parallel networks and perform progressive mutual teaching in feature layers and output layers. Specifically, we design a self-mutual attention learning (SMAL) module in the backbone architecture to transfer feature information. We deploy different attention components for the two networks to increase the dynamic of mutual learning. Then, we design a probability distribution distillation learning (PDDL) module in the classification head for the purpose of bidirectional class semantic interaction and mutual calibration, resulting in the performance improvement of emotion recognition. In summary, the main contributions of our work are summarized as follows:

- We propose a novel harmonious progressive mutual learning framework containing two parallel networks, which jointly utilize attention mechanisms and probability distributions.
- We construct a self-mutual attention module by using distinct attention components for two networks, which facilitates mutual learning of enhanced and supplementary features while preserving self-learning capabilities.
- In the classification head, we introduce bidirectional probability distribution distillation learning through KL loss, with an objective of mutual learning of class ambiguity and the calibration of the two networks.
- We demonstrate the effectiveness of our framework on three publicly available datasets, and our method reaches to state-of-the-art performance.

2 Related Works

In this section, we mainly review the most relevant works about mutual learning, including feature-based approaches and probability distribution-based approaches.

2.1 Feature-Based Approaches

Mutual learning has been received extensive research, where attention mechanisms have been widely applied. For examples, Ma et al. [13] fused global and local representations, and applied softmax function for attention learning on the fusion features to achieve implicit mutual teaching between local semantics and global long-term dependencies. Zhang et al. [14] employed L2 loss to quantify the disparity between cross-modality attention features across RGB modality images, IR modality images and the mixed-modality images. Liu et al. [15] viewed the shallow to deep layers of CNN as "experts" with different perspectives. The authors constructed attention images for different experts and input to other experts to achieve cross-layer mutual learning. However, there is a lack of methods that focus on mutual learning in the field of facial expression recognition. Some metric learning methods [16–20], which typically utilize feature metric functions to construct imitation losses, and can be considered as forms of mutual learning. However, these methods often use additional information, such as facial landmarks, action unit (AU), head pose and so on. Nevertheless, the accurate acquisition of this information itself is a significant challenge, and its improper utilization may detrimentally impact the effectiveness of emotion recognition.

2.2 Probability Distribution-Based Approaches

Probability distribution, which refers to the class posterior probability output by deep model, describes the class confidence levels for the input. By aligning probability distributions, mutual learning can be performed for the mutual teaching from the perspective of class semantics. Zhang et al. [21] constructed multiple peer networks that utilize the KL loss of probability distributions to force the consistency learning between different models, so as to train in a mutual learning manner. Bian et al. [11] constructed a handwritten mathematical expression recognition model consisting of a shared encoder and two parallel inverse decoders, where the two decoding branches align the output probability distributions at the time step through KL loss for mutual learning of complementary information. Xu et al. [22] utilized the probability distribution alignment method to achieve the mutual learning and adaptation of multi-source domain data. Qiao et al. [12] obtained cross-modal attention representations based on the softmax function, and then aligned the probability distributions for the mutual learning between global and local semantics. Wang et al. [23] designed a mutual learning network between overall and occluded images, which is achieved by aligning the probability distributions of the two types of images. However, the above methods primarily focus on mutual imitation or consistency learning, i.e., constraining the models to generate the same outputs, while ignoring the self-learning ability with differences. This can potentially result in a homogenization of knowledge among different models, which in turn make the mutual influence to be limited and disrupt the dynamics of mutual learning.

Inspired by existing works, this paper proposes a harmonious mutual learning method for facial emotion recognition, by combining attention mechanisms and the distillation learning of probability distributions. Unlike previous works, we design different attention components for the models in mutual learning to obtain the learning abilities of diverse features. This not only promotes mutual knowledge transfer between models, but also maintains diverse learning abilities and promotes the dynamics of mutual learning. At the same time, we design a bidirectional probability distribution alignment to facilitate the mutual transfer of class semantics, which is beneficial for handling the ambiguity of expressions and improving the recognition performance.

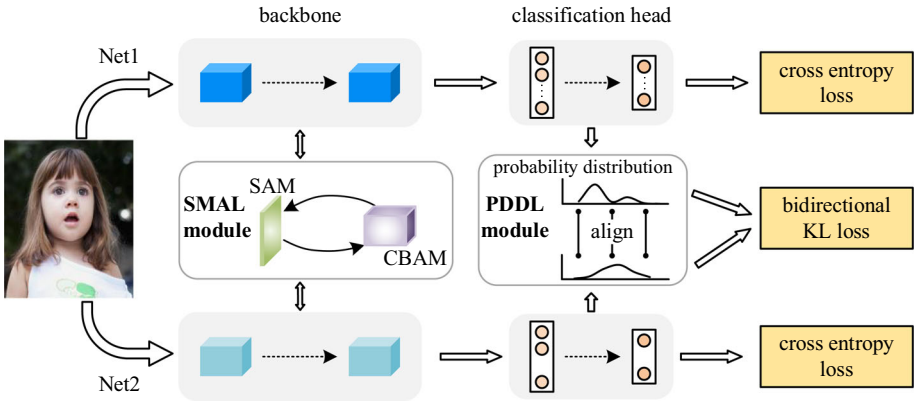


Fig. 1 The proposed framework. Our framework consists of two networks and conducts progressive mutual learning. In the backbone, SMAL module conducts enhanced and complementary interesting learning to capture more discriminative pattern, via utilizing different attention components: SAM and CBAM. In the classification head, PDDL module uses the probability distributions output by the peers to conduct class semantic interaction for correcting the learning of each other

3 Methodology

In this section, this paper proposes a new mutual learning method based on attention mechanisms and probability distributions. We capture the expression information of interest against irrelevant facial information through self-mutual attention learning. Further, by combining with probability distribution distillation learning, we can potentially calibrate the classification to increase emotion recognition performance.

3.1 The Overall Architecture

The proposed architecture contains two networks, Net1 and Net2, and two interactive modules, as shown in Fig. 1. We adopt resnet50 as the basic architecture for each network, producing a backbone with five convolutional blocks and a classification head with one fully-connected layer. The two interactive modules are self-mutual attention learning (SMAL) module and probability distribution distillation learning (PDDL) module. The former integrated in the backbone implements self-attention within networks and mutual attention between networks via embedding two submodules spatial attention module (SAM) and convolutional block attention module (CBAM) [24]. The latter implements bidirectional class semantic interaction using KL loss. More details will be described in the rest of this chapter.

3.2 Self-Mutual Attention Learning

In the mutual learning of the backbone, we aim to capture the importance of facial features while combining with enhanced and supplementary interesting learning to improve efficiency against irrelevant facial information. Therefore, the SMAL module mainly includes self-attention branches and mutual attention branches, as shown in Fig. 2. The self-attention captures saliency of the feature maps, and the mutual attention captures saliency with enhanced and complementary properties. Meanwhile, we hope to improve the diversity of the

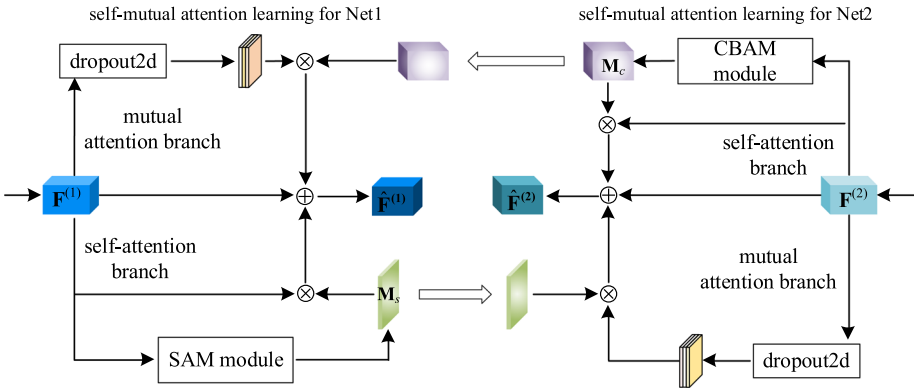


Fig. 2 The SMAL module

two networks in order to increase the dynamics in mutual learning. Therefore, we deploy two attention components for the two networks, namely SAM and CBAM modules. We integrate the SMAL module into the fourth and fifth blocks of the backbone.

SAM and CBAM are used for the self-attention learning of Net1 and Net2 respectively, and the information interaction is realized through the attention maps of SAM and CBAM. Formally, assuming that the feature map output by the convolution block of Net1 is expressed by $\mathbf{F}^{(1)} \in \mathbb{R}^{C\times H\times W}$. Here, C represents the number of channels. H and W represent the height and width of the feature maps. For Net1, self-mutual attention learning can be expressed by:

$$\hat{\mathbf{F}}^{(1)} = (1 + \mathbf{M}_s) \otimes \mathbf{F}^{(1)} + detach(\mathbf{M}_c) \otimes dropout2d(\mathbf{F}^{(1)}) \tag{1}$$

where $\mathbf{M}_s$ and $\mathbf{M}_c$ respectively indicate the self-attention weight map learned by SAM module and the mutual-attention weight map learned by CBAM module. $\otimes$ denotes element-wise multiplication. $dropout2d$ denotes randomly zeroing out some channels. $detach$ operator prevents gradient backpropagation toward to Net2. However, attention information from Net2 is not always applicable and potential. Therefore, we insert a $dropout2d$ operator for random suppressing, to provoke adaptive ability of learning from Net2. For Net2, self-mutual attention learning is implemented with the same approach, and can be expressed by:

$$\hat{\mathbf{F}}^{(2)} = (1 + \mathbf{M}_c) \otimes \mathbf{F}^{(2)} + detach(\mathbf{M}_s) \otimes dropout2d(\mathbf{F}^{(2)}) \tag{2}$$

where $\mathbf{F}^{(2)}$ is the feature map output by the convolution block of Net2.

SAM takes the feature map $\mathbf{F}^{(1)}$ as input. It first uses two parallel operators, namely average pooling and maximum pooling, to compress the channel information and generate two 2D maps: $\mathbf{F}^s_{avg} \in \mathbb{R}^{H\times W}$ and $\mathbf{F}^s_{max} \in \mathbb{R}^{H\times W}$. Then, the concat layer stacks the two maps along channel dimension. Next, a convolutional layer followed by a sigmoid activation serves as a fusion operator to product comprehensive spatial attention map $\mathbf{M}_s$. Mathematically, spatial attention can be expressed as follows:

$$\begin{aligned} \mathbf{M}_s &= \sigma(f^{3\times3}([avgpool(\mathbf{F}^{(1)}); maxpool(\mathbf{F}^{(1)})])) \\ &= \sigma(f^{3\times3}([\mathbf{F}^s_{avg}; \mathbf{F}^s_{max}])) \end{aligned} \tag{3}$$

where σ denotes the sigmoid function, $[\cdot]$ denotes concatenating the features along channel dimension, and $f^{3\times3}$ denotes a convolution operation with a filter size of 3×3.

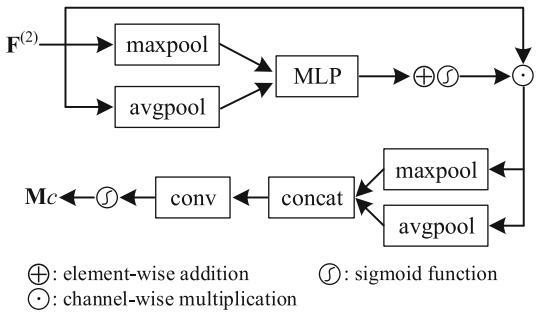


Fig. 3 The CBAM module

CBAM is the combination of channel attention and spatial attention, as shown in Fig. 3. CBAM first conducts max pooling and average pooling on $\mathbf{F}^{(2)}$ along the channel dimension respectively, obtaining channel descriptor: $\mathbf{v}_{max} \in \mathbb{R}^C$ and $\mathbf{v}_{avg} \in \mathbb{R}^C$. These pass a shared multilayer perceptron(MLP) with two fully-connected layers. At last, the channel attention map $\mathbf{v}^c \in \mathbb{R}^C$ can be obtained by sigmoid function, as shown below:

$$\begin{aligned} \mathbf{v}^c &= \sigma(MLP(avgpool(\mathbf{F}^{(2)})) + MLP(maxpool(\mathbf{F}^{(2)}))) \\ &= \sigma(W_1(W_0(\mathbf{v}_{avg})) + W_1(W_0(\mathbf{v}_{max}))) \end{aligned} \tag{4}$$

where W_1 and W_2 are the parameters of MLP. Then, channel attention features can be obtained by $\mathbf{F}^c \in \mathbb{R}^{C\times H\times W} = \mathbf{F}^{(2)} \odot \mathbf{v}^c$, where $\odot$ denotes channel-wise multiplication. $\mathbf{F}^c$ is taken as the input of spatial attention, so the attention map of CBAM module $\mathbf{M}_c$ can be computed by the following formula:

$$\begin{aligned} \mathbf{M}_c &= \sigma(f^{3\times 3}([avgpool(\mathbf{F}^c); maxpool(\mathbf{F}^c)])) \\ &= \sigma(f^{3\times 3}([\mathbf{F}^c_{avg}; \mathbf{F}^c_{max}])) \end{aligned} \tag{5}$$

In our architecture, the learnable parameters for SAM and CBAM are independent. Therefore, the two networks can learn both self-attention and conduct mutual attention, enabling interesting learning with enhanced and supplementary properties that can be understood intuitively in Sect. 4.6.

3.3 Probability Distribution Distillation Learning

Further, probability distribution distillation learning is integrated into the classification head, as shown in Fig. 4. Assume that in the forward propagation of the network, the output class score for a sample $\mathbf{x}_i$ is $[s_{i1}, s_{i2}, \ldots, s_{iK}]$, where K is the number of emotion classes. Then, in the learned probability distribution conditioned on $\mathbf{x}_i$ by the network, the probability of the sample that belongs to the j-th class d_{ij}, where we remove the superscript (1) or (2) for simplicity, can be obtained by softmax function with temperature parameter T:

$$d_{ij} = p(y_i \mid \mathbf{x}_i; \mathbf{W}) = \frac{exp(s_{ij}/T)}{\sum_{r=1}^{K} exp(s_{ir}/T)} \tag{6}$$

where $\mathbf{W}$ represents the learnable parameters. The larger the value of T, the smoother the probability distribution, and vice versa. Using formula (6), we can obtain the learned probability distributions for the two networks and denote them as $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$ respectively.

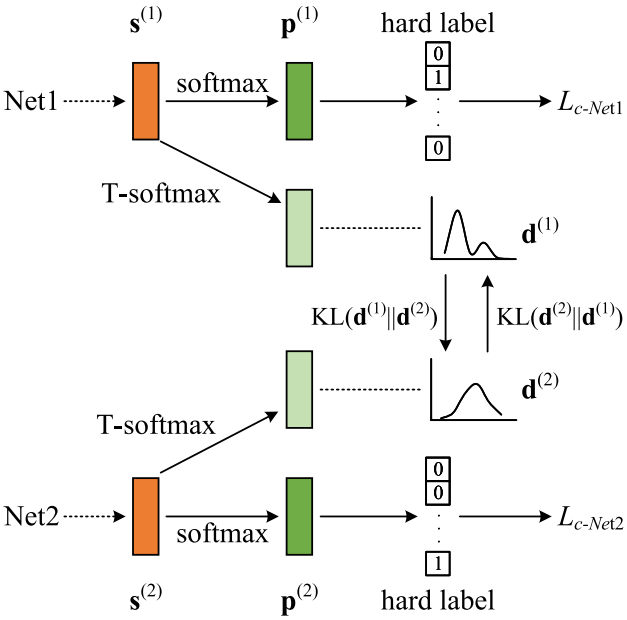


Fig. 4 Bidirectional probability distribution distillation learning. T-softmax denotes the softmax function with temperature parameter T

In this stage, the two networks take the probability distribution output by the other one as the latent ground-truth probability distribution. The optimization objective is to minimize the error between the two probability distributions. Therefore, bi-directional KL loss can be adopted to measure the error of mutual class knowledge distillation, as shown in the following:

$$\begin{aligned} L_{bi-kl} &= KL(\mathbf{d}^{(1)}\|\mathbf{d}^{(2)}) + KL(\mathbf{d}^{(2)}\|\mathbf{d}^{(1)}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(d_{ij}^{(1)}\log d_{ij}^{(1)} + d_{ij}^{(2)}\log d_{ij}^{(2)} - d_{ij}^{(1)}\log d_{ij}^{(2)} - d_{ij}^{(2)}\log d_{ij}^{(1)}) \end{aligned} \tag{7}$$

where N is the batch size. In emotion recognition system, one-hot label (hard label) is usually used for error measurement in training. Hard label can be regarded as a kind of supervision information with 100% confidence for one class, which may easily lead to over fitting especially in FER systems. Because there are great similarities between different emotions, as well as different emotional intensities in a facial image. Therefore, hard label will force the network to excessively fit one class with 100% confidence, which damages the generalization performance. By contrast, probability distribution is a class description degree vector, in which each element belongs to [0,1] and describes the class strength and the similarity between different classes. So, probability distribution can serve as a soft target and provide more transferable knowledge in the supervised learning. With the bidirectional probability distribution transfer, each of the two networks learns the latent ground-truth probability distributions and uses them as soft targets to mutually correct the supervised learning guided by the hard labels, thus mutual calibration can be achieved.

3.4 Emotion Recognition Loss

The cross entropy function is used as the loss function of emotion classification. For sample $\mathbf{x}_i$, assuming its one-hot label is $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iK}]$, the loss function for a batch can be expressed by:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij} \log p_{ij} \tag{8}$$

where p_{ij} is the output of softmax function. The value of y_{ij} is 0 or 1, which indicates whether $\mathbf{x}_i$ belongs to the j-th class:

$$y_{ij} = \begin{cases} 1, \mathbf{x}_i \text{ belongs to class } j \\ 0, \text{otherwise} \end{cases} \tag{9}$$

3.5 Multi-loss Strategy and Inference

We train the proposed architecture through a multi-loss strategy, including emotion recognition losses for Net1 and Net2, as well as the bi-directional KL loss. The final optimization objective is as follows:

$$L = L_{bi-kl} + L_{c-Net1} + L_{c-Net2} \tag{10}$$

Multi-loss strategy allows the two networks train jointly. In this way, each network performs supervised learning with hard label and accepts the soft target from the other one to achieve compromise training, thus the generalization can be boosted.

In the inference phase, we use the fusion of Net1 and Net2 as the prediction, which can be formulated as:

$$y = arg\max_{j^*} \frac{1}{2}\left(s_j^{(1)} + s_j^{(2)}\right) \tag{11}$$

where $s^{(1)}$ and $s^{(2)}$ are class scores of Net1 and Net2 respectively.

4 Experiment

In this section, we first describe three public datasets, data preprocessing methods and experimental configurations. Then, we compare with state-of-the-art methods to show superiority of the proposed method. Finally, we show the visualization and perform ablation study to further demonstrate the effectiveness of the proposed framework.

4.1 Dataset

RAF-DB [25]: It is a popular emotion recognition dataset. There are 7 emotion classes: surprise, fear, disgust, happiness, sadness, anger and neutral. It contains 15,339 real-world facial images collected from the Internet. This dataset has provided protocol for model optimization and performance evaluation, in which 12,271 images are divided as training set and 3,068 as test set, and our experiments follow this protocol.

Fig. 5 Some samples from RAF-DB, FER2013 and SFEW datasets

FER2013 [26]: It is a large-scale dataset collected by the Google search engine. It contains 7 emotion classes, and the total number of images is 35,888, where 28,709 images for training, 3,589 images for validation and 3,589 images for testing. In this dataset, image size is 48×48 pixels.

SFEW [27]: It is a dataset extracted from video. There are 1766 images in the dataset, which are divided into three parts, namely 958 for training, 436 for validation and 372 for testing. Since the labels of the test set are used for competition and remain private, we train on the training set and report the recognition accuracy on the validation set while following the previous methods being compared in our experiment analysis.

In our experiment, facial images in RAF-DB and SFEW datasets are aligned using the landmarks provided by the dataset and the detected landmarks by detector [28] respectively, and face regions are extracted and resized to size 256×256 pixels. For FER2013 dataset, the above preprocessing is elided since the provided images have been well addressed. Data augmentation is applied to extend datasets, such as random cropping and random rotation. The experimental images are shown in Fig. 5, and are resized to 224×224 pixels to fit the input size of the network.

4.2 Experimental Setting

Our framework is implemented using PyTorch and deployed on a Titan RTX GPU. The momentum, weighted decay and batch size are set to 0.9, 0.0001 and 96, respectively. We use Adam optimizer, and adopt an initial learning rate 0.0002. On the RAF-DB and FER2013 datasets, the learning rate is decayed by 0.1 every 30 epochs, and total number of training epochs is 90. The number of images in SFEW dataset is relatively small, and the training fluctuation is large. So, we apply two Adam optimizers and deploy them on Net1 and Net2 respectively. The optimizer for Net1 decays learning rate every 14 epochs, the optimizer for Net2 decays learning rate every 7 epochs, and the total training epochs are 20. In loss L_{bi-kl}, T is set as 1.2. On the RAF-DB and FER2013 datasets, the two networks are initialized by Ms-celeb-1 m [29] and ImageNet [30], respectively. On SFEW dataset, we use the Ms-

Table 1 Accuracy (%) of different methods on RAF-DB dataset

Method	Year	Accuracy
gACNN [31]	2018	85.07
Gan et al. [36]	2019	86.31
LDL-ALSG [9]	2020	85.53
RAN [32]	2020	86.90
LBAN-IL [33]	2021	85.89
Ruan et al. [37]	2021	89.47
PAT-CNN [34]	2022	88.43
Indolia et al. [8]	2023	81.06
WeiCL [35]	2023	86.96
DAN [6]	2023	89.70
Le et al. [10]	2023	90.51
Ours		90.71

celeb-1 m model and a model pre-trained on RAF-DB dataset to initialize the two networks respectively.

4.3 Results on RAF-DB Dataset

Table 1 presents FER performance of the proposed method on RAF-DB dataset. In Table 1, we also compare with the approaches utilizing attention mechanisms and probability distributions. The methods in [6, 8, 31–33] perform attention learning for facial images, an attempt to enhance emotion-related information and suppress the others. Specifically, Wen et al. [6] embedded spatial attention and channel attention among three parallel networks, and enhanced the attention through fusion operation. They achieved accuracy of 89.70% on the RAF-DB dataset. Indolia et al. [8] established a self-attention module in the convolutional block of resnet by using softmax activation to handle intra-class variation and inter-class similarity of expressions, and the accuracy is 81.06%. Li et al. [31] and Wang et al. [32] proposed a region attention network, which aims to explore the key facial patches for emotion recognition. To improve the FER performance, Li et al. [33] constructed a local binary attention module for specific regions to obtain the real emotion hidden under the face. Cai et al. [34] proposed a probabilistic attribute tree convolutional neural network to deal with the influence of identity-related attributes and achieved accuracy of 88.43%. Xi et al. [35] proposed a novel weighted contrastive objective function to measure positive and negative samples labeled by pseudo labels, in order to reduce the intra-class variation while enlarging the distance among different instances. This method obtains accuracy of 86.96%. RAF-DB is an in-the-wild dataset, in which there exist great changes in background, illumination and the attributes on face are more emotion-irrelevant. Our method performs self-mutual attention that is beneficial for the enhanced and complementary learning of interesting features and obtains accuracy of 90.71%, which is superior than the above methods by obvious performance gaps. Besides our method also shows better recognition performance than other methods listed in Table 1 that utilize probability distributions [9, 10, 36].

Table 2 Accuracy (%) of different methods on FER2013 dataset

Method	Year	Accuracy
MTCNN [41]	2017	60.70
ECNN [40]	2017	69.96
MRMREP [39]	2018	70.66
Shao et al. [42]	2019	56.64
Gan et al. [36]	2019	73.73
Minaee et al. [38]	2021	70.02
LBAN-IL [33]	2021	73.11
PAT-CNN [34]	2022	73.28
Indolia et al. [8]	2023	64.89
SSA-Net [7]	2023	67.57
WeiCL [35]	2023	71.42
Ours		74.06

4.4 Results on FER2013 Dataset

The evaluation on FER2013 dataset is reported in Table 2 for comparison with state-of-the-art methods. As can be seen, our method obtains accuracy of 74.06%, which is superior to [8, 34–36] and is consistent with the comparison on the RAF dataset. The methods in [7, 33, 38], which aim to relieve the influence of irrelevant facial information, also explore attention mechanisms, but obtain inferior performance than ours. We also make comparison with some methods that use multiple networks for emotion recognition. Among them, Li et al. [39] proposed redundancy reduction method for 35 deep models, and obtained accuracy of 70.66% by fusing multiple networks. Wen et al. presented a probability-based method for the prediction fusion of deep models, and the corresponding accuracy is 69.96% [40]. Our method performs mutual learning of the interesting features and mutual calibration for the two networks, resulting in better fusion result. Moreover, our architecture is superior than some methods that potentially utilize the ability of mutual learning by multi-task framework or diversified representation. Concretely, Xiang et al. designed a two-stream framework to exert the auxiliary ability of face detection [41]. Shao et al used handcraft features to utilize the collaborative learning of different emotion features from extra extractor [42]. By comparison, our method is 13.36% and 17.42% higher than these two methods.

4.5 Results on SFEW Dataset

Table 3 summarizes the results achieved on SFEW dataset. Meng et al. presented a multi-task framework to jointly recognize emotion and identity, an attempt to relieve the identity affect for FER [16]. With the same purpose of suppressing identity facial information, Liu et al. develop (N+M)-tuplet cluster loss that is expression-sensitive, and obtain accuracy of 54.19% [43]. Besides, some methods introduce attention learning [6, 7, 31, 32] and obtain accuracies of 53.18%, 50.00%, 51.47% and 56.40% respectively. Our method obtains accuracy of 60.18%. The contrasts show that our method achieves obviously better result. At the same time, the results indicate the effectiveness compared to methods [9, 10, 36] that introduce probability distributions. Furthermore, our method is superior to methods [33, 34, 36] that report emotion recognition performance on both RAF-DB and FER2013 datasets.

Table 3 Accuracy (%) of different methods on SFEW dataset

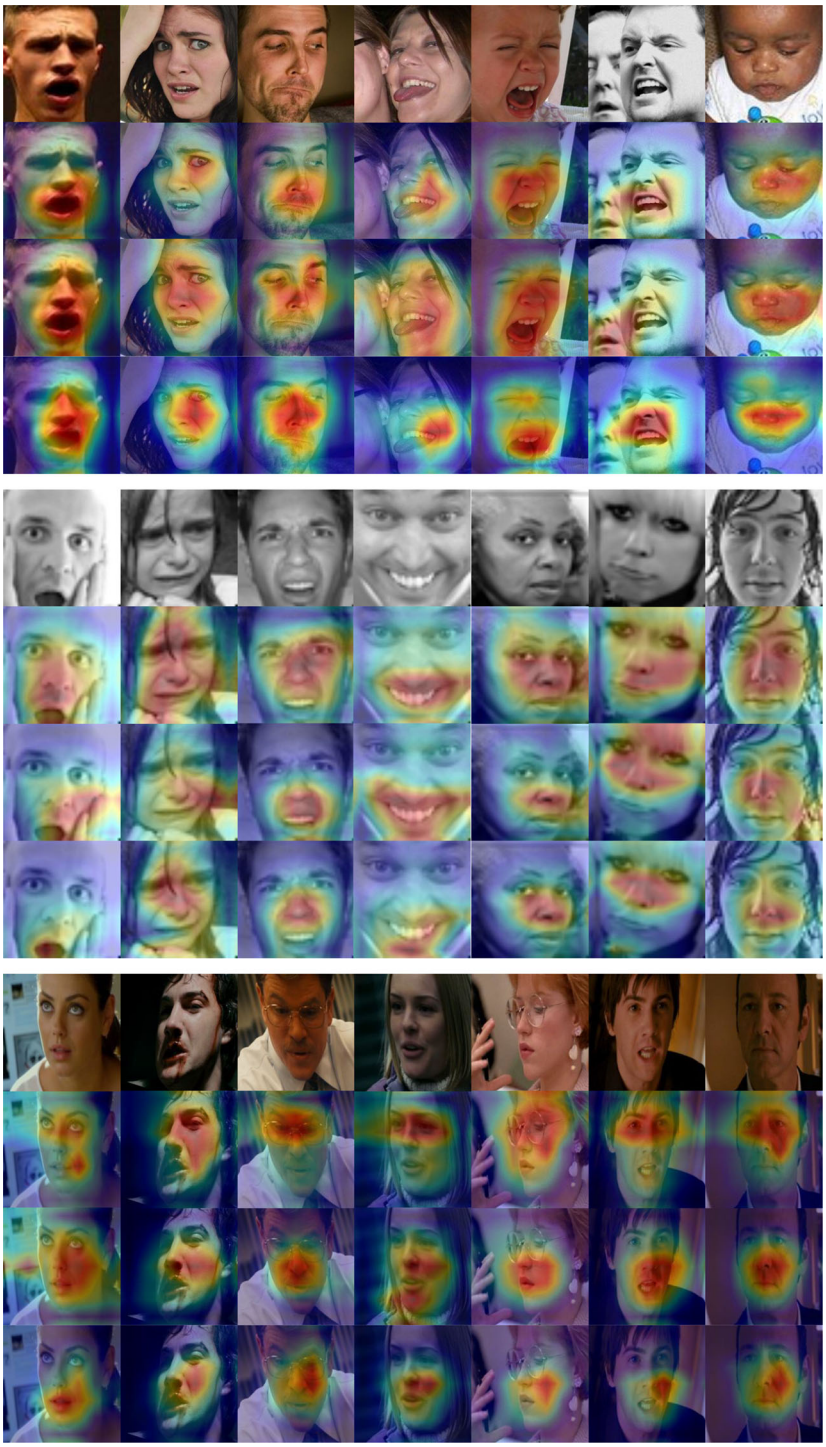
Method	Year	Accuracy
IACNN [16]	2017	50.98
Liu et al. [43]	2017	54.19
gACNN [31]	2018	51.47
Gan et al. [36]	2019	55.73
RAN [32]	2020	56.40
LDL-ALSG [9]	2020	56.50
LBAN-IL [33]	2021	55.28
PAT-CNN [34]	2022	57.57
SSA-Net [7]	2023	50.00
DAN [6]	2023	53.18
Le et al. [10]	2023	59.90
Ours		60.18

4.6 Visualization

Figure 6 shows the heatmap visualization of the outputs of different methods on RAF-DB, FER2013 and SFEW datasets, to give a clear and visual understanding of the proposed framework. Resnet+SAM and resnet+CBAM are the architectures of Net1 and Net2 respectively, so we regard them as baseline methods and visualize their heatmaps. As can be seen, for the same image, these two methods always show attentions that have both similarities and differences. For example, they have relatively large response values in some key areas that around nose, mouth and eyes. Meanwhile, we can also observe that there are obvious differences between them. The above observation inspires us to deploy the two networks with SAM and CBAM respectively, and perform mutual learning of attention. Meanwhile, we can also see that there are obvious differences between them. The above observation inspires us to deploy the two networks with SAM and CBAM respectively to perform mutual attention learning. In our method, the attention information from Net1 or Net2 selectively pass to the other network, which is benefit for interesting learning with enhanced and complementary properties and encode discriminative features for performance improvement.

4.7 Ablation Study

Finally, ablation study is performed to provide further insight into the key modules, and the experimental results are listed in Table 4. We report performance of the single network with different settings as baselines, i.e., resnet+initialization1, resnet+initialization2, resnet+SAM and resnet+CBAM. Specifically, resnet+initialization1 and resnet+initialization2 represent the methods that adopt resnet50 as the network architectures initialized by the pretrained Ms-celeb-1 m and ImageNet models respectively, which correspond to Net1 and Net2 without any additional modules. Resnet+SAM and resnet+CBAM represents adding SAM and CBAM modules into the above two baselines respectively, which indicates that the models only perform self-attention without mutual attention. Besides, we summarize the performance of the key modules in the proposed framework through the resnet+SMAL method. Resnet+SMAL represents adding SMAL module between Net1 and Net2. As can be seen, on SFEW dataset, the effectiveness of attention mechanisms is relatively obvious, since the accuracy contrasts



◀**Fig. 6** The heatmap visualizations for different methods on RAF-DB, FER2013 and SFEW datasets (three image blocks from top to bottom). The images from leftmost column to rightmost column correspond to surprise, fear, disgust, happiness, sadness, anger and neutral, respectively. For each image block, the images from top to bottom correspond to the original image, and the heatmap visualizations of resnet+SAM, resnet+CBAM and ours, respectively

Table 4 Ablation study on the RAF-DB, FER2013 and SFEW datasets

Method	Accuracy		
	RAF	FER2013	SFEW
Resnet+initialization1	88.77	70.68	50.23
Resnet+initialization2	87.71	70.03	54.76
Resnet+SAM (Net1)	88.79	71.21	52.25
Resnet+CBAM (Net2)	88.03	70.49	56.42
Ours (Resnet+SMAL)			
Net1	89.18	72.75	54.68
Net2	88.27	71.05	57.16
Fusion	90.51	73.50	58.72
Ours (resnet+SMAL+PDDL)			
Net1	90.05	73.42	56.65
Net2	89.76	72.81	59.54
Fusion	90.71	74.06	60.18

referring to with or without attention learning component are 52.25% vs 50.23% and 56.42% vs 54.76%. On RAF-DB and FER2013 datasets, attention mechanisms tend to play small role. On the three datasets, the Net1 of resnet+SMAL shows consistent superiority compared to resnet+SAM, since the accuracies increase by 0.39%, 1.54%, 2.43%, respectively. Performance gains can also be found in Net2. These results elaborate the contribution of SMAL module, which enables mutual attention learning with enhanced and complementary properties. In addition, the PDDL module is also essential, since certain performance gain can be found by comparing our method with resnet+SMAL. The results on the three datasets demonstrate that our key modules, including SMAL and PDDL, are effective to improve emotion recognition performance.

5 Conclusion

In this paper, we present a novel mutual learning method for emotion recognition, which tends to be harmonious because it can increase the dynamics of mutual teaching. Specifically, we construct a framework with two emotion recognition networks and perform progressive mutual learning in the backbone and the classification head, through utilizing attention mechanisms and probability distributions. The self-mutual attention learning module is integrated into the convolutional block of the backbone, allowing to encode discriminative facial features by the learning of interesting information with enhanced and complementary properties. In this module, we introduce SAM and CBAM submodules for the two networks, which can preserve the self-learning capability to promote mutual teaching. Further, we conduct mutual distillation learning in classification head, enabling mutual calibration for emotion recogni-

tion. Experimental results on three public datasets show that the proposed method achieves state-of-the-art performance.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 62167001, 62106054, 62307009), and Guangxi Universities Young and Middle-aged Teachers Basic Ability Improvement Project (Grant No. 2022KY0050).

References

1. Wilhelm T (2019) Towards facial expression analysis in a driver assistance system. In: 2019 14th IEEE international conference on automatic face and gesture recognition (FG 2019), pp 1– 4 . IEEE
2. Sajjad M, Zahir S, Ullah A, Akhtar Z, Muhammad K (2020) Human behavior understanding in big multimedia data using cnn based facial expression recognition. Mobile Netw Appl 25(4):1611–1621
3. Savchenko AV, Savchenko LV, Makarov I (2022) Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Trans Affect Comput 13(4):2132–2143
4. Li B, Mehta S, Aneja D, Foster C, Ventola P, Shic F, Shapiro L (2019) A facial affect analysis system for autism spectrum disorder. In: 2019 IEEE international conference on image processing (ICIP), pp 4549–4553 . IEEE
5. Gan Y, Chen J, Yang Z, Xu L (2020) Multiple attention network for facial expression recognition. IEEE Access 8:7383–7393
6. Wen Z, Lin W, Wang T, Xu G (2023) Distract your attention: multi-head cross attention network for facial expression recognition. Biomimetics 8(2):199
7. Liu Y, Peng J, Dai W, Zeng J, Shan S (2023) Joint spatial and scale attention network for multi-view facial expression recognition. Pattern Recognit. 139:109496
8. Indolia S, Nigam S, Singh R (2023) A framework for facial expression recognition using deep self-attention network. J Ambient Intell Human Comput 14(7):9543–9562
9. Chen S, Wang J, Chen Y, Shi Z, Geng X, Rui Y (2020) Label distribution learning on auxiliary label space graphs for facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13984– 13993
10. Le N, Nguyen K, Tran Q, Tjiputra E, Le B, Nguyen A (2023) Uncertainty-aware label distribution learning for facial expression recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 6088– 6097
11. Bian X, Qin B, Xin X, Li J, Su X, Wang Y (2022) Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning. Proc the AAAI Conf Artif Intell 36:113–121
12. Qiao Y, Jing L, Song X, Chen X, Zhu L, Nie L (2023) Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 37, pp 9507– 9515
13. Ma Q, Yu L, Tian S, Chen E, Ng WW (2019) Global-local mutual attention model for text classification. IEEE/ACM Trans Audio Speech Lang Process 27(12):2127–2139
14. Zhang D, Zhang Z, Ju Y, Wang C, Xie Y, Qu Y (2022) Dual mutual learning for cross-modality person re-identification. IEEE Trans Circuits Syst Video Technol 32(8):5361–5373
15. Liu D, Zhao L, Wang Y, Kato J (2023) Learn from each other to classify better: cross-layer mutual attention learning for fine-grained visual classification. Pattern Recognit 140:109550
16. Meng Z, Liu P, Cai J, Han S, Tong Y (2017) Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE international conference on automatic face and gesture recognition (FG 2017), pp 558– 565 . IEEE

17. Zhao X, Liang X, Liu L, Li T, Han Y, Vasconcelos N, Yan S (2016) Peak-piloted deep network for facial expression recognition. In: Proceedings of the European conference on computer vision (ECCV), pp 425–442. Springer
18. Zhang K, Huang Y, Du Y, Wang L (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process 26(9):4193–4203
19. Liu X, Kumar BV, Jia P, You J (2019) Hard negative generation for identity-disentangled facial expression recognition. Pattern Recognit 88:1–12
20. Liu Y, Dai W, Fang F, Chen Y, Huang R, Wang R, Wan B (2021) Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition. Inf Sci 578:195–213
21. Zhang Y, Xiang T, Hospedales TM, Lu H (2018) Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4320– 4328
22. Xu Y, Kan M, Shan S, Chen X (2022) Mutual learning of joint and separate domain alignments for multi-source domain adaptation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1890– 1899
23. Wang Y, Wang L, Zhou Y (2023) Bi-level deep mutual learning assisted multi-task network for occluded person re-identification. IET Image Process 17(4):979–987
24. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3– 19
25. Li S, Deng W, Du J (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, pp 2852–2861
26. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H, et al (2013) Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing, pp 117– 124. Springer
27. Dhall A, Ramana Murthy O, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: emotiw 2015. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp. 423– 426
28. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
29. 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
30. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255 . IEEE
31. 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
32. 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
33. Li H, Wang N, Yu Y, Yang X, Gao X (2021) Lban-il: a novel method of high discriminative representation for facial expression recognition. Neurocomputing 432:159–169
34. Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y (2023) Probabilistic attribute tree structured convolutional neural networks for facial expression recognition in the wild. IEEE Trans Affect Comput 14(3):1927–1941
35. Xi Y, Mao Q, Zhou L (2023) Weighted contrastive learning using pseudo labels for facial expression recognition. Vis Comput 39(10):5001–5012
36. Gan Y, Chen J, Xu L (2019) Facial expression recognition boosted by soft label with a diverse ensemble. Pattern Recognit Lett 125:105–112
37. Ruan D, Yan Y, Lai S, Chai Z, Shen C, Wang H (2021) Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7660–7669
38. Minaee S, Minaei M, Abdolrashidi A (2021) Deep-emotion: facial expression recognition using attentional convolutional network. Sensors 21(9):3046
39. Li D, Wen G (2018) Mrmr-based ensemble pruning for facial expression recognition. Multimedia Tools Appl 77(12):15251–15272
40. Wen G, Hou Z, Li H, Li D, Jiang L, Xun E (2017) Ensemble of deep neural networks with probability-based fusion for facial expression recognition. Cognit Comput 9(5):597–610
41. Xiang J, Zhu G (2017) Joint face detection and facial expression recognition with mtcnn. In: 2017 4th International conference on information science and control engineering (ICISCE), pp 424– 427 . IEEE
42. Shao J, Qian Y (2019) Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing 355:82–92

43. Liu X, Vijaya Kumar B, You J, Jia P (2017) Adaptive deep metric learning for identity-aware facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 20–29