Received: 11 September 2021 | Revised: 13 December 2021 | Accepted: 20 January 2022

IET Image Processing

DOI: 10.1049/ipr2.12441

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Real-time facial expression recognition based on iterative transfer learning and efficient attention network

Yinghui Kong[1,2] | Shuaitong Zhang[1,2] | Ke Zhang[1,2] | Qiang Ni[3] | Jungong Han[4]

[1] Department of Electronics and Communication Engineering, North China Electric Power University, Baoding, Hebei, China

[2] Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding, Hebei, China

[3] School of Computing and Communications, Bailrigg, Lancaster University, Lancaster, UK

[4] WMG Data Science, University of Warwick, Coventry, UK

Correspondence
Yinghui Kong, Department of Electronics and Communication Engineering, North China Electric Power University, Baoding 071003, Hebei, China.
Email: kongyhbd2015@ncepu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 62076093, 61871182; Fundamental Research Funds for the Central Universities, Grant/Award Numbers: 2020YJ006, 2020MS099; S&T Program of Hebei, Grant/Award Number: SZX2020034

Abstract

Real-time facial expression recognition is the basis for computers to understand human emotions and detect abnormalities in time. To effectively solve the problems of server overload and privacy information leakage, a real-time facial expression recognition method based on iterative transfer learning and efficient attention network (EAN) for edge resource-constrained scenes is proposed in this paper. Firstly, an EAN is designed with its parameter number and computation amount strictly limited by depth separable convolution and local channel attention mechanism. Then, the soft labels of facial expression data were obtained by EAN based on the idea of knowledge distillation, so as to provide more supervision information for the training process. Finally, an iterative transfer learning method of teacher-student (T-S) network was proposed; it refines the soft labels of the teacher network and further improves the recognition accuracy of the student network. The tests on the public datasets, FER2013 and RAF-DB, show that this method can significantly reduce the model complexity and achieve high recognition accuracy. Compared with other advanced methods, the proposed method strikes a good balance between complexity and accuracy, and well meets the real-time deployment requirements of facial expression recognition technology for edge resource-constrained scenes.

1 | INTRODUCTION

As an important way for human to express emotion and convey intention, the efficient recognition of facial expressions has found a wide application. In 1978, Ekman et al. [1] divided the facial expressions into six basic categories (anger, disgust, fear, happiness, sadness and surprise), taking the lead in the expression recognition research on the controlled scenes of laboratory. Twenty years later, Lyons et al. [2] added "neutral" expression into the basic expression categories, and constructed JAFFE dataset. In 2010, CK+ dataset [3] greatly expanded the scale of expression data in laboratory scenes. After 2013, the expression datasets sampled in the wild, such as FER2013 [4] and RAF-DB [5], were built successively, further promoting the research and deployment of expression recognition technology in the fields such as medical care, fatigue monitoring and smart home.

Early research on expression recognition [6–8] relied on hand-designed features, which were efficient in execution but could not adequately adapt to the facial data from various scenarios. With the improvement of hardware performance and data volume, convolution neural networks have been widely used in facial expression recognition tasks. In many studies [9–11] multi-task and multi-scale information was integrated on the basis of classical networks, such as VGG (Visual Geometry Group) [12] and ResNet (Residual Network) [13], so as to improve the recognition accuracy. Qin [14] and Pramerdorfer et al. [15] further ensembled multiple convolution neural networks, greatly surpassing the average human recognition level on FER2013. The neural network methods above enhanced the adaptability and accuracy of the models by their huge number of parameters, but most of them are suitable for centralized processing on high-performance servers, resulting in high data

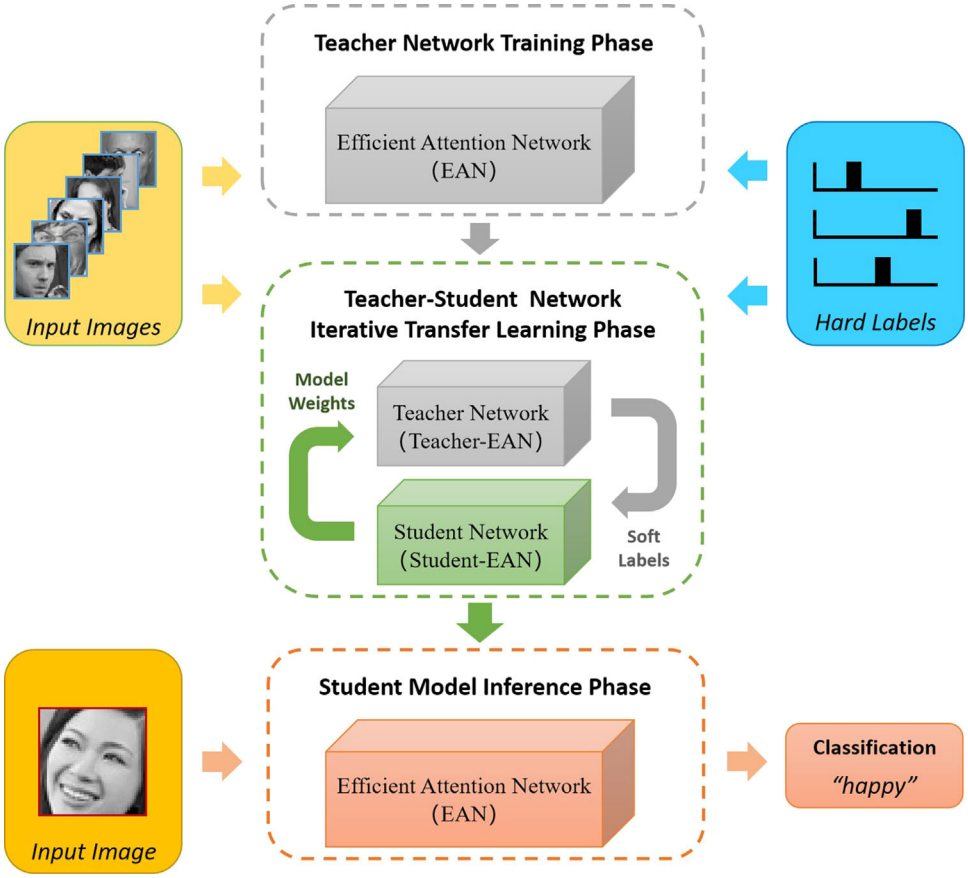

FIGURE 1 Real-time expression recognition framework based on iterative transfer learning and efficient attention network

transmission cost and privacy leakage risk. It is difficult for them to deploy directly on edge devices with resource-constrained scenes (such as mobile terminals and embedded equipment). In this case, an effective real-time expression recognition method is required. In recent years, research on real-time facial expression recognition using lightweight networks with fewer parameters and lower computation amount has attracted extensive attention, and a considerable amount of researchers [16–20] have greatly compressed the network parameter number through lightweight structure design methods, especially depth separable convolution and multi-channel reuse. However, too compacted parameters also decrease the fitting ability of the network, making it hard to guarantee the recognition accuracy in the wild.

Recently, the attention mechanism has been used to enhance the recognition performance of light weight networks [21–24]. It can dynamically adjust the feature map of convolution layer for each data sample with only a few parameters, so as to improve the fitting ability. In addition, the method of knowledge distillation [25, 26] changes the training schemes of the network without introducing additional parameters, providing an effective solution. In [25], the soft labels predicted by teacher networks were utilized to provide the similarity between expression categories for student networks. In [26], through the teacher network trained with full face data, the feature information was supplemented for the student network trained with half face data. However, they still relied on the complex network structure of VGG. Gan et al. [25] modified the output number of the last fully connected layer of the VGG-16 to 7, and Georgescu et al. [26] fine-tuned VGG-f [27] and VGG-face [28] on expression datasets, making it impossible to deploy directly on edge devices.

To meet the requirements of high precision and real-time expression recognition on edge devices, we propose a real-time expression recognition method that combines iterative transfer learning and efficient attention mechanism. The main work and contributions of this paper are as follows:

1. Inspired by the idea of knowledge distillation, we propose an iterative transfer learning method for teacher-student (T-S) network. The soft labels of the teacher network are used to assist the training of the student network without introducing any additional parameters. By refining the soft labels of the teacher network in an iterative way, the recognition accuracy of the student network is improved significantly.
2. We propose an efficient attention network which is a combination of the depth separable convolution and local channel attention mechanism; it strictly controls the parameter number and calculation amount so as to ensure the real-time inference performance.

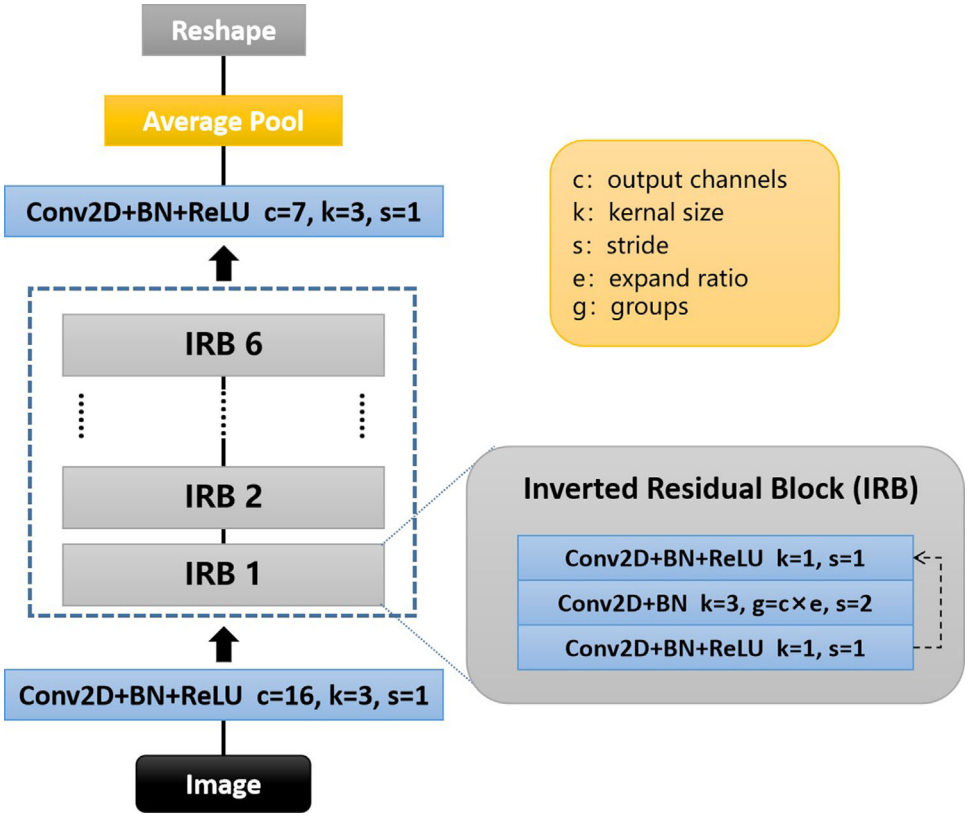

FIGURE 2 Lightweight depth separable network structure

3. Tests on public benchmark expression datasets FER2013 and RAF-DB show that our method requires fewer parameters than other state-of-the-art lightweight methods, while the recognition accuracies are similar, thus strikes a balance between the complexity and accuracy.

2 | RELATED WORK

2.1 | Lightweight expression recognition method based on CNN

Lightweight convolution neural networks can improve the inference speed of expression recognition models and reduce their storage space and power consumption at the same time. Wang et al. [16] used the depth separable convolution to replace regular convolution in VGG [12]; it preserves the pre-training knowledge of shallow convolution layers and cuts down the network parameter number. Arriaga et al. [17] used the average pooling layer instead of fully-connected layers and further diminish the layers of Xception [29] to reduce the model parameter number. Cotter et al. [18] presented a MobiExpressNet based on MobileNet [30]. The number of parameters was controlled by adjusting the channel number of inverted residual blocks. Zhou et al. [19] combined the depth separable convolution with maximum pooling, thus reducing the parameter and computation amount within convolution blocks.

By using the depth separable convolution, the conventional convolution layer is decomposed into point-wise and depth-wise convolution. The model parameters are thus reduced and the model fitting ability is decreased as well. Inspired by the channel reuse idea of DenseNet [31], Zhao et al. [20] multiplexed the multi-scale channel features of conventional convolution to lighten the model. It controls the parameter number and ensures the network fitting ability as well, but fails to solve the problem of high computation amount. Tang et al. [32] lowered the computation amount by sampling the low-frequency information of the input image, but the parameter number of model was still very large.

2.2 | Expression recognition method based on attention mechanism

The attention mechanism can weaken the influence of irrelevant features on the network and has been widely used in expression recognition. To increase the weight of salient frames, Meng et al. [33] and Kumar et al. [34] learned the relationship of expression sequence frames through the fully-connected layer. Li et al. [35] defined 24 facial expression significant regions based on prior knowledge, and predicted the importance of each region by pooling, convolution and fully-connected layers, so as to decrease the interference of occluded regions. Wang et al. [36] cropped the face into multiple regions to weaken the

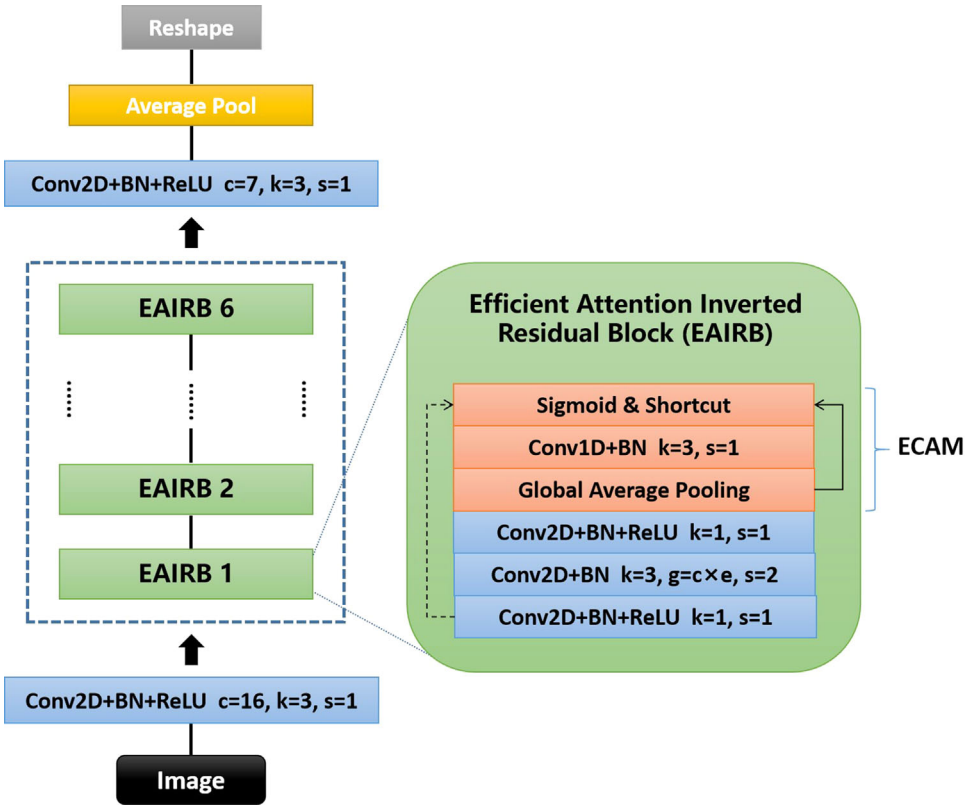

FIGURE 3 Structure of efficient attention network

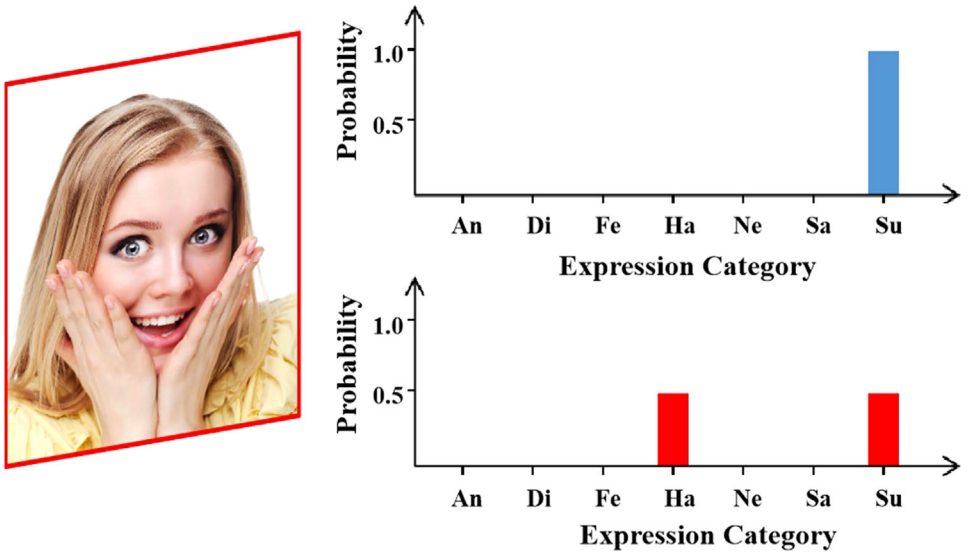

FIGURE 4 Comparison of hard (top) and soft (bottom) labels

negative impact of the occluded regions on recognition through a two-level attention mechanism. Li et al. [37] combined spatial attention with channel attention to improve the recognition accuracy of the backbone network.

The expression recognition methods based on attention mechanism mainly learn the attention weights through the fully-connected or two-dimension convolution layer. However, their high computation complexity hampers the real-time inference performance. Wang et al. [38] recently proposed a method of learning local channel attention weights through one-dimension convolution, and achieved good efficient recognition performance in common object classification task.

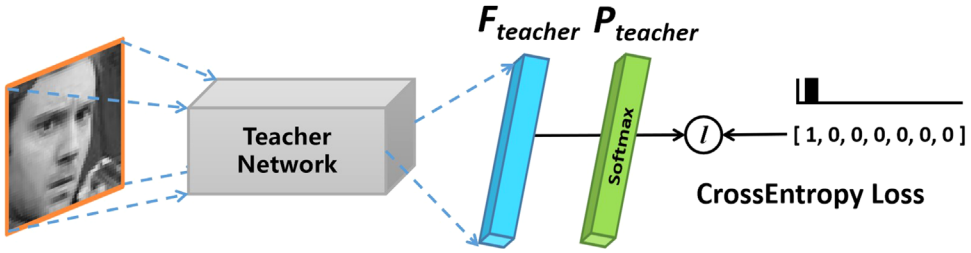

FIGURE 5 Process of teacher network training stage

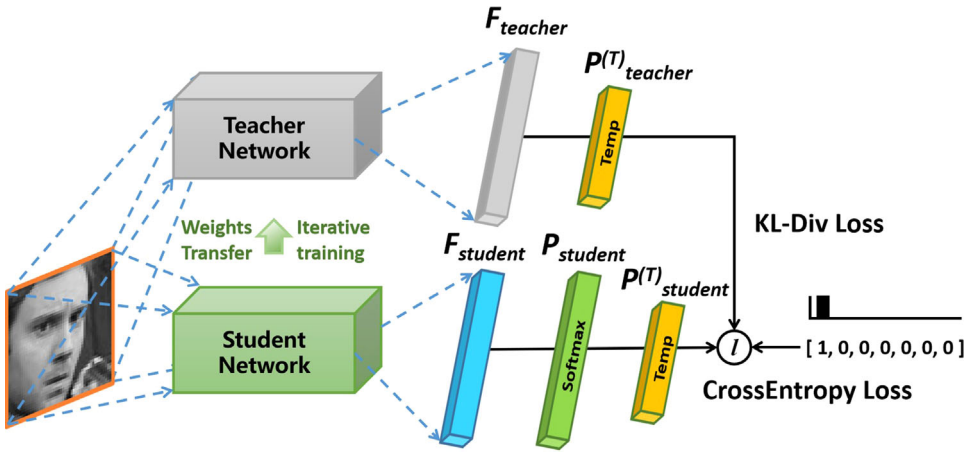

FIGURE 6 Process of T-S network iterative transfer learning phase

FIGURE 7 Samples from FER2013 (first line) and RAF-DB (second line) datasets

TABLE 1 Performance comparison of different network architecture on FER2013 and RAF-DB

Models	Output channels (c) and expansion ratio (e) [c_{head}, $c_1(e_1)$, $c_2(e_2)$, $c_3(e_3)$, $c_4(e_4)$, $c_5(e_5)$, $c_6(e_6)$, c_{tail}]	Parameters (M)	FLOPs (M)	FER2013 Accuracy (%)	RAF-DBAccuracy (%)
model_a	[32, 16(6), 24(6), 24(6), 32(6), 32(6), 64(6), 7]	0.077	20.892	69.74	82.93
model_b	[32, 16(3), 32(3), 32(3), 64(3), 64(3), 128(1), 7]	0.069	24.834	69.70	82.89
model_c	[16, 16(6), 24(6), 24(6), 32(6), 32(6), 64(6), 7]	0.069	9.446	69.41	82.73
model_d	[32, 16(3), 24(3), 24(3), 48(3), 48(3), 128(1), 7]	0.053	11.723	69.41	82.64
model_e	[16, 8(3), 24(6), 24(6), 48(3), 48(3), 128(1), 7]	0.053	11.157	69.46	82.60
model_f	[16, 8(3), 24(3), 24(3), 48(3), 48(3), 64(3), 7]	0.052	5.126	69.58	82.33
EAN	[16, 8(3), 16(6), 16(6), 32(6), 32(6), 64(3), 7]	0.044	5.054	69.74	82.56

FIGURE 8 Image preprocessing of our method

2.3 | Knowledge distillation method based on soft labels

The knowledge distillation methods define knowledge as the probability of each category predicted by the teacher network. Under this definition, knowledge distillation methods based on soft labels can realize knowledge transfer through lightweight student network imitating the outputs of complex teacher network. Compared with traditional one-hot labels, soft labels have corresponding classification probability in each category and can provide more inter-class similarity information to the student network, so as to reduce the difficulty in lightweight network training [39].

The quality of soft labels usually depends on the selection of teacher networks. The study in [40] indicated that the performance of the student network is not positively correlated with the teacher network when their network capacities are quite different. Hence, the strategy of "teacher network early stop" was proposed to reduce the accuracy loss of the student network. Shen et al. [41] trained and integrated multiple complex teacher networks, and elaborately designed the structure selector. In each training round, the structure selector searches the optimal teacher network to guide the training of lightweight model. In [42], a mutual learning approach was proposed to synchronize the training of the teacher network on the student network, and it effectively avoided the pre-training process of the teacher network and enhanced the knowledge transfer between models. Xie et al. [43] combined knowledge distillation with semi-supervised learning and generated pseudo labels for extra unlabelled samples by using the teacher model. To improve the recognition accuracy, this method allows a more complex student model to be trained repeatedly on a variety of data with

TABLE 2 Performance of the iterative transfer learning method with different hyperparameters on FER2013 and RAF-DB

Methods	α	T	FER2013 Accuracy (%)	RAF-DB Accuracy (%)
EAN	—	—	69.74	82.56
iter_trans_a	0.3	5	70.08	84.90
iter_trans_b	0.3	10	70.44	84.19
iter_trans_c	0.5	5	70.63 (+0.89)	85.30 (+2.74)
iter_trans_d	0.5	10	70.47	84.94
iter_trans_e	0.7	5	70.24	84.52
iter_trans_f	0.7	10	70.16	84.62

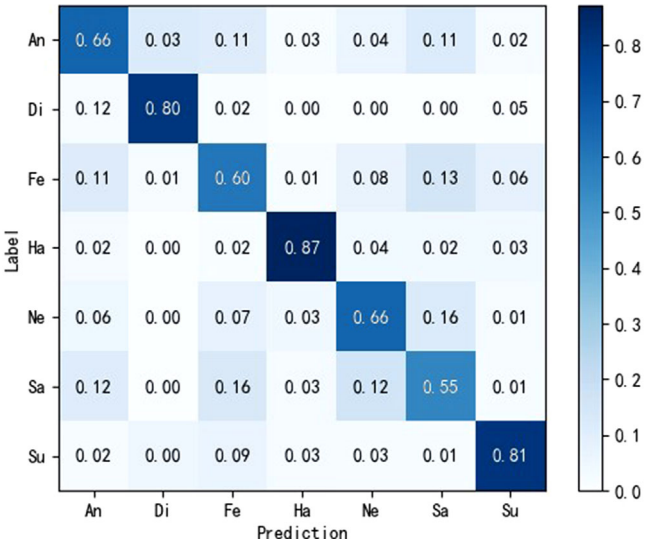

FIGURE 9 Confusion matrix on FER2013 dataset

TABLE 3 Performance comparison of various lightweight methods on FER2013

Methods	Parameters	Accuracy (%)
BLOCK-FERNET [32]	9,000,000	64.41
VGG-retrain-v1 [16]	2,400,000	69.50
VGG-retrain-v2 [16]	1,000,000	69.80
DenseNet-1 [20]	95,263	70.91
MobiExpressNet [18]	75,079	67.96
LCNN [19]	58,423	67.00
Proposed method	43,843	70.63

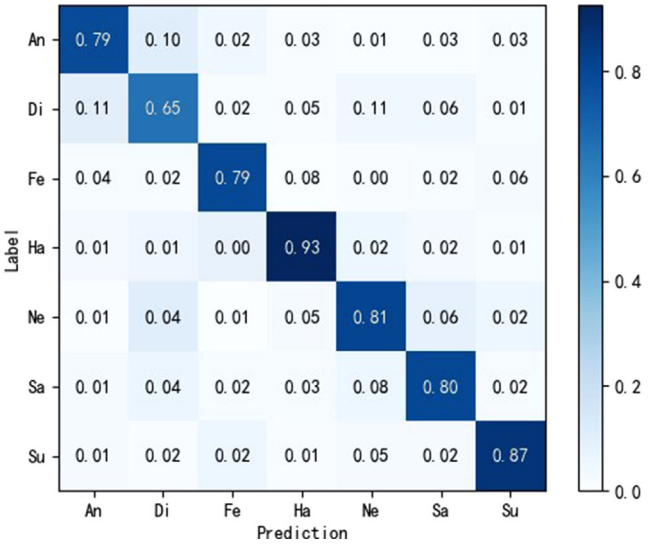

FIGURE 10 Confusion matrix on RAF-DB dataset

TABLE 4 Performance comparison of various methods on RAF-DB

Methods	Backbone	Accuracy (%)
DLP-CNN [5]	VGG-8	82.74
MRE-VGG [9]	VGG-16	82.63
PAT-ResNet [10]	ResNet-34	84.19
DSAN-RES-RACE [11]	ResNet-50	85.27
DSAN-VGG-RACE [11]	VGG-16	85.37
Soft Label [25]	VGG-16	85.20
Proposed method	EAN	85.30

TABLE 5 Deployment performance comparison with mainstream backbone networks

Backbones	Parameters (M)	Memory (MB)	FLOPs (M)	Latency (ms)
VGG-16 [12]	33.634	140.529	557.136	11.952
ResNet-18 [13]	11.172	65.460	1101.272	12.255
MobileNetV2 [30]	2.233	15.040	16.796	6.010
SqueezeNet [48]	0.733	18.881	101.243	4.104
ShuffleNetV2 [49]	0.126	6.284	10.164	2.485
mini-Xception [17]	0.057	27.998	92.184	7.483
Proposed method	0.044	3.902	5.054	1.985

noises. However, the training process is quite complex and the final number of model parameters is also larger.

3 | METHODOLOGY

The proposed framework for real-time expression recognition based on iterative transfer learning and efficient attention network (EAN) is shown in Figure 1. The framework mainly includes three phases, namely the teacher network training, the T-S network iterative transfer learning, and the student model inference, which are all implemented by the proposed EAN.

It is worth noting that EAN is constructed by depth separable convolution and local channel attention block. Such a lightweight network not only greatly improves the recognition speed of the model, but also has high recognition accuracy with the addition of attention mechanism. EAN makes the process

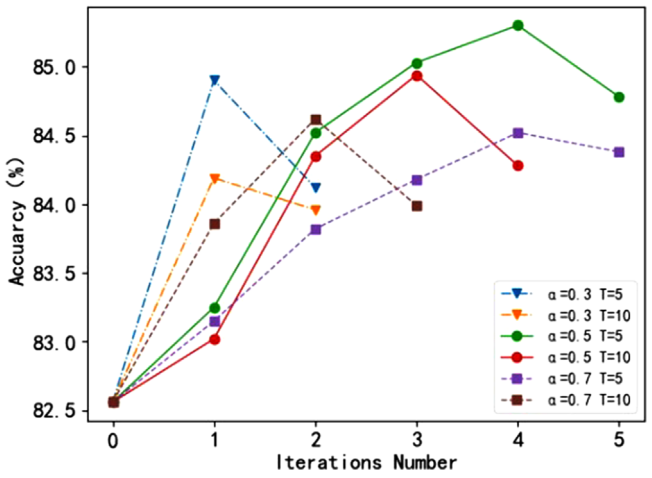

FIGURE 11 Accuracy comparison of EAN under different iteration times and hyperparameters on RAF-DB

of recognition very efficient, so it is called efficient attention network.

In the teacher network training phase, EAN firstly learns original one-hot labels from the datasets and gets the soft label information preliminarily conformed to the facial expression relationship. After that, the soft labels output by the teacher network are refined in the iterative transfer learning phase, so as to constantly improve the recognition accuracy. Finally, the optimal iterative student network model is used to infer and predict independently in the inference phase, so the real-time prediction speed of the model on the edge device can be ensured.

TABLE 6 Evaluation of our method on RAF-DB dataset

Local attention	T-S iteration	Weight transfer	Parameter number	Accuracy
×	×	×	43,822	81.54%
√	×	×	43,840	82.56%
√	√	×	43,840	84.13%
√	√	√	43,840	85.30%

3.1 | Construction of EAN

Inspired by the MobilenetV2 proposed in [30], the depth separable network can greatly reduce the parameters and calculation amount of conventional convolution. It can be

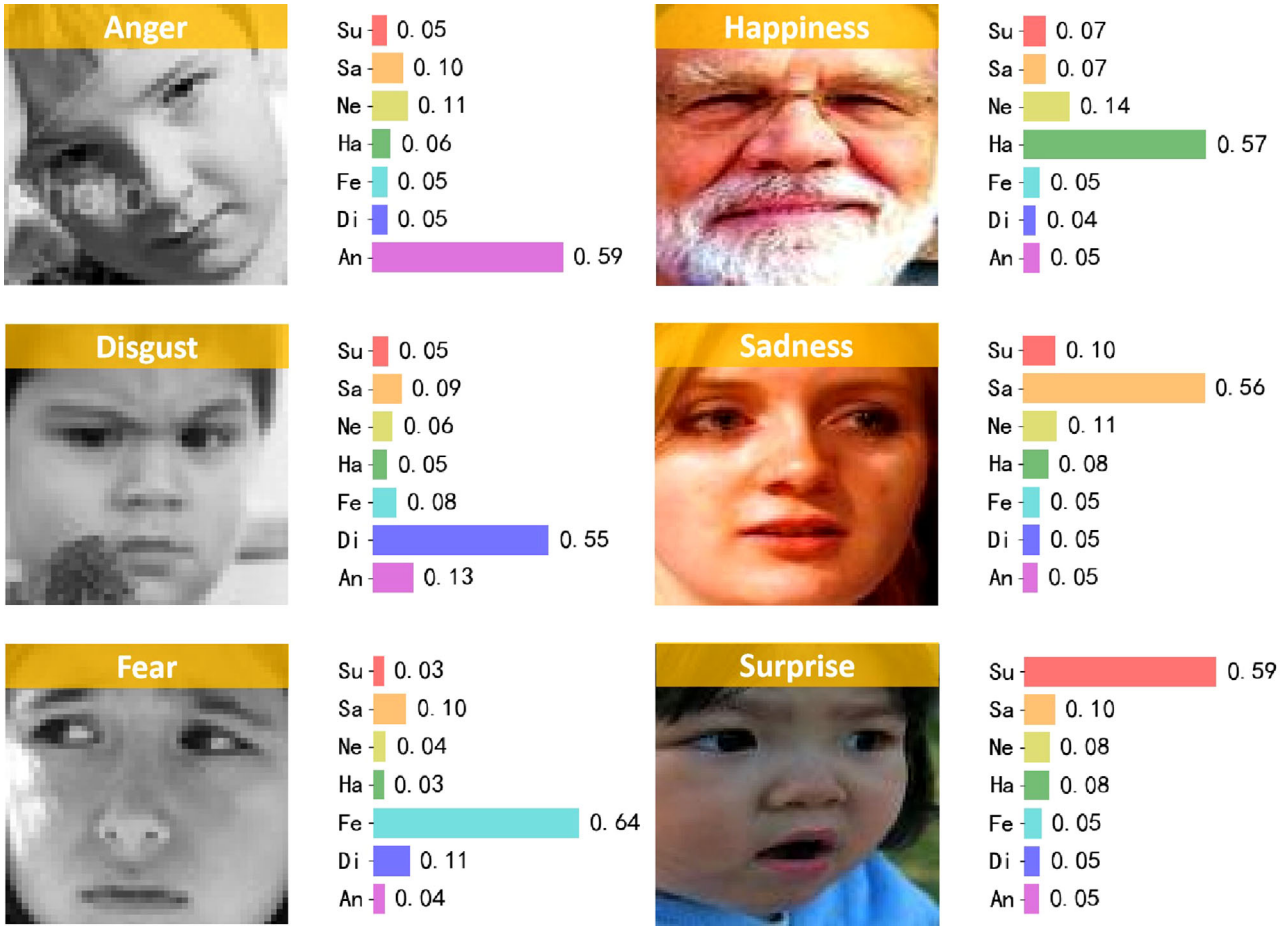

FIGURE 12 Visualization results of supervisory signals

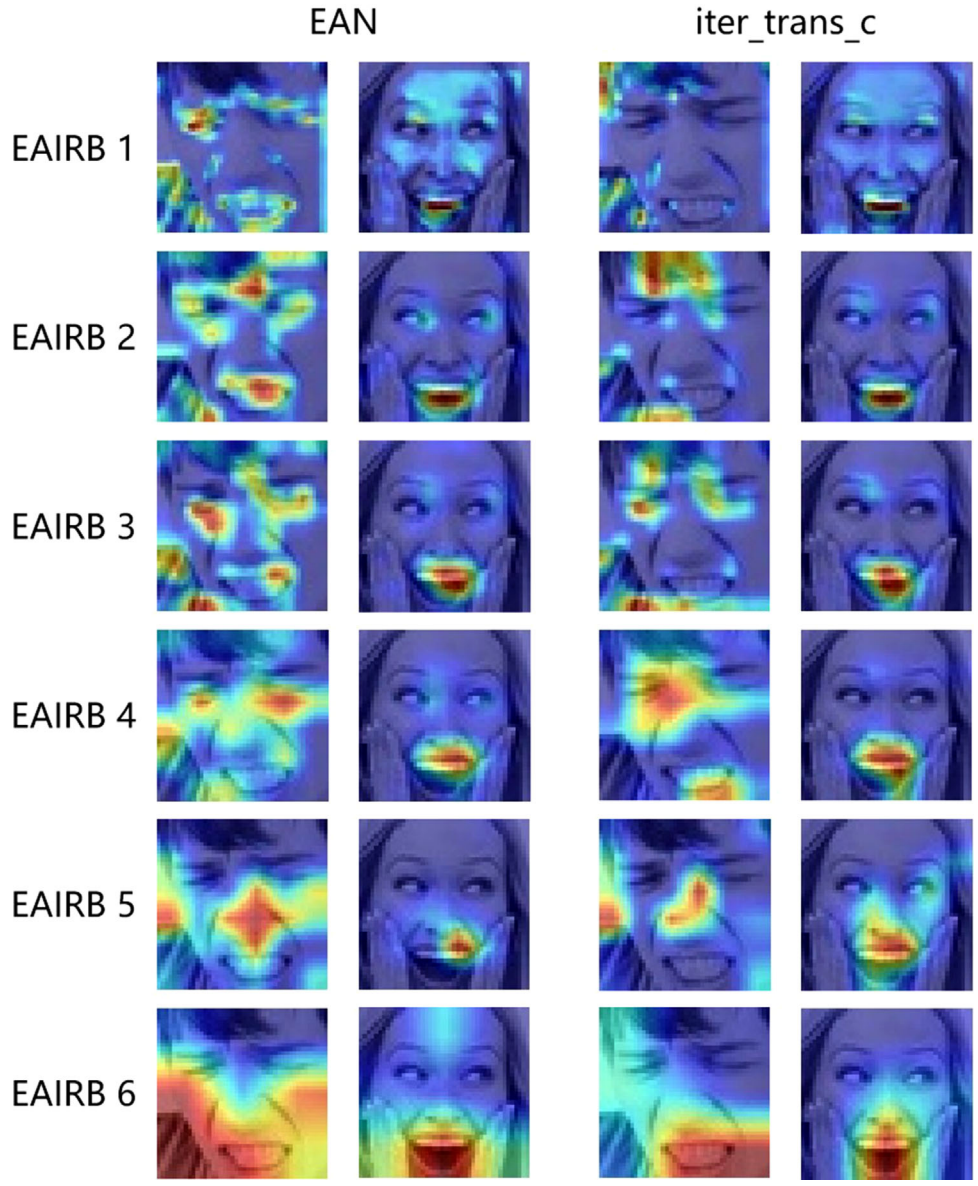

FIGURE 13 Grad-CAM visualization on FER2013

found in many researches [9–11, 25] that to utilize the previous architecture continually, some scholars up-sampled the input images, which multiplies the computation amount without any substantial increase in information indeed. Therefore, we propose an extremely lightweight depth separable network for low-resolution facial images (44 × 44). As shown in Figure 2, The main part of the network includes six cascaded inverted residual blocks (IRBs), and the down sampling is performed with the step size of two at the first, second, fourth and sixth IRB to save the memory consumption. For the third and fifth IRB, we also add short-cuts to assist gradient propagation.

The final recognition performance of the model depends on the ability of feature extraction, mainly reflected in the setting of the convolution kernel size k and output channels number c. The lower the input resolution, the smaller k should be selected to avoid the feature loss caused by the rapid growth of receptive field. For this reason, we fix the kernel size of the regular convolution and depth-wise convolution in each inverted residual block to 3 × 3. In the meanwhile, the number of output channels c and the expansion ratio of each inverted residual block e also determine the number of features extracted by the network, which directly determine the amount of model parameters and calculation. These two hyperparameters depend primarily on empirical selection and experimental verification, which will be discussed further in Section 4.3.

Moreover, the batch normalization (BN) is used to smooth the output distribution between convolution layers. Although its regularization effect is not as good as the recent methods of [44] and [45], it can avoid extra time delay of model

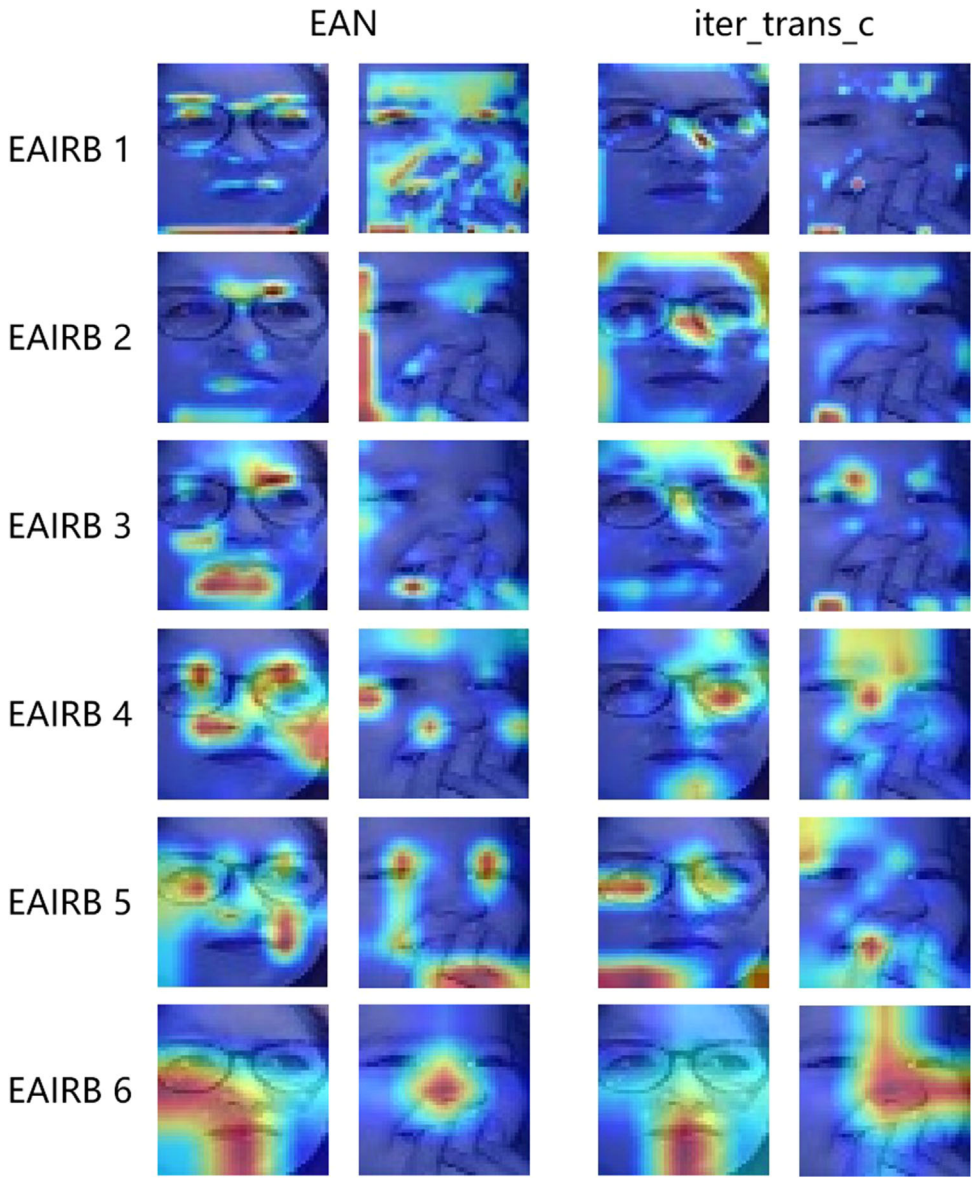

FIGURE 14 Grad-CAM visualization on RAF-DB

inference, for the parameters of BN layer are combined with the weight of the convolution layer. The output of BN layer is activated by ReLU function. Compared with other activation functions such as ReLU6, PReLU and Swish [46], although there is a certain degree of information loss, it is friendly supported by the hardware and saves more inference time.

Furthermore, to compensate for the decline of model fitting ability caused by depth separable convolution, we construct efficient attention inverted residual blocks (EAIRB) by introducing the local channel attention mechanism [38]. The structure of EAN is formed and shown in Figure 3, containing 43,840 parameters only. Its memory consumption is about 4.07 MB, and its single inference time is 1.985 ms in the ONNX-CPU runtime.

3.2 | Teacher network training

There is a large inclusion relationship between different facial expressions [1]. For example, the compound expression "excited" contains both "surprise" and "happy". A comparison of the hard and soft labels is shown in Figure 4. Previous datasets completely labelled this kind of data as the "surprise", so the "happy" component following with "surprise" was completely ignored. Soft labels can better present the similarity relationship between expression categories, providing more effective supervision information for the training network. Due to the complexity of marking soft labels manually, we adopt the semi-supervised learning approach in [25, 43] and take the trained and softened teacher network outputs as the soft labels of samples.

The training process of the teacher network is almost similar to that of the common classification model. The softmax loss is utilized here to demonstrate the difference between the model output and the original label of the datasets. The one-hot label (hard label) of the datasets is used as the supervision signal of the network, and the model parameters are optimized by backward propagation, as shown in Figure 5.

Considering the small capacity of the current expression datasets, for complex networks such as VGG-16 and ResNet-50, overfitting may emerge when fitting the one-hot labels in the training set [19]. So, we select the lightweight EAN as the teacher network and take the predicted probability as the soft label. Specifically, we firstly use the hard labels to optimize the outputs of EAN ($F_{teacher}$) on the training set, soften the model output with the distillation temperature T and perform the Softmax transform to obtain the probability distribution value of the i-th expression ($\hat{p}_{teacher}^{(T,\,i)}$), as shown in formula (1).

$$\hat{p}_{teacher}^{(T,\,i)} = \frac{\exp\left(\frac{F_{teacher}^{(i)}}{T}\right)}{\sum_{j=0}^{6}\exp\left(\frac{F_{teacher}^{(j)}}{T}\right)} \tag{1}$$

where $F_{teacher}^{(i)}$ represents the output value of EAN for the i-th expression of the input tensor without softmax.

To train the model sufficiently, we also augment the expression data by a random clipping method and use a less initial learning rate lr_1 to optimize the cross entropy loss $L_{tea}(\theta_{teacher})$ between the model outputs and the hard labels, as shown in formula (2).

$$L_{tea}\left(\theta_{teacher}\right) = \sum_{i=0}^{6} CrossEntropy\left(l_{hard}^{(i)},\ \hat{p}_{teacher}^{(i)}\right) \tag{2}$$

where $\theta_{teacher}$ represents the parameter weight of the teacher network, $l_{hard}^{(i)}$ represents the i-th probability of the hard label, and $\hat{p}_{teacher}^{(i)}$ is the probability of model prediction without softening for the i-th expression, a special form when the distillation temperature T equals 1, defined as follows:

$$\hat{p}_{teacher}^{(i)} = \frac{exp\left(F_{teacher}^{(i)}\right)}{\sum_{j=0}^{6}\ exp\left(F_{teacher}^{(j)}\right)} \tag{3}$$

3.3 | Iterative transfer learning for T-S network

The soft labels obtained by the teacher network contain the probability of each expression category, and it can supplement a similarity relationship between categories for the student network training. In such way, the training difficulty for the lightweight network can be decreased and the model accuracy can be improved as well. In addition, the recognition accuracy of the student model trained by soft labels is better, so it can be used as a teacher network for iterative training to provide soft label information with higher quality for the next student network. The proposed T-S iterative transfer learning method is shown in Figure 6.

The proposed method of T-S iterative transfer learning selects EAN as both teacher network and student network. The iterative training process of the student network is guided by both soft and hard labels. In the first round, the teacher model parameters are fixed as $\theta_{teacher}$ from the teacher network training phase. The teacher network is no longer trained in the iteration process, and its parameters are only updated by the weight transfer of the trained student network. Additionally, the updated soft label will change the original data distribution, so we adjust the learning rate to lr_2 to optimize the network parameter in a larger range until the accuracy of the student network is no longer improved.

The whole loss function of the iterative process is the weighted sum of KL divergence and softmax loss, as shown in formula (4).

$$L_{tea\&stu}\left(\theta_{student}\right) = \alpha \cdot T^2 \cdot L_{KL} + (1-\alpha)\cdot L_{CE} \tag{4}$$

where $\theta_{student}$ represents the parameter weight of the student network; α is the hyperparameters of the soft label proportion; T is the hyperparameters of distillation temperature in the experiment; L_{KL} and L_{CE} correspond to the difference between the predicted output of the student network and the soft and hard labels respectively, defined as follows:

$$L_{KL} = \sum_{i=0}^{6} KLDiv\left(\hat{p}_{teacher}^{(T,\,i)}, \hat{p}_{student}^{(T,\,i)}\right) \tag{5}$$

$$L_{softmax} = \sum_{i=0}^{6} CrossEntropy\left(l_{hard}^{(i)}, \hat{p}_{student}^{(T,\,i)}\right) \tag{6}$$

$$\hat{p}_{student}^{(T,\,i)} = \frac{\exp\left(\frac{F_{student}^{(i)}}{T}\right)}{\sum_{j=0}^{6}\exp\left(\frac{F_{student}^{(j)}}{T}\right)} \tag{7}$$

To avoid the feature loss after iterations, the parameters of the student network remain unchanged during the iteration. After a single iteration, the weight of teacher network parameters is updated to the training results ($\theta_{student}$), and L_{KL} is changed to the KL divergence loss between the previous and current round of student model outputs, as shown in formula (8).

$$L_{KL} = KLDiv\left(\hat{p}_{last-student}^{(T)},\ \hat{p}_{student}^{(T)}\right) \tag{8}$$

4 | EXPERIMENTS

4.1 | Datasets

FER2013 is the competition dataset designated by International Conference on Machine Learning, collected by Google face recognition interface from the Internet. The samples are 48×48 grey images, and they are divided into the training set (28,708 images), public test set (3589 images) and private test set (3589 images). The existence of amount of noises in the dataset makes the recognition rather difficult. The average human recognition accuracy is about 65%, and some examples are shown in Figure 7.

RAF-DB is also collected and constructed by the Internet, annotated by crowdsourcing. It contains a total of 15,339 diverse facial images, all 100×100 RGB images. These samples are divided into a training set (12,271 images) and a test set (3068 images). Most of them are aligned facial images with relatively small pose changes, as shown in Figure 7.

4.2 | Implementation settings

The experiments in this paper were conducted based on PyTorch deep learning framework. We used NVIDIA GeForce RTX 2080S for training and Intel Core i7 10875H for CPU inference test in Win10.

The learning rates of lr_1 and lr_2 were 0.01 and 0.1 in the teacher network training phase and T-S iterative transfer learning phase, respectively; the batch size was set to 128. When the loss function dropped to the plateau period, the learning rate was adjusted to 0.9 times of the original. Besides, we also adopted a pre-processing strategy similar to [47]. The images were cropped and flipped horizontally from the top-left, top-right, centre, bottom-left and bottom-right, as shown in Figure 8.

4.3 | Results and analysis

To verify the effectiveness of the proposed EAN and iterative transfer learning method, we selected the widely used public datasets, FER2013 and RAF-DB, to test the accuracy.

We explored several depth separable network structures with the parameter number of about 0.05 M to ensure the extremely lightweight characteristics of the model ultimately used for recognition. Referring to the hyperparameter selection of mainstream depth separable networks, such as MobileNetV2 [30] and EfficientNet [22], we fine-tuned the output channels and expansion ratio of the convolution block to obtain different network structures (models_a-f) and tested the basic recognition performance on facial expression recognition datasets. The results are listed in Table 1.

In the table, for these models each layer has different number c of convolution output channels. The first and last layers are marked with subscripts of head and tail respectively. Similarly, the EAIRBs 1–6 shown in Figure 3 are marked as subscript 1–6, and each block has a different number e of channel expansion ratio. It can be seen that the proposed EAN structure has the minimum parameters and FLOPs (Floating Point of Operations) when the recognition accuracy is at a similar level.

Given different soft label proportions α and distillation temperature T, the iterative transfer learning method corresponding to iter_trans(a-f) and the recognition results based on EAN are listed in Table 2. As can be seen, the different proportion of soft labels and distillation temperature has some impact on the recognition accuracy of EAN on two data sets, and they were clearly improved. The iter_trans_c (corresponding to the proportion of soft labels $\alpha = 0.5$ and distillation temperature $T = 5$) had the best recognition effect. Compared with the initial EAN, the recognition accuracies on two datasets were improved by 0.89% and 2.74% respectively, and the improvement effect of iterative transfer learning method on RAF-DB was more significant than that on FER2013. This is because there are much noise and label errors in FER2013, leading to the relatively low quality of teacher model outputs.

Figures 9 and 10 are the confusion matrices on FER2013 and RAF-DB. The recognition effects of the proposed method on each expression category are demonstrated. The recognition effect of the model for "happy" and "surprised" was better, and the accuracy was over 80% on both two datasets.

4.4 | Performance comparison

A performance comparison of our method and other lightweight methods on FER2013 is shown in Table 3. Notably, there were many parameters in [32], but the model computation was reduced significantly by sampling the low-frequency information of the input image, and the FLOPs was about 23 M. Although the study in [20] reduced the number of model parameters through a large amount of channel reuse, the calculation amount was still four times that of [32]. The accuracy of our method was slightly lower than [20], but the calculation required by our model was greatly lowered and its FLOPs was only 5.054 M.

The performance comparison between our and other mainstream methods on RAF-DB is shown in Table 4. It can be seen that the recognition accuracy of the light-weight expression recognition method based on iterative transfer learning is very close to the recognition accuracy improved on the complex backbone network VGG and ResNet in [11].

It is worth noting that both the method in [25] and our method focus on using teacher networks to learn soft label knowledge and supplement the relationship among expression categories for the training of student network. However, our study is based on an extremely lightweight network and we aim to obtain more reasonable soft labels through multiple iterations and avoid the feature loss between iterations through weight transfer. In spite of the similar and competitive accuracies of these two methods, our model has significant advantages in deployment. As shown in Table 5, nearly 33 M parameters,

136 MB memory and 552 M FLOPs were reduced by our model compared with VGG-16 used in [25]. Besides, our work does not need to supplement other additional datasets, thus avoiding the complex and time-consuming pre-training process.

We compare the deployment performance of the proposed method with other backbone networks (such as VGG, ResNet, MobileNet, SqueezeNet etc.) in parameters, memory, FLOPs and latency, as shown in Table 5. For the sake of fairness, all the values were tested in the same image resolution (44 × 44). The inference latency is the average single frame time required by the model to infer 1000 times. It can be seen that the inference speed of our method was up to 1.985 ms/frame on the CPU, and the required memory was only 3.902 MB, which better meets the real-time requirements of local expression recognition than other mainstream backbone networks.

4.5 | Ablation experiments

To explore the impact and contribution of each part of our method, we also conducted ablation experiments on RAF-DB dataset. The test results are listed in Table 6, using local attention, T-S iteration and weight transfer (true-√ or false-×).

As can be seen from Table 6, the local channel attention mechanism efficiently collects the correlation features between channels by one-dimensional convolution. Only 18 additional training parameters were added, significantly improving the recognition accuracy of the backbone network. The T-S iteration introduced the soft label knowledge into the lightweight model, while the weight transfer strengthened the features during the T-S iteration. Both the T-S iteration and weight transfer methods can improve the accuracy of backbone without introducing additional parameters. Combining these three parts, the recognition accuracy of the network was the highest, which verifies the effectiveness of the work in this paper.

It should be noted that the number of T-S iterations and weight transfer required under different hyperparameters are not fixed. In Figure 11, the recognition accuracy variation on the RAF-DB dataset is demonstrated. The number of iterations corresponds to different hyperparameter schemes, that is, soft label α and distillation temperature T. It can be seen that the overall accuracy shows a tendency of first rising and then falling, and the hyperparameters corresponding to the middle position is more conducive to the overall accuracy improvement of EAN.

4.6 | Visualization experiments

To further verify the effect of the proposed method, we visualized the supervisory signals (α% soft labels and (1-α)% hard labels) of the optimal hyperparametric scheme (iter_trans_c) in the last iteration, as shown in Figure 12.

In Figure 12, the original dataset label of the sample is on the top of the image, while the right side is the weighted results of the corresponding soft and hard labels. It can be seen that the similarity relationships between facial expression categories were obtained by our method as expected, and the maximum value of the weighted result was consistent with the original annotation of the dataset. For the facial expressions such as "anger", "happiness" and "sadness", the changes are small, so the "neutral" expressions labelled by the teacher model has a relatively high value. For the "disgust" expression, the frown is more obvious and it is more similar to the "anger" expression; the movements of eyebrows, eyes and mouth in the "fear" expression are more similar to "sadness" and "disgust"; while the gestures of the eyebrows, eyes and mouth in the "fear" expression are more similar to "sadness" and "disgust"; the "surprise" expression shown in the figure is not similar to any of the individual expressions, so the probability distribution is relatively uniform.

In addition, according to the method in [50], six EAIRBs of the model were visualized by gradient weighting, and the gradient-weighted class activation mapping (Grad-CAM) visualization are shown in Figures 13 and 14. The left is the convolution blocks' thermodynamic diagrams of the original EAN, while the right corresponds to the best hyperparametric scheme (iter_trans_c) of iterative transfer learning.

By comparing the two EAIRB 6 in Figures 13 and 14, it can be seen that the significant regions obtained by EAN fitted more irrelevant information such as clothing background and hand occlusion. However, the model after iterative transfer learning concentrates more on the real face area, and the Grad-CAM is far away from the image edge and the occlusion of glasses and fingers. The model focused on the real face, thus reducing the overfitting degree of the model. The main facial regions for different postures and datasets were accurately located by the proposed method. With the deepening of the network, the significant regions are gradually fused, and the final classification is performed based on the most discriminating position in the sample.

5 | CONCLUSION

In this paper, a real-time expression recognition method is proposed for resource-constrained devices. On one hand, the efficient attention network ensures the fitting ability and deployment performance of the model. On the other hand, the T-S iterative transfer learning method improves the recognition accuracy of the lightweight model without introducing additional parameters. The experimental results on the mainstream datasets FER2013 and RAF-DB show that the proposed method achieves the balance between complexity and accuracy and meets the requirements in practical situations as well.

In the future, we plan to combine some compression technologies such as pruning, quantization and weight aggregation with ensemble learning to further improve the performance of efficient attention network, and design some applications for edge devices under resource-constrained scenes.

ACKNOWLEDGEMENTS

This work was supported by: National Natural Science Foundation of China (Nos. 62076093, 61871182), Fundamental Research Funds for the Central Universities (Nos. 2020YJ006, 2020MS099).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest to this work. The authors declare that they do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in FER2013 at https://doi.org/10.1016/j.neunet.2014.09.005, reference number [4].

The data that support the findings of this study are openly available in RAF-DB at https://doi.org/10.1109/CVPR.2017.277, reference number [5].

ORCID

Shuaitong Zhang https://orcid.org/0000-0003-1193-3159
Ke Zhang https://orcid.org/0000-0003-3271-3585

REFERENCES

1. Ekman, P., Friesen, W.V.: Facial action coding system (FACS): A technique for the measurement of facial actions. Riv. Psichiatr. 47, 126–138 (1978)
2. Lyons, M.J., Akamatsu, S., et al.: Coding facial expressions with Gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. Nara, Japan (2002)
3. Lucey, P., Cohn, J.F., et al.: The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops. San Francisco (2010)
4. Goodfellow, I.J., Erhan, D., et al.: Challenges in representation learning: A report on three machine learning contests. In: International Conference on Neural Information Processing. Daegu, pp. 117–124 (2013)
5. Li, S., Deng, W., et al.: 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. Honolulu, pp. 2852–2861 (2017)
6. Klaser, A., Marszałek, M., et al.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008–19th British Machine Vision Conference. Leeds (2008)
7. Scovanner, P., Ali, S., et al.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia. Augsburg, pp. 357–360 (2007)
8. Mohseni, S., Kordy, H.M., et al.: Facial expression recognition using DCT features and neural network based decision tree. In: Proceedings ELMAR-2013. Zadar, pp. 361–364 (2013)
9. Fan, Y., Lam, J.C., et al.: Multi-region ensemble convolutional neural network for facial expression recognition. In: International Conference on Artificial Neural Networks. Rhodes, pp. 84–94 (2018)
10. Cai, J., Meng, Z., et al.: Probabilistic attribute tree in convolutional neural networks for facial expression recognition. arXiv preprint arXiv:1812.07067 (2018)
11. Fan, Y., Li, V., et al.: Facial expression recognition with deeply-supervised attention network. IEEE Trans. Affective Comput. 4, 1–16 (2020)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. He, K. & Zhang, X. et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, pp. 770–778 (2016)
14. Qin, Z., Wu, J.: Visual saliency maps can apply to facial expression recognition. arXiv preprint arXiv:1811.04544 (2018)
15. Pramerdorfer, C., Kampel, M.: Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903 (2016)
16. Wang, Y., Wu, J., et al.: Lightweight deep convolutional neural networks for facial expression recognition. In: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP). Kuala Lumpur, pp. 1–6 (2019)
17. Arriaga, O., Valdenegro-Toro, M., et al.: Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557 (2017)
18. Cotter, S.F.: MobiExpressNet: A deep learning network for face expression recognition on smart phones. In: 2020 IEEE International Conference on Consumer Electronics (ICCE). Las Vegas, NV, pp 1–4 (2020)
19. Zhou, N., Liang, R., et al.: A lightweight convolutional neural network for real-time facial expression detection. IEEE Access 9, 5573–5584 (2020)
20. Zhao, G., Yang, H., et al.: Expression recognition method based on a lightweight convolutional neural network. IEEE Access 8, 38528–38537(2020)
21. Howard, A., Sandler, M., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, pp. 1314–1324 (2019)
22. Tan, M., Le Efficientnet, Q.: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. Faridabad, pp. 6105–6114 (2019)
23. Tan, M., Chen, B., et al.: Mnasnet: Platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, pp. 2820–2828 (2019)
24. Tan, M., Le, Q.V.: EfficientNetV2: Smaller Models and Faster Training. arXiv preprint arXiv:2104.00298 (2021)
25. Gan, Y., Chen, J., et al.: Facial expression recognition boosted by soft label with a diverse ensemble. Pattern Recognit. Lett. 125, 105–112 (2019)
26. Georgescu, M.-I., Ionescu, R.T.: Teacher-student training and triplet loss for facial expression recognition under occlusion. In: 2020 25th International Conference on Pattern Recognition (ICPR). Milan, pp. 2288–2295 (2021)
27. Chatfield, K., Simonyan, K., et al.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
28. Parkhi, O.M., Vedaldi, A., et al.: Deep face recognition. In: Proceedings of BMVC. Swansea, pp. 6–17 (2015)
29. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, pp. 1251–1258 (2017)
30. Sandler, M., Howard, A., et al.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, pp. 4510–4520 (2018)
31. Huang, G., Liu, Z., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, pp. 4700–4708 (2017)
32. Tang, Y., Zhang, X., et al.: Facial expression recognition using frequency neural network. IEEE Trans. Image Process. 30, 444–457(2020)
33. Meng, D., Peng, X., et al.: Frame attention networks for facial expression recognition in videos. In: 2019 IEEE International Conference on Image Processing (ICIP). Taipei, pp. 3866–3870 (2019)
34. Kumar, V., Rao, S., et al.: Noisy student training using body language dataset improves facial expression recognition. In: European Conference on Computer Vision. Glasgow, pp. 756–773 (2020)
35. Li, Y., Zeng, J., et al.: Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Process. 28, 2439–2450 (2018)
36. Wang, K., Peng, X., et al.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. 29, 4057–4069 (2020)
37. Li, Y., Lu, G., et al.: Facial expression recognition in the wild using multi-level features and attention mechanisms. IEEE Trans. Affective Comput. 10, 1–12 (2020)

38. Wang, Q., Wu, B., et al.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, pp. 11531–11539 (2020)
39. Hinton, G., Vinyals, O., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
40. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, pp. 4794–4802 (2019)
41. Shen, Z., He, Z., et al.: Meal: Multi-model ensemble via adversarial learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, HI, pp. 4886–4893 (2019)
42. Zhang, Y., Xiang, T., et al.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, pp. 4320–4328 (2018)
43. Xie, Q., Luong, M.-T., et al.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, pp. 10687–10698 (2020)
44. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV). Munich, pp. 3–19 (2018)
45. Ulyanov, D., Vedaldi, A., et al.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
46. Ramachandran, P., Zoph, B., et al.: Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941 7, 1 (2017)
47. Miao, S., Xu, H., et al.: Recognizing facial expressions using a shallow convolutional neural network. IEEE Access 7, 78000–78011 (2019)
48. Iandola, F.N., Han, S., et al.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360 (2016)
49. Ma, N., Zhang, X., et al.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). Munich, pp. 116–131 (2018)
50. Selvaraju, R.R., Cogswell, M., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, pp. 618–626 (2017)

How to cite this article: Kong, Y., Zhang, S., Zhang, K., Ni, Q., Han, J.: Real-time facial expression recognition based on iterative transfer learning and efficient attention network. IET Image Process. 1–15 (2022). https://doi.org/10.1049/ipr2.12441