Received: 26 June 2022 | Revised: 12 November 2022 | Accepted: 17 March 2023

DOI: 10.1049/ipr2.12798

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

A lightweight method for face expression recognition based on improved MobileNetV3

Xunru Liang[1,2,3] | Jianfeng Liang[1,2,3] | Tao Yin[1,2,3] | Xiaoyu Tang[1,2,3]

[1]School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou, China

[2]School of Electronics and Information Engineering, South China Normal University, Foshan, China

[3]National Demonstration Center for Experimental Physics Education, Guangzhou, China

Correspondence
Xiaoyu Tang, School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou, China.
Email: tangxy@scnu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 62001173; Project of Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds), Grant/Award Number: pdjh2021a0126

Abstract

Facial expression recognition plays a significant role in the application of man–machine interaction. However, existing models typically have shortcomings with numerous parameters, large model sizes, and high computational costs, which are difficult to deploy in resource-constrained devices. This paper proposes a lightweight network based on improved MobileNetV3 to mitigate these disadvantages. Firstly, we adjust the channels in the high-level network to reduce the number of parameters and model size, and then, the coordinate attention mechanism is introduced to the network, which enhances the attention of the network with few parameters and low computing cost. Furthermore, a complementary pooling structure is designed to improve the coordinate attention mechanism, which enables it to assist the network in extracting salient features sufficiently. In addition, the network with the joint loss consisting of the softmax loss and centre loss is trained, which can minimize the intra-class gap and improve the classification performance. Finally, the network is trained and tested on public datasets FERPlus and RAF-DB, with the best accuracy of 87.5% and 86.6%, respectively. The FLOPs, parameters, and the memory storage size are only 0.19GMac, 1.3 M, and 15.9 MB, respectively, which is lighter than most state-of-the-art networks. Code is available at https://github.com/RIS-LAB1/FER-mobilenet.

1 | INTRODUCTION

Facial expressions are one of the primary expressions of the rich emotional activities of humans [1]. According to the Mehrabian theorem, facial expressions can express 55% of the total information of human emotional communication, while voice and verbal content account for 38% and 7%, respectively, which indicates that human emotional communication relies mainly on the judgment of the individual's facial expression with whom they are communicating. As an important research topic in the field of computer vision and artificial intelligence, facial expression recognition is widely used in a range of applications, such as human-computer interaction [2], mental health analysis [3], and fatigue driving detection [4].

In practical applications, the models of face expression recognition are usually deployed on embedded mobile devices with poor computing power and limited storage capacity, which can reduce the cost of the application. As a result, it is necessary to keep models as simple as possible to reduce computing cost. Nevertheless, most exiting face expression recognition algorithms are implemented based on large-size models, such as the series networks of VGG and ResNet. Although these networks obtain high accuracy by adding complex structures or numerous parameters, the computing cost and the complexity of the networks are also significantly increased. In addition, these models require lots of storage resources.

Aiming at the problems of numerous parameters, large model size, and high computing cost, we proposed a new lightweight network based on the MobileNetV3 [5]. Specifically, the channels in the high-level network were modified to reduce the parameters, FLOPs, and model size. And then, in order to enhance the attention of the network to the relevant regions of expression with fewer parameters, the coordinate attention mechanism [6] was introduced into the network. Nevertheless,

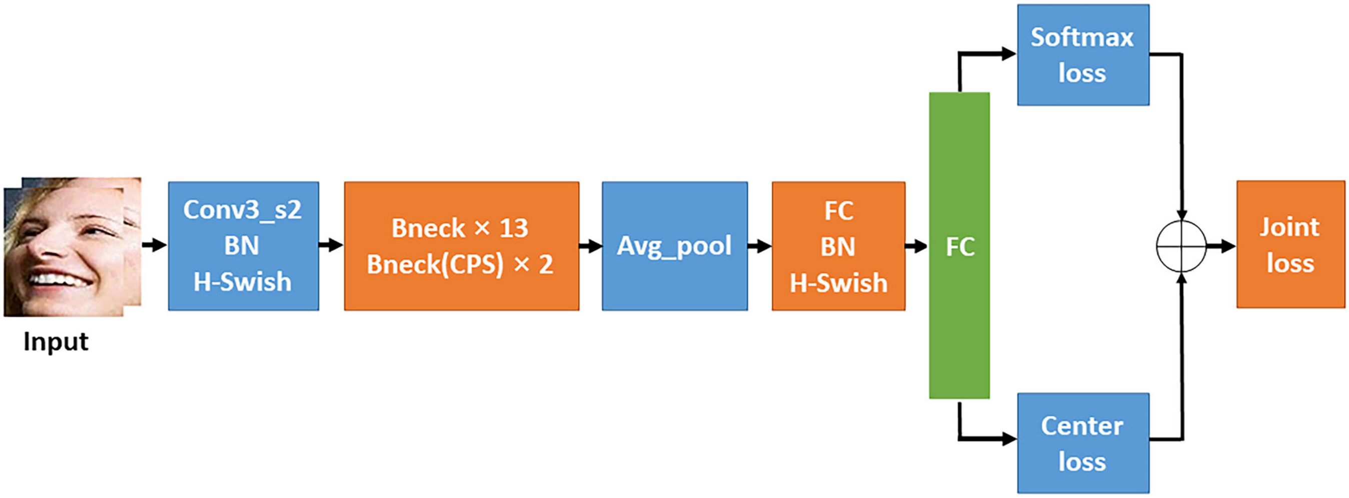

FIGURE 1 The proposed network structure is based on the MobileNetV3, with CPSCA module and joint loss added, and channels adjusted.

the coordinate attention mechanism has the shortcoming of inadequate extraction of the salient features because of the average pooling for 1D feature encoding in both W and H directions. Here, we propose a complementary pooling structure (CPS) to improve the coordinate attention mechanism. The CPS has two pooling groups, and each group has both the maximum pooling and the average pooling. After going through the group pooling, the feature vectors from each group are stitched along the spatial dimension, respectively, and then they are fused by adding. The experimental results demonstrate that the improved coordinate attention mechanism can assist the network in highlighting critical salient features, thus improving the recognition accuracy. In addition, the improved network was trained under the supervision of the joint loss consisting of softmax loss and centre loss [7], which improved the classification performance of the network. The structure diagram of the proposed network is shown in Figure 1.

The proposed methods can be summarized as follows:

(i) To reduce the number of parameters and the model size of the network, we adjust the channels in the high-level network.
(ii) We introduce the coordinate attention mechanism to the network to enhance the attention of the network to the expression region and thus improve the accuracy. Furthermore, we propose a complementary pooling structure named CPS to improve the coordinate attention mechanism, which can assist the network in extracting the prominent features.
(iii) The proposed network is trained under the supervision of the joint loss, which minimizes the intra-class gap and improves the classification performance of the network. Compared with some state-of-the-art models, the proposed network has a smaller model size, fewer parameters, and lower computing cost, while the recognition accuracy reaches the level of some state-of-the-art models, which is suitable for resource-constrained devices.

2 | RELATED WORK

Traditional methods for facial expression recognition generally include three steps: face detection, feature extraction, and feature classification. The traditional methods for feature extraction are local binary patterns (LBP) [8], Gabor wavelet [9], histogram of oriented gradient (HOG) [10] etc. When the face image is well lit and unobstructed, these methods perform well. However, when the scene changes, the key features of the face cannot be extracted automatically, making the recognition ineffective. As a result, the traditional facial expression methods are limited by the application environment.

With the development of deep learning, people have studied facial expression recognition by deep learning methods. Unlike the traditional methods, deep learning neural networks can extract valuable features automatically without manual extraction and learn the corresponding expression features automatically according to different scenarios, which performs better than the traditional methods in natural environments. Cheng et al. [11] optimized the structure and parameters of the network based on VGG-19 and addressed the lack of training samples through migration learning. Minaee et al. [12] proposed an attentional convolutional network to recognize expressions. They searched for the regions that had an essential impact on the classification results through visualization techniques. Tian et al. [13] proposed a discriminative facial expression recognition network (DFER-Net) to recognize facial expressions in the wild. Based on the traditional deep convolutional neural network, a new quadruple mean loss layer was introduced to expand the intra-class similarity and inter-class variability and thus improve the recognition ability of the network.

Although more complex and deeper network models can obtain better classification performances, the FLOPs and the number of parameters of the models are also increased, which results in consuming more storage space and computational resources. In practical application scenarios, such as healthcare [14] and transportation [15], deep learning models are usually deployed on embedded platforms with limited resources to achieve their functions. Therefore, we have to consider the number of parameters and the calculation cost of the model. Common methods of network compression include pruning, quantization, knowledge distillation, and neural structure search [16, 17].

Zhou et al. designed a real-time lightweight system, which used multi-task cascaded convolutional networks (MTCNN) to perform the task of facial expression detection [18]. In this method, a Global Average Pooling layer was used instead of a fully connected layer, and the residual modulus and depthwise separable convolution were combined, which skilfully reduced quantities of parameters and made the network structure simpler. The accuracy of this method on FER-2013 data set reached 67%. To solve the problems of single scale facial features and excessive network parameters in facial expression recognition, Hu et al. proposed a lightweight multi-scale model [19]. A lightweight convolutional neural network Xception was chosen as the backbone, and the convolutional block attention module was embedded to learn the key facial features. In addition, the depthwise separable convolution of multi-scale convolution kernel was used to expand the receptive field and extract more feature information. This method achieved the best accuracy of 68.14% on the FER2013 data set. Although the above two methods have achieved the goal of light weight, their accuracy still needs to be improved. Ale et al. proposed a lightweight facial expression recognition method based on the MobileNetV2 [20]. This model extracted features from the input images by an Inception block, and then used the backbone bottleneck layers for model compressing and fine-grained learning. After that, another CNN block expanded the learned features for a fully-connected layer to classify the facial expressions. The experimental results showed that this method could lighten the weight while keeping a high accuracy. Nevertheless, the computational cost and the complexity can be further reduced.

Here, we proposed a lightweight network for facial expression recognition. This network drastically reduces the FLOPs and parameters of the network with high accuracy, which satisfies the requirements of the resource-constrained embedded mobile devices.

3 | PROPOSAL METHOD

MobileNetV3 [5] is a lightweight network proposed by the Google team. Unlike other networks, MobileNetV3 adopts depth-wise separable convolution instead of traditional convolution, which results in a significant decrease in the parameters and a substantial reduction in computing costs. Besides, the network introduced complementary search techniques and SENet simultaneously to improve accuracy. MobileNetV3 has good performance in image classification with high recognition accuracy and fewer computing resources, making it ideal for resource-constrained mobile devices. Therefore, we propose a lightweight network based on improved MobileNetV3 for facial expression recognition. And the improvements are introduced in this section.

3.1 | Adjust channels

The bneck structure shown in Figure 2 is a unique structure in MobileNetV3, and it is the fundamental unit that comprises MobileNetV3. In this work, we modified bneck structures to lightweight the network. Since the bneck structures in the high-level network in MobileNetV3 have many output channels and large expanding sizes, the network is able to extract more high-level semantic features, thus improving accuracy. However, there are many parameters in this part, which increases the complexity and computing cost of the model.

To further satisfy the requirements of embedded devices with limited computing resources, we cut down the number of channels in the last two bneck structures by half, including input channels, expand channels, and output channels. And at the same time, we also reduced the number of output channels in the penultimate fully connected layer by half. Through these modifications, the number of parameters can be reduced significantly and the size of the model becomes smaller. Apart from that, we replaced the ReLU functions in the bneck structure with the h-swish function. Because the h-swish function can reduce the number of memory accesses and latency costs [5] and its non-linear fitting ability is better than ReLU. And the full specification of the improved networks is shown in Table 1.

3.2 | Coordinate attention and complementary pooling structure

In the original MobileNetV3, the channel attention SENet [21] is introduced to some bneck structures to improve the accuracy. The SENet module assigns different weights to each channel based on its contribution to the output, ensuring that the network prioritizes the most valuable channels. However, the SENet neglects the relationships between positions, which is a disadvantage for the facial expression recognition tasks. Since only the expression area contributes significantly to the accuracy rate, such as the mouth, cheeks, and eyes, and the non-expression area contributes little, such as the hair, forehead etc. Hou et al. [6] proposed the coordinate attention mechanism (CA), which can capture cross-channel information and obtain direction-aware and position-sensitive information, improving the accuracy of models. Besides, the CA block contains fewer parameters than the SENet block, which is beneficial for the lightweight optimization of the network. The structure of the CA block is shown in Figure 3.

The CA block uses average pooling for 1D feature encoding in both W and H directions, which can obtain a large perceptual field. However, when the average pooling results of the neighbouring rows and columns of the effective features are similar,

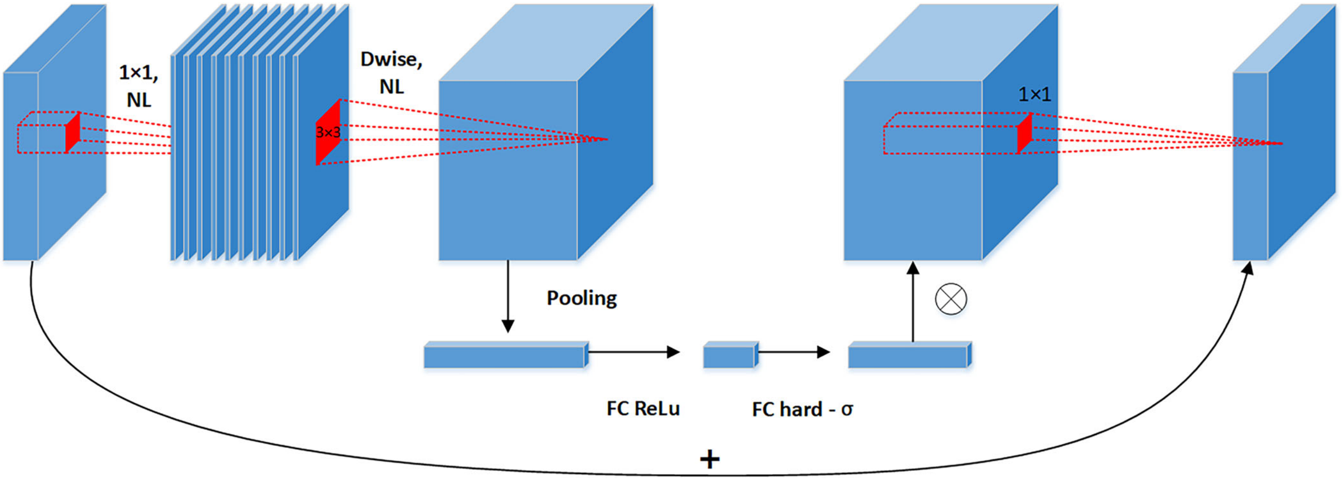

FIGURE 2 The bneck structure is the basic unit block of the MobileNetV3, including (a) the inverted residual with linear bottleneck. (b) Depthwise separable convolutions. (c) Lightweight attention module.

TABLE 1 Specification for the improved network. NL denotes the type of nonlinearity and HS denotes H-Swish. NBN denotes no batch normalization. s denotes stride.

Input	Operator	Exp size	out	Attention	NL	s
$224^2 \times 3$	Conv2d	–	16	–	HS	2
$112^2 \times 16$	Bneck,3 × 3	16	16	–	HS	1
$112^2 \times 16$	Bneck,3 × 3	64	24	–	HS	2
$56^2 \times 24$	Bneck,3 × 3	72	24	–	HS	1
$56^2 \times 24$	Bneck,5 × 5	72	40	SENet	HS	2
$28^2 \times 40$	Bneck,5 × 5	120	40	SENet	HS	1
$28^2 \times 40$	Bneck,5 × 5	120	40	SENet	HS	1
$28^2 \times 40$	Bneck,3 × 3	240	80	–	HS	2
$14^2 \times 80$	Bneck,3 × 3	200	80	–	HS	1
$14^2 \times 80$	Bneck,3 × 3	184	80	CPSCA	HS	1
$14^2 \times 80$	Bneck,3 × 3	184	80	CPSCA	HS	1
$14^2 \times 80$	Bneck,3 × 3	480	112	CA	HS	1
$14^2 \times 112$	Bneck,3 × 3	672	112	CA	HS	1
$14^2 \times 112$	Bneck,5 × 5	672	80	CA	HS	2
$7^2 \times 80$	Bneck,5 ×5	480	80	CA	HS	1
$7^2 \times 80$	Bneck,5 ×5	480	80	CA	HS	1
$7^2 \times 80$	Conv2d,1×1	–	960	–	HS	1
$7^2 \times 960$	Pool, 1 × 1	–	–	–	–	1
$1^2 \times 960$	Conv2d 1 × 1, NBN	–	640	–	HS	1
$1^2 \times 640$	Conv2d 1 × 1, NBN	–	k	–	–	1

the prominent features will receive the same level of attention as the surrounding unremarkable features, which prevents the network from clearly distinguishing the salient features from the unremarkable features and thus failing to extract the salient features adequately. When the maximum pooling is applied in both W and H directions to form a new block named MAXCA, the network cannot obtain a global perceptual view of rows and columns. Although the prominent features can receive a higher level of attention than the surrounding unremarkable features, the network is prone to locate inaccurately because the pooling results are only determined by the maximum values of rows and columns, which ignores other detailed features.

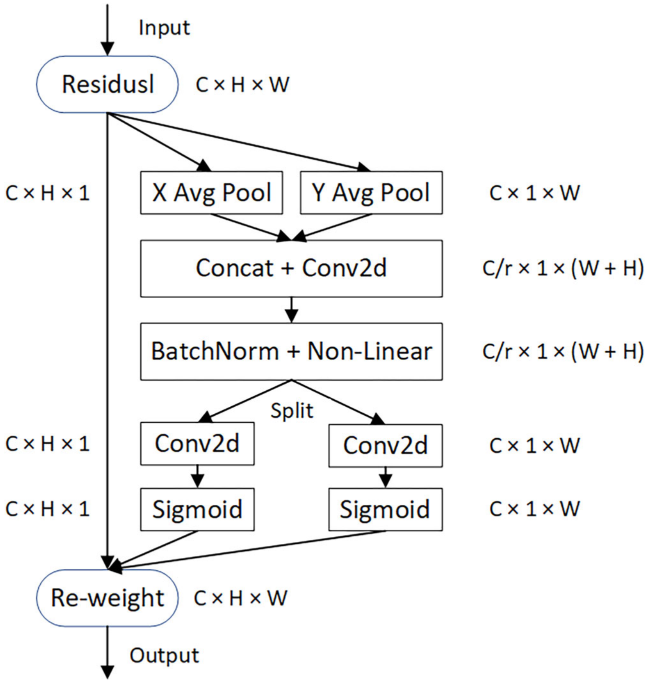

FIGURE 3 The structure of coordinate attention. Here, "X Avg Pool" and "Y Avg Pool" refer to 1D horizontal global pooling and 1D vertical global pooling, respectively.

In order to highlight the prominent features and retain detail features, a new complementary pooling structure named CPS is proposed here, which is compatible with the advantages of both the CA and the MAXCA. The structure of the CPS is shown in Figure 4.

When the first group utilizes maximum pooling in the *W*-direction, the pooling result is the value of the most prominent feature in each row. And the larger the value, the more prominent the feature is, and the more attention the row receives. Thus, the prominent features receive more attention than the adjacent unremarkable features in the *H*-direction. In the meanwhile, the average pooling is utilized in the *H*-direction, which enables the network to obtain global perceptual fields of view for each column and thus retains more detail features. Through the average pooling, only the columns with more prominent features can be given more attention, which enables the network to localize more accurately. The second group utilizes average pooling in the *W*-direction and maximum pooling in the *H*-direction. Similarly, the prominent features receive more attention than the adjacent unremarkable features in the *W*-direction. And it obtains a global perceptual field of view for each row and retains more features at the same time, which enables the network to locate accurately. The pooling results of the two groups are summed and fused after the convolution layer, which can increase the value of the prominent features and makes them more conspicuous. Although the value of unremarkable features also increases, the increase of unremarkable features is smaller than the increase of the prominent features. Consequently, the prominent features are distinguished from the surrounding unremarkable features after the batch normalization layer. In this work, the CPS was introduced into the original CA block to form a new block named Complementary Pooling Structure Coordinate Attention (CPSCA), and its structure is shown in Figure 5.

Since the middle-level network has more valuable features and fewer interference features than the low-level network, and has fewer valuable features and more interference features than the high-level network. Therefore, a block that can highlight prominent critical features is required in the middle-level network. Consequently, the CPSCA block was introduced to the middle-level network. While in the high-level network, there are more valuable features and fewer interference features. Since CPSCA adopts maximum pooling, some valuable features will be discarded when the CPSCA is introduced into the high-level network. However, the CA block applied average pooling, which can retain more valuable features. As a result, the CA blocks were introduced in the high-level network to enable the network to pay more attention to the valuable features.

3.3 | Joint loss

Softmax loss is commonly used in image classification tasks to measure the difference between the predicted probability distribution and the probability distribution of the actual labels by cross-entropy. During the training of the model, the optimal global solution is obtained to have a better model by calculating the minimum of the loss function iteration by iteration. The formula for softmax loss is as (1) and (2).

$$L_{\mathrm{s}} = -\frac{1}{\mathrm{m}}\left[\sum_{\mathrm{i}=1}^{m}\sum_{j=1}^{k}\delta\log\frac{\exp(W_j^{\,T}x_i + b_j)}{\sum_l^k \exp(W_l^{\,T}x_i + b_l)}\right], \tag{1}$$

$$\delta = \begin{cases} 1, \mathrm{y_i} = \mathrm{j} \\ 0, \mathrm{y_i} \neq \mathrm{j} \end{cases} \tag{2}$$

where x_i, y_i, k, and m stand for features, labels, the number of classification categories, and the number of training samples. W_j, W_l, b_j, and b_l are the convolutional network parameters.

The centre loss [7] can improve the separability of features by training to learn the depth feature centres of each class, and shrink the distance between the depth features of the samples and their corresponding class centres. The formula for the central loss is as (3).

$$L_c = \frac{1}{2}\sum_{i\,=\,1}^{m}\left\|x_i - c_{y_i}\right\|_2^2, \tag{3}$$

where m, x_i, and c_{y_i} stand for the number of samples in a batch, the feature vector of the ith training sample, and the y_ith class center of the deep features.

The softmax loss separates the deep features of different classes, while the centre loss pulls the deep features of the same

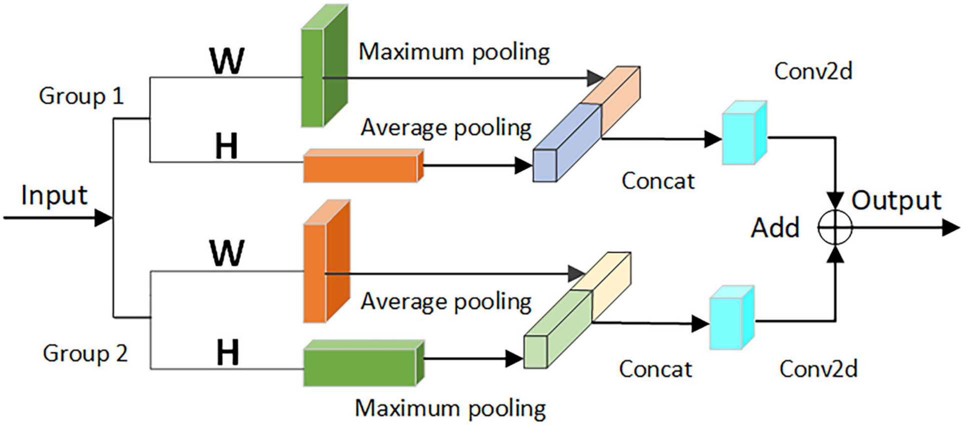

FIGURE 4 The Complementary Pooling Structure (CPS). The first group utilizes maximum pooling in the *W*-direction and utilizes average pooling in the *H*-direction. The second group utilizes average pooling in the *W*-direction and utilizes maximum pooling in the *H*-direction. The two groups form the CPS.

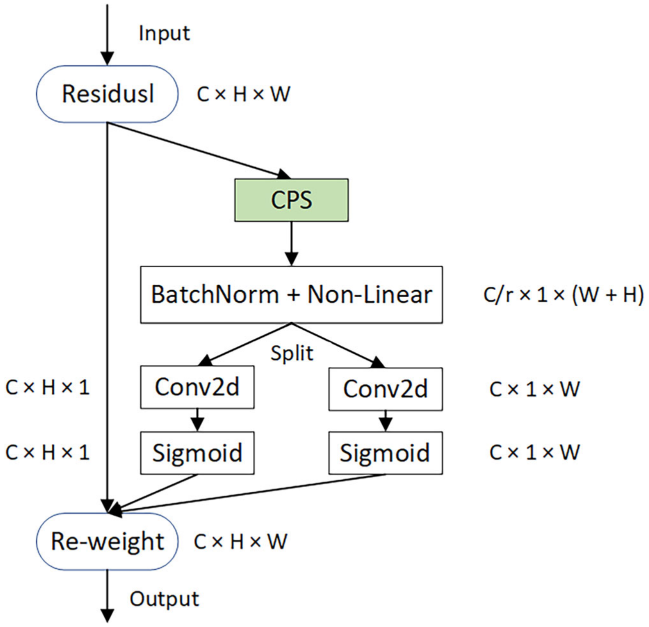

FIGURE 5 The CPS was added to the original CA module, forming CPSCA module.

class to their centres [7]. When the network is trained with only softmax loss, the performance is not good enough because of the large distance in a class. Therefore, the proposed network was trained with the joint loss consisting of the softmax loss and the centre loss, which enlarged the differences of inter-class and reduced the variation of the intra-class, thus enhancing the performance of the discriminative power of the network. The formula of the joint loss function is as (4).

$$L_j = L_s + \lambda L_c, \tag{4}$$

where λ is a hyperparameter and it can balance the contribution of softmax loss and centre loss.

4 | EXPERIMENT AND RESULTS

In this section, the experimental environment and data sets of this work will be described at first. Then, the improvements of the proposed network will be demonstrated through a series of comparative experiments.

4.1 | Experimental environment and data sets

In this work, the experimental conditions are as follows: the hardware environment is Intel(R) Core (TM) i7-9700K, 3.6 GHz, NVIDIA GeForce RTX3060 31.2GiB. The software environment for the experiment is ubuntu18.04, cuda10.2, cudnn8.3.0, python3.7.5, and Pytorch framework. We conducted the experiments based on the MobileNetV3-large model from torchvision PyTorch.

In this work, we used the Adam optimization algorithm, and the initial learning rate was set to 0.01. The dropout rate was set to 0.4 and the learning rate was adjusted in an adaptive way that the learning rate would reduce to 0.1 times of the last one when the loss value of the validation set did not decrease for ten consecutive times. The batch size was set to 64, and the epoch was set to 100.

The experiments were conducted on the publicly available data set FERPlus [22] and RAF-DB [23]. The FERPlus data set is obtained by extending the FER2013 data set. The dataset contains eight categories of expressions: surprise, fear, disgust, happiness, sadness, anger, contempt, and neutrality. It is a large data set of face expressions close to the real world, with 25,060 images in the training set, 3,199 images in the validation set, and 3,153 images in the test set. Each image was resized to 48 × 48 size after face alignment.

The RAF-DB data set has 30,000 real images containing seven categories of expressions: Surprise, fear, disgust, happiness, sadness, anger, and neutral, annotated with basic or compound expression by 40 trained human coders. In this experiment, we used the basic emotion images, with 12,271

TABLE 2 Comparison with different methods on RAF-DB data set.

Method	FLOPs (GMac)	Parameters (M)	Accuracy	Model size (MB)
VGG16	15.5	138.7	81.7%	1740.8
Resnet18	1.8	11.7	86.7%	140.8
Resnet50	4.1	25.6	86.5%	301.7
MobileNetV3	0.23	4.1	86.4%	50.1
MobileNetV3+Modify channels (Ours)	0.19	1.9	85.9%	23.5
MobileNetV3+Modify channels+H-Swish (Ours)	0.19	1.9	86.4%	23.5
Improved MobileNetV3 (Ours)	0.19	1.3	86.6%	15.9

images as the training set and 3,068 images as the test set. The size of each image was 100 × 100.

In order to enhance the robustness of the network model and to prevent overfitting, we used the online data augmentation method to extend data. The specific procedures were to randomly flip the image, translate it, rotate it at random plus or minus 10°, adjust the contrast randomly, and add Gaussian noise. The pictures for each epoch were not the same, which enabled the network to learn the effective features of faces so that the robustness and the generalization will be enhanced.

4.2 | Effectiveness of the proposed network

To verify the effectiveness of the proposed network, comparative experiments were conducted on the RAF-DB data set with different methods. The input pictures were resized to 224 × 224 before training and testing to eliminate the impact of different sizes of pictures. Apart from accuracy, we also used FLOPs, Parameters, and Model Size as comparison criteria, which were often used to measure the lightweight degree of models.

As can be seen from Table 2, the accuracy of our method is higher than that of VGG16, RestNet50, and original MobileNetV3, while slightly lower than that of ResNet18. However, compared with ResNet18, the FLOPs, the number of parameters, and the model size of the improved network are decreased by 89.6%, 88.9%, and 88.7%, respectively, while the accuracy is decreased by 0.1%. When compared with the original MobileNetV3, the FLOPs, the number of parameters, and the model size are all reduced significantly after modifying channels, while the accuracy drops slightly. When replacing the ReLU function with the H-Swish function, the accuracy is improved to some extent. The reason is that the non-linear fitting ability of the H-Swish function is better than ReLU. And after the improvements that replacing parts of SENet structures in bneck structures with CA and CPSCA blocks and training with the joint loss, the number of parameters and the model size of the network are further decreased, while the accuracy is improved. Comparing the results of different methods, the improved MobileNetV3 has optimal performance with fewer parameters, less calculation cost, and low storage cost, which is more suitable for mobile devices with limited resources.

TABLE 3 Comparison with some state-of-art methods on the RAF-DB data set.

Method	FLOPs (GMac)	Parameters (M)	Accuracy	Model size (MB)
gACNN [24]	>15.5	>134.3	85.1%	–
Separate-Loss [25]	1.8	11.2	86.4%	–
SCN [26]	1.8	11.2	87.0%	45
LDL-ALSG [27]	4.1	23.5	85.5%	–
RAN [28]	14.6	11.2	86.9%	45
CVT [29]	–	51.8	88.2%	–
Our network	0.19	1.3	86.6%	15.9

TABLE 4 Comparison with some of the latest methods on the FERPlus data set.

Method	FLOPs (GMac)	Parameters (M)	Accuracy	Model Size (MB)
SCN [26]	1.8	11.2	88.0%	45
Ensemble-based CNN [30]	–	–	87.2%	–
TAMNet [31]	–	51.7	80.6%	–
RAN [28]	14.6	11.2	88.6%	45
DICNN [32]	–	1.1	85.2%	5.4
Our network	0.19	1.3	87.5%	15.9

To further evaluate and verify the performance of the proposed network, we compared the improved network with some existing state-of-the-art models on the RAF-DB data set, such as gACNN [24], Separate-Loss [25], SCN [26], LDL-ALSG [27], RAN [28], and CVT [29]. The results are shown in Table 3.

In addition, we also compared the proposed network with some state-of-art methods on FERPlus, such as SCN [26], RAN [28], Ensemble-based CNN [30], TAMNet [31], and DICNN [32].

As shown in Tables 3 and 4, some state-of-the-art models achieve higher accuracy on the data set RAF-DB and FERPlus. However, they have shortcomings of numerous parameters, high FLOPs, and large model sizes. The accuracy of the proposed network reaches the level of existing state-of-the-art models, while the FLOPs, the number of parameters, and model sizes are all significantly smaller than that of the state-of-the-art models. The comparison results demonstrate that the improved network balances the trade-off between recognition accuracy and resource consumption. And the improved network makes it possible to achieve more efficient and accurate facial expression recognition on mobile devices with constrained resources.

4.3 | Effectiveness of CPSCA

To evaluate the effectiveness of CPSCA in the middle-level network, we removed the attention mechanism from the middle

TABLE 5 Comparisons with different attentions.

Attention	FERPlus	RAF-DB
Baseline	86.6%	85.4%
CA	87.2%	86.0%
MAXCA	87.1%	86.5%
CPSCA	87.5%	86.6%

layer bneck structure in the improved MobileNetV3 to form the baseline network, and then added CA, MAXCA, and CPSCA respectively for comparison experiments. The experiments are conducted on the FERPlus and RAF-DB data set, and the experimental results are shown in Table 5.

The experimental results in Table 5 demonstrate that the accuracy of our method is 0.9% and 1.2% higher than the baseline on the FERPlus and RAF-DB data set, and it is also better than the CA block and the MAXCA block. It can be seen that the CA block can effectively extract the channel information and location information of facial expressions, thus improving the accuracy. However, since the CA block adopts the average pooling method in both the W and H directions, it is susceptible to background interference, causing the valuable features to be submerged in the unremarkable features. And when the maximum pooling method is utilized in both the W and H directions, the network tends to overlook the relatively ordinary data, resulting in the loss of feature information. The CPSCA adopted the complementary pooling method, which allowed the maximum pooling operation in both W and H directions to execute in parallel with the average pooling operation, so that it can identify expression regions more accurately and mitigates the loss of feature information. Through the comparison results, it can be seen that the CPSCA block performs better in prominent feature extraction and mitigate the loss of feature information, which is significant for improving the accuracy of facial expression recognition.

4.4 | Effectiveness of joint loss

Due to the different data sets, the weight value of λ in joint loss is also different. In order to find the best λ value to improve the recognition accuracy, experiments with different λ values were carried out in this paper. And the experiment result is shown in Figure 6.

It can be seen from Figure 6 that the proposed network achieves the highest accuracy on the FERPlus and RAF-DB data set when the λ value is 0.4 and 0.1, respectively. That is because the FERPlus data set contains eight categories, where some contempt images are similar to the neutral images, the weight of centre loss needs to be increased to reduce the intra-class gap and then distinguish the contempt images from the neutral images. While the RAF-DB data set contains seven categories and the differences between each category are obvious, so the weight of central loss is smaller.

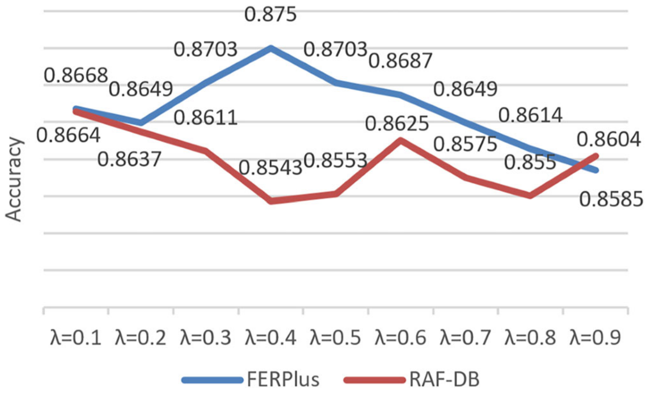

FIGURE 6 The accuracies of the network with different values of λ. When the λ value is set to 0.4 and 0.1, the proposed network achieves the highest accuracy on the FERPlus and the RAF-DB data set, respectively.

TABLE 6 Comparison with softmax loss and joint loss.

Loss	FERPlus	RAF-DB
softmax	87.2%	86.4%
centre	86.4%	85.6%
Triplet [33]	87.0%	86.0%
softmax + center	87.5%	86.6%
softmax + Triplet	87.1%	86.5%

We conducted comparative experiments on FERPlus data set and RAF-DB data set to verify the effectiveness of joint loss. Compared with centre loss, the proposed method had obvious advantages on both data sets, and the accuracy was improved by 1.1% and 1%, respectively. Compared with softmax loss, joint loss was also slightly ahead. Besides, Triplet loss [33] was also a common method in facial recognition tasks, which can reduce the distance between positive samples and anchor while expanding the distance between negative samples and anchor. However, it can be seen from Table 6 that the performance of triplet loss is not as good as that of joint loss. The experimental results indicates that the joint supervision of softmax loss and centre loss can enhance the separability of facial expression depth features and improve the accuracy more effectively than a single loss function.

5 | CONCLUSION

This paper proposed a lightweight network based on improved MobileNetV3-large for facial expression recognition. We adjusted the channels in the high-level network to reduce parameters and model size, and then, we introduced the coordinate attention mechanism to enhance the attention of the network. Aiming at the shortcomings of the coordinate attention mechanism, we proposed a complementary pooling structure to improve it. Finally, we trained and tested the network with the supervision of the joint loss. The experimental results demonstrate that the proposed lightweight network achieves high accuracy with few parameters, low FLOPs, and small model size, which is suitable for resource-constrained devices.

In the future, we will produce an embedded facial expression recognition system and deploy it on mobile devices in combination with the proposed network to achieve real-time facial expression recognition.

AUTHOR CONTRIBUTIONS

Xunru Liang: Formal analysis, Investigation, Methodology, Software, Validation, Writing—original draft, Writing—review and editing. Jianfeng Liang: Formal analysis, Investigation, Methodology, Software, Validation, Writing—original draft. Tao Yin: Formal analysis, Investigation, Validation, Writing—original draft. Xiaoyu Tang: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review and editing.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 62001173), the Project of Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) (Grant No. pdjh2021a0126).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Xiaoyu Tang https://orcid.org/0000-0002-6038-9623

REFERENCES

1. Bartlett, M.S., et al.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, pp. 568–573 (2005). https://doi.org/10.1109/CVPR.2005.297
2. Barentine, C., et al.: A VR teleoperation suite with manipulation assist. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. HRI' 21 Companion, pp. 442–446. Association for Computing Machinery, Boulder, CO, USA (2021). https://doi.org/10.1145/3434074.3447210
3. Fei, Z., et al.: Deep convolution network based emotion analysis towards mental health care. Neurocomputing 388, 212–227 (2020). https://www.sciencedirect.com/science/article/pii/S0925231220300783
4. Joshi, A., et al.: In-the-wild drowsiness detection from facial expressions. In: 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, pp. 207–212 (2020)
5. Howard, A., et al.: Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 1314–1324 (2019). https://doi.org/10.1109/ICCV.2019.00140
6. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 13708–13717 (2021). https://doi.org/10.1109/CVPR46437.2021.01350
7. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds). Computer Vision - ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9911. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
8. Li, X., et al.: Heat kernel based local binary pattern for face representation. IEEE Signal Process Lett. 17(3), 308–311 (2010). https://doi.org/10.1109/LSP.2009.2036653
9. Jones, J.P., Palmer, L.A.: An evaluation of the two- dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58(6), 1233–1258 (1987) PMID: 3437332, https://doi.org/10.1152/jn.1987.58.6.1233. eprint
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). pp. 886–893 (2005) https://doi.org/10.1109/CVPR.2005.177
11. Cheng, S., Zhou, G.: Facial expression recognition method based on improved VGG convolutional neural network. Int. J. Pattern Recognit. Artif. Intell. 34(7), 2056003 (2020) https://doi.org/10.1142/S0218001420560030
12. Minaee, S., Minaei, M., Abdolrashidi, A.: Deep-Emotion: Facial expression recognition using attentional convolutional network. Sensors 21(9), 3046 (2021). https://www.mdpi.com/1424-8220/21/9/3046
13. Tian, Y., Li, M., Wang, D.: DFER-Net: Recognizing facial expression in the wild. In: 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, pp. 2334–2338 (2021) https://doi.org/10.1109/ICIP42928.2021.9506770
14. Sertic, P., et al.: Intelligent real-time face-mask detection system with hardware acceleration for COVID-19 mitigation. Healthcare 10(5), 857 (2022). https://www.mdpi.com/2227-9032/10/5/873
15. Naimi, H., Akilan, T., Khalid, M.A.S.: Fast traffic sign and light detection using deep learning for automotive applications. In: 2021 IEEE Western New York Image and Signal Processing Workshop (WNYISPW), Rochester, NY, USA, pp. 1–5 (2021). https://doi.org/10.1109/WNYISPW53194.2021.9661284
16. Wang, C.-H., et al.: Lightweight deep learning: An overview. IEEE Consum. Electron. Mag. 1–12 (2022), https://doi.org/10.1109/MCE.2022.3181759
17. Liu, D., et al.: Bringing AI to edge: From deep learning's perspective. Neurocomputing 485, 297–320 (2022). https://www.sciencedirect.com/science/article/pii/S0925231221016428
18. Zhou, N., Liang, R., Shi, W.: A lightweight convolutional neural network for real-time facial expression detection. IEEE Access 9, 5573–5584 (2021). https://doi.org/10.1109/ACCESS.2020.3046715
19. Ale, L., Fang, X., Chen, D., Wang, Y., Zhang, N.: Lightweight Deep Learning model for facial expression recognition. In: 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, pp. 707–712 (2019). https://doi.org/10.1109/TrustCom/BigDataSE.2019.00100
20. Hu, Z., Yan, C.: Lightweight multi-scale net-work with attention for facial expression recognition. In: 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). pp. 695–698 (2021) https://doi.org/10.1109/AEMCSE51986.2021.00143
21. Hu, J., et al.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42(8), 2011–2023 (2020). https://doi.org/10.1109/TPAMI.2019.2913372
22. Barsoum, E., et al.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. ICMI '16. Tokyo, Japan, pp. 279–283 (2016). https://doi.org/10.1145/2993148.2993165
23. Li, S., Deng, W., Du, J.: Reliable crowd-sourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2584–2593 (2017). https://doi.org/10.1109/CVPR.2017.277
24. Li, Y., et al.: Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Process. 28(5), 2439–2450 (2019). https://doi.org/10.1109/TIP.2018.2886767

25. Li, Y.-J., et al.: Separate loss for basic and compound facial expression recognition in the wild. In: Proceedings of The 11th Asian Conference on Machine Learning, ACML 2019, Nagoya, Japan, pp. 897–911 (2019)
26. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 6896–6905 (2020). https://doi.org/10.1109/CVPR42600.2020.00693
27. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 13981–13990 (2020). https://doi.org/10.1109/CVPR42600.2020.01400
28. Wang, K., et al.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. 29, 4057–4069 (2020). https://doi.org/10.1109/TIP.2019.2956143
29. Ma, F., Sun, B., Li, S.: Facial expression recognition with visual transformers and attentional selective fusion. IEEE Trans. Affective Comput. (2021). https://doi.org/10.1109/TAFFC.2021.3122146
30. Siqueira, H., Magg, S., Wermter, S.: Efficient facial feature learning with wide ensemble-based convolutional neural networks. arXiv: 2001.06338. https://arxiv.org/abs/2001.06338
31. Shao, J., Luo, Y.: TAMNet: Two attention modules-based network on facial expression recognition under uncertainty. J. Electron. Imaging 30(3), 033021 (2021). https://doi.org/10.1117/1.JEI.30.3.033021
32. Saurav, S., et al.: Dual integrated convolutional neural network for real-time facial expression recognition in the wild. Visual Computer 38, 1083–1096 (2022). https://doi.org/10.1007/s00371-021-02069-7
33. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering.h In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682

How to cite this article: Liang, X., Liang, J., Yin, T., Tang, X.: A lightweight method for face expression recognition based on improved MobileNetV3. IET Image Process. 17, 2375–2384 (2023). https://doi.org/10.1049/ipr2.12798