Hindawi
International Journal of Aerospace Engineering
Volume 2022, Article ID 6523234, 15 pages
https://doi.org/10.1155/2022/6523234

Research Article

Multiscale Feature Fusion Attention Lightweight Facial Expression Recognition

Jinyuan Ni,[1] **Xinyue Zhang**,[2] **and Jianxun Zhang**[1]

[1]*College of Computer Science and Engineering, Chongqing University of Technology, 400054, China*
[2]*Sydney Smart Technology College, Northeastern University, 066004, China*

Correspondence should be addressed to Jianxun Zhang; zjx@cqut.edu.cn

Received 27 June 2022; Revised 19 July 2022; Accepted 2 August 2022; Published 24 August 2022

Academic Editor: Qing Gao

Facial expression recognition based on residual networks is important for technologies related to space human-robot interaction and collaboration but suffers from low accuracy and slow computation in complex network structures. To solve these problems, this paper proposes a multiscale feature fusion attention lightweight wide residual network. The network first uses an improved random erasing method to preprocess facial expression images, which improves the generalizability of the model. The use of a modified depthwise separable convolution in the feature extraction network reduces the computational effort associated with the network parameters and enhances the characterization of the extracted features through a channel shuffle operation. Then, an improved bottleneck block is used to reduce the dimensionality of the upper layer network feature map to further reduce the number of network parameters while enhancing the network feature extraction capability. Finally, an optimized multiscale feature lightweight attention mechanism module is embedded to further improve the feature extractability of the network for human facial expressions. The experimental results show that the accuracy of the model is 73.21%, 98.72%, and 95.21% on FER2013, CK+ and JAFFE, respectively, with a covariance of 10.14 M. Compared with other networks, the model proposed in this paper has faster computing speed and better accuracy at the same time.

1. Introduction

In recent years, with the rapid development of space technology, human-robot interaction in on-orbit service (OOS) space robots has become an important research area in space technology [1–3]. Although the intelligence of space robots is limited, space human-computer interaction plays an important role in space mission applications. Space robots can replace or assist astronauts in various on-board/off-board activities, and it is particularly important for space robots to recognise astronaut commands [4]. Facial expression recognition by astronauts is a widely used method of human-robot interaction in space that does not rely on the highly intelligent capabilities of space robots, and it can effectively combine the decision-making capabilities of humans with the precise operational capabilities of space robots to improve their operational capabilities [5–7]. The accuracy of the astronaut's facial expression recognition and the size of the expression recognition model are important indications of the increased efficiency of space robots.

Early on in the research process, the features of facial expressions were basically extracted manually, but the recognition accuracy was not high because the facial expressions in the natural environment were easily affected by many factors, such as occlusion, background, and pose [8]. In recent years, deep learning has achieved major breakthrough results in image recognition. Sun et al. [9] designed a facial expression recognition system combining shallow and deep features with an attention mechanism and proposed an attention mechanism model based on the relative positions of facial feature points and textural features of local regions of faces for better extraction of shallow features. Wenmeng and Hua [10] proposed a new end-to-end coattentive multitasking convolutional neural network that consists of a channel coattentive module and a spatial coattentive module. Their approach demonstrates better performance relative to single tasking and multitasking. Shi et al. [11]

proposed a facial expression recognition method based on a multibranch cross-connected convolutional neural network, which was built based on residual connections, network-in-network, and tree structure combined; it also added fast cross-connections for the summation of the convolutional output layer, which makes the data flow between networks smoother and improves the feature extractability of each sensory domain. Kong et al. [12] proposed a lightweight facial expression recognition method based on an attention mechanism and key region fusion, and to reduce the computational complexity, a lightweight convolutional neural network was used as the basic recognition model for expression classification, which reduces the computational effort of the network to some extent. Zhou et al. [13] designed a lightweight convolutional neural network that uses a multitask cascaded convolutional network to accomplish face detection and combines a residual module and a depthwise separable convolutional module to reduce a large number of parameters of the network and make the model more portable.

Although most of the above studies were able to extract features and lighten the model to some extent, there are still shortcomings. For example, the face acquisition process is susceptible to factors, such as lighting, background, and pose, resulting in a reduced learning ability of the model when training the face sample set and insufficient feature extractability. The number of network layers of the deep learning model also affects the accuracy of classification recognition to a certain extent, i.e., as the number of network layers increases, the phenomenon of gradient disappearance occurs, causing a decrease in recognition accuracy. To solve the above problems, this paper proposes a multiscale feature fusion attention lightweight network, making the following main contributions.

First, during the image preprocessing stage, a random erasing method based on data labels is used to mask the facial expression images to expand the training set samples and improve the robustness of the model.

Second, to further extract the deep features of facial expressions, an improved convolutional block attention module (CBAM) is embedded in the model, which rerepresents the features of facial expressions in both channel and spatial dimensions.

Third, to solve the problem of model redundancy caused by too many convolutional layers, the improved bottleneck layer is used to reduce the dimensionality of the network, which saves the computation of the network and increases the nonlinear expression capability of the model.

Fourth, to lighten the model, an improved depthwise separable convolution module is added to reduce the number of parameters computed by the network while speeding up the network operations.

Finally, through comparison with different network models, it can be verified that the model proposed in this paper has higher accuracy and lightness.

2. Related Work

2.1. Spatial/Channel Attention Mechanism [14, 15]. CBAM is a lightweight module that combines channel attention and spatial attention to dramatically improve model performance while requiring a small amount of computation and a small number of parameters. The channel attention mechanism [16–18] focuses on which channel features are meaningful using global average pooling and global maximum pooling to obtain two feature maps and then feeds them sequentially into a weight-sharing multilayer perceptron with a 1×1 convolution to better fuse channel information. The spatial attention mechanism [19, 20] focuses on spatial features, mainly on the part of the input image that is richer in effective information. One of the pooling operations is performed along the channel axis, i.e., each pooling compares values between different channels rather than values in different regions of the same channel.

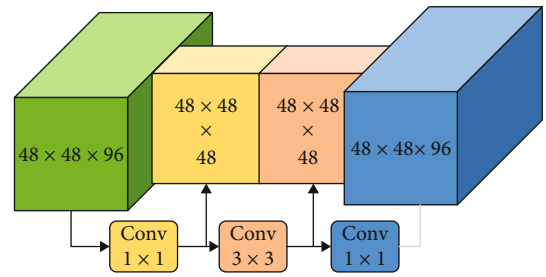


Figure 1: Structure of the bottleneck.

2.2. BottleNeck Layer. The bottleneck layer [21] is the core structure of the residual network [22], which mainly contains three convolutional layers, as shown in Figure 1. The size of the convolution kernel in the first layer is 1×1, which is mainly aimed at reducing the dimensionality of the feature map and thus the number of network parameters. The size of the convolution kernel in the second layer is 3×3, and the main purpose is to extract deeper semantic information without enhancing the number of network parameters. The convolutional kernel size of the third layer is 1×1, and the main purpose is to updimension the feature map to obtain the desired dimension size.

The formula for calculating the number of parameters during the conventional convolution operation is shown in

$$\text{Params} = C_{\text{in}} \times k \times k \times C_{\text{out}}, \tag{1}$$

where Params represents the number of parameters in the convolution process, C_{in} and C_{out} represent the number of channels of the input and the output feature map, respectively, and k represents the size of the convolutional kernel.

Assuming that the size of the intermediate feature map channels is C_{mid}, the number of parameters during the bottleneck operation is shown in

$$P_{\text{BottleNeck}} = C_{\text{in}} \times 1 \times 1 \times C_{\text{mid}} + C_{\text{mid}} \times 3 \times 3 \times C_{\text{mid}} + C_{\text{mid}} \times 1 \times 1 \times C_{\text{out}}. \tag{2}$$

From the calculation of the input and the output feature map sizes in Figure 1, the number of parameters

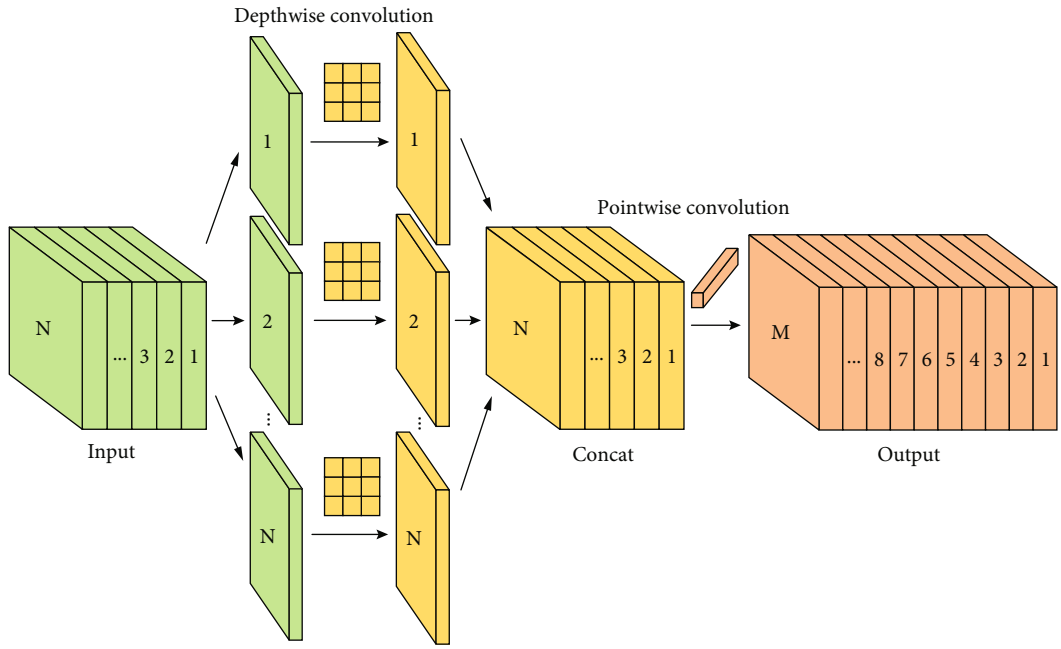


Figure 2: Structure of depthwise separable convolution.

generated by the regular convolution operation process can be obtained:

$$P_{\text{Conv}} = 96 \times 3 \times 3 \times 96. \tag{3}$$

The number of parameters generated by the bottleneck layer is:

$$P_{\text{BottleNeck}} = 96 \times 48 + 48 \times 9 \times 48 + 48 \times 96. \tag{4}$$

By comparing the two, the number of parameters generated during the bottleneck operation is greatly reduced.

2.3. Depthwise Separable Convolution. Depthwise separable convolution [23] is the core structure of the lightweight network MobileNet [24, 25], which is a combination of two parts: depthwise convolution and pointwise convolution. The specific structure is shown in Figure 2. Depthwise separable convolution contains a lower number of parameters and lower computational cost than the conventional convolution operation process. The number of convolution kernels in depthwise convolution is the same as the number of channels in the previous layer, and one convolution kernel is responsible for one channel. The number of channels in the feature map generated by this process is the same as the number of input channels, which cannot extend the dimensionality of the feature map, and the convolution operation for each channel independently cannot effectively use the feature information of different channels at the same spatial location. Pointwise convolution [26] mainly uses a 1×1 convolution to combine the feature maps obtained in the previous step in a weighted manner in the depth direction.

We assume that the input feature map size is $W_i \times H_i \times C$; W_i, H_i, and C represent the width, height, and the number of channels of the input feature map, respectively. The standard convolution size is $W_c \times H_c \times C \times H$, which denotes the width, height, the number of channels, and the number of convolution kernels of the conventional convolution, respectively, and the size of the output feature map after the conventional convolution operation is $W_o \times H_o \times N$. Then, the computation of the regular convolution is:

$$F_{\text{Conv}} = W_o \times H_o \times N \times W_c \times H_c \times C. \tag{5}$$

The depthwise separable convolution first uses a convolution size of $W_c \times H_c \times C \times 1$ convolution for depthwise convolution and then uses $1 \times 1 \times C \times N$ of the convolution for pointwise convolution. The depthwise separable convolution is computed as:

$$F_{\text{DSC}} = W_o \times H_o \times 1 \times W_c \times H_c \times C + W_o \times H_o \times N \times 1 \times 1 \times C. \tag{6}$$

The ratio of the two is:

$$\frac{F_{\text{DSC}}}{F_{\text{Conv}}} = \frac{1}{N} + \frac{1}{W_c \times H_c}. \tag{7}$$

If the size of the input feature map is $48 \times 48 \times 96$ and the size of the convolution kernel is $3 \times 3 \times 96$, then the ratio of the parameter computation is $(1/96) + (1/9)$. Therefore, if the depthwise separable convolution is used instead of the regular convolution, the computation is reduced by a factor of nearly 9.

2.4. Wide Residual Neural Network Model. To resolve the problem of gradient disappearance caused by increasing depth in deep neural networks, a residual learning unit is

introduced to more easily optimize deep networks by adjusting the relationship between the input and output through constant mapping. In the ResNet residual learning unit, the neural network input is x, while the best mapping is $H(x)$, $F(x)$ denotes ResNet Function, after the nonlinear convolution layer to achieve $F(x) = H(x) - x$, the constant mapping of itself is expressed as $H(x) = F(x) + x$. This constant mapping can then reduce the complexity and the computation of the model and, to a certain extent, mitigate problems such as gradient disappearance caused by stacking with the number of layers. However, the deep residual network pursues network depth too much, and the performance of the model does not improve considerably as the number of modules increases. The Wide ResNet residual learning module [27] adds a factor k to the original residual module to widen the number of convolution kernels [28], which reduces the number of layers, where k denotes the number of multiples of filters in the convolution layer. However, it does not reduce the model parameters, and it speeds up the computation, making it easier for the stacking layer to learn new features from the input image features.

The residual learning unit is shown in Figure 3, where dropout regularization prevents overfitting of the model and ReLU denotes the activation function; a is the ResNet residual learning module, and b is the Wide ResNet residual learning module.

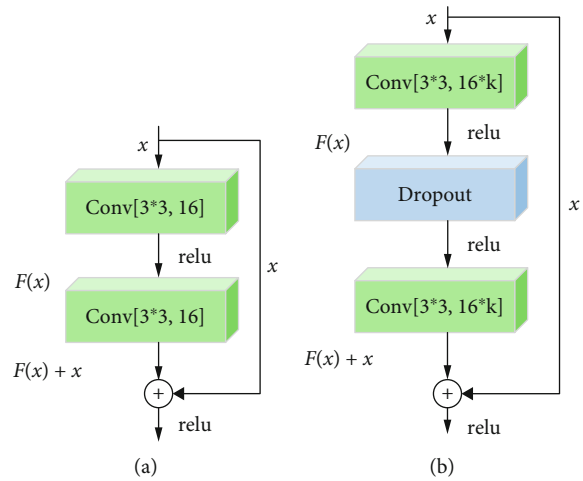


Figure 3: Learning module.

3. Methods

3.1. Overall Architecture. Since too few layers of a fully connected neural network will lead to insufficient feature representation of facial expressions in the model, too many layers will increase the computation of the network and cause the problem of network redundancy. This paper combines the above problems and designs a multiscale feature fusion attention lightweight facial expression recognition network. In the image preprocessing stage, noise is added to the training set by an improved random erasing method, which enhances the robustness of the model while enriching the entire dataset.

Then, the preprocessed facial expression images are passed into the network. First, the number of parameters of the model is reduced by the depthwise separable shuffle module to speed up the computing speed of the network. The SCAM is embedded in the middle, and then, the network is characterized by the grouping bottleneck module to reduce the dimensionality of the network, which saves the computation of the network and increases the nonlinear expression capability of the model. Then, it passes through the depthwise separable shuffle module and finally enters the Softmax layer to classify the output results. The overall architecture of the model is shown in Figure 4.

The input object image size of this network is 48×48, and the number of channels is 3. Each convolutional layer is followed by a BN layer and a ReLU activation function layer. The BN layer accelerates the training and convergence of the network and prevents the gradient from disappearing to a certain extent. To improve the feature representability of the network, a SCAM is provided behind each packet bottleneck module and depthwise separable shuffle module. After entering the grouping bottleneck module, the dimensionality of the output is halved, the number of channels is doubled, the risk of overfitting is reduced, and the parameters of the computation are reduced by the Global-Ave-Pooling layer. Finally, the pictures are classified by the Softmax layer, and the categories contain a total of 7 categories: angry, disgusted, scared, happy, sad, surprised, and neutral. The model parameters are shown in Table 1.

3.2. Image Preprocessing. Data enhancement is a common method in the image preprocessing stage, which mitigates the overfitting of the model and improves its generalizability to a certain extent. This paper expands the training set samples and enhances the robustness of the model by adding a small amount of noise to the images through an improved random erasing method [29].

First, in the preprocessing stage, the probability of random erasing of the object image is set as p, the area of the original image is set as S, the minimum and the maximum thresholds of the random erasing image are set as S_l and S_h, respectively, the aspect ratio of the occlusion matrix is set as r_e, the area of random erasing is set as S_e, the height of the area of the random erasing matrix is set as H_e, and the width of the area of the random erasing matrix is set as W_e. An example of the random erasing formula is as follows:

$$S_e = S \times \text{Random}(S_l, S_h), \tag{8}$$

$$H_e = \sqrt{S_e \times r_e}, \tag{9}$$

$$W_e = \sqrt{\frac{S_e}{r_e}}. \tag{10}$$

Among them, the specific parameters of random erasing are set, as shown in Table 2.

A randomly selected point $P_e = (x_e, y_e)$ on the image, x_e and y_e, is bounded by the following example, where W

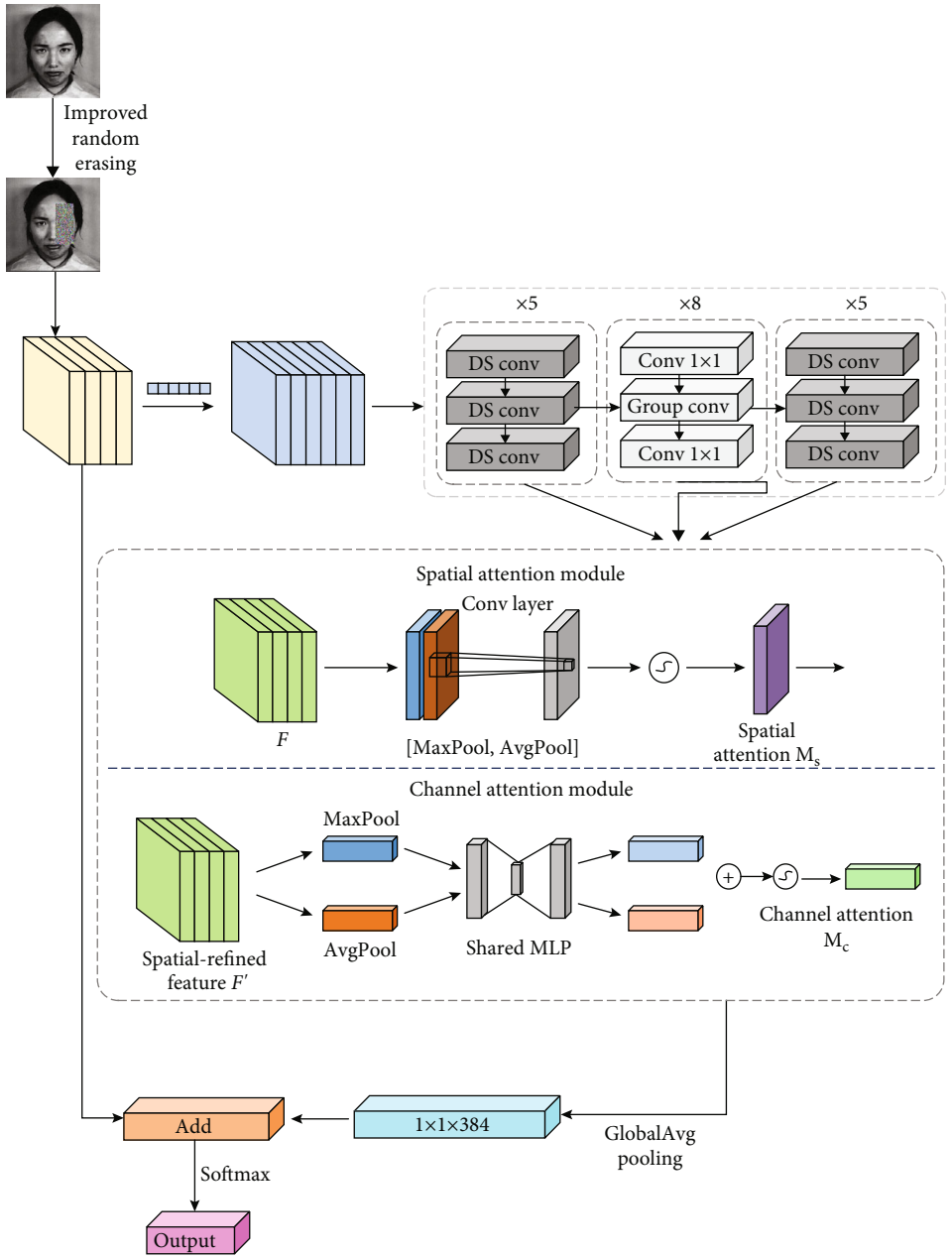


Figure 4: Overall architecture.

indicates the width of the image and H indicates the height of the image.

$$\begin{Bmatrix} x_e = \text{random}(0, W) \\ y_e = \text{random}(0, H) \end{Bmatrix}. \tag{11}$$

Since the background noise of the facial expression pictures affects the accuracy of recognition and the random erasing processing does not necessarily cover the facial expression region, causing redundancy in the original dataset, the random erasing method is improved to ensure that the random erasing region must be at the face location, and the coordinate values of x_e and y_e are requalified, for example, as follows:

$$\begin{Bmatrix} x_e = \text{random}(R_x, R_{xl}) \\ y_e = \text{random}(R_y, R_{yl}) \end{Bmatrix}, \tag{12}$$

where R_x, R_y, R_{xl}, and R_{yl} denote the true coordinate values of the upper left vertex and the upper right vertex of the face image range, respectively. By limiting the selection range of the random point P_e points so that each random erasing can cover the facial expression range, the random erasing

Table 1: Model parameters.

Light-NTWRN				
Type	Filters	Size	Output	Repetition
Input	—	—	48 × 48 × 3	—
Conv	3 × 3	16	48 × 48 × 16	—
BN+ReLU	—	—	48 × 48 × 16	—
DS-1	3 × 3	96	48 × 48 × 96	5
BN+ReLU			48 × 48 × 96	
DS-2	3 × 3	96	48 × 48 × 96	
BN+ReLU+dropout			48 × 48 × 96	
DS-3	3 × 3	96	48 × 48 × 96	
BN+ReLU+SCAM			48 × 48 × 96	
Conv-1	1 × 1	192	24 × 24 × 192	8
BN+ReLU			24 × 24 × 192	
GConv-1	3 × 3	192	24 × 24 × 192	
BN+ReLU+dropout			24 × 24 × 192	
Conv-2	1 × 1	192	24 × 24 × 192	
BN+ReLU+SCAM			24 × 24 × 192	
DS-4	3 × 3	384	12 × 12 × 384	5
BN+ReLU			12 × 12 × 384	
DS-5	3 × 3	384	12 × 12 × 384	
BN+ReLU+dropout			12 × 12 × 384	
DS-6	3 × 3	384	12 × 12 × 384	
BN+ReLU+SCAM			12 × 12 × 384	
GlobalAvg pooling	—	—	1 × 1 × 384	—
Softmax	—	—	1 × 1 × 7	—

Table 2: Random erasing parameters.

Parameter	Value
p	0.5
S_l	0.05
S_h	0.3
r_e	0.3

method and the improved method are compared, as shown in Figure 5.

As seen from Figure 5, the improved method can ensure that each random erasing is within the range of facial expressions, artificially extends the dataset of training samples, improves the robustness of the model, and effectively reduces the risk of model overfitting.

3.3. Spatial Channel Attention Module (SCAM). To further extract the deep features of different facial expressions and improve the accuracy of facial expression recognition, this paper improves the lightweight attention module (convolutional block attention module) proposed by Woo et al. [30]. This is a simple and effective attention module for convolutional neural networks. Given an intermediate feature map, our module sequentially generates attention maps along two separate dimensions, channel and space, and then multiplies the attention map into the input feature map for adaptive feature refinement. Because SCAM is a lightweight, general-purpose module, it can be seamlessly integrated into any CNN architecture with negligible computational cost. Since convolutional operations extract information features by mixing cross-channel and spatial information, we use our modules to emphasise features that are meaningful in these two main dimensions: the channel and the spatial axis. To achieve this, we apply the channel and spatial attention modules in turn, so that each branch can learn what and where to pay attention to on the channel and spatial axes, respectively. Our modules thus effectively aid the flow of information in the network by learning which information needs to be emphasised or suppressed. The features of the object image are represented in two dimensions, spatial and channel, first by the spatial attention module and then by the channel attention module, and finally, the generated features are obtained. The structure of the SCAM proposed in this paper is shown in Figure 6.

The proposed SCAM contains two independent submodules, the spatial attention module and the channel attention module, which perform feature extraction on space and channels. The input feature map F is passed through the two attention modules first, and then, the final features are output F'', $M_s(F)$ indicates that the feature map F has passed the spatial attention mechanism, $\otimes$ is multiplied by the corresponding element, and F' indicates the output feature map after passing the spatial attention mechanism; $M_c(F')$ indicates that the feature map F has passed the channel attention mechanism, and F'' indicates the output feature map after passing the SCAM attention mechanism, as shown in the following example:

$$F' = M_s(F) \otimes F, \tag{13}$$

$$F'' = M_c\left(F'\right) \otimes F'. \tag{14}$$

(1) Spatial attention module

In the process of facial expression recognition, different expressions are associated with specific regions. Moreover, an overall facial expression consists of several regions, and more attention needs to be paid to the local features with the highest expression relevance. The SCAM is shown in Figure 7.

First, the input feature map will perform global max pooling and global average pooling, followed by a CONCAT operation based on the channel and a 7×7 convolutional dimensionality reduction, and finally, it will generate the spatial attention feature by sigmoid normalization, where $\mathrm{MaxPool}(F)$ denotes the global max pooling, $\mathrm{AvgPool}(F)$ denotes the global average pooling, $f^{7\times7}$ denotes the convolution kernel for 7×7 size, σ is the sigmoid function, and

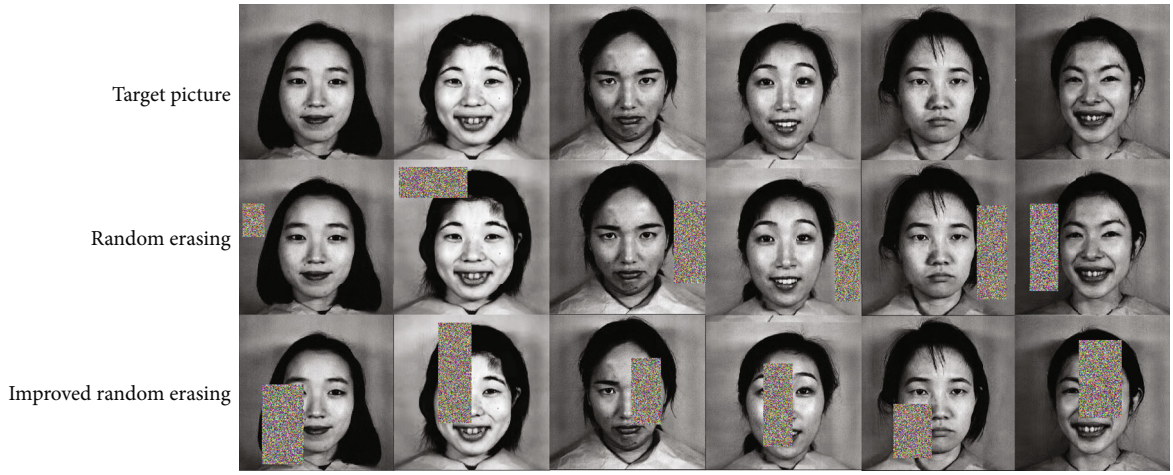


Figure 5: Comparison chart of the random erasing experiment.

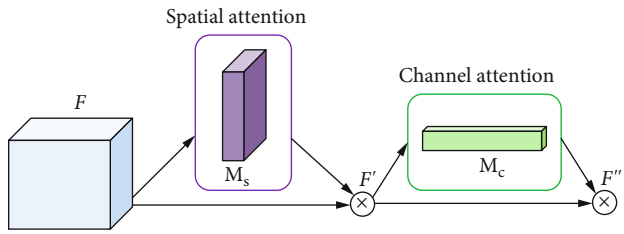


Figure 6: Scam overall attention module.

$M_s(F)$ is the output feature map after passing the spatial attention mechanism. The example is as follows:

$$M_s(F) = \sigma\left(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right). \quad (15)$$

(2) Channel attention module

To represent the feature information of facial expressions in multiple dimensions, the feature maps output by the spatial attention module are used as the input of this module, based on global max pooling and global average pooling of width and height, respectively, and the two obtained features are fed into a neural network composed of hidden layers and a multilayer perceptron (MLP). Then, the final features are merged and output using element-by-element summation, as follows:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))). \quad (16)$$

3.4. Grouping Bottleneck Method. In this paper, the grouping bottleneck is improved based on the group convolution method, and its specific structure is shown in Figure 8, which consists of 1×1 and 3×3 convolutions, where the number of convolution kernels in the first layer of 1×1 is half of the number of input feature map channels, and the reduction in the number of convolution kernels can reduce the number of network parameters. The size of the input feature map of the bottleneck block is $48\times48\times96$, and the size of the feature map is $24\times24\times192$ after the $1\times1\times192$ convolution. A 1×1 convolution reduces the number of parameters of the network by half while deepening the network to extract deep semantic information, which substantially reduces the subsequent convolution computation. The second layer of the bottleneck block is a 3×3 convolution as a group convolution layer, the number of convolution kernels of group convolution is the number of channels of the input feature map, the feature map of $24\times24\times192$ is divided into 192 feature maps of $24\times24\times1$ by channel, the features are extracted using 192 convolution kernels of 3×3, the corresponding element positions of the input and output feature maps are summed in pairs to obtain the final feature map, and the method of summing corresponding elements can solve the network degradation problem to some extent. Since the second layer of the original structure is a 3×3 convolutional structure changed to a 3×3 grouped convolutional structure, it can reduce the number of parameters of the network, reduce the complexity of the model, and improve the computational speed of the network, because when the ordinary convolutional operation is performed, the input feature map size is $C\times H\times W$ and there are N convolutional kernels, then the output feature map, and the number of convolutional kernels. The size of each convolutional kernel is $C\times K\times K$, and the total number of parameters of N convolutional kernels is $N\times C\times K\times K$.

Grouped convolution groups the input feature maps and then convolves each group separately, if the input feature map size is $C\times H\times W$ and the number of output feature maps is N; if we set to divide into G groups, the number of input feature maps of each group is C/G, then the number of output feature maps of each group map is N/G, the size of each convolutional kernel is $(C/G)\times K\times K$, the total number of convolutional kernels is still N, the number of convolutional kernels in each group is N/G, the convolutional kernels only convolve with the input map of the same

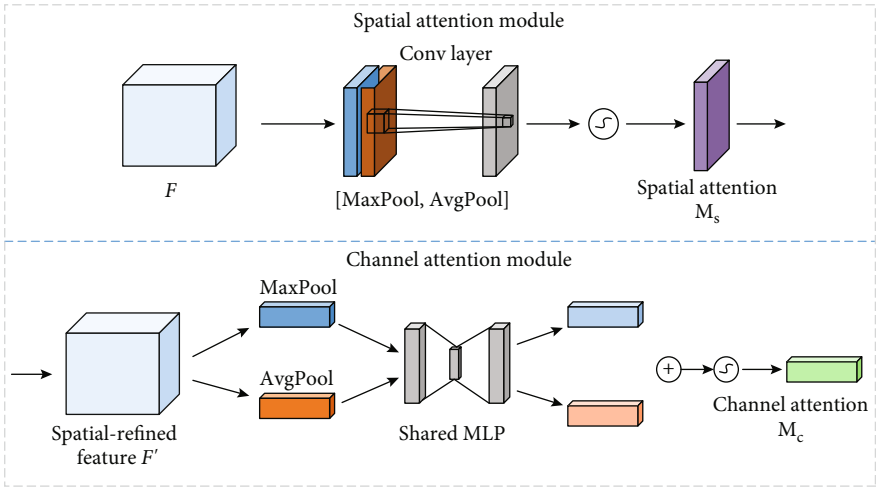


Figure 7: SCAM.

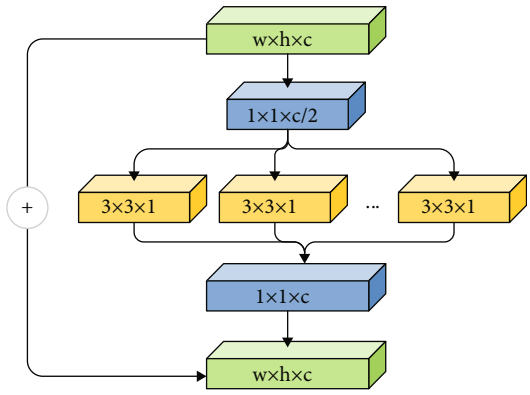


Figure 8: Grouping bottleneck.

group with them, and the total number of convolutional kernels is $N \times (C/G) \times K \times K$, so the total number of parameters is the original $1/G$.

The number of parameters of the bottleneck and the grouping bottleneck is shown in Table 3, and it was found that the grouping bottleneck block has a substantial decrease in the number of parameters compared with the original bottleneck block, with a ratio of nearly 1/10. The nonlinear expression capability of the model is increased.

3.5. Depthwise Separable Shuffle Method. In this paper, channel shuffling is used to improve the depthwise separable convolution [31, 32], and its structure is shown in Figure 9. The depthwise separable convolution first uses depthwise convolution to process the input feature map, and different channels use different convolution operations and then use the CONCAT method for channel stitching. Thus, the final output features are derived from only part of the input channel features, and there is no information exchange between the different channels, which leads to the limited characterizability of the extracted features. Although the depthwise separable convolution uses pointwise convolution to further

Table 3: Comparison of the number of participants before and after bottleneck improvement.

Module	Number of participants (pcs)	Ratio
3×3Conv	580327	1.78
Bottleneck	324873	1
Grouping bottleneck	33257	0.10

increase the dimensionality of features, which can enhance the communication of spatial feature information to a certain extent, the increase in dimensionality leads to an increase in the number of network parameters. Deep separable convolution is divided into deep convolution operation and point-by-point convolution operation. In the deep convolution operation, if the input feature dimension is $D_F \times D_F \times M$, M is the number of channels, and the parameter of the convolution kernel is $D_k \times D_k \times 1 \times M$, the output feature dimension after deep convolution is $D_F \times D_F \times M$. Each channel only corresponds to one convolution kernel when convolving, so the FLOPs are $M \times D_F \times D_F \times D_k \times D_k$. In the point-by-point convolution operation, the input is the feature after deep convolution, the dimension is $D_F \times D_F \times M$, the parameter of convolution kernel is $1 \times 1 \times M \times N$, the output dimension is $D_F \times D_F \times M$, the convolution process does 1×1 standard convolution for each feature, and the FLOPs are $N \times D_F \times D_F \times M$. In this paper, the point-by-point convolution operation is replaced by the channel shuffle method, which reduces the number of parameters of the point-by-point convolution operation. The channel shuffle method has the same function as the point-by-point convolution method; however, the number of parameters of the whole network does not increase because its dimensionality does not change, while the characterizability of the features is enhanced, which reduces the complexity of the network to some extent improves the training speed of the model, which can improve the whole network's face recognition accuracy.

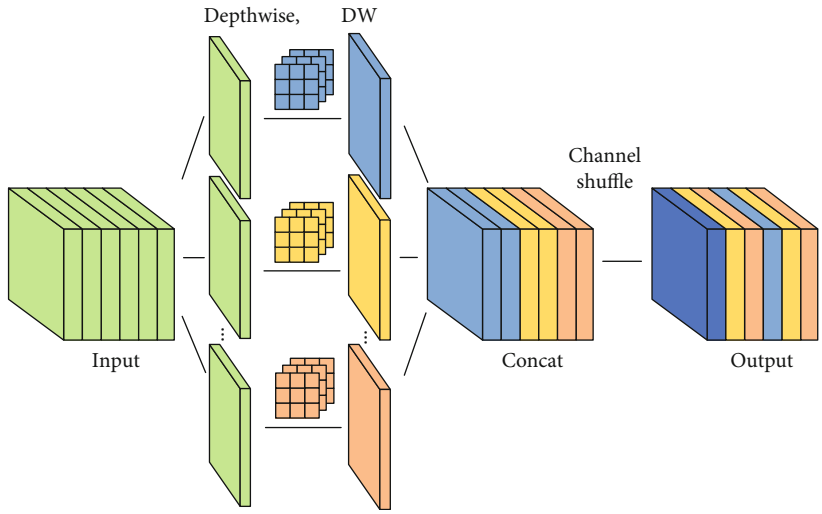


Figure 9: Depthwise separable shuffle method.

4. Experiment and Analysis

4.1. Experiment Preparation. To verify the accuracy and the effectiveness of the Light-NTWRN network model proposed in this paper, the light-NTWRN network model is subjected to comparative ablation experiments on the FER2013, CK+ dataset, and JAFFE dataset. The experiment is based on the TensorFlow deep learning framework for training, and testing is conducted on Pycharm with the following hardware environment configuration: Win10 operating system, Intel Core i7-10700F with 2.9 GHz CPU and 16 G RAM and NVIDIA GeForce RTX 3070 (8 GB) graphics card. During the experiments, 70% of the facial expression images are randomly selected as the training set, and 30% of the facial expression images are randomly selected as the test set. Additionally, the experimental parameters are set as shown in Table 4.

Table 4: Experimental parameter settings.

Parameter	FER2013	CK+	JAFFE
Optimizer	SGD	SGD	SGD
Momentum	0.9	0.9	0.9
Batch size	30	20	40
Learning rate	0.01	0.01	0.01
Learning rate decay	0.5/50	0.5/50	0.5/50
Loss function	Cross entropy	Cross entropy	Cross entropy
Epochs	300	300	300

4.2. Facial Expression Dataset. The FER2013 facial expression dataset consists of 35,886 facial expressions, and the dataset is expanded to 80,000 by an improved random erasing method where the training set contains 56,000 and the test set contains 24,000, and each image is composed of a grayscale image with a fixed size of 48×48, which contains a total of 7 expressions, namely, angry, disgusted, fear, happy, sad, surprised, and neutral. The facial expression images of FER2013 are more difficult to recognize because of the interference of occlusion, pose, low contrast, and background.

CK+ is expanded from the Cohn-Kanda dataset, which contains a total of 123 participants, 593 image sequences, and a total of 7 expressions. The CK+ dataset acquisitions are all collected under the same lighting background, the acquisition environment is better, and the dataset is expanded to 1500 images through an improved random erasing method, with 70% of the training set and 30% of the test set.

The JAFFE dataset was selected from 10 Japanese female students who each made 7 different expressions, consisting of a total of 213 photos, which were expanded to 3408 photos by rotation, flip, contrast enhancement, panning, cropping, scaling, and improved random erasing methods.

4.3. Ablation Experiment. To verify the effectiveness of the Light-NTWRN network model proposed in this paper, ablation experiments are conducted for each module, and the experimental results are shown in Table 5. WRN denotes the improved wide residual network, RE denotes the improved random erasing method, SCAM denotes the improved attention mechanism module, GBN denotes the grouping bottleneck method, and DS denotes the depthwise separable shuffle method, where WRN+RE+SCAM+GBN+DS denotes the Light-NTWRN network proposed in this paper.

First, the facial expression images are input into the model after the improved random erasing operation, and for the model to acquire more local features of facial expressions, the improved SCAM is embedded into the network to reassign the feature weights of facial expressions from both the channel and space dimensions. The grouping bottleneck method is improved to solve the problem of model redundancy caused by too many convolutional layers. To reduce the number of parameters computed by the network and speed up the network operation, an improved depthwise

Table 5: Light-NTWRN network ablation experiments.

Model	FER2013 (%)	CK+ (%)	JAFFE (%)	Parameter (M)
WRN	69.60%	95.38%	91.42%	18.35
WRN+RE	71.05%	97.54%	93.88%	18.35
WRN+RE+SCAM	72.27%	98.35%	94.28%	23.21
WRN+RE+SCAM+GBN	72.58%	98.60%	94.87%	15.72
Light-NTWRN (ours)	73.21%	98.72%	95.21%	10.14

separable shuffle method is added. To verify the effectiveness of each improved module, the Light-NTWRN network ablation experiments are shown in Table 5.

The ablation experiments are shown in Figure 10, where part a represents the FER2013 ablation experiment, part b represents the CK+ ablation experiment, and part c represents the JAFFE ablation experiment. According to the ablation experiments of the FER2013 dataset in part a, we can see that Light-NTWRN has the fastest convergence rate, and the model recognition accuracy grows slowly when trained to 100 epochs and gradually levels off when trained to 210 epochs. The accuracy gradually levels off, and the highest accuracy reaches 73.21%.

From the ablation experiments of the CK+ dataset in part b of Figure 10, it can be seen that the accuracy of the model increases rapidly at the beginning of training, and the accuracy of the model recognition oscillates up and down from the 50th epoch to the 100th epoch. When the training reaches 150 epochs, the accuracy tends to be stable, and the highest accuracy can reach 98.72%. From the ablation experiments of the JAFFE dataset in part c, we can see that the accuracy of the model also grows faster at the beginning of training, and when, the training reaches 180 epochs, the accuracy tends to be stable, and the highest accuracy can reach 95.21%. From the dataset, it was found that the accuracy of the model is improved after adding SCAM, but there is a slight loss of its network operation speed. GBN and DS can effectively reduce the number of network parameters and improve the accuracy of the model. Furthermore, the accuracy of the model proposed in this paper on the three datasets FER2013, CK+, and JAFFE is improved by 3.61%, 3.34%, and 3.79%, respectively, compared with that of the original model, and the number of parameters is reduced by 44.74% compared to that for the original network, which proves that the proposed model has better effectiveness and faster computing speed.

To further verify the effectiveness and the robustness of the proposed model in this paper, the confusion matrix experiments are shown in Figure 11, where part a represents the confusion matrix on the FER2013 dataset, part b represents the confusion matrix on the CK+ dataset, and part c represents the confusion matrix on the JAFFE dataset.

From the confusion matrix on the FER2013 dataset in part a, we can see that the recognition accuracy of the three categories of anger, fear, and sadness is low because the activities of these three categories of facial expressions are less obvious, and the feature points are difficult to extract. The recognition performance of each category on the CK+ dataset is better, and the accuracy is higher. On the JAFFE dataset, the recognition accuracy of the anger and disgust categories is lower because the misidentified samples all belong to the negative category of emotions, which are more similar, facial features are difficult to extract, so recognition is more challenging.

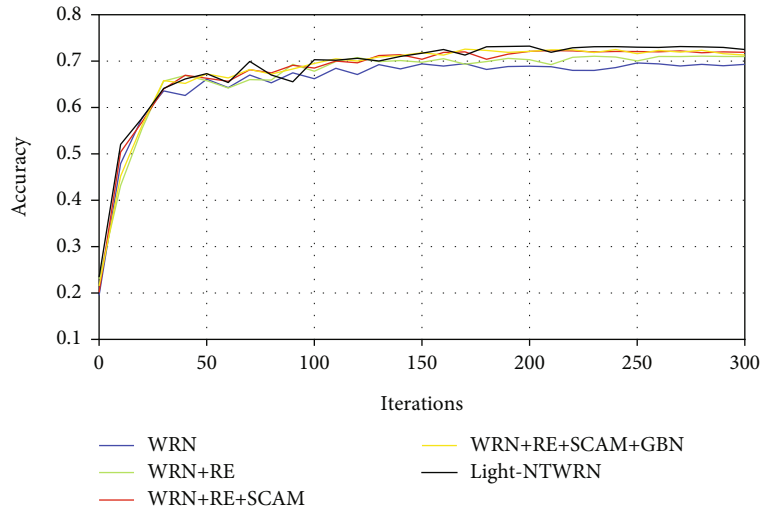
4.4. Mainstream Algorithm Comparison Experiment. To verify the effectiveness of the Light-NTWRN algorithm proposed in this paper for facial expression recognition, comparison experiments are conducted with five mainstream algorithms, mainly AlexNet, VGG16, VGG19, ResNet18, and ResNet50, to compare the size of the number of parameters and the specific recognition accuracy on the three datasets, and the specific results are shown in Table 6.

The Light-NTWRN algorithm proposed in this paper has the highest accuracy for facial expression recognition on the FER2013 dataset, with an improvement of nearly 2% compared to the ResNet50 model in the mainstream algorithm. From the experimental results on the CK+ dataset, the recognition accuracy of the VGG16 model is the highest among the mainstream networks, while the recognition accuracy of the model proposed in this paper is improved by 3.26% compared to the VGG16. The recognition accuracy on the JAFFE dataset is as high as 95.21%.

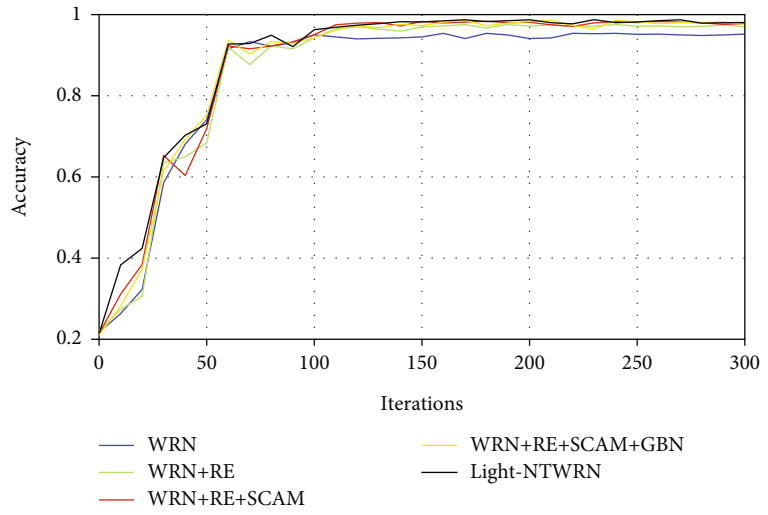
This can further verify the effectiveness of the three improved methods proposed in this paper, which can improve the recognition accuracy. Compared with the other five mainstream algorithms, the Light-NTWRN algorithm proposed in this paper has the highest accuracy and the best algorithm performance in terms of facial expression recognition and has strong generalization.

In terms of the number of model parameters, the number of model parameters of the network proposed in this paper is 10.14 M, which is the lowest compared with the other five mainstream algorithms and can maintain high recognition accuracy, which verifies the advanced and excellent model. It also further verifies the effectiveness of the three improvement methods proposed in this paper for model lightweighting.

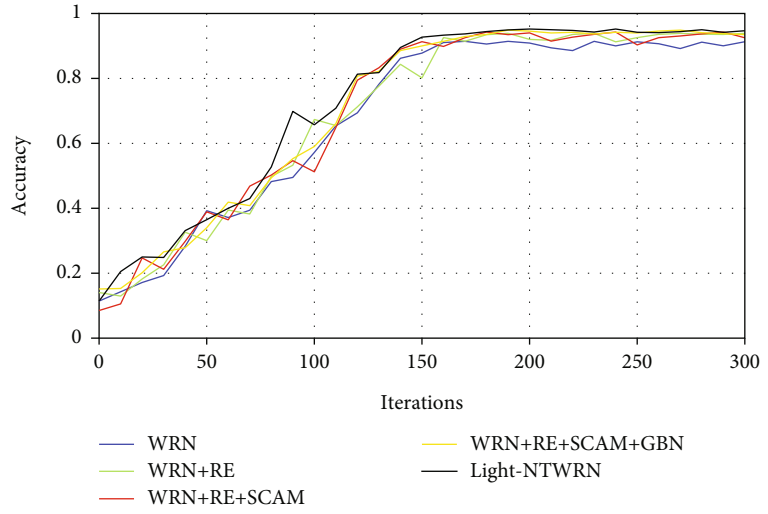
We compare the proposed method in this paper with other existing methods. The existing more advanced methods are MANet [33], a model that obtains key region features by adaptively learning weights; Minaee [34], a model that assigns residual blocks to spatial mask information; WMDCNN [35], a model that mixes two-channel weighting of static images; and APRNET50 [36], a model that uses multiscale feature extraction blocks instead of residual units. The



(a) FER2013 ablation experiment

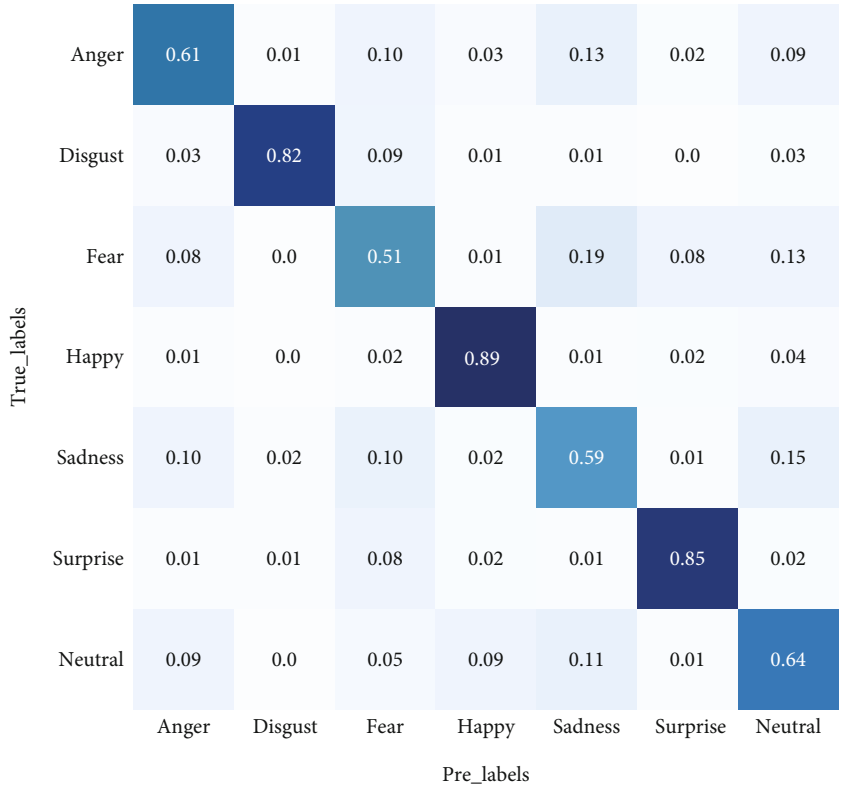


(b) CK+ ablation experiment



(c) JAFFE ablation experiment

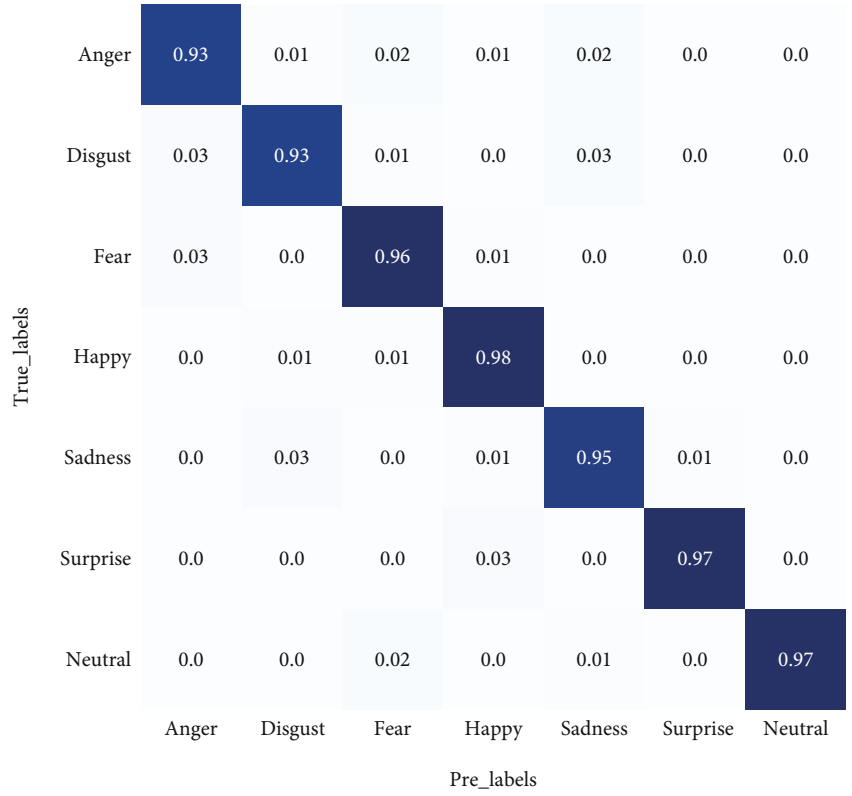
Figure 10: Ablation experiment.



(a) Confusion matrix on the Fer2013 dataset

(b) Confusion matrix on the CK+ dataset

Figure 11: Continued.



(c) Confusion matrix on the JAFFE dataset

Figure 11: Confusion matrix.

Table 6: Comparison experiments of mainstream algorithms.

Model	FER2013 (%)	CK+ (%)	JAFFE (%)	Parameter (M)
AlexNet	67.51	87.59	89.83	60.92
VGG16	68.89	95.46	91.04	14.75
VGG19	68.53	92.18	90.37	20.06
ResNet18	70.09	89.39	92,55	11.69
ResNet50	71.26	92.46	93.08	25.56
Light-NTWRN (ours)	73.21	98.72	95.21	10.14

Table 7: Recognition rates of various algorithms on the facial expression dataset.

Model	FER2013 (%)	CK+ (%)	JAFFE (%)
MANet [33]	69.46	96.28	—
Minaee [34]	70.20	98.00	92.80
WMDCNN [35]	—	98.50	92.30
APRNET50 [36]	73.00	94.95	94.80
Light-NTWRN (ours)	73.21	98.72	95.21

comparison is performed on the FER2013, CK+, and JAFFE datasets. It can be seen in Table 7 that the model proposed in this paper has the highest accuracy, and the effectiveness of the model proposed in this paper can be proven by the above experiments.

5. Conclusion

This paper proposes a multiscale feature fusion attention lightweight facial expression recognition method that effectively suppresses the influence of irrelevant feature information on the model while slowing the gradient disappearance caused by too many layers of the neural network, thus reducing the number of parameters computed by the network and improving the computational speed of the model. The improved SCAM module focuses more on feature information to speed up the convergence of the model and improve its performance. The improved random erasing method expands the training set while enhancing the robustness of the model to noise. The grouping bottleneck method reduces the dimensionality of the target image while increasing the nonlinear expression capability of the model. In addition, the depthwise separable shuffle method reduces the number

of parameters computed by the network while speeding up the computational speed of the network. The accuracy of the proposed model (Light-NTWRN) is 73.21% on the FER2013 dataset, 98.72% on the CK+ dataset, and 95.21% on the JAFFE dataset, while having a lower number of parameters, and the experimental results are better than many current mainstream algorithms, showing better effectiveness and robustness. However, the recognition accuracy is still not high enough in the case of obscured facial expressions, and more attention should be given to the recognition performance of these datasets in the future.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest

There is no conflict of interest regarding the publication of this paper.

Acknowledgments

The work was supported by the Key Research Project of Science and Technology of Chongqing Education Commission (no. kjzd-k201801901) and Chongqing Postgraduate Research and Innovation Project (CYS22663).

References

[1] P. E. Jianqing, W. U. Haoxuan, L. I. Tianliang, and H. A. Yu, "Workspace, stiffness analysis and design optimization of coupled active-passive multilink cable-driven space robots for on-orbit services," *Chinese Journal of Aeronautics*, 2022.

[2] K. Hambuchen, J. Marquez, and T. Fong, "A review of NASA human-robot interaction in space," *Current Robotics Reports*, vol. 2, no. 3, pp. 265–272, 2021.

[3] Q. Gao, X. Zhang, and W. Pang, "Fast and accurate hand visual detection by using a spatial-channel attention SSD for hand-based space robot teleoperation," *International Journal of Aerospace Engineering*, vol. 2022, Article ID 3396811, 11 pages, 2022.

[4] L. Yingxiao, H. Ju, M. Ping, and R. Jiang, "Target localization method of non-cooperative spacecraft on on-orbit service," *Chinese Journal of Aeronautics*, 2022.

[5] X. L. Ding, Y. C. Wang, Y. B. Wang, and K. Xu, "A review of structures, verification, and calibration technologies of space robotic systems for on-orbit servicing," *SCIENCE CHINA Technological Sciences*, vol. 64, no. 3, pp. 462–480, 2021.

[6] R. R. Santos, D. A. Rade, and I. M. da Fonseca, "A machine learning strategy for optimal path planning of space robotic manipulator in on-orbit servicing," *Acta Astronautica*, vol. 191, pp. 41–54, 2022.

[7] P. Rousso, S. Samsam, and R. Chhabra, "A mission architecture for on-orbit servicing industrialization," in *2021 IEEE Aerospace Conference (50100)*, pp. 1–14, Big Sky, MT, USA, 2021.

[8] J. Xing and J. Zhong, "MiniExpNet: a small and effective facial expression recognition network based on facial local regions," *Neurocomputing*, vol. 462, pp. 353–364, 2021.

[9] X. Sun, P. Xia, and F. Ren, "Multi-attention based deep neural network with hybrid features for dynamic sequential facial expression recognition," *Neurocomputing*, vol. 444, pp. 378–389, 2021.

[10] Y. Wenmeng and X. Hua, "Co-attentive multi-task convolutional neural network for facial expression recognition," *Pattern Recognition*, vol. 123, p. 108401, 2022.

[11] S. Cuiping, T. Cong, and W. Liguo, "A facial expression recognition method based on a multibranch cross-connection convolutional neural network," *IEEE ACCESS*, vol. 9, pp. 39255–39274, 2021.

[12] Y. Kong, Z. Ren, K. Zhang, S. Zhang, Q. Ni, and J. Han, "Lightweight facial expression recognition method based on attention mechanism and key region fusion," *Journal of Electronic Imaging*, vol. 30, no. 6, article 063002, 2021.

[13] N. Zhou, R. Liang, and W. Shi, "A lightweight convolutional neural network for real-time facial expression detection," *IEEE Access*, vol. 9, pp. 5573–5584, 2020.

[14] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[15] Y. Chen, L. Liu, V. Phonevilay et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, vol. 51, no. 7, pp. 4367–4380, 2021.

[16] H. Zhang, G. Peng, Z. Wu, J. Gong, D. Xu, and H. Shi, "MAM: a multipath attention mechanism for image recognition," *IET Image Processing*, vol. 16, no. 3, pp. 691–702, 2022.

[17] L. Yao, S. He, K. Su, and Q. Shao, "Facial expression recognition based on spatial and channel attention mechanisms," *Wireless Personal Communications*, vol. 125, no. 2, pp. 1483–1500, 2022.

[18] H. Wang and H. Zhang, "Adaptive target tracking based on channel attention and multi-hierarchical convolutional features," *Pattern Analysis and Applications*, vol. 25, no. 2, pp. 305–313, 2022.

[19] Z. Qiu, S. I. Becker, and A. J. Pegna, "Spatial attention shifting to emotional faces is contingent on awareness and task relevancy," *Cortex*, vol. 151, pp. 30–48, 2022.

[20] C. Chen, D. Gong, H. Wang, Z. Li, and K. Y. K. Wong, "Learning spatial attention for face super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 1219–1231, 2021.

[21] Z. Xue, T. Li, S. T. Peng, C. Y. Zhang, and H. C. Zhang, "A data-driven method to predict future bottlenecks in a remanufacturing system with multi-variant uncertainties," *Journal of Central South University*, vol. 29, no. 1, pp. 129–145, 2022.

[22] S. Panigrahi and U. S. N. Raju, "Pedestrian detection based on hand-crafted features and multi-layer feature fused-Res Net model," *International Journal on Artificial Intelligence Tools*, vol. 30, no. 5, article 2150028, 2021.

[23] C. Sekhar Vorugunti, V. Pulabaigari, P. Mukherjee, and A. Sharma, "DeepFuseOSV: online signature verification using hybrid feature fusion and depthwise separable convolution neural network architecture," *IET Biometrics*, vol. 9, no. 6, pp. 259–268, 2020.

[24] R. F. Rachmadi, S. Nugroho, and I. Purnama, "Lightweight residual network for person re-identification," *IOP Conference Series Materials Science and Engineering*, vol. 1077, no. 1, article 012046, 2021.

[25] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-MobileNet: an approach of facial expression recognition," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4435–4444, 2022.

[26] C. F. Xception, "Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, America Hawaii, 2017.

[27] A. I. Mohammed and A. A. Tahir, "A new optimizer for image classification using wide ResNet (WRN)," *Academic Journal of Nawroz University*, vol. 9, no. 4, pp. 1–13, 2020.

[28] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2017, https://arxiv.org/abs/1605.07146.

[29] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, 2017.

[30] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, Munich Germany, 2018.

[31] L. Wang and D. He, "Image super-resolution reconstruction algorithm based on channel shuffle," in *2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, pp. 225–229, Shenyang, China, 2021.

[32] X. Y. Zhang, K. Zhao, T. Xiao, M. M. Cheng, and M. H. Yang, "Structured sparsification with joint optimization of group convolution and channel shuffle," in *Uncertainty in Artificial Intelligence. PMLR*, pp. 440–450, America New York, 2021.

[33] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020.

[34] A. Abdolrashidi, "Deep-emotion: facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[35] H. Zhang, B. Huang, and G. Tian, "Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture," *Pattern Recognition Letters*, vol. 131, pp. 128–134, 2020.

[36] C. Jiamin and X. Yang, "Expression recognition based on attention pyramid convolution residual network," Computer engineering and application: 1-11 [2022-04-26] http://kns.cnki.net/kcms/detail/11.2127.TP.20210702.1749.004.html.