applied sciences

MDPI

Article

Lightweight Facial Expression Recognition Based on Class-Rebalancing Fusion Cumulative Learning

Xiangwei Mou [1,2,*], Yongfu Song [1], Rijun Wang [2], Yuanbin Tang [2] and Yu Xin [2]

[1] College of Electronic and Information Engineering/Integrated Circuits, Guangxi Normal University, Guilin 541004, China; yfsong@stu.gxnu.edu.cn
[2] Teachers College for Vocational and Technical Education, Guangxi Normal University, Guilin 541004, China; rijunwang@mailbox.gxnu.edu.cn (R.W.); gxnutyb@mailbox.gxnu.edu.cn (Y.T.); 15829309591@163.com (Y.X.)
* Correspondence: xwmou@mailbox.gxnu.edu.cn

Abstract: In the research of Facial Expression Recognition (FER), the inter-class of facial expression data is not evenly distributed, the features extracted by networks are insufficient, and the FER accuracy and speed are relatively low for practical applications. Therefore, a lightweight and efficient method based on class-rebalancing fusion cumulative learning for FER is proposed in our research. A dual-branch network (Regular feature learning and Rebalancing-Cumulative learning Network, RLR-CNet) is proposed, where the RLR-CNet uses the improvement in the lightweight ShuffleNet with two branches (feature learning and class-rebalancing) based on cumulative learning, which improves the efficiency of our model recognition. Then, to enhance the generalizability of our model and pursue better recognition efficiency in real scenes, a random masking method is improved to process datasets. Finally, in order to extract local detailed features and further improve FER efficiency, a shuffle attention module (SA) is embedded in the model. The results demonstrate that the recognition accuracy of our RLR-CNet is 71.14%, 98.04%, and 87.93% on FER2013, CK+, and RAF-DB, respectively. Compared with other FER methods, our method has great recognition accuracy, and the number of parameters is only 1.02 MB, which is 17.74% lower than that in the original ShuffleNet.

Keywords: facial expression recognition; feature extraction; class-rebalancing; lightweight

Citation: Mou, X.; Song, Y.; Wang, R.; Tang, Y.; Xin, Y. Lightweight Facial Expression Recognition Based on Class-Rebalancing Fusion Cumulative Learning. *Appl. Sci.* **2023**, *13*, 9029. https://doi.org/10.3390/app13159029

Academic Editor: Douglas O'Shaughnessy

Received: 5 May 2023
Revised: 4 August 2023
Accepted: 4 August 2023
Published: 7 August 2023

1. Introduction

Facial expressions, as expressions of emotions, play a major role in interpersonal communication. FER based on lightweight networks is very important for the implementation of human–computer interaction technology. With the continuous development of FER technology, FER is popularly applied in autonomous driving, criminal investigation, medical diagnosis, psychological assessment, and auxiliary classroom teaching and other fields [1].

Recently, FER methods of deep learning based on the convolutional neural network (CNN) have gained significant achievements [2,3]. Better results have been achieved for FER on the facial expression datasets CK+ [4], JAFFE [5], and so on taken under controlled conditions (laboratory, no complex background, etc.). However, it is difficult for natural expression datasets affected by noise, lighting changes, posture, and occlusion such as FER2013 [6] and RAF-DB [7] to achieve the expected effect in FER, and scholars have studied this accordingly. Wang et al. [8] showed an attention network based on facial regions, evaluated different region generation strategies, and adaptively integrated weighted features from regions and the entire face through the attention module, which significantly improved the performance of the network under occlusion and complex pose conditions. Hamid et al. [9] showed a method to apply depth histogram metric learning to FER in CNNs, which enhanced the accuracy of FER under uncontrolled conditions. Kim et al. [10] showed a new FER framework based on a support vector machine (SVM) classifier and

CNNs, which improved the ability of network feature extraction and expression classification. Gong et al. [11] showed a dual-branch multi-feature fusion network based on deep learning, in which one branch extracts multi-level facial features to enhance feature recognition capabilities and the other branch enhances the adaptability of learned features to direction and scope changes, improving the ability of network feature extraction. Although the above methods provide some effective research methods for FER in real scenes, there are still some issues: (1) Due to realistic factors such as complexity, subtlety, and occlusion of facial expression in real scenes, the robustness of the features extracted by shallow networks is poor, and it is difficult to capture the local detailed facial expressions' features. (2) The deep network model promotes the capacity of feature extraction and the accuracy of expression recognition, but simultaneously, it greatly increases the number of network parameters and computational costs. In practical applications, limited by the cost of computing hardware configuration, the method of deepening the network depth in exchange for recognition accuracy is not practical, which is not conducive to the further development of FER.

In response to the aforementioned problems, a lightweight and efficient method of FER is shown in this paper. The following is an overview of the main innovation points of this paper:

1. To resolve the issue of the uneven inter-class distribution of facial expression data, a dual-branch network (RLR-CNet) is designed in our paper.
2. To alleviate the issue of easily losing feature information and to cut down on the number of parameters of the RLR-CNet, a lighter Clip_K5_ShuffleNet inverted residual structure is proposed in the RLR-CNet.
3. To facilitate the transfer of facial expression feature information and promote the generalization ability of the RLR-CNet, the β-Mish activation function and the improved random masking method are used in the RLR-CNet, respectively.
4. To extract facial expressions' local key features and further enhance the accuracy of FER without significantly increasing computational complexity, a shuffle attention (SA) module is embedded into the RLR-CNet, which integrates spatial and channel attention.

2. Related Work

2.1. FER in Real-World

In the study of FER, whether through early traditional methods or popular deep learning methods, it contains three basic parts: facial recognition, feature extraction, and feature classification. In general, most methods are able to recognize facial expressions quickly and accurately in experimental situations, but the recognition performance in real scenes is greatly reduced [12–14], which brings great challenges to the practical application of FER.

In order to better recognize facial expressions in real scenes, more and more scholars have adopted deep learning methods such as CNN in the study of FER. Based on an existing pre-training model, Ng et al. [15] solved the problem of insufficient samples and poor characterization through multi-round fine-tuning. In the literature [16], with the complexity of the facial expression features in real scenes, a pseudo-tag generation strategy with a multi-area attention conversion network was proposed to promote the function of FER in the real scenes. Yao et al. [17] embedded a space and channel attention mechanism with HPMI in a VGG-16 network, which facilitates the flow of information between the image key information and network, and greatly solves the disappearance of facial expression feature information in real scenes. Siqueira et al. [18] designed network integration models with different structures for the datasets in the laboratory and real scenes to enhance the accuracy of FER. Because of the inter-class differentiation of facial expression data caused by various interference factors in real scenes, Shan et al. [19] proposed a method of cleverly integrating shallow features into deep features and established a local retention loss function to make the local parts in the expression class more compact and then promote

the capacity of FER. Pan et al. [20,21] adopted adversarial learning methods, and guided the network to improve the recognition of expressions containing obscured faces in real scenes by combining multiple loss functions.

2.2. FER Based on Lightweight Network

Deep learning methods are highly favored in the study of FER. Deeper and more abstract expression features are extracted through larger and deeper networks or fusion attention mechanisms, and then promote the recognition accuracy. However, it also greatly increases the parameters and computational complexity of the network, reduces the recognition efficiency of the network, and affects the practical application of FER.

Therefore, in the study of FER, while improving the recognition accuracy as much as possible, it is also essential to pay attention to the equipment configuration cost of model operation in practical applications, and minimize the number of computing costs. Mahmoudi et al. [22] offered a structure that decreases the parameter quantity of the network and promotes network performance by extending the classical linear convolution function to a higher-order kernel function, which has no other weights. Kong et al. [23] utilized lightweight networks and incorporated attention mechanisms for key feature extraction, which greatly reduced the model's computational complexity. Nan et al. [24] proposed a lightweight A-MobileNet that combines central loss and softmax loss functions to optimize the model parameters, which appreciably promotes the recognition accuracy over the original MobileNet without increasing the model parameters. Zhou et al. [25] offered a method with a multi-task cascaded network for facial recognition. To make the model more lightweight, they introduced depthwise separable convolution and residual modules into the network.

3. Models and Methods

Due to the problem of FER in real scenes, a lightweight model for FER based on class-rebalancing fusion cumulative learning is offered in this paper. Firstly, to promote the generalization ability of the network in real scenes, an improved random masking method is utilized to artificially introduce noise into the training dataset, while expanding the facial expression datasets. Then, the masked expression images are input into the dual-branch network (RLR-CNet); this network has a feature learning branch and class-rebalancing branch. For each branch, two data samples are obtained using a conventional uniform sampler and a reverse sampler, respectively, and are then fed into the corresponding branch. Under the control of cumulative learning, each branch goes through convolution and global average pooling to obtain feature vectors, which are then merged and weighted by channel fusion, and finally classified using a softmax layer. The main structure of our method is shown in Figure 1.

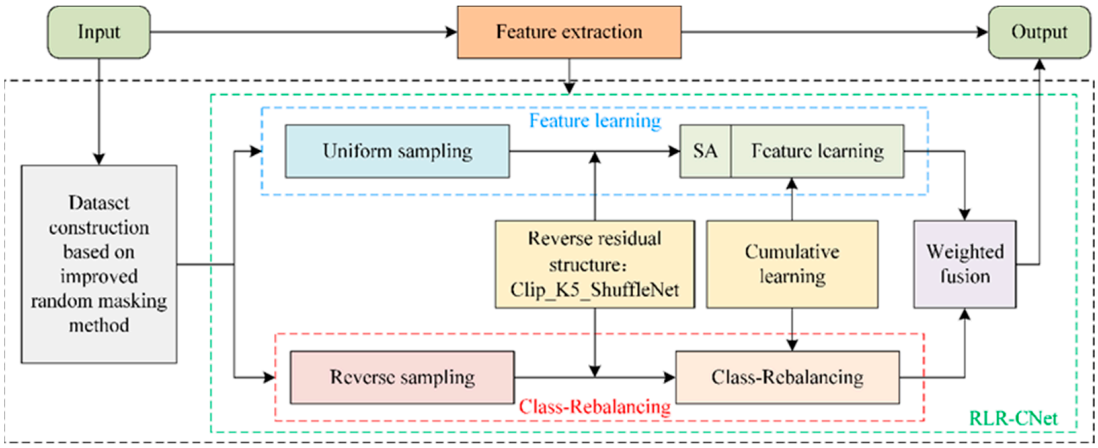

Figure 1. Overall architecture.

3.1. Dataset Construction Based on Improved Random Masking Method

In the dataset construction stage, it is a common method to promote the recognition accuracy and generalization ability of the network by optimizing the training datasets. In this research, the improved random masking method is used for randomly masking the expression images, artificially increasing the noise, expanding the training samples of the dataset, and making it closer to the real scene. Up to a point, the issue of overfitting is alleviated and the recognition efficiency of the model is enhanced [26].

Assuming the probability of masking is p, the area of the input facial expression image is $S = W \times H$, S_c is the area of random masking, the masking threshold is set as S_c/S, the range of the masking threshold is in the interval (s_l, s_h), r_c is the aspect ratio of the masking rectangle, the range of r_c is in the interval (r_1, r_2), and the height and width of the masking matrix area are set as H_c and W_c, respectively. The specific calculation formula is as follows:

$$S_C = Rand(s_l, s_h) \times S \tag{1}$$

$$r_c = Rand(r_1, r_2) \tag{2}$$

$$H_c = \sqrt{S_c \times r_c},\ W_c = \sqrt{S_c / r_c} \tag{3}$$

In general, a randomly selected point $P = (x_c, y_c)$ on the facial expression image, and $I_c = (x_c, y_c, x_c + W_c, y_c + H_c)$ is the area to be selected for masking. Point P can be determined at any point of the target image, but considering that most of the key information of facial expression features are concentrated in the top half of the image, the coordinates of point P are limited to the point at the top left 1/4 of the target image and the point at the top right 1/4 of the target image, which is recorded as $P' = (x'_c, y'_c)$. The masking area is updated as $I'_c = (x'_c, y'_c, x'_c + W_c, y'_c + H_c)$. The process of taking the coordinates of P and P' is as follows:

$$\begin{pmatrix} x_c = Rand(0, W) \\ y_c = Rand(0, H) \end{pmatrix} \rightarrow \begin{pmatrix} x'_c = Rand(x, x_i) \\ y'_c = Rand(y, y_j) \end{pmatrix} \tag{4}$$

The parameter settings of the image random masking improved algorithm in this paper are shown in Table 1. By improving the random masking method, the position of each masking can be covered as much as possible in the facial area, and the improved method is compared with the unimproved method, as shown in Figure 2.

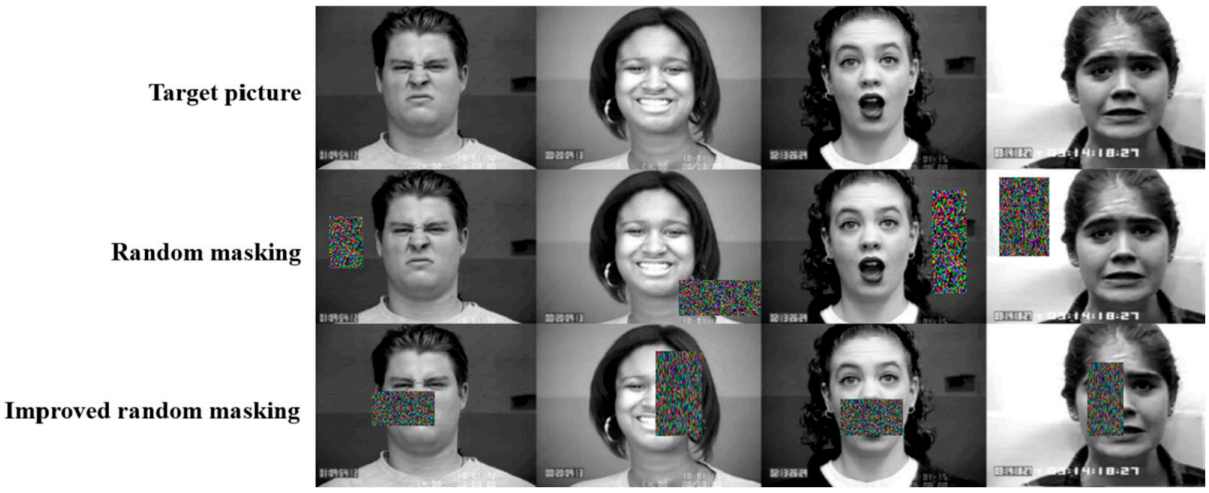

Figure 2. Comparison chart of the random masking experiment.

Table 1. Random masking parameters.

Parameter	Value
s_l	0.04
s_h	0.3
p	0.5
r_c	0.3

3.2. Design a Clip_K5_ ShuffleNet Inverted Residual Structure

3.2.1. Inverted Residual Structure

The general standard convolution has two main functions: one is to tune-up the size of the upper layer's feature map, and the other is to tune-up the number of the upper layer's feature maps (adjust the number of channels). Depth-separable convolution makes some changes to the standard convolution, which is composed of depthwise (DW) convolution and pointwise (PW) convolution. Given the same input as standard convolution, it outputs the same result as the standard convolution after two steps of sequential operation. But the number of parameters and the calculation cost are relatively decreased. Theoretically, the computational cost of the standard convolution is 8–9 times that of the depth-separable convolution, which largely contributes to the light weight of the network. The inverted residual structure is used in Mobilenet_V2 [27], as shown in Figure 3, which first increases the dimensionality of the input feature map using 1 × 1 convolution, performs a convolution operation using 3 × 3 DW convolution, and finally uses 1 × 1 convolution to reduce its dimensionality, to reduce the amount of convolution operations while also reducing the loss of feature information.

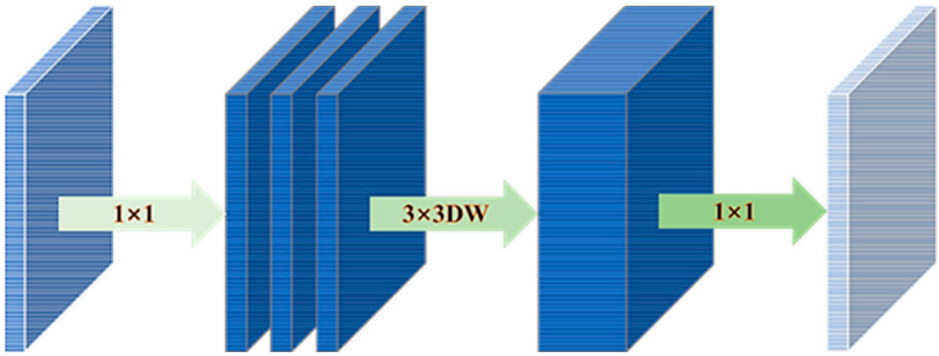

Figure 3. Reverse residual structure.

The inverted residual structure is continued in ShuffleNet [28], and some networks' parameters and calculations are reduced by stacking lightweight operators (DW convolution, etc.). And the inverted residual module in this network is divided into two structures: BasicBlock and DownBlock. There are 3 basic units (Stage2, Stage3, Stage4) in ShuffleNet, and each stage is repeatedly connected by a DownBlock and several BasicBlocks. As shown in Figure 4 (taking Stage2 as an example), the DownBlock needs to downsample and increase the dimension, so the input is copied into two copies, which are merged together after branch1 and branch2, respectively, and the channels are cross-rearranged to improve the information reuse of features. BasicBlock divides the input into two parts, and one part is directly merged and rearranged with the part of branch1 after the convolution extraction feature of branch2, to enhance the information communication between the two branches.

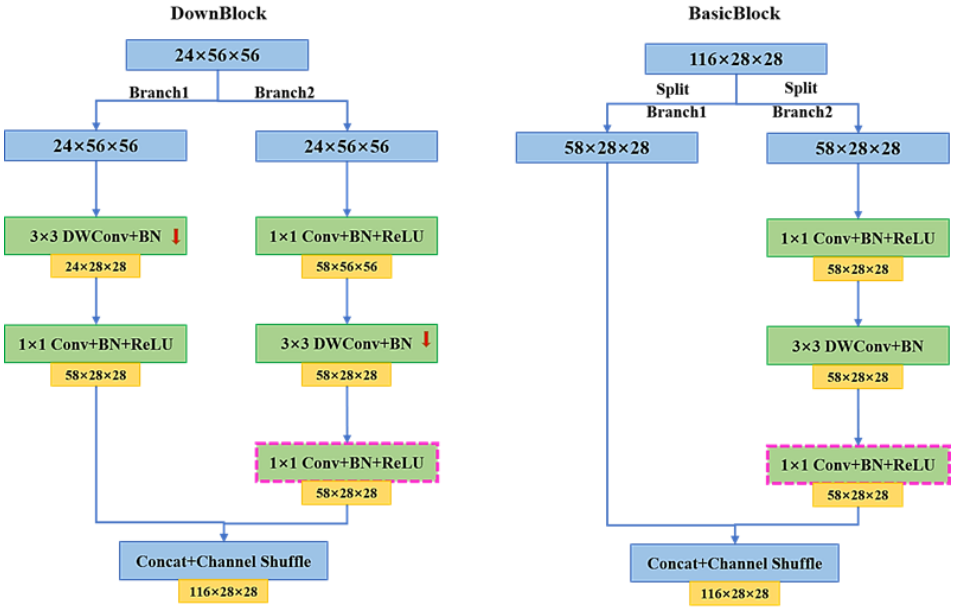

Figure 4. DownBlock and BasicBlock in Stage2.

3.2.2. Improved ShuffleNet_Block

In the inverted residual structure of ShuffleNet, to compensate for the lack of the inter-channel information fusion function in DW convolution, a 1 × 1 convolution is used before and after the DW convolution on branch2. There is no need for dimensionality enhancement or reduction, and there is no obvious effect on improving model efficiency. Therefore, to pursue a more lightweight network, the 1 × 1 convolution in the dashed box in Figure 4 is cropped to reduce the network redundancy and is designed to implement the Clip_ShuffleNet in this paper. Compared with the 1 × 1 convolution, DW convolution has low computational complexity in networks, and Peng et al. [29] found that large convolution kernels have a stronger characterization ability for feature information. So, the convolution kernels of DW convolution in ShuffleNet are expanded in this paper, as shown in Figure 5, all 3 × 3 DW convolutions are expanded to 5 × 5 DW convolutions on the basis of Clip_ShuffleNet, and then the Clip_K5_ShuffleNet is designed and implemented.

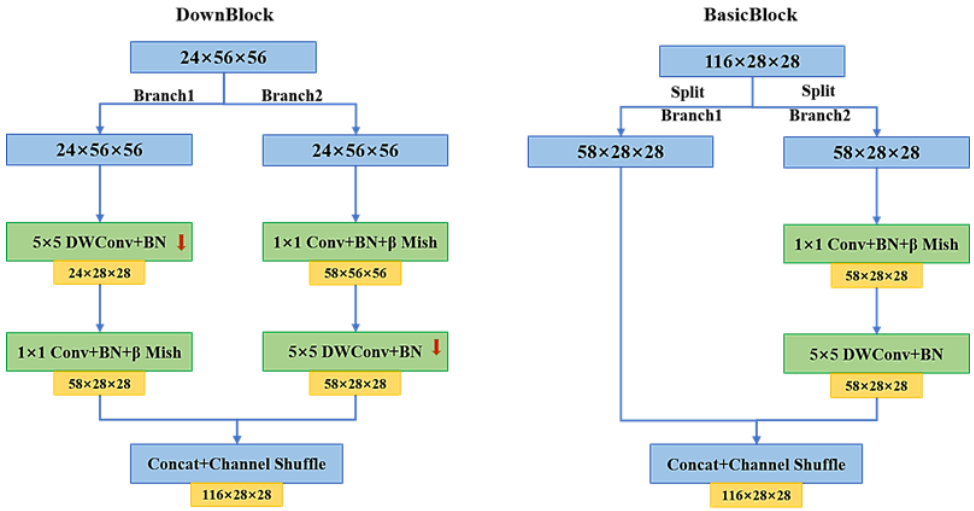

Figure 5. Improved DownBlock and BasicBlock.

Since the ReLU activation function will suppress the neurons when the input is negative, resulting in the model weights not being able to update, there is the issue of gradient disappearance, which affects the expression of networks. Therefore, the ReLU

function is replaced by the β-Mish activation function in our paper. The β-Mish function uses α and β factors to normalize the region below the boundary of the Mish [30] function. β-Mish is a smooth, continuous, and non-monotonic activation function that uses the self-gating property to retain some negative information while eliminating the hard zero boundary of ReLU. And the function also plays a good role in the flow of the gradient and enhances the nonlinear expressive ability of networks. To some extent, the use of this function reduces the overfitting phenomenon and promotes the recognition accuracy and generalization ability of the network. The definition of the β-Mish function is as follows, where the value of parameter α depends on parameter β, $\alpha/\beta = 1/5$, and the value of parameter β is between 1 and 200. In order to avoid saturation and slow down the training speed, this paper takes $\alpha = 7$ and $\beta = 35$. The β-Mish function formula is expressed as follows:

$$F(x) = x \cdot \tanh(\ln(1 + e^{\frac{\alpha x}{\sqrt{\beta + x^2}}})) \quad (5)$$

3.3. Shuffle Attention Module (SA)

The extraction of facial expression features is the most critical step in FER, which has a significant impact on recognition accuracy. The key to the extraction of facial expression features is the extraction of local detail features, such as the eyebrows, eyes, mouth, and other crucial parts that can best distinguish different expressions, and the human eye also pays more attention to these parts when discriminating expressions. The attention mechanism, as a theory proposed by a human-like cognitive behavior, focuses attention on more important information, and it can also quickly filter out more critical information from a lot of complex information and promote the efficiency of task processing. To further promote the accuracy of lightweight networks in FER, this paper introduces a lightweight and efficient Shuffle Attention module (SA) [31], which uses the replacement unit to efficiently fuse the channel and spatial attention mechanism, in exchange for a higher model recognition accuracy with a small increase in computational effort. Figure 6 shows the overall structure of the SA module, and the module mainly contains four points: Feature Grouping, Spatial Attention, Channel Attention, and Aggregation. Among them, Feature Grouping groups the input features, and each group of features is split into a Spatial Attention branch and Channel Attention branch along the channel dimension, which are used to learn channel features and spatial features, respectively. Aggregation fuses the two branches through 'Concat' and communicates feature information between groups through channel replacement operations to improve model performance.

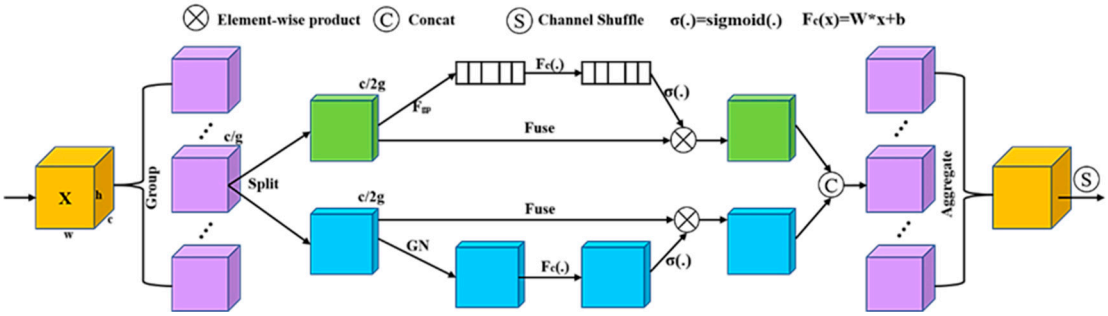

Figure 6. The structure of the SA module.

3.4. The Dual-Branch Network (RLR-CNet)

With the advancement of study in CNNs, there has been remarkable progress in image classification, which is inseparable from lots of scientific, reasonable, and high-quality datasets, such as MS COCO, ImageNet, and other datasets. These high-quality datasets have a relatively uniform distribution of sample sizes for each category, which is conducive to improving the representation ability of the network for feature extraction. However, most classification datasets (such as facial expression dataset) in real scenes have an uneven

distribution of sample sizes between classes, and exhibit a long-tail distribution [32] (a few classes have larger sample sizes and most classes have smaller sample sizes), to some extent, so the accuracy of model classification and recognition is affected. When extracting complex and subtle features such as facial expression features, it is especially critical to promote the quality of the dataset and the rationality of the sample distribution. Therefore, in order to resolve the issue of the uneven inter-class distribution of facial expression data and further promote the accuracy of FER in lightweight networks, RLR-CNet is proposed in this paper. This network has two branches based on fusion cumulative learning.

Specifically, the two branches proposed in this paper are called the "feature learning branch" and "class-rebalancing branch". Figure 7 shows a schematic diagram of the dual-branch network. The Clip_K5_ShuffleNet inverted residual structure is used into two branches [33], the lightweight ShuffleNet as the backbone network is used in the dual-branch network, and all blocks in Stage2, Stage3, and Stage4 are replaced with the improved Clip_K5_ShuffleNet modules in this paper. All weights of the network before Stage4 are shared between the two branches. In each convolutional layer, the input images are convolved with the same filter to achieve weight sharing and cut down the number of parameters. Good feature learning is beneficial for rebalancing the learning of class-rebalancing branches and can greatly reduce the network's computational load. For the two branches, the conventional uniform sampler and the reverse sampler are used, respectively, to obtain two data samples (x_1, y_1) and (x_2, y_2), which are used as the input of the feature learning branch and class-rebalancing branch, respectively, and the feature vectors f_1 and f_2 are obtained after the corresponding branch convolution and global average pooling. Considering that class-rebalancing can significantly boost classifier learning, it also impairs feature learning to some extent [34]. So, this paper introduces a specific method of cumulative learning, which staggers the learning "attention" of the two branches during network training. In other words, focusing on the learning of the feature learning branch in the early stage and the class-rebalancing branch in the later stage, to eliminate the effect of class-rebalancing on feature learning, achieves the goal of promoting the recognition accuracy of the network. Specifically, an adaptive trade-off parameter α is used to control the weights of f_1 and f_2, and the weighted feature vectors $\alpha * f_1$ and $(1-\alpha) * f_2$ are input into the classifiers W_1 and W_2, which are finally integrated together by channel merging. The output formula is as follows:

$$z = \alpha W_1^T f_1 + (1-\alpha) W_2^T f_2 \tag{6}$$

where z is the final output and the predicted probability for each class $i \in \{1, 2, \cdots, C\}$ can be calculated by the following formula:

$$\hat{p}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{7}$$

Notated as $\hat{p} = |\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_C|^T$, the final category with the maximum probability is used as the last recognition result of the dual-branch network.

The dual-branch network designed in our paper mainly has the following three characteristics:

1. The improved Clip_K5_ShuffleNet module is used in the whole of the dual-branch in this paper. Among them, the clipping of the 1×1 convolution reduces the model's parameters, and the use of 5×5 DW convolution is beneficial for extracting global features. Pairing the β-Mish activation function in each block enhances the flow of feature information. Moreover, in order to obtain the expressions' key features to ensure the validity of feature information, this paper incorporates the lightweight SA module in the feature learning branch but, in order not to add too many additional parameters as much as possible, only embeds the SA module between Stage 2 and

Stage 3. These designs enable the model to promote the accuracy of recognition and classification while ensuring light weight.

2. Weight sharing in the dual-branch network: The two branches share the weights of the network before Stage 4. On the one hand, the good learning of the feature learning branch is conducive to the learning of the class-rebalancing branch; on the other hand, the shared weights greatly reduce the computational complexity of the network and the speed of the model training, which in turn improves the recognition efficiency of the network.
3. The dual-branch network that combines cumulative learning: Parameter α is set to control the weights and loss functions for the two branches, and realize the transfer of learning "attention" between two branches. In this way, the influence of class-rebalancing on feature learning is eliminated, the recognition accuracy of small sample size classes is enhanced, and the recognition accuracy of the network is comprehensively improved. Where parameter α is adaptively adjusted according to the number of iterations for training, indirectly determined by the total training time T_s of the network and the current training time T, the formula is as follows:

$$\alpha = 1 - \left(\frac{T}{T_s}\right)^2 \tag{8}$$

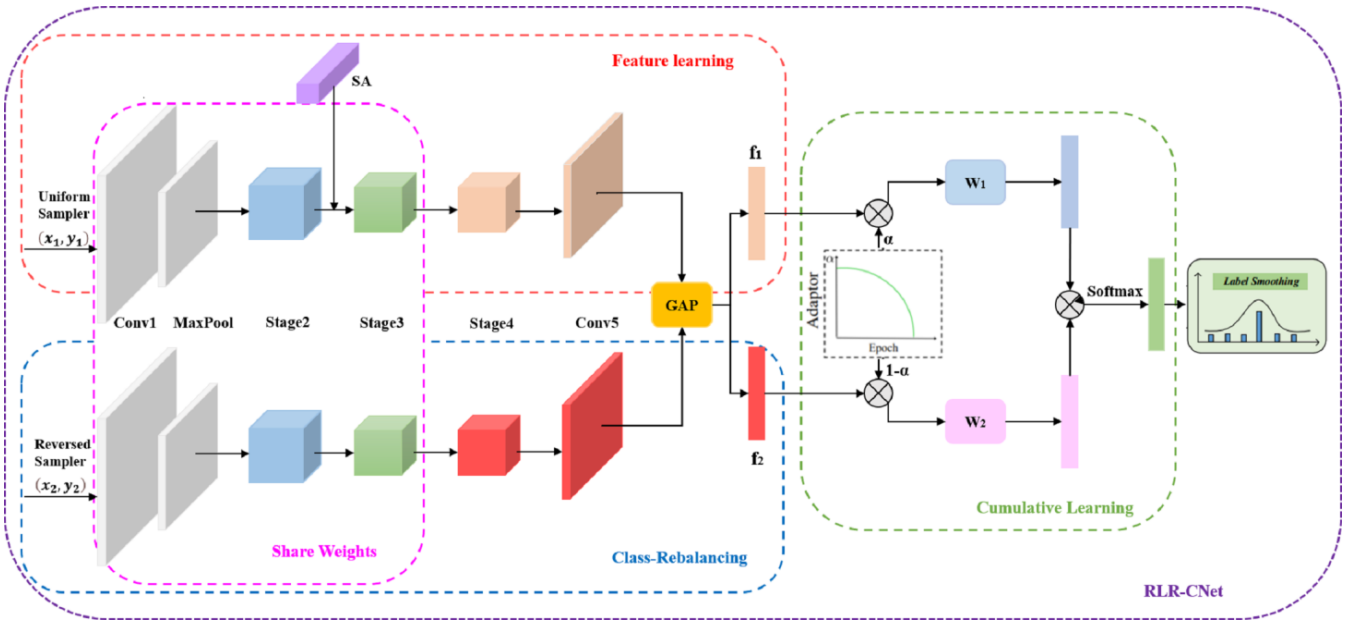

Figure 7. The structure of RLR-CNet.

4. Experiment and Analysis

4.1. Experimental Preparation and Evaluation Indicators

4.1.1. Experimental Preparation

To verify the accuracy and effectiveness of the proposed lightweight FER model based on class-rebalancing fusion cumulative learning, this paper conducts ablation and comparative experiments on the FER2013, CK+, and RAF-DB datasets. Especially, in order to ensure the repeatability of the method proposed in this paper, the ablation, comparative, and other experiments are conducted multiple times under the same experimental equipment and parameter conditions, and the error rates of the final experimental results are all below 0.009%, which better proves the repeatability of the method proposed in this paper. And in order to further prove the reusability of the method in this paper, all the experimental results are verified under the conditions of the same experimental configuration and differ-

ent experimental equipment, and the error rates of the experimental results are also kept below 0.01%, proving that the method in this paper has great reusability.

The experimental training and testing are based on the PyTorch deep learning framework on Pycharm. The configuration of the experimental server is as follows: Win10 operating system, Intel Core i5-12490F with 2.9 GHz CPU and 32 GB RAM, and NVIDIA GeForce RTX 3080 (10 GB) graphics card. Additionally, the settings of experimental parameters are shown in Table 2.

Table 2. Experimental parameter settings.

Parameter	FER2013	CK+	RAF-DB
Loss function	Cross Entropy	Cross Entropy	Cross Entropy
Learning rate	0.01	0.01	0.01
Optimizer	SGD	SGD	SGD
Batch size	16	16	16
Momentum	0.9	0.9	0.9
Learning rate decay	0.5/50	0.5/50	0.5/50
Epochs	300	300	300

4.1.2. Evaluation Indicators

The evaluation indicators commonly used by machine learning for classification models are as follows: accuracy and confusion matrix (also known as the error matrix). The accuracy refers to the proportion of the correct number of samples output by the model to the total number of samples, which can be expressed as

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

where TP, TN, FP, and FN represent the number of samples of True Positive, True Negative, False Positive, and False Negative, respectively. Obviously, the sum of these four is the total number of samples. The four indicators of TP, TN, FP, and FN are presented together in a table called a confusion matrix, which can analyze the misclassification of each category.

The size of the model is generally measured by the number of parameters and corresponds to spatial concepts and spatial complexity. Because of the large number of parameters in many models, they are usually measured in units of megabytes (MB). This can be represented as

$$param_{conv} = k_w * k_h * c_{in} * c_{out} \tag{10}$$

$$param_{fc} = n_{in} * n_{out} \tag{11}$$

where $k_w * k_h$, c_{in}, c_{out}, n_{in}, and n_{out} represent the size of the convolution kernel, the number of input channels, the number of output channels, the number of input channels of the dense layer, and the number of output channels of the dense layer, respectively. Finally, the parameter size of each layer is added together to obtain the total number of parameters.

4.1.3. Experimental Datasets

To verify the validity of the proposed model, we conduct experiments using FER2013, CK+, and RAF-DB datasets. These datasets contain different data scales and image complexity.

The FER2013 dataset contains a total of 35,886 facial expression samples. In this paper, the original sample is expanded to 60,000 based on the promoted random masking method. Among them, there are 46,000 training sets and 14,000 testing sets, the size of each sample is 48 × 48, and all samples are composed of grayscale images. And the dataset contains seven expressions of disgust, fear, anger, sad, happy, surprised, and neutral.

The CK+ dataset contains a total of 123 participants and 593 image sequences, all of which were collected under certain laboratory conditions. It is an upgraded version of the Cohn Kanda dataset with seven expressions. In this paper, the promoted random

masking method is used to expand its samples to 1500, including 1050 training sets and 450 testing sets.

The RAF-DB dataset contains a total of 29,672 facial expression samples, all of which were sourced from real-life facial expression datasets. This dataset contains seven basic emoticon labels and twelve composite emoticon datasets. This paper selects a basic expression dataset for the experiment, and expands the basic expression sample to 23,008 pieces through the promoted random masking method, including 16,106 training sets and 6902 testing sets.

4.2. Ablation Experiments

To show the availability and rationality of our method, ablation experiments are performed for each module, and the experimental results are shown in Table 3. RM denotes the promoted random masking method, Clip_K5_ShuffleNet denotes the improved inverted residual structure, SA denotes the shuffle attention mechanism module, RLR-CNet denotes the dual-branch network with cumulative learning, and RM + Clip_K5_ShuffleNet + SA + RLR-CNet denotes the lightweight network proposed in our paper.

Table 3. Performance comparison of different modules of network.

Model	FER2013	CK+	RAF-DB	Parameter
ShuffleNet	65.71%	95.23%	84.31%	1.24 MB
ShuffleNet + RM	67.87%	95.91%	85.11%	1.24 MB
ShuffleNet + RM + Clip_K5	67.84%	95.87%	84.98%	0.94 MB
ShuffleNet + RM + Clip_K5 + SA	69.35%	96.35%	86.44%	0.96 MB
RLR-CNet (ours)	71.14%	98.04%	87.93%	1.02 MB

Firstly, using ShuffleNet as the base network, the training samples are input into the network through a promoted random masking operation, and in order to prevent information loss and accelerate the speed of network operation, the lightweight ShuffleNet is improved to obtain a lighter Clip_K5_ShuffleNet network. To further extract local facial expression detail features, the SA module is integrated into the network to redistribute the feature weights of facial expressions from the two channel and spatial dimensions. To ensure the practicability and effectiveness of the network in real scenes, RLR-CNet is proposed in this paper. The effectiveness of each improvement module in the network is shown in Table 3.

In Table 3, it can be seen that after RM processing, the recognition accuracy of ShuffleNet on FER2013, CK+, and RAF-DB datasets increases by 2.16%, 0.68%, and 0.80%, respectively. The introduction of Clip_K5_ShuffleNet leads to a slight decrease in the recognition accuracy of the network, but on the other hand, it reduces the number of parameters, which is beneficial for accelerating the operation speed of the model. The introduction of the SA module further improves the recognition accuracy of the network.

Compared to the original network, our RLR-CNet promotes the recognition accuracy of the network by 5.43%, 2.81%, and 3.62%, and reduces the number of parameters by 17.74%. This shows that our method in the paper has some advantages in recognition accuracy and lightweight parameters.

The experimental training process of the proposed method in this paper is shown in Figure 8. For Figure 8a, the recognition accuracy growth of the model slows down at the 100th epoch, and the accuracy gradually becomes stable when it reaches the 200th epoch, with the highest accuracy reaching 71.14%. For Figure 8b, the recognition accuracy of the model rapidly increases at the beginning of the training, and when it reaches the 150th epoch, the accuracy tends to stabilize, with the highest accuracy reaching 98.04%. For Figure 8c, the recognition accuracy of the model continues to increase, and when it reaches the 180th epoch, the accuracy gradually becomes stable, with the highest accuracy reaching 87.93%. For Figure 8, the accuracy of the training set is always less than or equal to the accuracy of testing sets, and the accuracy of the two sets is generally very close, indicating

that the method proposed in this paper can better capture the data characteristics and fit the data, and has better generalization.

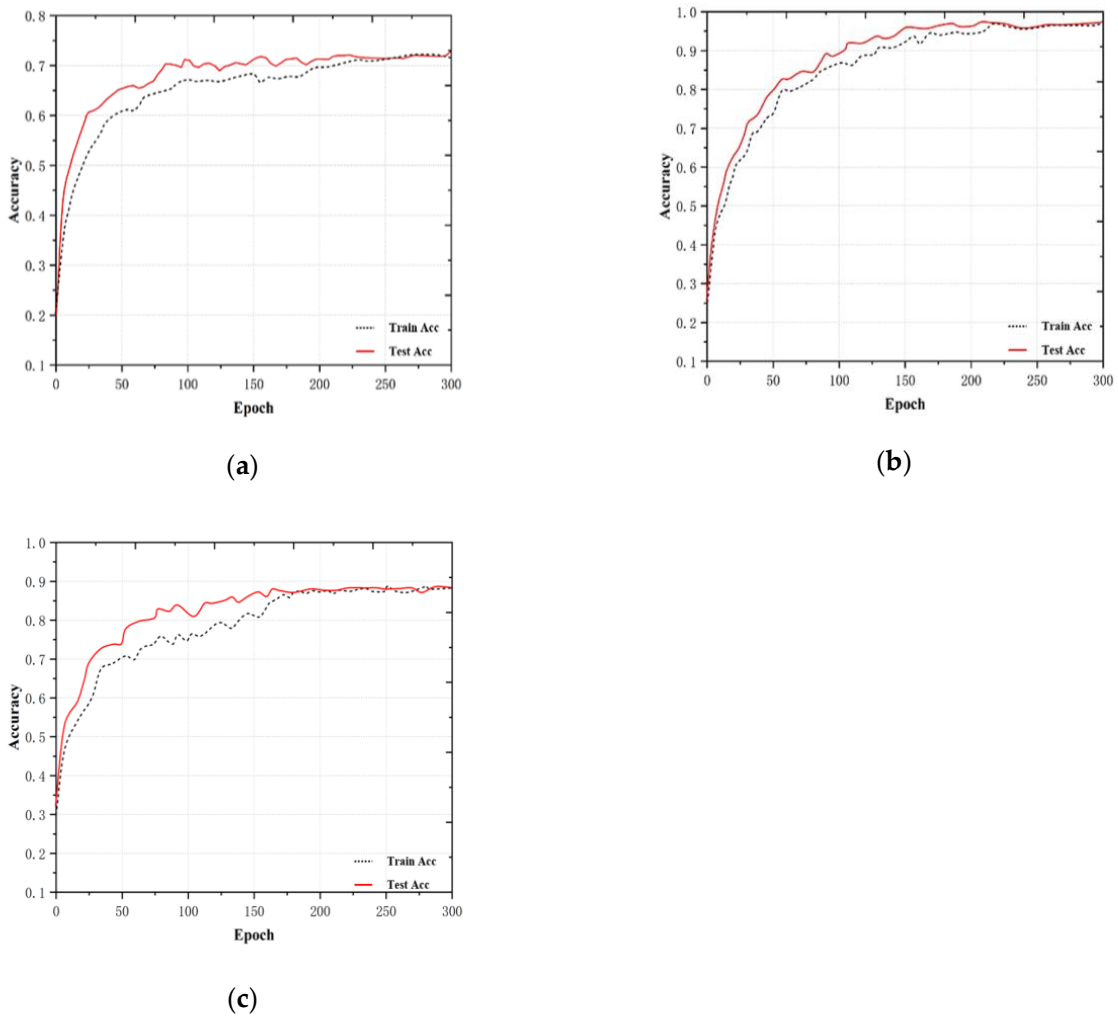

Figure 8. Training curve. (**a**) The experimental training process on the FER2013 dataset; (**b**) the experimental training process on the CK+ dataset; (**c**) the experimental training process on the RAF-DB dataset.

For exploring more accurately the recognition performance of our method for FER, the confusion matrices are drawn based on the experimental results on the FER2013, CK+, and RAF-DB datasets, as shown in Figure 9. For the FER2013 dataset, the recognition accuracy of the happy and surprised expressions is relatively high, both exceeding 80%. For the CK+ dataset, all facial expressions' recognition accuracies are above 95%. For the RAF-DB dataset, the happy, surprised, sad, and neutral expressions all have comparatively high recognition accuracy. Overall, on the three datasets, the recognition accuracy of happy and surprised expressions is relatively high, while the recognition accuracy of angry expressions is relatively low. This is because happy expressions often have significant features such as raised corners of the mouth and wrinkles around the eyes that are easy to recognize, and surprised expressions have recognized features such as an open mouth and wide eyes. Negative expressions such as anger and fear have strong similarities, which make it difficult to distinguish subtle changes, resulting in low recognition accuracy. Overall, the proposed method achieves good recognition performance for various facial expressions on the FER2013, CK+, and RAF-DB datasets.

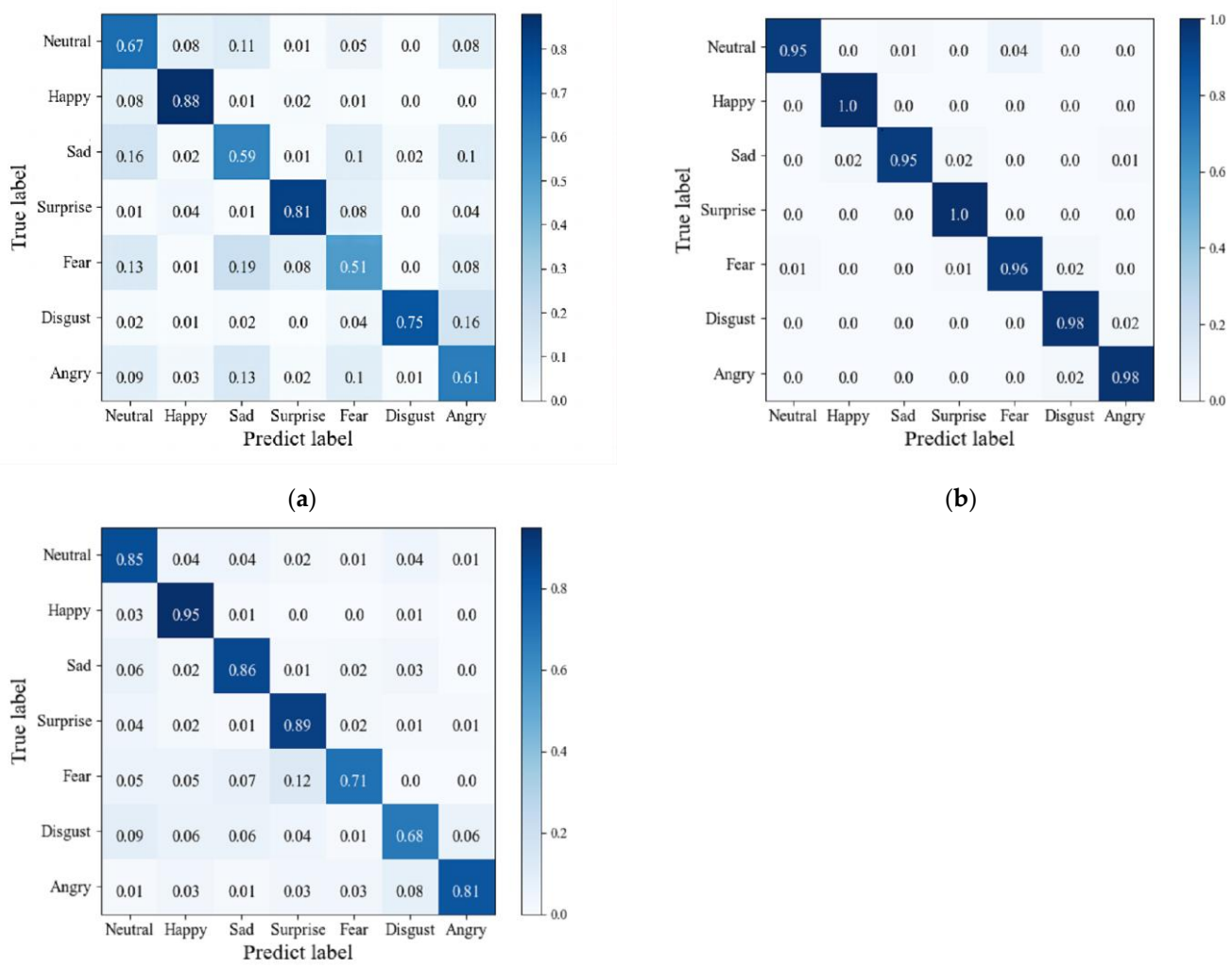

Figure 9. Confusion matrix. (**a**) Confusion matrix on the FER2013 dataset; (**b**) confusion matrix on the CK+ dataset; (**c**) confusion matrix on the RAF-DB dataset.

4.3. Comparative Experiment of Mainstream Algorithms

To prove the effectiveness of the RLR-CNet proposed in this paper for FER, a comparative experiment is carried out with several mainstream algorithms such as ResNet18, ResNet50, VGG16, VGG19, and AlexNet from the aspects of the number of parameters of the model and the recognition accuracy. The experimental results are shown in Table 4.

Table 4. Comparison experiments of mainstream algorithms.

Model	FER2013 (%)	CK+ (%)	RAF-DB (%)	Parameter (MB)
ResNet18	70.09	89.39	84.10	11.69
ResNet50	71.26	92.46	86.01	25.56
VGG16	68.89	95.46	81.68	14.75
VGG19	68.53	92.18	81.17	20.06
AlexNet	67.51	87.59	55.60	60.92
ours	71.14	98.04	87.93	1.02

For the FER2013 dataset of the facial expressions' recognition accuracy, the proposed method in our paper achieves the highest accuracy among all the models, with an improvement of approximately 1.05% compared to ResNet18. For the CK+ dataset, the recognition accuracy of VGG16 is higher than those of other mainstream networks, while the proposed

method in our paper achieves an improvement of 2.58% compared to VGG16. For the RAF-DB dataset, the recognition accuracy of the proposed method in our paper reaches 87.93%. The improvement in the recognition accuracy further verifies the effectiveness and strong generalization ability of the proposed method. In terms of model parameters, the parameter size of the proposed model is 1.02 MB, which is lowest among the mainstream algorithms. In general, while achieving a lightweight model, our method ensures great recognition accuracy, and then it demonstrates the efficiency and superiority of this paper's method.

To further prove the effectiveness of our method, we conduct comparative experiments with some existing advanced methods on FER2013, CK+, and RAF-DB datasets. The advanced methods mainly include DCN and Inception V4, which are very novel in recent years, as well as lightweight networks such as Mini-Xception, MFN, and MANet, and classification networks embedded with modules such as SE and CBAM attention mechanisms. The experimental results are shown in Tables 5–7, respectively. It can be observed from the experimental results that, for the FER2013 dataset, the recognition accuracy is above 65%, while our method achieves an accuracy of 71.14%. For the CK+ dataset, the recognition accuracy is above 94%, while our method achieves an accuracy of 98.04%. For the RAF-DB dataset, the recognition accuracy is above 75%, while our method achieves an accuracy of 87.93%. This further proves that our method can ensure great accuracy in FER under the condition of lightweight implementation.

Table 5. Performance comparison of different methods on FER2013 dataset.

Model	Accuracy (%)
MANet [35]	69.46
Inception V4 [36]	66.80
DCN [37]	69.30
Minace [38]	70.20
ours	71.14

Table 6. Performance comparison of different methods on CK+ dataset.

Model	Accuracy (%)
DeRL [39]	97.30
APRNET50 [40]	94.95
ResMasking [41]	98.46
DTAGN [42]	97.25
ours	98.04

Table 7. Performance comparison of different methods on RAF-DB dataset.

Model	Accuracy (%)
Mini-Xception [43]	76.26
MFN [44]	85.39
SCN [45]	87.03
LA-Net [46]	87.00
ours	87.93

5. Conclusions

Aiming at the problems of a large number of parameters and a low accuracy of the current FER model, a lightweight model of FER based on class-rebalancing fusion cumulative learning is proposed in this paper. Through the proposed RLR-CNet, the problem where the model's recognition accuracy is affected, which is due to the imbalanced inter-class distribution of facial expression data, is solved, and the reduction in the model's parameters also speeds up the operation speed of the model to a certain degree. And the embedding of the SA module enhances the ability of the model to extract local details of

facial expressions, and then improves the accuracy of FER on the lightweight network as a whole. According to the experimental results, the accuracy of the proposed method is 71.14% on the FER2013 dataset, 98.04% on the CK+ dataset, and 87.93% on the RAF-DB dataset. The model contains fewer parameters and achieves great recognition accuracy while implementing a lightweight network, and its accuracy is better than most current mainstream algorithms, demonstrating better effectiveness and applicability. In this paper, the object of study is the single-label expression dataset. The single-label expression can well represent the emotions contained in various expressions and has a good recognition accuracy for common expression categories, but in real scenes, there are still some minor expression categories that are not taken into account, such as tension and pride. The recognition of these minor expressions needs to be achieved by studying composite multi-label expressions. Therefore, in future research, more attention should be paid to the study of compound multi-label facial expression recognition.

Author Contributions: Conceptualization, X.M. and Y.S.; methodology, X.M., Y.S. and R.W.; experimental test, Y.S.; writing—original draft preparation, Y.S. and X.M.; writing—review and editing, X.M., Y.S., R.W., Y.T. and Y.X.; supervision, X.M., R.W. and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation Project of Guangxi Normal University (Grant No.: 2021JC012); Science and Technology Planning Project of Guangxi Province, China (No. 2022AC21012); the industry-university-research innovation fund projects of China University in 2021 (No. 2021ITA10018); the fund project of the Key Laboratory of AI and Information Processing (No. 2022GXZDSY101).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The sources of the datasets in this article are as follows: FER2013: "https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge"; CK+: "http://www.pitt.edu/~emotion/ck-spread.htm"; RAF-DB: "http://www.whdeng.cn/RAF/model1.html". Readers can apply based on the above URL.

Acknowledgments: The authors thank Haiying Xia and Lintao Chen for providing theoretical and technical support during the experiments. We also thank other partners in the laboratory, such as Hongyang Chen, Juan Hu, and Mengchen Yan, for their help during the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [CrossRef]
2. Canedo, D.; Neves, A.J.R. Facial Expression Recognition Using Computer Vision: A Systematic Review. *Appl. Sci.* **2019**, *9*, 4678. [CrossRef]
3. Shahzad, H.M.; Bhatti, S.M.; Jaffar, A.; Akram, S.; Alhajlah, M.; Mahmood, A. Hybrid Facial Emotion Recognition Using CNN-Based Features. *Appl. Sci.* **2023**, *13*, 5572. [CrossRef]
4. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010.
5. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998.
6. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the 20th International Conference on Neural Information Processing (ICONIP), Daegu, Republic of Korea, 3–7 November 2013.
7. Li, S.; Deng, W.; Du, J.P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
8. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef]
9. Sadeghi, H.; Raie, A.A. HistNet: Histogram-based convolutional neural network with Chi-squared deep metric learning for facial expression recognition. *Inf. Sci.* **2022**, *608*, 472–488. [CrossRef]

10. Kim, J.C.; Kim, M.H.; Suh, H.E.; Naseem, M.T.; Lee, C.S. Hybrid Approach for Facial Expression Recognition Using Convolutional Neural Networks and SVM. *Appl. Sci.* **2022**, *12*, 5493. [CrossRef]
11. Gong, W.; Wang, C.; Jia, J.; Qian, Y.; Fan, Y. Multi-feature Fusion Network for Facial Expression Recognition in the Wild. *J. Intell. Fuzzy Syst.* **2022**, *42*, 4999–5011. [CrossRef]
12. Ge, H.; Zhu, Z.; Dai, Y.; Wang, B.; Wu, X. Facial expression recognition based on deep learning. *Comput. Methods Programs Biomed.* **2022**, *215*, 106621. [CrossRef] [PubMed]
13. Bian, J.; Mei, X.; Xue, Y.; Wu, L.; Ding, Y. Efficient hierarchical temporal segmentation method for facial expression sequences. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 1680–1695. [CrossRef]
14. Hassaballah, M.; Aly, S. Face recognition: Challenges, achievements and future directions. *IET Comput. Vis.* **2015**, *9*, 614–626. [CrossRef]
15. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9 November 2015.
16. Chun, C.; Ryu, S.K. Road Surface Damage Detection Based on Semi-supervised Learning Using Pseudo Labels. *J. Korea Inst. Intell. Transp. Syst.* **2019**, *18*, 71–79. [CrossRef]
17. Yao, L.; He, S.; Su, K.; Shao, Q. Facial expression recognition based on spatial and channel attention mechanisms. *Wirel. Pers. Commun.* **2022**, *125*, 1483–1500. [CrossRef]
18. Siqueira, H.; Magg, S.; Wermter, S. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New York Hilton Midtown, New York, NY, USA, 7–12 February 2020.
19. Li, S.; Deng, W. A deeper look at facial expression dataset bias. *IEEE Trans. Affect. Comput.* **2020**, *13*, 881–893. [CrossRef]
20. Pan, B.; Wang, S.; Xia, B. Occluded facial expression recognition enhanced through privileged information. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 15 October 2019.
21. Xia, B.; Wang, S. Occluded Facial Expression Recognition with Step-Wise Assistance from Unpaired Non-Occluded Images. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12 October 2020.
22. Mahmoudi, M.A.; Chetouani, A.; Boufera, F.; Tabia, H. Kernel-based convolution expansion for facial expression recognition. *Pattern Recognit. Lett.* **2022**, *160*, 128–134. [CrossRef]
23. Kong, Y.; Ren, Z.; Zhang, K.; Zhang, S.; Ni, Q.; Han, J. Lightweight facial expression recognition method based on attention mechanism and key region fusion. *J. Electron. Imaging* **2021**, *30*, 063002. [CrossRef]
24. Nan, Y.; Ju, J.; Hua, Q.; Zhang, H.; Wang, B. A-MobileNet: An approach of facial expression recognition. *Alex. Eng. J.* **2022**, *61*, 4435–4444. [CrossRef]
25. Zhou, N.; Liang, R.; Shi, W. A lightweight convolutional neural network for real-time facial expression detection. *IEEE Access* **2020**, *9*, 5573–5584. [CrossRef]
26. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York Hilton Midtown, New York, NY, USA, 7–12 February 2020.
27. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton New Orleans Riverside, New Orleans, LA, USA, 2–7 February 2018.
28. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet V2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
29. Peng, C.; Zhang, X.; Yu, G.; Luo, J.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
30. Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv* **2019**, arXiv:1908.08681.
31. Zhang, Q.L.; Yang, Y.B. SA-Net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021.
32. Horn, G.V.; Perona, P. The devil is in the tails: Fine-grained classification in the wild. *arXiv* **2017**, arXiv:1709.01450.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
34. Zhou, B.; Cui, Q.; Wei, X.S.; Chen, Z.M. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
35. Gan, Y.; Chen, J.; Yang, Z.; Xu, L. Multiple attention network for facial expression recognition. *IEEE Access* **2020**, *8*, 7383–7393. [CrossRef]
36. Momeny, M.; Neshat, A.A.; Jahanbakhshi, A.; Mahmoudi, M.; Ampatzidis, Y.; Radeva, P. Grading and fraud detection of saffron via learning-to-augment incorporated Inception-v4 CNN. *Food Control* **2023**, *147*, 109554. [CrossRef]
37. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016.
38. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef] [PubMed]
39. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.

40. Chen, J.; Xu, Y. Expression recognition based on the convolution residual network of attention pyramid. *Comput. Eng. Appl.* **2022**, *58*, 123–131.
41. Pham, L.; Vu, T.H.; Tran, T.A. Facial Expression Recognition Using Residual Masking Network. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021.
42. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015.
43. Arriaga, O.; Valdenegro, T.M.; Plöger, P. Real-time convolutional neural networks for emotion and gender classification. *arXiv* **2017**, arXiv:1710.07557.
44. Tang, H.; Xiang, J.; Chen, H.; Lu, R.; Xia, Z. Lightweight facial expression recognition method based on multi-region fusion. *Laser Optoelectron. Prog.* **2023**, *60*, 0610006.
45. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
46. Ma, H.; Celik, T.; Li, H.C. Lightweight attention convolutional neural network through network slimming for robust facial expression recognition. *Signal Image Video Process.* **2021**, *15*, 1507–1515. [CrossRef]