An Expression Recognition Model Combining Attention and Residual Networks

Jiale Gu[*], *Xiaohong Jin*

Department of Financial Technology, Suzhou Industrial Park Institute of Services Outsourcing, China
gujl@siso.edu.cn, jinxh@siso.edu.cn

Abstract

How to integrate multi-scale features and establish interdependencies between remote channels is a challenge for expression recognition networks. Based on this, the authors put forward a residual network on PSA (pyramid split attention)-ResNet, which replaces the 3 x 3 convolution in the ResNet50 residual module with PSA to draw multi-scale features and enhance the cross-channel information's correlation availably. At the same time, to reduce the differences between similar expressions and expand the distance between different types of expressions, a joint loss function about Island Loss and SoftMax Loss has been introduced to optimize the parameters. The method proposed in this paper carried out simulation experiments with 2 datasets, Fer2013 and CK+, increasing precision rates to 74.26% and 98.35% respectively. This result further confirms that the method has yielded a great result on expression recognition compared with the cutting-edge algorithms.

Keywords: Pyramid split attention, Residual network, Channel attention, Group convolution

1 Introduction

Facial expressions contain abundant emotional information. The variables of facial expressions reflect the fluctuations on psychological emotions in interpersonal communication [1-2]. According to statistics, 55% of information in daily communication is transmitted through facial expressions [3]. Facial expression recognition is a further development of visual detection technology. How to enable computers to correctly recognize the expression information on faces is a significant and challenging task in the computer field. Recently, facial expression recognition has spread widely in the fields of smart classrooms, investigation and interrogation, safe driving, and medical diagnosis, gradually becoming a research hotspot in academia and industry. A complete facial expression recognition process is composed of four steps: facial expression acquisition, facial expression preprocessing, facial expression feature abstraction, and facial expression classification [4]. Among them, feature extraction is the key to facial expression recognition technology, which directly affects the final recognition effect. It mainly includes two different methods: traditional manual feature extraction and deep learning based automatic feature learning.

Among traditional facial expression recognition methods, manually designed special operators have been utilized for feature extraction, and then the feature vectors are supplied into classifiers like SVM (Support Vector Machine) and KNN (K-Nearest Neighbor Algorithm) to output the expression recognition results. Common feature extraction operators include LBP (Local Binary Pattern) for extracting local texture features, HOG (Histogram of Oriented Gradient) for extracting edge features, ASM (Active Shape Models) for deriving geometric features, and Color Moments for deriving color features. Traditional feature derivation algorithms already get certain results but manually designed feature extractors are still limited by the designer's own experience and knowledge. They extract comparatively low-level features, which have an inaccuracy effect in classification for the high-level semantic information.

In tandem with the growth of deep learning, convolutional neural networks have been broadly applied in computer vision, and neural network models such as VGGNet, GoogleNet, ResNet have successively emerged. Unlike traditional feature extraction algorithms, convolutional neural networks extract deep features through multi-layer convolution and nonlinear transformations, and they use error back propagation algorithms to continuously optimize network parameters, thereby automatically extracting semantic information from images and improving model recognition accuracy. In the field of expression recognition, by using ResNet [5] to train the two datasets FER2013 and CK +, the model's precision was as high as 69.50% and 97.21% respectively. Although deep learning has yielded fair results in expression recognition, with the network layers multiplies, the model parameters become increasingly large, and the computational complexity increases exponentially. Nevertheless, in practical classification tasks, the productive expression information usually allocates only in the local regions of the image. A great deal of non-expression information gives rise to feature redundancy, thus influencing the model recognition's accuracy.

Some researchers have utilized the attention mechanism into facial expression recognition to strengthen the ability to extract key features to represent fruitful

[*]Corresponding Author: Jiale Gu; Email: gujl@siso.edu.cn

DOI: https://doi.org/10.70003/160792642025092605012

feature information. Hu [6] et al. proposed a SENet (Squeeze and Excitation Networks) based on channel attention, which adjusts the weights of diverse channels in feature map to increase weights of salient feature channels while suppressing the features of non-salient channels. Jaderberg [7] et al. designed a Spatial Attention (SA) mechanism to assign weights to different positions in an image, which focuses on specific areas of the image to discover the relationships between pixels in different positions. Woo [8] et al. offered a CBAM (Convolutional Block Attention Module), linking channel attention and spatial attention to obtain more detailed information that needs attention. Practice has shown that although the use of channel attention [9-13], spatial attention [14], or their fusion [15-16] can significantly improve the performance on expression recognition, it ignores the extraction of expression features at different levels and granularities, and constructs multi-scale feature maps to understand the interdependence between cross channel information. Based on these research results, this article presents an expression recognition network that integrates pyramid segmentation attention and joint loss, and its main contributions are as follows:

(1) We have introduced the PSA (Pyramid Split Attention) module [17] by utilizing a multi-scale pyramid convolutional structure to integrate feature information of various scales onto every channel level feature map, using channel attention weights to enhance the correlation of multi-scale cross channel information, and establishing remote channel dependency relationships.

(2) We have replaced the 3*3 convolution in the residual module of the ResNet50 network with PSA, forming a Pyramid Split Attention Based on ResNet (PSA-ResNet). The multi-scale features extracted by this network can improve the accuracy of expression recognition and provide better pixel level attention.

(3) To narrow the distance between similar expressions and reduce the probability of model misjudgment, we have trained the network by using the combined loss function of SoftMax Loss and Island Loss. We carried out a large number of experiments on the two datasets, FER2013 and CK +, which verifies the validity of the methods.

2 Construction of an Expression Recognition Model

To improve the cross-channel representation ability of multi-scale key features and capture subtle changes between different expressions, this article proposes a residual network model based on pyramid segmentation attention (PSA-ResNet) for recognizing various facial expressions. The entire network architecture is shown in Figure 1.

Firstly, we obtain the initial weight parameters of the network by prior training the model of ResNet50 based on ImageNet dataset. Then, we transfer it to the PSA-ResNet model to have it learn rich multi-scale features, integrate information from different scales into channel level feature maps, recalibrate the cross channel attention weights of multi-scale features, and enhance the weight values of regions with significant facial expression changes. Finally, we use a joint function of SoftMax Loss and Island Loss for parameter optimization to reduce the probability of misjudgment in facial expression recognition to upgrade the accuracy of model recognition during the training process.

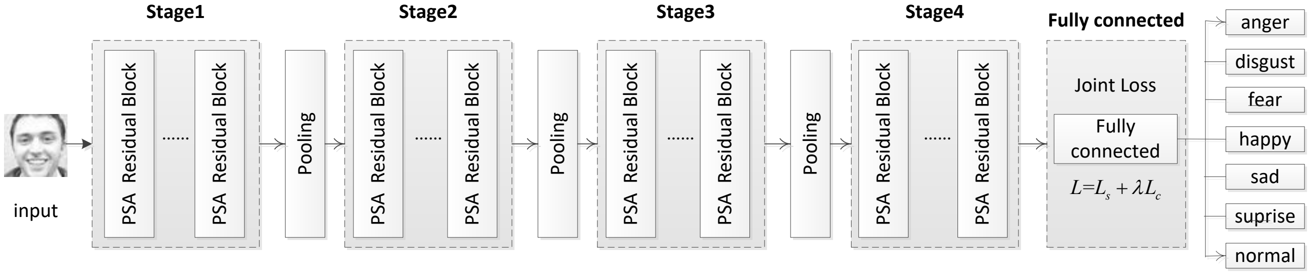

Figure 1. PSA-ResNet framework

2.1 PSA Residual Block

Convolutional neural networks have witnessed an increasing number of layers to extract high-level semantic information, such as the 8-layer AlexNet, 19-layer VGGNet, and 22-layer GoogLeNet. However, merely increasing the network layer count does not improve the model's feature learning ability. When models reach a certain depth, network degradation may happen, resulting in a decline in accuracy. ResNet can address the issues of gradient explosion and vanishing in deep neural networks by means of residual learning and identity mapping, thus upgrading the training efficiency of neural networks. Among the ResNet series, ResNet34, ResNet50, and ResNet101 are widely used. Compared with ResNet34, ResNet50 adopts three-layer residual blocks rather than two-layer ones. It can maintain model accuracy while significantly reducing the number of parameters. As for ResNet101, it has too many layers and pays too much attention to semantic information while neglecting detailed features. Consequently, this paper selects ResNet50 as the baseline model for facial expression recognition feature extraction.

The ResNet network is a network structure proposed by the Microsoft Research Institute to solve the phenomenon of gradient explosion and gradient disappearance that occurs in deep neural networks training. This model effectively improves the training efficiency and accuracy of the neural network through residual learning and identity mapping. As shown in Figure 2, where x represents the input of the neural network, residual mapping F(x) is composed of two weight layers and a ReLU activation function. H(x) = F(x) + x is the ideal mapping, which is the output after adding the input x. By converting the fitting objective function from H(x) to F(x) and turning the output into the superposition of the input and the residual mapping, the network is made more sensitive to the slight changes of the output H(x) and the input x.

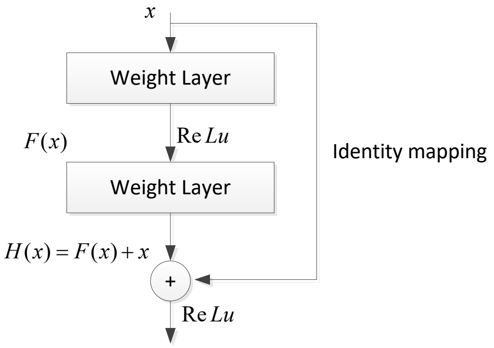

Figure 2. ResNet residual learning and identity mapping

As an important member of the ResNet lines, ResNet50 consists of 3, 4, 6, and 3 residual blocks in layer2-layer5 respectively. The ResNet residual block structure is demostrated in Figure 3(a). We replace the site in accordance with the 3*3 convolution with PSA to obtain a new PSA residual block based on pyramid segmentation attention, as shown in Figure 3(b). Correspondingly, the PSA residual blocks are stacked according to the ResNet50 network style to obtain a new PSA-ResNet based network model, which is used as the backbone network for facial expression recognition.

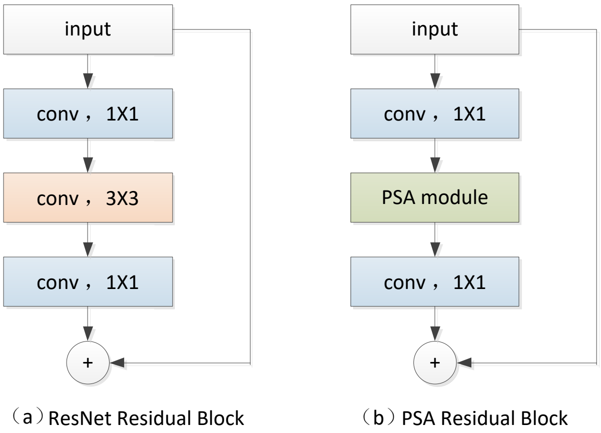

Figure 3. ResNet and PSA residual block

2.2 Pyramid Segmentation Attention PSA Module

The attention mechanism enables neural networks to allocate different weights to different parts of the input data, thereby selecting the most critical information for the current task. To establish an efficient attention mechanism, this paper introduces the PSA module, which enhances the message correlation between multi-scale features and cross channels by using polyscale pyramid convolutional structures and channel attention mechanisms. It captures expression information at different levels and granularities, thereby improving the accuracy of expression prediction.

The PSA module structure is presented in Figure 4. The implementation process is divided into the following four steps:

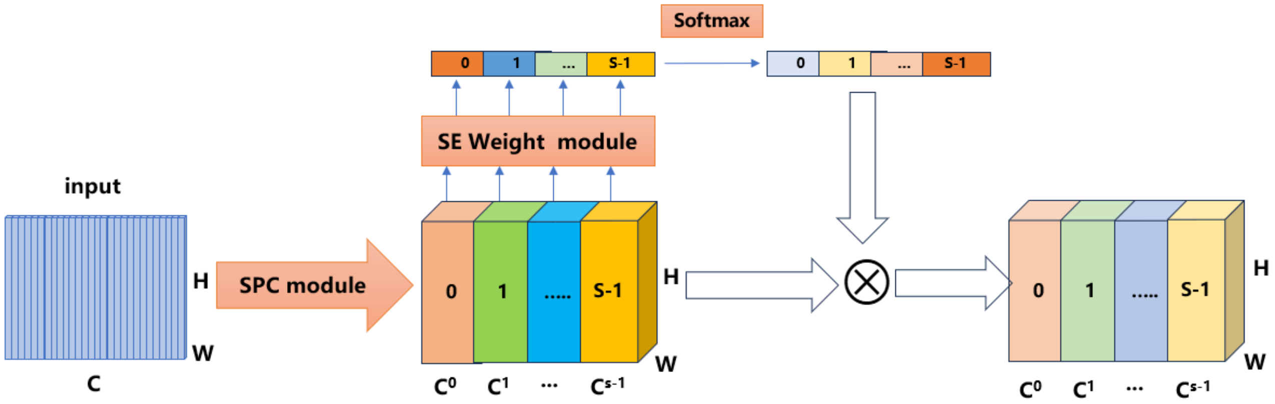

Figure 4. PSA structure

(1) Use the segmentation and fusion module (SPC) to segment the input feature vector from the channel direction into several groups and use the convolutional kernels of various scales in pyramid structure to extract feature information containing multiple scales on different channel levels.

(2) Send the SPC module output to the SE Weight module to calculate the weight values of various channels and acquire the attention vector of every channel feature map.

(3) Normalize the attention vector through the SoftMax function, readjust the attention weight of channels, and get new cross-channel attention weights.

(4) Perform point multiplication on the obtained polyscale spatial information and multi-channel attention weights element, and output the fine-grained features with more affluent polyscale feature information.

2.2.1 Segmentation Fusion SPC Module

The SPC module calculation process is presented in Figure 5. It mainly realizes multi-scale feature extraction in a multi-branch manner. Each branch processes the input feature map in parallel using different-sized convolution kernels to obtain different scales of receptive fields, draw spatial information in various scales, and obtain more affluent feature representations.

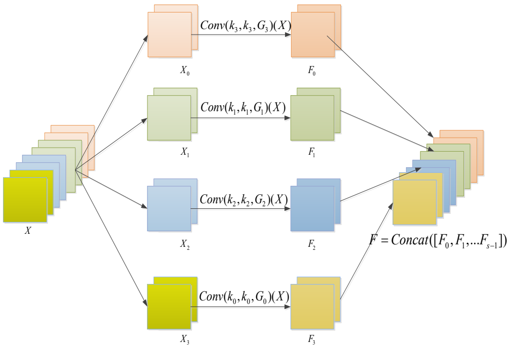

Figure 5. Calculation process of segmentation fusion module of SPC

First, the feature map of input is separated into S (S = 4) parts in terms of channels, which is conveyed as the channels' number $[X_0, X_1 ..., X_{s-1}]$ to meet each part. The different-sized convolution kernels are used respectively, and the size of every distributed convolution kernel satisfies the following:

$$k_i = 2\times(i+1)+1 (i = 0,1,...,S-1) \tag{1}$$

With the increase of the convolution kernel's size, it causes a rapid increase in the calculation amount. Based on this, in the SPC module, the features of each portion segmented are applied with grouped convolutions, which effectively avoids the increase in the calculation amount caused by the increase in the convolution kernel's size. The groups and the convolution kernel's size meet the following relationship formula:

$$G_i = 2^{\frac{k_i - 1}{2}} \tag{2}$$

Therefore, the polyscale feature map's generation function can be expressed as:

$$F_i = Conv(k_i \times k_i, G_i)(X), i = 0,1,...,S-1 \tag{3}$$

Finally, multi-scale feature maps are stitched in the channel direction to obtain an overlaid multi-channel feature, and the stitching function is shown as follows:

$$F\text{=}Cat([\mathrm{F}_0, \mathrm{F}_1, ..., \mathrm{F}_{s\text{-}1}]) \tag{4}$$

2.2.2 Channel Attention SE Module

The channel attention mechanism enables the network to calculate the significance of every channel of the feature map, with the aim that the neural network can fix the feature channels, which are effective in task, and suppress the feature channels that are meaningless for the current task to achieve the goal of enhancing the performance of the entire network. The SE module structure is presented in Figure 6, with two parts: squeeze and excitation.

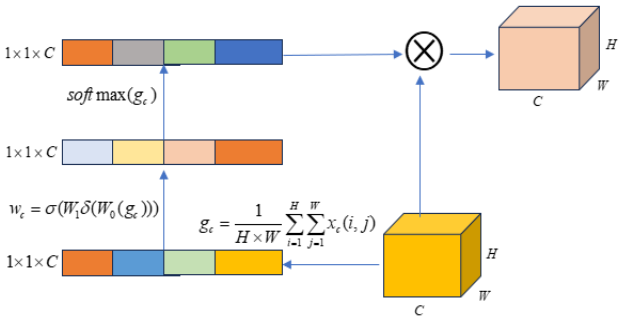
$1\times1\times C$
$soft\max(g_c)$
$1\times1\times C$
$w_c = \sigma(W_1\delta(W_0(g_c)))$
$g_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j)$
$1\times1\times C$
$\otimes$
H
W
C
H
W
C

Figure 6. SE module structure

First, channel dimension of the input feature map is compressed using global average pooling to obtain a global feature map, whose dimension is 1x1xc. If the input of the channel c is x, the calculation formula for global average pooling is:

$$g_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j) \tag{5}$$

Second, the interrelationship between different channels is calculated using the completely connected layer and the initiation function to gain the attention weights. The attention weight of the channel c can be expressed as:

$$w_c = \sigma(W_1\delta(W_0(g_c))) \tag{6}$$

In this formula, σ and δ represent the operations of the ReLU and sigmoid activation functions respectively, and the weights of the two fully connected layers are represented by W_0 and W_1 respectively. The completely connected layer introduces non-linear transformation through the activation function by linearly combining the input data with the weight matrix, and it more effectively models the interdependent relationship between the feature channels.

Finally, the attention weight vector is normalized using the softmax function, and it is fused with the output of the polyscale feature abstraction module to better express the image features.

2.3 Joint Loss Function

Although the traditional SoftMax Loss function expands the distance between categories, ensuring that different categories are clearly separable, it does not consider the differences within the categories. In the expression recognition task, the difference between different people making the same expression may be much greater than the difference between a person making different expressions. If only SoftMax Loss is used, it may lead to misjudgment of the expression, thereby affecting the correct recognition of the expression by the model. To obtain a better effect of expression classification, in the design of the loss function, it is necessary to consider how to narrow the distance within the same type of expression and expand the distance between different types of expressions.

Therefore, this paper introduces the joint loss function of SoftMax Loss and Island Loss. The ability of Center Loss to narrow the distance within the class is strong, making the data of the same type of expression perform more tightly, which is beneficial to improve the classification effect. When introduced into the network model, the calculation process is as follows:

$$L_{IL} = L_c + \lambda_1 \sum_{c_j \in N} \sum_{\substack{c_k \in N \\ c_k \neq c_j}} \left(\frac{c_k \bullet c_j}{\| c_k \|_2 \| c_j \|_2} + 1 \right) \quad (7)$$

Here, N is the set of expression labels, c_k and c_j are the centers of L2 norm for k-class samples and j-th class samples. $(\cdot)$ represents dot product. The first item is the gap between the penalty sample and its corresponding core, and the second item is the similarity between the penalty expressions. λ_1 is used to balance these two items. The joint loss function of the PSA-ResNet model can be expressed with the following formula:

$$L = L_s + L_{IL} \quad (8)$$

Here, L_s represents the Softmax Loss function. By minimizing Island loss, the recognition performance between different faces with the same expression and different expressions on the same person's face can be improved.

3 Experiments and Analyses

3.1 Experimental Dataset

To evaluate the expression recognition model proposed in this paper effectively, training and testing are carried out on the FER2013 and CK+ datasets, and comparative experiments are conducted with the current mainstream methods.

(1) FER2013 dataset: This dataset is composed of 35,886 gray-scale images with a resolution of 48x48. 28,708 are in the training set and 3,589 are in both the validation set and the test set. Each image is labeled with a right category, which includes 7 types of expressions with the numbers 0-6. The Chinese and English labels for the homologous labels are as listed: 0-anger, 1-disgust, 2-fear, 3-happy, 4-sad, 5-surprise, and 6-normal. In the training set, the image that appeared most times was the happy expression image, up to 7,215, and the number of disgust expression images was only 436. This uneven data distribution, as well as the presence of some label noise and images in non-face areas, may influence the training of the model. This paper implements following processing: a) Perform radiographic transformation, random flipping, rotation, and other data enhancement methods on the disgust expression images to expand the number of such images and balance the overall data distribution of the training set. b) Delete the images with some label errors and non-face areas. After processing a) and b), re-execute the operation in a) to enhance the data. The distribution before and after data enhancement is presented in Figure 7.

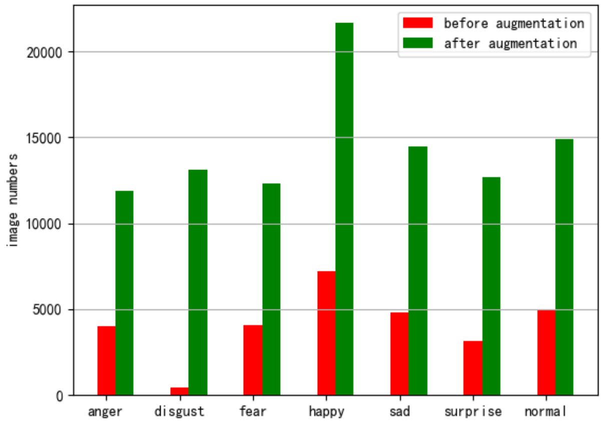

Figure 7. Distribution of FER2013 before and after data augmentation

(2) CK+ dataset: This is an extension of the CK dataset, which includes 327 labeled facial expression sequences of 123 objects. In this paper, four frames are randomly extracted from each sequence to form 1,308 labeled expression images as the training set. In order to improve the summary ability and robustness, a series of enhancement operations have also been implemented on the dataset, including random rotation, flipping, brightness adjustment, and color adjustment. The enhanced images are separated into the training set and validation set with a ratio of 2:1 for each type of expression. The distribution of the training set is presented in Figure 8.

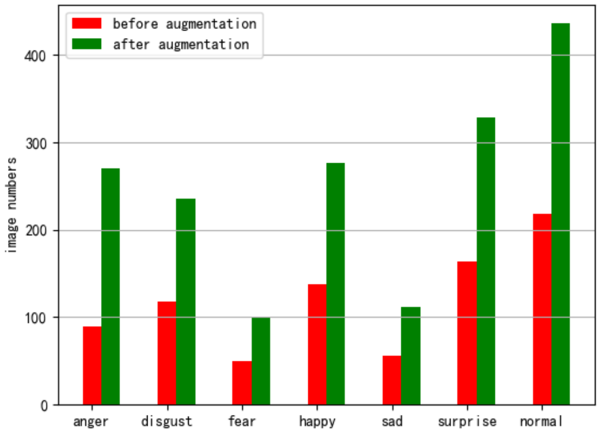

Figure 8. Distribution of CK+ before and after data augmentation

3.2 The Setting of Environment and Parameter

The software platform settings are as follows: Python3.8 is used as the program language. PyTorch2.0 is used to build the deep learning network framework. The operating system is 64-bit Microsoft Windows 10. The CPU is the i7-9700K with a memory of 128G. The model of the graphics card is NVIDIA GTX 2080Ti. The corresponding experimental environment parameters are shown as below.

Table 1. Settings of environment and parameter

Hardware information	Related configuration
Operating system	Microsoft Windows 10
Python	Python 3.8
Pytorch	PyTorch2.0
CPU	Intel(R) Xeon(R) CPU E5-2680 v4
GPU	RTX 3080 Ti(12GB)
Cuda	11.4

The hyperparameters setting of this experiment is as follows: The training batch is set to 200 rounds, 32 images per batch, with an initial learning rate of 0.001. The joint loss function is adopted, and the Adam optimizer is used in the experimental process to optimize the training process. When the validation loss function does not decrease within the 30 batches, the learning rate is reduced at a rate of 10 times. The model's hyperparameters are shown in Table 2.

Table 2. Model hyperparameter settings

Hyperparameter	Value
Epoches	200
Batchsize	32
Learning rate	0.001
Optimizer	Adam

3.3 Experimental Results Analyses

3.3.1 Experimental Results of the Dataset

Figure 9 presents the experimental results on Fer2013. We can see that by increasing the training batches, accuracy on both the training set and validation set steadily increased. During the earlier steps of training, the accuracy increases greatly. When training at the 30th to 50th round, due to the different data distributions on the training and validation sets, there is a fluctuation in the accuracy. However, after 100 rounds, the recognition accuracy of the model becomes very balanced. This can be explained that by deleting of some incorrect samples and non-face areas in the training set, the model shows a higher accuracy on it.

Figure 10 demonstrates the accuracy of the model of the CK+ dataset. We see that during the earlier steps of model training, the accuracy also maintains a relatively high growth rate. Between the 25th and 75th rounds, the accuracy is still continuously increasing. Since the data distributions of the sets are consistent, growth trends in both datasets remain consistent. After the 100th round, the model recognition rate becomes very stable. After introducing the pyramid segmentation attention and the joint loss function, the accuracy on both Fer2013 and CK++ steadily increased. There was no overfitting or under fitting phenomena. This verifies the effectiveness of the methods in this paper.

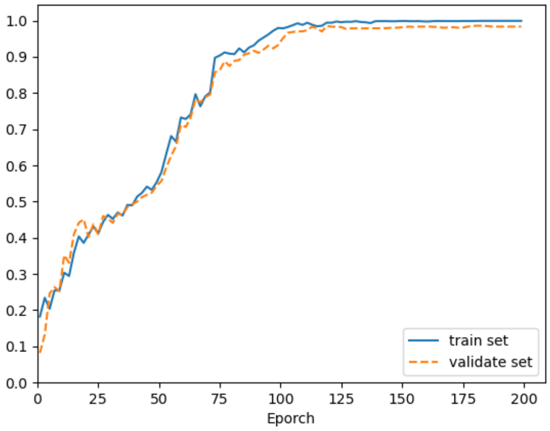

Figure 9. Accuracy on Fer2013

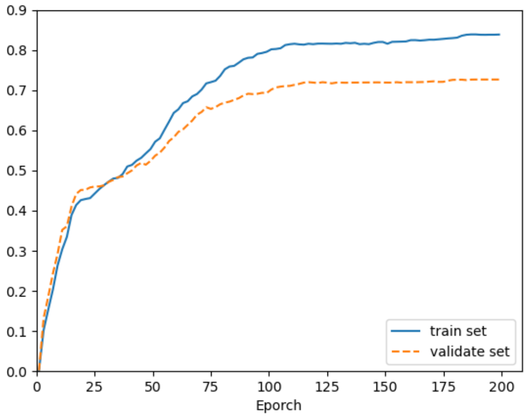

Figure 10. Accuracy on CK++

To prove the superiority of pyramid segmentation attention mechanism, we mark the ResNet50 model without the attention mechanism as Basic. We sequentially embedded the SE, CBAM, ECA, and AFF attention modules, and use the CK+ dataset for comparative experiments. From Table 1, we can see that the recognition accuracy from the network model that is integrated with the pyramid segmentation attention module is the highest.

Table 3. Comparison of accuracy from different attention mechanisms

Attention	Accuracy %
Basic	94.80
SE	96.80
CBAM	97.62
ECA	97.82
AFF	98.07
PSA	98.35

To further analyze the accuracy of various types of expression recognition, this paper uses the disorder matrix to visualize predicted performance of the model on different categories of expressions. From the disorder matrix on the Fer2013 validation in Figure 11, the method has a better recognition rate for happiness and surprise, reaching 0.84 and 0.82 respectively. At the same time, the accuracy of recognizing anger, sadness, and fear is "not good". The main reason is because the two expressions of happiness and surprise have distinct facial features, while the expressions of sadness, anger, and fear have the actions of frowning and opening the mouth, which is not easy to set apart, resulting in a lower recognition degree for these three types of expressions.

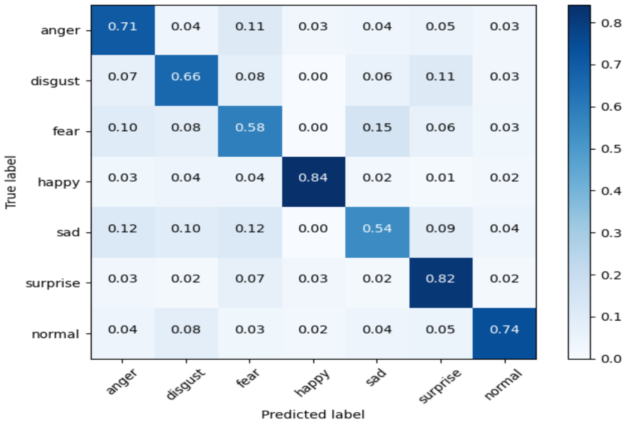

Figure 11. Fer2013 validation set disorder matrix

Figure 12 is the disorder matrix of the recognition rate of the seven categories of expression on the CK+ validation set. It is not difficult to see that compared with Fer2013, the accuracy of each type of expression recognition has been greatly improved. The reason is that the CK+ is obtained by participants posing designated expressions under laboratory conditions, and the data collection is rigorous and reliable, which greatly reduces the interference of human factors and the environment, thus making the detection performance of the model on the CK+ dataset more accurate.

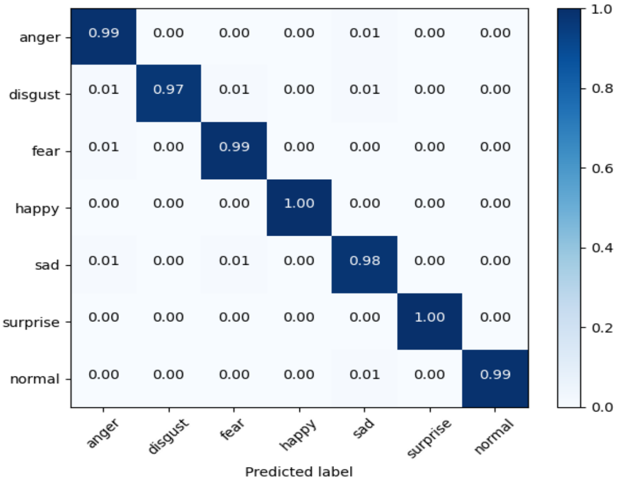

Figure 12. CK+ validation set confusion matrix

3.3.2 Ablation Experiments

To evaluate the importance of each module, this paper proves the effectiveness of the pyramid segmentation attention module and joint loss function module in Table 2. Among them, the ResNet50 basic network without adding any module is called Basic, PSA represents the network model that integrates the pyramid segmentation attention, and each module is attached to the basic network in turn for comparative experiments. The results show that while each module can increase the correctness of model recognition, using both modules into the model can obtain the best recognition result. In contrast with the basic model, our method increased accuracy by 3.5% on the Fer2013 dataset and 4.3% on the CK+ dataset.

From Table 4, it can be seen that after adding the joint loss function, the accuracy on the CK+and Fer datasets increased by 0.63% and 1.45%, respectively. The analysis found that only a small number of similar expressions in the CK+dataset had significant differences, so the improvement after adding the Center Loss loss function was relatively small. After adding the PSA module to the basic model, the accuracy improvement reached 1.52% and 2.53%, respectively. This is because the module effectively extracts multi-scale features, enhances the correlation of cross channel information, generates better pixel level attention, and improves the model's ability to discriminate expressions. The results show that integrating both modules into the base model can achieve the best recognition results. Compared with the base model, the method used in this paper improved by 4.3% on the CK+dataset and 3.5% on the Fer2013 dataset.

Table 4. Ablation experiments

Method	Accuracy %	
	CK+	Fer2013
Basic	94.80	69.92
Basic + joint loss	95.43	71.67
Basic + PSA	96.32	72.45
Basic+ PSA + joint loss	98.35	74.26

3.3.3 Comparison with Other Algorithms

To verify the advancement of the PSA-ResNet algorithm, comparative experiments were carried out with network models such as VGGNet [16], ResNet [18], InceptionV4 [19], MobileNetV2 [20], and DenseNet [21] on selected dataset. The results of each algorithm are demonstrated in Table 3. The accuracy of the proposed algorithm in this paper reached 74.26% and 98.35% in the Fer2013 and CK+ datasets, respectively, which are higher in contrast with other algorithms, indicating that the network model based on pyramid segmentation attention and joint loss can efficaciously abstract key features and greatly enhance the effect on expression recognition.

As shown in Table 5, VGGNet replaces large-sized convolution kernels with a series of small-sized ones, which facilitates the extraction of subtle facial expression features and achieves human recognition level results. However, the model has too many parameters and requires a lot of training time. ResNet uses two 1x1 convolutions to increase and decrease the dimensionality of channels, which not only maintains the accuracy of the model but also reduces network parameters, resulting in an accuracy improvement of 1.07% and 1.14%, respectively. MobileNetv2 introduces Inverted Residual structures to reduce computational complexity and memory consumption. However, the Inverted Residual structures expand and compress channels, which to some extent limit direct interaction between features and information fusion.DenseNet significantly reduces the number of parameters through dense connections and feature reuse, with a reduction of nearly 50% compared to ResNet at similar levels. However, feature reuse requires a large amount of memory, making the training cycle too long. In conclusion, the model proposed in this paper can achieve a high recognition rate while maintaining a small number of parameters, which verifies the progressiveness of the model.

Table 5. Comparison on precision of different algorithms

Method	Accuracy %	
	CK+	Fer2013
VGGNet	97.26	68.79
ResNet	97.80	69.01
MobileNetV2	98.16	70.80
DenseNet	98.05	72.30
Ours	98.35	74.26

4 Conclusion

This paper brings up an expression recognition network on pyramid segmentation attention and joint loss, using SPC to achieve poly scale feature abstraction and SE to enhance information correlation between cross-channels, which improves the accuracy of expression edges and long-distance predictions. To expand the distance of different types of expressions and reduce the distance of the same type of expressions, SoftMax Loss and Island Loss were used to optimize the network model to additionally improve the recognition result. The model has a simple structure and a stable training process, and there is no underfitting or overfitting phenomenon. As seen in the experiments, in contrast with the cutting-edge algorithms, this algorithm achieved better precision. However, the precision of the model for the recognition of some types of expressions is still not ideal, which is the direction that needs to be optimized in the future work.

Acknowledgements

China University Research Innovation Fund: Research on the Construction of Ubiquitous Service System Based on Data Center (2021KSA02004).

References

[1] V. Diaz, W. E. Wong, Z. Chen, Enhancing Deception Detection with Exclusive Visual Features using Deep Learning, *International Journal of Performability Engineering*, Vol. 19, No. 8, pp. 547-558, August, 2023. https://doi.org/10.23940/ijpe.23.08.p7.547558

[2] N. Shelke, D. Sale, S. Shinde, A. Kathole, R. Somkunwar, A Comprehensive Framework for Facial Emotion Detection using Deep Learning, International *Journal of Performability Engineering*, Vol. 20, No. 8, pp. 487-497, August, 2024. https://doi.org/10.23940/ijpe.24.08.p3.487497

[3] M. Pantic, L. J. M. Rothkrantz, Expert System for Automatic Analysis of Facial Expressions, *Image and Vision Computing*, Vol. 18, No. 11, pp. 881-905, August, 2000. https://doi.org/10.1016/S0262-8856(00)00034-2

[4] S. Li, W. Deng, Deep facial expression recognition: A Survey, *IEEE Transactions on Affective Computing*, Vol. 13, No. 3, pp. 1195-1215, July-September, 2022. https://doi.org/10.1109/TAFFC.2020.2981446

[5] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, 2015, pp. 1-14.

[6] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 8, pp. 2011-2023, August, 2020. https://doi.org/10.1109/TPAMI.2019.2913372.

[7] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial Transformer Networks, *NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, Montreal Canada, 2015, pp. 2017-2025, February, 2020.

[8] S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: Convolutional Block Attention Module, *Proceeding of the European Conference on Computer Vision*, Munich, Germany, 2018, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[9] K. Wang, X. Y. Peng, J. F. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, *IEEE Transactions on Image Processing*, Vol. 29, pp. 4057-4069, January, 2020. https://doi.org/10.1109/TIP.2019.2956143

[10] W. Wei, Q. X. Jia, Y. L. Feng, G. Chen, M. Chu, Multi-modal facial expression feature based on deep-neural networks, *journal on multimodal user interfaces*, Vol. 14, No. 1, pp. 17-23, March, 2020. https://doi.org/10.1007/s12193-019-00308-9

[11] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, ShuffleNet V2: practical guidelines for efficient CNN architecture design, *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018, pp. 122-138. https://doi.org/10.1007/978-3-030-01264-9_8

[12] A. Swaminathan, A. Vadivel, M. Arock, FERCE: Facial Expression Recognition for Combined Emotions Using FERCE Algorithm, *IETE Journal of Research*, Vol. 68, No.

5, pp. 3235-3250, 2020.
https://doi.org/10.1080/03772063.2020.1756471

[13] P. Zhang, W. Kong, J. B. Teng, Facial expression recognition based on multi-scale feature attention mechanism, *Computer Engineering and Applications*, Vol. 58, No. 1, pp. 182-189, January, 2022.
https://doi.org/10.3778/j.issn.1002-8331.2106-0174

[14] Y. L. Ji, Y. H. Hu, Y. Yang, H. T. Shen, Region attention enhanced unsupervised cross-domain facial emotion recognition, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 4, pp. 4190-4201, April, 2023.
https://doi.org/10.1109/TKDE.2021.3136606

[15] M. J. Yu, H. C. Zheng, Z. F. Peng, J. Dong, H. Du, Facial expression recognition based on a multi-task global-local network, *Pattern Recognition Letters*, Vol. 131, pp. 166-171, March, 2020.
https://doi.org/10.1016/j.patrec.2020.01.016

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1-9.
https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298594

[17] H. Zhang, K. K. Zu, J. Lu, Y. Zou, D. Y. Meng, EPSANet: An Efficient Pyramid Squeeze Attention Block on convolutional neural network, *Computer Vision – ACCV 2022: 16th Asian Conference on Computer Vision*, Macao, China, 2022, pp. 541-557.
https://doi.org/10.1007/978-3-031-26313-2_33

[18] S. Phawinee, J. F. Cai, Z. Y. Guo, H.-Z. Zheng, G.-C. Chen, Face recognition in an intelligent door lock with ResNet model based on deep learning, *Journal of Intelligent & Fuzzy Systems*, Vol. 40, No. 4, pp. 8021-8031, April, 2021.
https://doi.org/10.3233/JIFS-189624

[19] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, V. K. Asari, Inception recurrent convolutional neural network for object recognition, *Machine Vision and Applications*, Vol. 32, No. 1, Article No. 28, January, 2021.
https://doi.org/10.1007/s00138-020-01157-3

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, *CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, pp. 4510-4520.
https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00474

[21] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261-2269.
https://doi.org/10.1109/CVPR.2017.243

Biographies

Jiale Gu, male, Jan 1981 in Wuxi, Jiangsu Province, master degree, associate professor professional in big data and image recognition. The author has been engaged in research in the field of big data and image recognition for a long time, and has published more than ten related papers and two monographs. Hosted 5 related projects at or above the city level, including 1 project funded by the Ministry of Education’s Industry University Research Fund. Approved as an outstanding young backbone teacher in Jiangsu Province’s “Qing Lan Project” and a subject leader in Suzhou Industrial Park.

Xiaohong Jin, Title: Professor. Research Interests: English Writing Instruction, Bilingual Lexicography, Chinese Taoist Culture.