

## RESEARCH ARTICLE

# A Novel Attention Residual Network Expression Recognition Method

HUI QI<sup>1,2</sup>, XIPENG ZHANG<sup>1</sup>, YING SHI<sup>1,3</sup>, AND XIAOBO QI<sup>1</sup><sup>1</sup>School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China<sup>2</sup>Shanxi Key Laboratory of Intelligent Optimization Computing and Blockchain Technology, Jinzhong 030619, China<sup>3</sup>School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

Corresponding author: Hui Qi (qihui@tynu.edu.cn)

This work was supported in part by the Shanxi Patent Transformation Special Programs under Grant 202302009 and Grant 202302012, in part by the Basic Research Program (Free Exploration) of Shanxi Province under Grant 20210302123334, and in part by the Taiyuan Normal University Achievement Transformation and Technology Transfer Base under Grant 2023P003.

**ABSTRACT** Expressions serve as intuitive reflections of a person's psychological state, making the extraction of effective features for accurate facial expression recognition a crucial research problem. However, when facial information is incomplete, the existing convolutional neural networks face some challenges in extracting features. To address this issue, this paper introduces a pyramidal convolutional attention residual network(PCARNet) based on the ResNet18. PCARNet combines the pyramidal convolution module and an improved convolutional attention mechanism to effectively extract expression features and achieve high-precision facial expression recognition. The proposed model utilizes pyramidal convolution to extract facial expression features at multiple scales, capturing both global and local information of the face. Grouped convolution is employed to reduce the computational complexity and the number of parameters. Additionally, to avoid the adverse effects of channel dimensionality reduction on the attention mechanism and enhance the capacity for information exchange across channels, the Share MLP module within the convolutional attention mechanism was replaced by a one-dimensional convolution with adaptive kernel size. The improved convolutional attention mechanism assigns weights to the extracted multiscale features based on both channel and spatial dimensions, enhancing the representation of crucial facial features. Experimental results demonstrate the high recognition accuracy of the proposed method on public datasets such as Fer2013, RAF-DB, and CK+. The accuracies achieved are 73.725%, 87.516%, and 95.455%, respectively. Compared to other methods, the proposed approach shows improvements of at least 1.4%, 2.4%, and 0.25% on the respective datasets, confirming its high reliability and performance.


**INDEX TERMS** Residual network, pyramidal convolution, attention mechanism, facial expression recognition.

## I. INTRODUCTION

People commonly utilize facial expressions to convey emotions and information in everyday communication. With the continuous advancement of deep learning, computer vision has progressively achieved remarkable performance in the field of image analysis. The research on expression recognition has been deepening, and its applications have expanded to various domains such as intelligent transportation, medical services, intelligent education, and criminal

deception detection, demonstrating significant potential for further development.

The process of face expression recognition can be divided into several steps, including image acquisition, face detection, feature extraction, and expression matching and recognition. Among these steps, feature extraction plays a crucial role in the effectiveness of facial expression recognition. This is due to the subtle differences between expressions, which cannot be discerned without extracting the key features of expressions to distinguish expressions accurately. Currently, feature extraction methods are primarily categorized into two types: traditional methods and deep learning methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil .

Traditional expression feature extraction techniques primarily rely on hand-designed approaches [1], [2]. These methods encompass Principal Component Analysis (PCA) [3], Local Binary Patterns (LBP) [4], and Histogram of Oriented Gradient (HOG) [5]. However, traditional methods are difficult to handle face expression recognition tasks involving complex backgrounds and diverse angles, exhibiting limited generalization ability. Therefore, an automatic feature extraction method is necessary to enhance the model's generalization ability.

Deep learning methods have gained widespread adoption in various computer vision domains due to their exceptional feature extraction capabilities. Convolutional neural networks (CNNs) have surpassed manual methods in feature extraction, offering autonomous learning and improved accuracy and stability in recognizing facial expressions. Most applications of deep learning in expression recognition are based on network models such as VGGNet [6], ResNet [7], and MobileNet [8], which are primarily structured as Deep Convolutional Neural Networks (DCNNs). Wang et al. [9] proposed an enhanced MobileNet network model, incorporating depth separable convolution to reduce network computational complexity, and employing Support Vector Machines (SVM) for facial expression classification, resulting in a lighter model. Lan et al. [10] introduced a joint regularization strategy based on the ResNet18 residual network, mitigating model overfitting and enhancing accuracy by integrating multiple regularization methods into the network. This approach improves both the model's generalization ability and overall performance. Zhou et al. [11] designed a lightweight CNN by employing a multi-task cascade convolutional network for face detection and combining residual and depth-separable convolutional modules to reduce network parameters, thus increasing model portability. Furthermore, they screened shallow features of facial expressions using techniques like global average pooling and fused shallow and deep features for expression recognition. Shi et al. [12] proposed a multi-branch cross-connected CNN-based facial expression recognition method. Their approach combines residual connections, network of networks, and tree structure to improve feature extraction. Additionally, they introduced fast cross-connections to the convolutional output layer, enhancing data flow between networks and feature extraction abilities in each sensory domain.

Attention mechanisms represent a category of optimization modules that have yielded substantial advancements in recent years within deep neural network research. These mechanisms label pivotal features in an image by readjusting their weights, aiming to mitigate background interference and discern the information-rich regions within the image. Wang et al. [13] introduced a regional attention network that integrates both a self-attention module and a relational attention module. This framework aims to comprehensively learn attention weights for individual regions in an end-to-end

approach, consequently enhancing the accuracy of facial expression recognition. Li et al. [14] fused LBP features with an attention model in facial expression recognition by integrating the attention mechanism within the network architecture. This integration facilitated the network in assigning diverse attention weights to distinct segments of the input, thereby directing the network's focus towards more relevant features. Guo et al. [15] proposed an enhanced method for facial expression recognition incorporating an attention mechanism and Involution operator. This method utilizes VGG19 as the foundational network and introduces the attention mechanism at the front-end to extract highly relevant expression features. Additionally, it enhances the distribution of feature data through a joint regularization strategy, resulting in improved accuracy in expression recognition.

The aforementioned methods have somewhat enhanced the accuracy of facial expression recognition. Literatures [9], [10], [11], and [12] primarily focus on refining the base model and enhancing its feature extraction capability. In real-world scenarios, images comprise not only facial details but also extraneous background information. Thus, literatures [13], [14], and [15] integrate the attention mechanism into the network architecture, leveraging it to mitigate irrelevant features. This approach significantly improves the model's generalization ability. However, in the pursuit of further enhancing expression recognition accuracy for practical application, several challenges remain that need addressing as follows:

(1) In datasets with incomplete facial information, such as faces with partial occlusion or deflection, the existing model lack the ability to extract facial expression features, resulting in low recognition accuracy.

(2) Most existing attention mechanisms employ channel dimensionality reduction operations to reduce model complexity in the information extraction process. However, this approach disrupts the direct correspondence between channels and their weights, impacting the information extraction capability of the attention mechanism.

To address these issues, this paper proposes a solution that involves the following: Extracting multi-scale features from images using pyramid convolution, allowing for the capture of global and local facial information. Avoiding the adverse effects of channel dimensionality reduction on the attention mechanism while enhancing information fusion between channels using an improved convolutional attention mechanism. Enhancing key features within the multi-scale features extracted from pyramid convolution. As a result, a Pyramid Convolutional Attention Residual Network (PCARNet) is designed. The contributions of this paper are as follows:

(1) Replacing the initial  $3 \times 3$  convolution in ResNet18 with a pyramid convolution structure to extract multi-scale feature information. This modification enables the capture of global and local facial information

by utilizing multiple convolution kernels of varying scales.

(2) Introducing the Efficient Convolutional Block Attention Module (ECBAM) attention mechanism module to enhance information fusion among multi-scale channels. The ECBAM attention effectively highlights key feature regions for expression recognition in both the channel and spatial dimensions, while avoiding the negative impact of channel dimensionality reduction.

(3) Conducting experiments on Fer2013 [16], RAF-DB [17], and CK+ [18] datasets, resulting in accuracies of 73.725%, 87.516%, and 95.455%, respectively. These experiments demonstrate the model's ability to improve expression recognition accuracy and highlight its significant advantage.

## II. RELATED WORK

### A. PYRAMID CONVOLUTION

Pyramid Convolution(PyConv) [19] employs convolution kernels of varying sizes, as illustrated in Fig. 1. Inspired by the concept of grouped convolution in Fig. 2, PyConv achieves the utilization of different-sized convolution kernels at each level by partitioning the input data into distinct groups and assigning different-sized convolution kernels to each group. This approach enables PyConv to process the input data using multiple convolution kernels of diverse scales. Specifically, it utilizes small convolution kernels, such as  $1 \times 1$  and  $3 \times 3$ , to capture detailed image information, while employing larger convolution kernels, such as  $5 \times 5$  and  $7 \times 7$ , to learn semantic information across larger receptive fields. By capturing information at different scales and in various environments, PyConv enhances the model's performance compared to standard convolution. Importantly, PyConv achieves this improvement without introducing additional parameters. Qiao and Zhang [20] proposed a Pyramid Hourglass Network based on Pyramid Convolution. The use of multiple convolutional kernels of varying sizes increases the model's ability to extract features and represent them and improves the accuracy of prediction. It can be seen that pyramid convolution does improve the prediction accuracy to a certain extent, but it can be seen in Figure 1 that the different sizes of convolution operations in pyramid convolution are computed within their respective groups, which leads to the data information existing only inside this group. There is no interaction between the channels, and information has been blocked. So we need a method to allow communication between different channels of information to avoid information blocking.

### B. CBAM

CBAM [21], short for Convolutional Block Attention Module, is an attention mechanism that enhances the optimization of adaptive features more efficiently. It operates by weighting the input features using information from both the spatial and channel dimensions of a given intermediate feature map. The

weighted features are then multiplied with the input features to achieve adaptive optimization. Importantly, CBAM can be trained end-to-end alongside the underlying convolutional neural network model without significantly increasing computational complexity. Compared to SENet [22], which solely focuses on channel attention mechanism, CBAM combines both spatial and channel attention mechanisms. As a result, CBAM outperforms SENet in terms of performance. The structure of the CBAM module is illustrated in Fig. 3. However, the channel attention in CBAM employs an MLP layer for information extraction and incorporates channel dimensionality reduction to reduce computational complexity. Nevertheless, the channel dimensionality reduction operation in the channel attention module projects the channel features into a low-dimensional space and then maps them back to the feature map, resulting in an indirect correspondence between the channel and its weights. This indirect correspondence negatively impacts the attention mechanism and consequently affects recognition accuracy. Thus, it is essential to avoid the channel dimensionality reduction operation in the attention mechanism to achieve more efficient information extraction.

## III. APPROACH

Existing models that struggle to extract facial expression features adequately when facing with incomplete facial information, leading to diminished recognition accuracy. Thus, this paper introduces a novel algorithm, PCARNet, designed to extract both global and local facial information, aiming to address the challenge of low recognition accuracy encountered by models in scenarios involving incomplete facial information. The proposed algorithm consists of three main steps. Firstly, the facial expression dataset is subjected to preprocessing to enhance the data quality. Secondly, facial expression features are extracted by utilizing the Pyramid Convolutional Attention Residual Network (PCARNet). PCARNet comprises of four PCAR Blocks, which each block utilizes PyConv to extract multi-scale features from the image, enabling the capture of both local and global facial information. Meanwhile, ECBAM is used to facilitate channel-wise fusion of information and emphasize crucial facial features. Finally, classify with a fully connected layer. The workflow of the proposed approach is shown in Fig. 4.

### A. DATASET PREPROCESSING

In this thesis, we select the Fer2013, CK+, and RAF-DB expression datasets. The Fer2013 dataset is initially provided in CSV format, so we utilize pandas to parse and extract the dataset into images. To enhance the quantity and quality of the samples and improve the model's generalization ability, we perform data augmentation on three datasets through random cropping and level-flipping techniques. In Fig 4, the data processing is illustrated using the fer2013 dataset as an example.

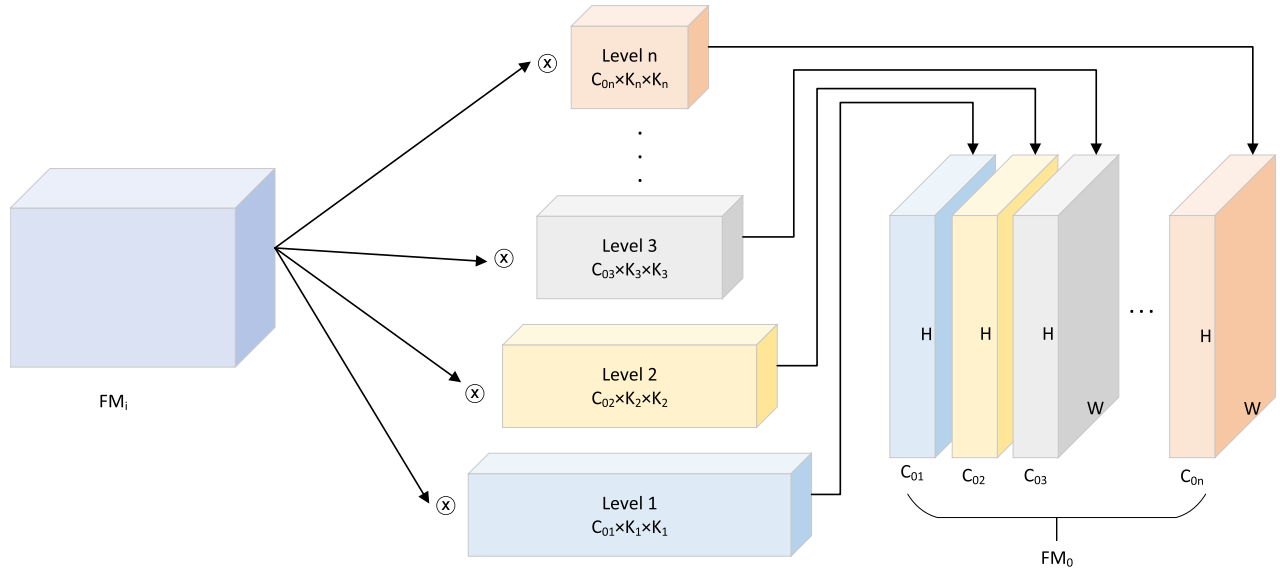


FIGURE 1. Pyramid convolution.

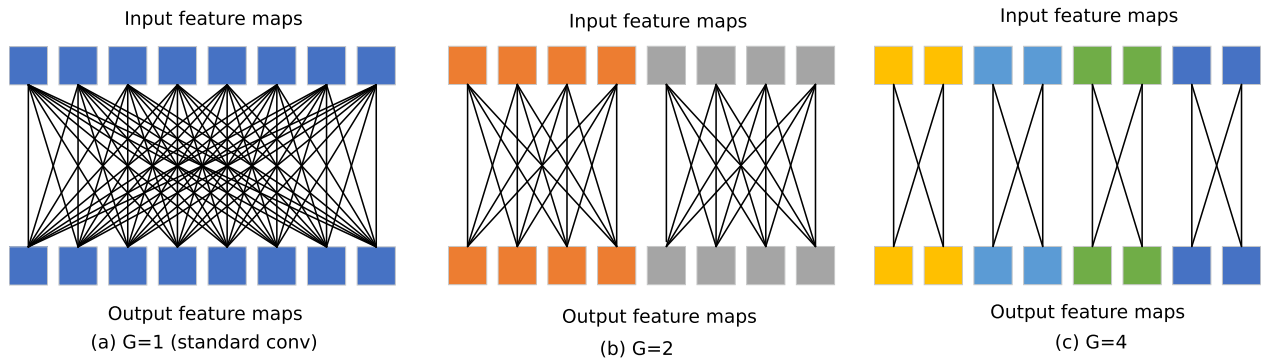


FIGURE 2. Group convolution.

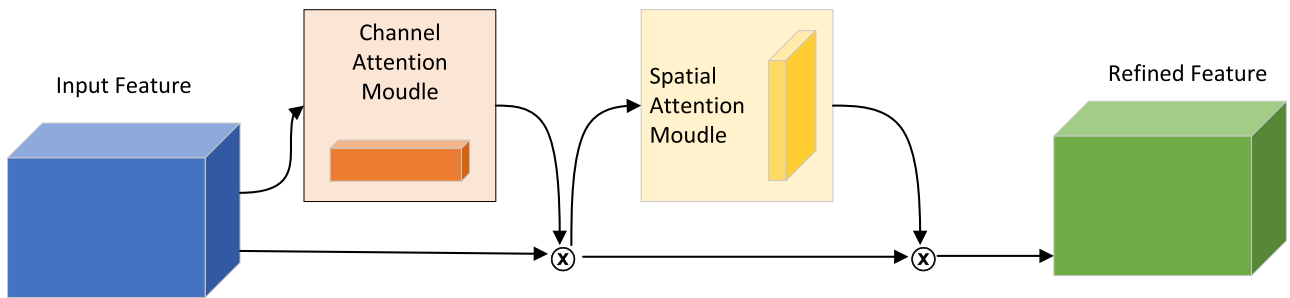


FIGURE 3. CBAM module.

**B. PCARNET**

The network architecture comprises a  $3 \times 3$  convolutional layer, followed by PCAR Block0, PCAR Block1, PCAR Block2, PCAR Block3, a Global Average Pooling (GAP) layer, and a fully connected layer. The overall structure of the PCARNet network is illustrated in Fig. 5, where “C” represents the number of channels.

The model starts by performing feature extraction from the input image using a  $3 \times 3$  convolutional layer. Following the convolutional layer, there are four interconnected blocks at different depths, each containing two improved

ResNet18 residual blocks. In these modified blocks, the first  $3 \times 3$  convolution is replaced with PyConv. The pyramid convolution utilizes distinct parameters to extract multi-scale features of facial expressions, enhancing the network’s feature extraction capability. Subsequently, an improved attentional mechanism referred to as ECBAM is employed to allocate attentional weights to the extracted multiscale features across both channel and spatial dimensions. This attention mechanism highlights key facial expression features, resulting in improved recognition accuracy. Finally, the second  $3 \times 3$  convolution in the ResNet18 residual block

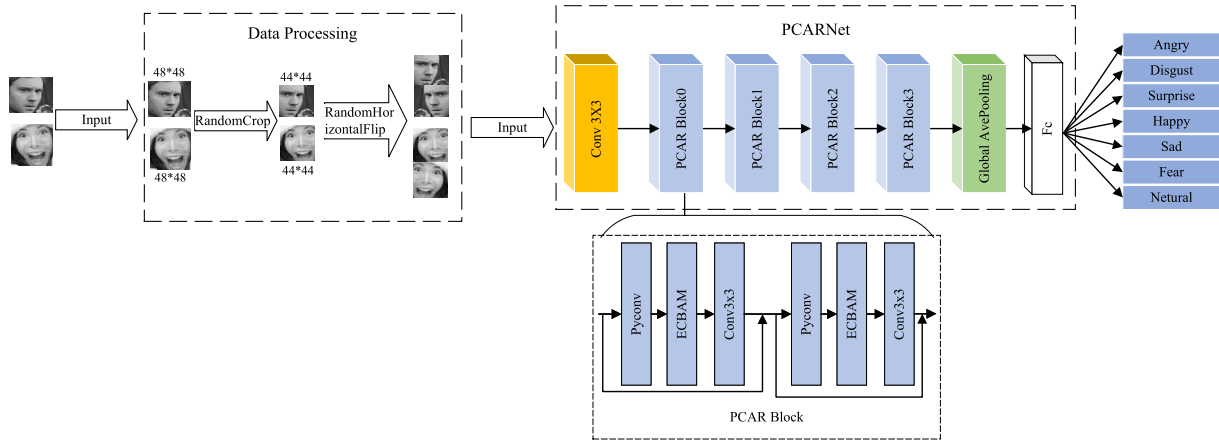


FIGURE 4. Workflow of the proposed method.

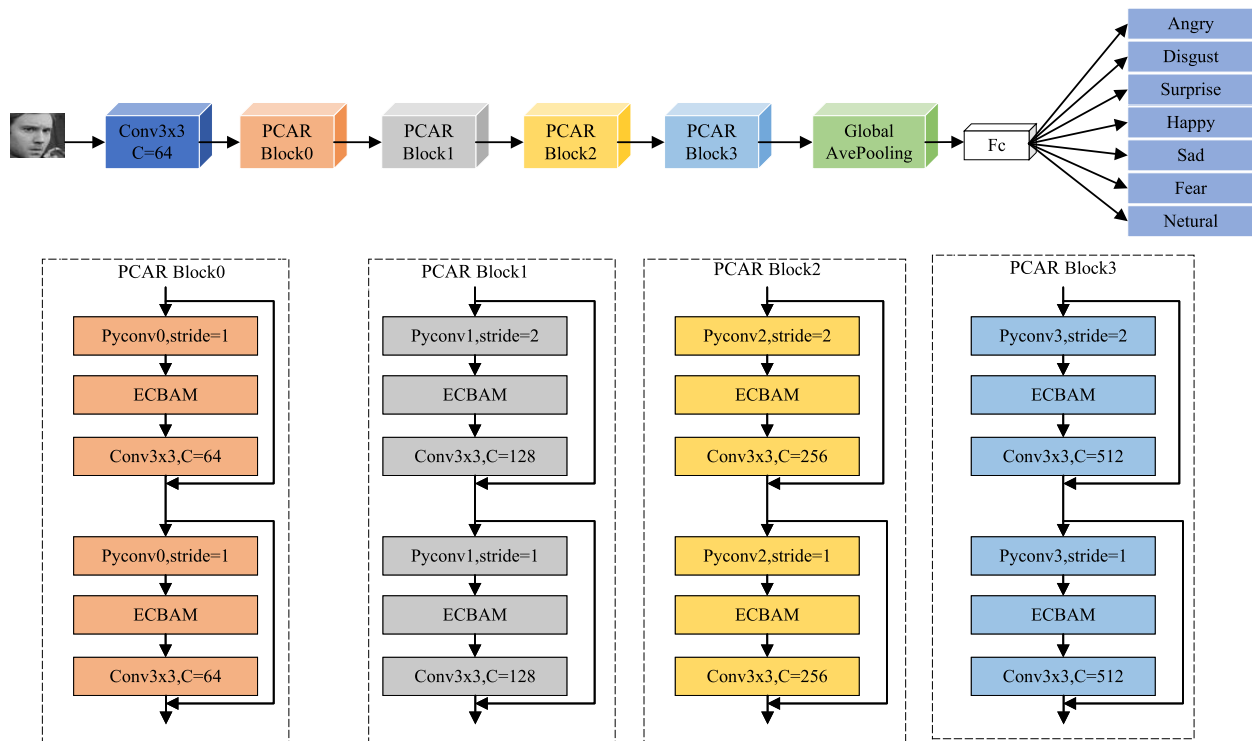


FIGURE 5. Structure of the proposed method.

is used to integrate the extracted information. The model incorporates a GAP layer after the PCAR Block3 to alleviate overfitting. A fully connected layer is used for classification.

### 1) PYCONV

To address the challenges posed by datasets containing incomplete facial information, such as occlusion or deflection, this paper proposes the use of pyramid convolution for extracting multi-scale features that capture both global and local information about the face. Particularly in scenarios with incomplete facial information, global information plays a crucial role in facial recognition. Pyramid Convolution (PyConv) employs multiple convolution kernels of varying scales to process the input data. To enable the utilization of distinct kernel at varying levels within PyConv, the input

feature maps undergo segmentation into separate groups. Subsequently, independent kernel applications are performed for each group of input feature maps. It utilizes smaller kernels to capture fine-grained details of the image and larger kernels to capture global information across a broader receptive field. Due to the increased computational effort associated with larger convolution kernels, PyConv utilizes a grouped convolution operation. PyConv increasing the number of groups as the size of the convolution kernel increases. By doing so, PyConv is able to extract multi-scale information without significantly increasing the computational workload. This approach enhances the model's performance. As the model's depth increases, the size of the feature map decreases. To account for this, the paper uses four PyConv layers with distinct parameters based on the

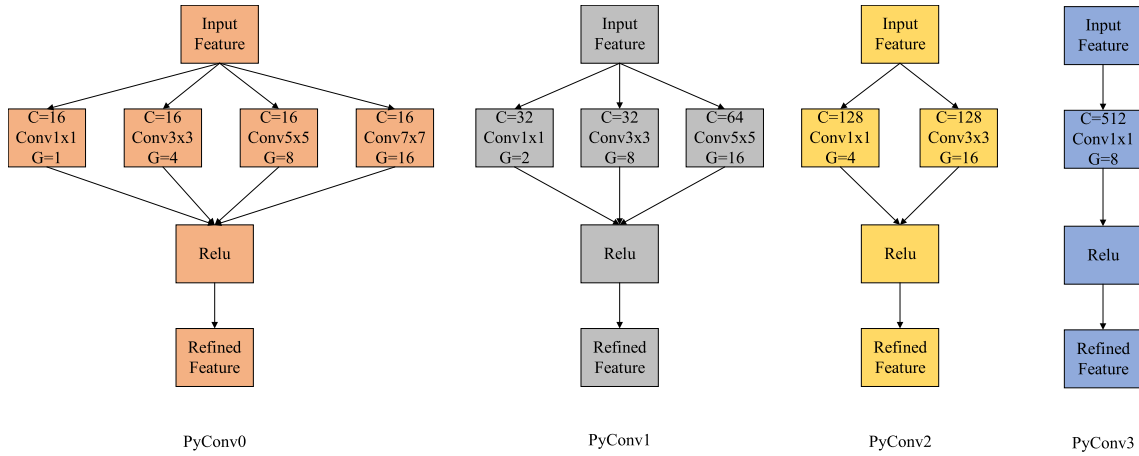


FIGURE 6. PyConv layers.

feature map size. The Relu activation function is applied after each PyConv layer to enhance the neural network’s nonlinear fitting capability and improve the model’s expressiveness. The architecture of the four PyConv layers is depicted in Fig. 6, where “C” represents the number of channels and “G” denotes the number of groups used for grouping the convolutions.

2) ECBAM

When generating channel attention, the CBAM module first downsizes the feature map and then extracts channel features by mapping each feature using a fully connected layer. To capture larger spatial feature information, a 7×7 convolution kernel is used in the spatial attention to extract spatial features from the samples. However, the PyConv module utilizes grouped convolution, which can result in limited information exchange between channels due to data being confined within each group, resulting in information blocking. To enhance channel interaction across different scales, this study draws inspiration from the literature [23] and replaces the fully connected layer with 1D-convolution operations using adaptive kernel sizes. This approach strengthens the exchange of feature information in the channel dimension by aggregating information from neighboring channels. Moreover, it avoids the negative impact of channel dimensionality reduction on the attention mechanism. To account for the low image resolution of the dataset utilized in this paper, two 3×3 convolutions are employed in spatial attention as a replacement for the original 7×7 convolution. This modification allows for the extraction of spatial features effectively. The improved module proposed in this study is called ECBAM.

The improved channel attention module first aggregates the multi-scale feature information through global pooling and maximum pooling operations. This generates global pooled features and maximum pooled features. Next, the two features undergo 1D-convolution with an adaptive convolution kernel size. The resulting features are then added together and passed through the Sigmoid function and

the resulting output is expanded to match the size of the input feature map. Finally, the expanded result is multiplied element-wise with the input feature map to obtain the feature map after the addition of channel attention. The improved channel attention mechanism is depicted in Fig. 7.

Specifically, the process of calculating channel attention can be represented as follows:

$$M_C(F) = \delta \left( f_{1D}^k (Avgpool(F)) + f_{1D}^k (Maxpool(F)) \right) = \delta \left( f_{1D}^k (F_{Avg}^C) + f_{1D}^k (F_{Max}^C) \right) \quad (1)$$

where  $\delta$  represents the Sigmoid activation function and  $f_{1D}^k$  represents the adaptive one-dimensional convolution with a convolution kernel size. The size of the adaptive convolution kernel  $k$  is determined by the equation mentioned in the literature [23]:

$$k = \left\lfloor \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right\rfloor \quad (2)$$

C represents the number of channels in the input feature map. The coefficients  $\gamma$  and  $b$  are adjustable parameters for customization. In this paper, we use the parameters that gave the best results in the literature [23],  $\gamma$  is assigned a value of 2, and  $b$  is assigned a value of 1. The notation  $\lfloor t \rfloor_{odd}$  denotes the nearest odd number to  $t$ .

The improved spatial attention module conducts global pooling and maximum pooling operations on the feature map after channel attention. This process generates global pooled features and maximum pooled features. The two spatial features are then concatenated along the channel axis. Next, two 3×3 convolutions are applied to assign weights to different regions of the spatial features, and the resulting features from these convolutions undergo Sigmoid activation function operations. The generated output is subsequently expanded to match the size of the input feature map. Finally, the expanded output is element-wise multiplied with the input feature map to obtain the feature map with added spatial attention. The improved spatial attention module is illustrated in Fig. 8.

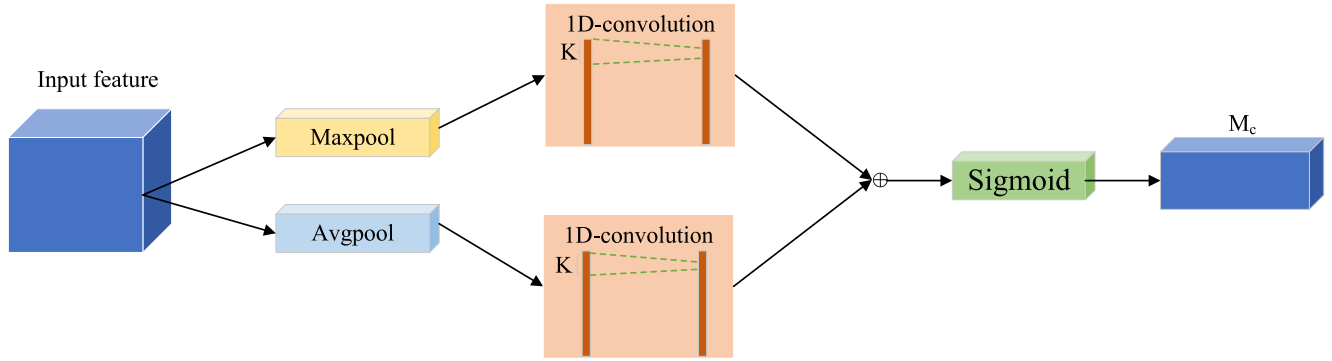


FIGURE 7. Improved channel attention.

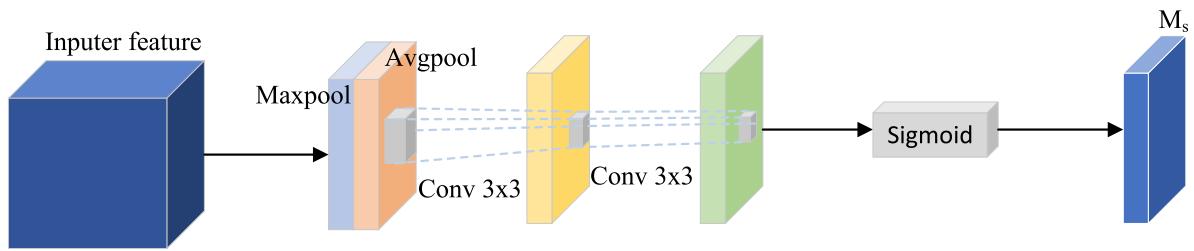


FIGURE 8. Improved spatial attention.

The spatial attention calculation process can be represented as follows:

$$\begin{aligned}
 M_S(F) &= \delta \left( f_{con}^3 \left( f_{con}^3 [Avgpool(F); Maxpool(F)] \right) \right) \\
 &= \delta \left( f_{con}^3 \left( f_{con}^3 [F_{Avg}^S; F_{Max}^S] \right) \right) \quad (3)
 \end{aligned}$$

where  $\delta$  represents the Sigmoid activation function and  $f_{con}^3$  represents a convolution with a kernel size of 3.

#### IV. ANALYSIS OF EXPERIMENTAL RESULTS

##### A. DATA SET

In this paper, we conducted experiments using three publicly available face expression datasets: Fer2013, RAF-DB dataset, and CK+ dataset.

The Fer2013 dataset, obtained from a Kaggle competition project, consists of 7 expression categories. It includes a training set with 28,709 images and two test sets, each containing 3,589 images. As shown in Fig. 9.

The RAF-DB dataset comprises approximately 30,000 images collected from the internet. These images have been manually annotated and reliably estimated to provide samples with 7 basic expression labels and 11 composite expression labels. For this experiment, we utilized the 7 basic facial expressions from the RAF-DB dataset. The training set consists of 12,271 images, and the test set contains 3,068 images. As shown in Fig. 10.

The CK+ dataset is an extension of the CK dataset and consists of 593 videos of 123 sampled individuals, encoded using peak frames. From the CK+ dataset, we selected 327 emotion sequences. As shown in Fig. 11.



FIGURE 9. Example of Fer2013 dataset.

##### B. EXPERIMENTAL SETUP

The experimental setup for this study is as follows: a Windows 10 system with an Intel Core i5-12400F CPU, 16GB of RAM, and an NVIDIA GeForce RTX3060 (12GB) GPU. The deep learning framework used for training is Pytorch. The weight updates are performed using stochastic gradient descent with a momentum of 0.9. The loss function is the cross-entropy loss function. All three datasets were trained using a batch size of 32

For the Fer2013 dataset, the random cropping size is set to  $44 \times 44$ . The initial learning rate is 0.008, and it decreases by a factor of 0.9 every 5 iterations after 50 iterations. The total number of iterations is set to 250. In this experiment,



FIGURE 10. Example of RAF-DB dataset.



FIGURE 11. Example of CK+ dataset.

the public test set of the Fer2013 dataset is used for model selection, while the private test set is used to evaluate the selected model.

For the RAF-DB dataset, the random cropping size is set to  $92 \times 92$ . The initial learning rate is 0.01, and it decreases by a factor of 0.9 every 5 iterations after 40 iterations. The total number of iterations is set to 200.

As for the CK+ dataset, the random cropping size is  $44 \times 44$ . The initial learning rate is 0.008, and it decreases by a factor of 0.9 every 5 iterations after 20 iterations. The total number of iterations is set to 100. Due to the limited number of samples in the CK+ dataset, a 10-fold cross-validation approach is used during testing to calculate the final recognition rate of the model.

Furthermore, during the testing phase, an ensemble approach is employed to reduce the impact of interfering factors. Specifically, 10 images are collected for each test sample. The top left corner, bottom left corner, top right corner, bottom right corner, and center of each image are cropped to obtain 5 images. These cropped images are then horizontally flipped to obtain 5 additional images. The final decision is made by averaging the results of these 10 processed images, thereby reducing the classification error.

C. EXPERIMENTAL RESULTS ON DATASETS

This paper conducts experiments on three datasets, namely Fer2013, RAF-DB, and CK+ to validate the effectiveness of the proposed face expression recognition method. The experimental results are then visualized.

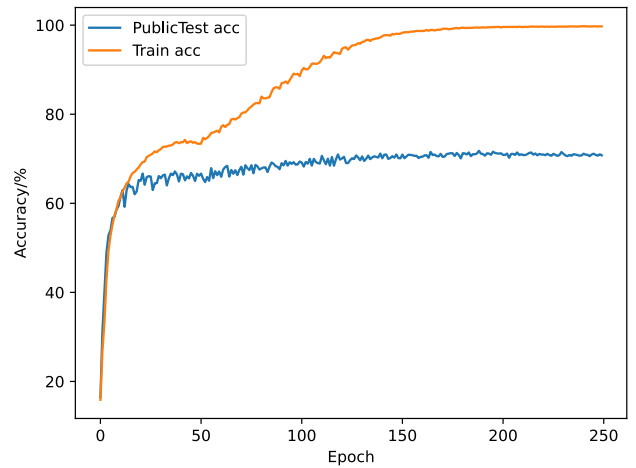


FIGURE 12. Fer2013 training accuracy curve.

Fig. 12 illustrates the training accuracy curve of the model on the Fer2013 dataset. Upon analyzing the graph, it can be deduced that the curve exhibits an overall upward trend as the number of iterations increases, indicating an improvement in training accuracy. In the initial 20 epochs, the training accuracy shows the fastest improvement rate. Subsequently, the curve’s rate of ascent begins to decelerate, resulting in a gradual decrease in slope. However, around the 50th epoch, the curve experiences an accelerated rise once again. After approximately 150 epochs, the curve becomes smooth, and the training accuracy reaches a state of relative stability, with no significant further changes. Meanwhile, during the early stages of training, before the 20th epoch, the test accuracy demonstrates the most rapid improvement. However, thereafter, the overall direction of the curve gradually stabilizes, the slope diminishes, and the rate of accuracy improvement slows down until convergence is ultimately achieved.

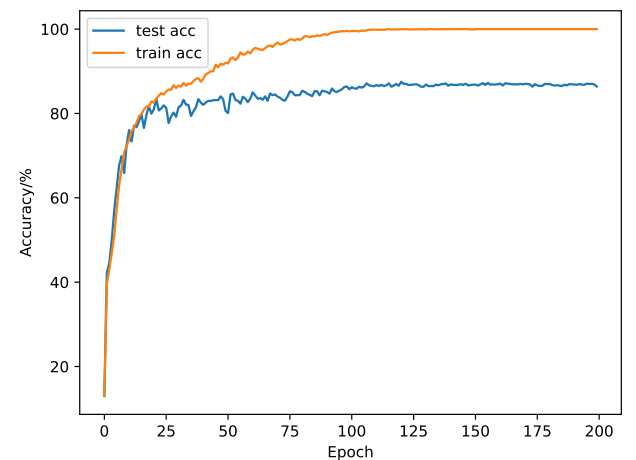


FIGURE 13. RAF-DB training accuracy curve.

Fig. 13 depicts the training accuracy curve derived from the model on the RAF-DB dataset. Upon careful analysis of the graph, it can be inferred that the curve exhibits a consistent upward trend as the number of iterations increases, indicating



a constant improvement in training accuracy. The most significant improvement in training accuracy is observed in the initial stages of the training phase, specifically before 25 epochs. Subsequently, between 20 and 100 epochs, the rate of increase in accuracy begins to diminish, and the slope of the curve gradually decreases. After 100 epochs, the curve becomes smoother, and the training accuracy ceases to exhibit significant changes until it ultimately converges. The rate of increase in test accuracy is highest in the early stages of training, i.e., the first 20 epochs. However, as the curve stabilizes, the slope gradually decreases, and the rate of improvement in accuracy slows down until convergence.

To further analyze the recognition accuracy of the model for different types of expressions, this paper presents the confusion matrices of the recognition results on three datasets: Fer2013, RAF-DB, and CK+.

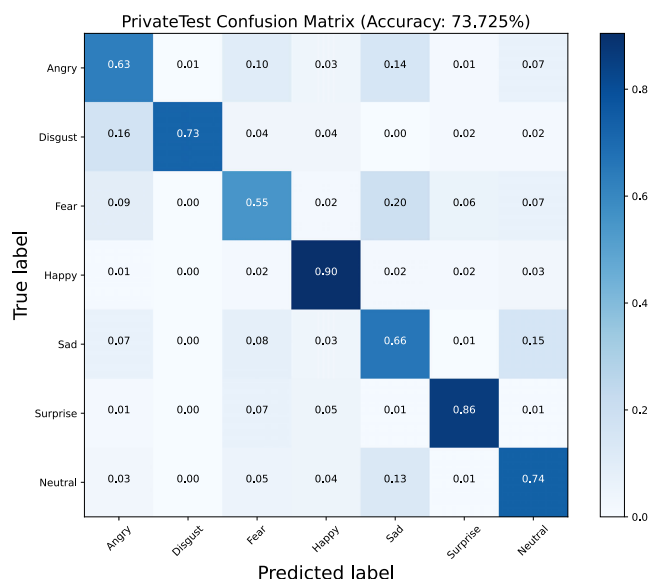


FIGURE 14. Fer2013 confusion matrix.

The confusion matrix depicting the performance of the proposed method on the Fer2013 test set is presented in Fig. 14. It is evident from the matrix that the proposed method attains superior recognition accuracy specifically for happy and surprised facial expressions, with accuracies of 0.90 and 0.86, respectively. However, the recognition rates for the fear, anger, and sadness expressions are comparatively lower. This observation can be attributed to the fact that the facial features associated with happy and surprised expressions are more distinct, enabling the neural network to extract relevant features more easily and achieve higher recognition rates. For instance, facial expression images of happy individuals tend to exhibit clear differentiation in feature information, such as the upward movement of the corners of the mouth and the presence of lines at the corners of the eyes. Similarly, surprised expressions are characterized by widened eyes and an open mouth.

In contrast, the recognition rates for fear, anger, and sadness expressions are relatively low. Among these, the fear category is particularly challenging to recognize due

to the similarity between fear and surprise expressions. Both expressions involve mouth opening, albeit with slightly different magnitudes, leading to potential confusion between the two categories. Furthermore, the remaining two categories, anger and sadness are both negative emotional expressions, sharing similar facial features and exhibiting minimal differences in key facial points. Consequently, there is a high likelihood of confusion, resulting in relatively low recognition accuracy for these three categories of expressions.

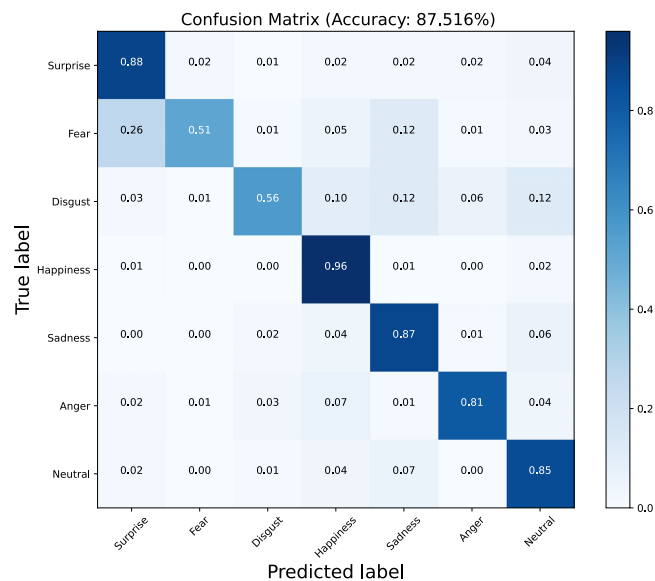


FIGURE 15. RAF-DB confusion matrix.

The confusion matrix of the RAF-DB test set, as depicted in Fig. 15, reveals that the algorithm achieves a higher average recognition rate on the RAF-DB dataset compared to the Fer2013 dataset. This disparity can be attributed to the larger dissimilarity between similar samples in the RAF-DB dataset, as well as the higher pixel value of the RAF-DB data (100×100) in comparison to the Fer2013 dataset’s 48×48, resulting in clearer images. Consequently, the algorithm exhibits a higher average recognition rate on the RAF-DB dataset than on the Fer2013 dataset.

Furthermore, the confusion matrix of the CK+ test set, illustrated in Fig. 16, demonstrates a significant improvement in the accuracy of recognizing the seven expressions when compared to both the previous Fer2013 test set and the RAF-DB dataset. This improvement is attributed to the CK+ dataset being captured under controlled laboratory conditions, minimizing the impact of environmental and human factors. Consequently, the CK+ dataset ensures higher image quality, leading to even better recognition accuracy for the model.

#### D. ABLATION EXPERIMENT

To assess the efficacy of each component of the pyramidal convolutional attention residual network (PCARNet) proposed in this study, ablation experiments were conducted on

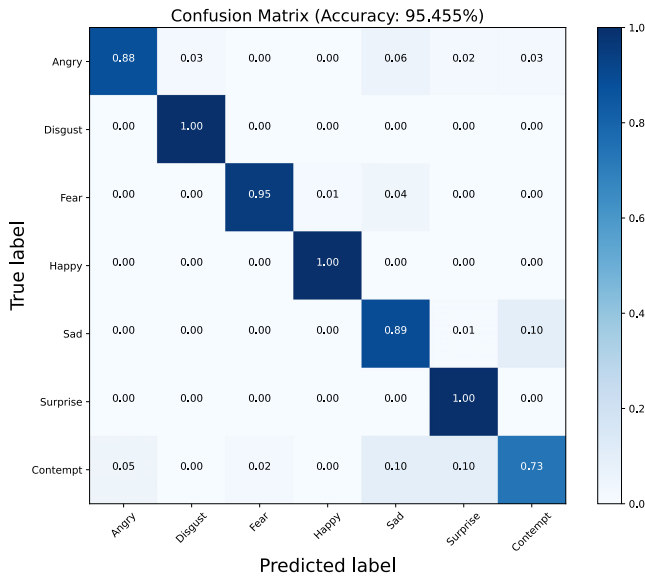


FIGURE 16. CK+ Confusion Matrix.

each module. The PyConv module represents the pyramidal convolutional structure, while the ECBAM module represents the optimized convolutional attention mechanism. Initially, facial expression images underwent data augmentation before being fed into the network. Subsequently, the improved pyramidal convolutional structure facilitated the extraction of facial expression features at multiple scales, thereby enhancing the model’s representational capacity. Furthermore, the ECABM attention mechanism facilitated information fusion across channels, enabling the extraction of deeper facial expression features and improving the accuracy of network recognition. The results of the PCARNet ablation experiments are presented in Table 1.

TABLE 1. PCARNet ablation experiment.

Baseline	PyConv	ECBAM	Fer2013	RAF-DB	CK+
+	-	-	72.471%	86.441%	94.040%
+	+	-	72.973%	87.256%	94.444%
+	-	+	72.889%	87.060%	95.152%
+	+	+	<b>73.725%</b>	<b>87.516%</b>	<b>95.455%</b>

The PyConv module represents the pyramidal convolutional structure, while the ECBAM module represents the optimized convolutional attention mechanism.

As depicted in Table 1, this study utilizes pyramid convolution for feature extraction, resulting in accuracy improvements of 0.5%, 0.8%, and 0.4% on the Fer2013, CK+, and RAF-DB datasets, respectively, compared to the baseline method. The classification performance is enhanced. Consequently, the utilization of pyramid convolution enhances the model’s feature extraction capability by utilizing convolution kernels of various sizes to capture both global and local facial information. The impact of this enhancement is more pronounced on the Fer2013 and RAF-DB datasets compared to the CK+ dataset. This discrepancy arises from the fact

that the CK+ dataset is obtained under controlled laboratory conditions, providing more complete facial information. Conversely, the Fer2013 and RAF-DB datasets represent real-world scenarios, where some of facial information is incomplete. Pyramid convolution effectively addresses this issue by extracting both local and global information. Furthermore, the integration of the ECBAM module proves to be valuable in improving expression recognition accuracy. The attention module facilitates the extraction of key features within the focus area, enhancing the discriminative power of the network. The combination of the ECBAM module and pyramid convolution synergistically enhances the extraction of multi-scale information while promoting effective information exchange across different channels. This comprehensive approach yields the best results for expression recognition.

V. COMPARISON EXPERIMENT

To assess the efficacy of the PCARNet algorithm proposed in this paper for facial expression recognition on human faces, we conduct a comparative analysis with the latest expression recognition algorithms from recent years.

TABLE 2. Comparison experiments with other methods.

Data	Method	Accuracy	Precision	Recall	F1
Fer2013	Zhou[24]	70.97%	-	-	-
	Liang[25]	70.30%	70.28%	70.49%	70.38
	RMFANet [26]	70.298%	68.95%	70.44%	69.69
	MIANet [27]	72.28%	71.43%	<b>72.53%</b>	71.98
	PCARNet(ours)	<b>73.725%</b>	<b>74.66%</b>	72.4%	<b>73.51</b>
RAF-DB	LDL-ALSG [28]	85.330%	-	-	-
	VGG+SE[29]	83.280%	77.26%	<b>78.66%</b>	77.95
	Li[30]	84.420%	-	-	-
CK+	WS-LGAN[31]	85.070%	77.06%	78.33%	77.69
	PCARNet(ours)	<b>87.516%</b>	<b>83.87%</b>	77.92%	<b>80.79</b>
	DNN[32]	93.24%	68.85%	70.82%	69.82
CK+	Gabor[33]	94.20%	92.19%	<b>93.11%</b>	92.65
	LBF[34]	93.77%	88.98%	91.4%	90.17
	DMTL[35]	95.20%	87.6%	88.11%	87.85
	PCARNet(ours)	<b>95.455%</b>	<b>93.17%</b>	92.14%	<b>92.65</b>

Specifically, we compare our method with the ones mentioned in literatures [24], [25], [26], and [27] using the Fer2013 dataset. In literature [24], the Softmax loss is improved by exploring different cosine values and input feature lengths. This enables the learning of more discriminative features and leads to better performance through an optimized loss function. In literature [25], a lightweight convolutional neural network structure is employed to extract

expression features. Additionally, a compression excitation module is incorporated to enhance the network's feature extraction capability for facial expressions, utilizing different compression rates. Literature [26] introduces a feature filtering module and proposes a residual multiscale feature fusion attention mechanism model. This model is based on feature filtering combined with improved convolution techniques. Literature [27] introduces a multi-scale integrated attention mechanism network (MIANet) that incorporates an Inception structure for extracting multi-scale feature information from an image. The network utilizes an efficient channel attention (ECA) mechanism to highlight regions associated with facial expressions and suppress irrelevant background regions, thereby enhancing the representation of crucial facial features.

Moving on to the RAF-DB dataset, we compare our method with the literatures [28], [29], [30], and [31]. Literature [28] presents a novel approach called Label Distribution Learning on Assisted Label Space Graphs (LDL-ALSG). This method utilizes label topology information from related but distinct tasks to improve facial expression recognition. In literature [29], a multiple attention network is proposed to recognize facial expressions. This network incorporates attention-enhanced features to improve accuracy. Literature [30] proposes an adaptively adjusted confidence threshold for semi-supervised deep facial expression recognition. This method aims to enhance recognition performance by dynamically adjusting the confidence threshold. Literature [31] introduces a weakly supervised local-global attention network (WS-LGAN). This network employs the attention mechanism to address partial location and feature fusion challenges in facial expression recognition.

At last, we compare the results of literatures [32], [33], [34], and [35] with the proposed method on the CK+ dataset. Literature [32] employs an extended deep neural network for expression recognition, while literature [33] uses a Gabor filter and a genetic algorithm. Literature [34] combines a mildly boosted decision tree with a neural network, and literature [35] uses adaptive pooling in a CNN for expression recognition. Table 2 displays the results of the comparisons.

Based on the experiments conducted, it is evident that the pyramid convolutional attention residual network proposed in this paper achieves the highest accuracy and precision rate in facial expression recognition. Due to the fact that Precision and recall metrics involve a trade-off, the recall of the method presented in this paper may not be the highest. However, the F1 metric for this proposed method is the highest. The F1 metric takes into account both precision and recall, and the table demonstrates that the proposed method in this paper is superior overall and exhibits higher quality.

## VI. CONCLUSION

To address the issue of insufficient ability to extract facial expression features in the presence of incomplete facial information, leading to low recognition accuracy, this paper proposes a novel algorithm called PCARNet for

face expression recognition. The proposed model utilizes pyramid convolution to extract both global and local features of the face. Furthermore, an adaptive one-dimensional convolution is applied in the attention mechanism to avoid the negative impact of channel dimensionality reduction on feature weights. This approach also promotes information fusion among multiscale features, enhancing key features within the extracted multi scale features from the pyramid convolution and improving the accuracy of expression recognition. To evaluate the effectiveness of the proposed model, comparative experiments are conducted in an end-to-end manner on publicly available face expression datasets, namely Fer2013, RAF-DB, and CK+. The results demonstrate that the proposed method significantly enhances the accuracy of expression recognition. It also indicates the importance of considering both global and local information in facial expression recognition. However, it is worth noting that this method does not take into account the relationship between facial expressions and ethnicity or age. Future work would focus on integrating these factors with expression features and applying them in real-world scenarios to further enhance the accuracy and practicality of expression recognition.

## REFERENCES

- [1] A. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognit.*, vol. 47, no. 3, pp. 1282–1293, Mar. 2014.
- [2] Y. Tong, R. Chen, J. Yang, and M. Wu, "Robust facial expression recognition based on local tri-directional coding pattern," in *Proc. Conf. Complex, Intell. Softw. Intensive Syst. (CISS)*. Cham, Switzerland: Springer, 2018, pp. 606–614.
- [3] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, Aug. 1987.
- [4] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: [10.1109/TPAMI.2002.1017623](https://doi.org/10.1109/TPAMI.2002.1017623).
- [5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 886–893, doi: [10.1109/CVPR.2005.177](https://doi.org/10.1109/CVPR.2005.177).
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, vol. 16, pp. 770–778, 2016, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
- [9] W. X. Wang, X. Zhou, and X. H. He, "Face expression recognition based on improved MobileNet network," *Comp. Appl. Softw.*, vol. 37, no. 4, pp. 137–144, 2020.
- [10] L. Q. Lan, X. Li, and Q. Y. Liu, "A face expression recognition method based on joint regularization strategy," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 46, no. 9, pp. 1797–1806, 2020, doi: [10.13700/j.bh.1001-5965.2020.0073](https://doi.org/10.13700/j.bh.1001-5965.2020.0073).
- [11] N. Zhou, R. Liang, and W. Shi, "A lightweight convolutional neural network for real-time facial expression detection," *IEEE Access*, vol. 9, pp. 5573–5584, 2021, doi: [10.1109/ACCESS.2020.3046715](https://doi.org/10.1109/ACCESS.2020.3046715).
- [12] C. Shi, C. Tan, and L. Wang, "A facial expression recognition method based on a multibranch cross-connection convolutional neural network," *IEEE Access*, vol. 9, pp. 39255–39274, 2021, doi: [10.1109/ACCESS.2021.3063493](https://doi.org/10.1109/ACCESS.2021.3063493).

- [13] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020, doi: [10.1109/TIP.2019.2956143](https://doi.org/10.1109/TIP.2019.2956143).
- [14] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, Oct. 2020.
- [15] J. Guo, Y. Dong, X. Liu, and S. Lu, "Facial expression recognition improved by attention mechanism and involution operator," *Comput. Eng. Appl.*, vol. 59, no. 23, pp. 95–103, 2023.
- [16] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, Daegu, Korea, 2013, pp. 117–124.
- [17] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019, doi: [10.1109/TIP.2018.2868382](https://doi.org/10.1109/TIP.2018.2868382).
- [18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101, doi: [10.1109/CVPRW.2010.5543262](https://doi.org/10.1109/CVPRW.2010.5543262).
- [19] I. Cosmin Duta, L. Liu, F. Zhu, and L. Shao, "Pyramidal convolution: Rethinking convolutional neural networks for visual recognition," 2020, *arXiv:2006.11538*.
- [20] J. Q. Qiao and L. Zhang, "X-ray target detection based on pyramidal convolution and band pooling," *Adv. Laser Optoelectron.*, vol. 59, no. 4, pp. 217–228, 2022.
- [21] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2018, pp. 3–19.
- [22] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: [10.1109/TPAMI.2019.2913372](https://doi.org/10.1109/TPAMI.2019.2913372).
- [23] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539, doi: [10.1109/CVPR42600.2020.01155](https://doi.org/10.1109/CVPR42600.2020.01155).
- [24] J. Zhou, X. Jia, L. Shen, Z. Wen, and Z. Ming, "Improved softmax loss for deep learning-based face and expression recognition," *Cognit. Comput. Syst.*, vol. 1, no. 4, pp. 97–102, Dec. 2019.
- [25] H. G. Liang and Y. X. Lei, "Research on facial expression recognition enhancing separable convolution channel features," *Comput. Eng. Appl.*, vol. 58, no. 2, pp. 184–192, Jan. 2021.
- [26] H. Zheng and G. Z. Zhao, "Facial expression recognition based on improved convolution multi-scale," *Comput. Eng. Appl.*, vol. 60, no. 2, pp. 231–243, 2024.
- [27] S. S. Luo, M. J. Li, and M. Chen, "Multi-scale integrated attention mechanism for facial expression recognition network," *Comput. Eng. Appl.*, vol. 59, no. 1, pp. 1–18, 2023, doi: [10.3778/j.issn.1002-8331.2203-0170](https://doi.org/10.3778/j.issn.1002-8331.2203-0170).
- [28] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 13981–13990, doi: [10.1109/CVPR42600.2020.01400](https://doi.org/10.1109/CVPR42600.2020.01400).
- [29] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020, doi: [10.1109/ACCESS.2020.2963913](https://doi.org/10.1109/ACCESS.2020.2963913).
- [30] H. Li, N. Wang, X. Yang, X. Wang, and X. Gao, "Towards semi-supervised deep facial expression recognition with an adaptive confidence margin," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 4156–4165, doi: [10.1109/CVPR52688.2022.00413](https://doi.org/10.1109/CVPR52688.2022.00413).
- [31] H. Zhang, W. Su, and Z. Wang, "Weakly supervised local-global attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 37976–37987, 2020, doi: [10.1109/ACCESS.2020.2975913](https://doi.org/10.1109/ACCESS.2020.2975913).
- [32] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, Apr. 2019.
- [33] A. Boughida, M. N. Kouahla, and Y. Lafifi, "A novel approach for facial expression recognition based on Gabor filters and genetic algorithm," *Evolving Syst.*, vol. 13, no. 2, pp. 331–345, Jul. 2021.
- [34] I. Gogić, M. Manhart, I. S. Pandžić, and J. Ahlberg, "Fast facial expression recognition using local binary features and shallow neural networks," *Vis. Comput.*, vol. 36, no. 1, pp. 97–112, Jan. 2020.
- [35] X. F. Wang, K. Wang, and X. Liu, "Adaptive re-weighted pooling deep multi-task learning for facial expression recognition," *Comput. Eng. Des.*, vol. 43, no. 4, pp. 1111–1120, Apr. 2022, doi: [10.16208/j.issn1000-7024.2022.04.029](https://doi.org/10.16208/j.issn1000-7024.2022.04.029).



**HUI QI** received the master's degree in computer science from Shanxi University, in 2009. She is currently a Professor with Taiyuan Normal University. Her research interests include machine learning, data mining, and image processing.



**XIPENG ZHANG** is currently pursuing the master's degree with Taiyuan Normal University. His research interests include image processing and deep learning.



**YING SHI** received the master's degree in computer science from Shanxi University, in 2015. She is currently a Lecturer with Taiyuan Normal University. Her research interests include machine learning and image processing.



**XIAOBO QI** received the Ph.D. degree in computer science from Shanxi University, in 2021. She is currently a Lecturer with Taiyuan Normal University. Her research interests include image processing, distributed computing, and data analysis.