Multimedia Tools and Applications (2023) 82:14287–14306
https://doi.org/10.1007/s11042-022-13799-8

TRACK 3: BIOMETRICS AND HCI

Facial expression recognition based on strong attention mechanism and residual network

Zhizhe Qian[1] · Jing Mu[1] · Feng Tian[2] · Zhiyu Gao[1] · Jie Zhang[1]

Received: 26 October 2021 / Revised: 15 July 2022 / Accepted: 5 September 2022 /
Published online: 28 September 2022

Abstract

Most facial expression recognition (FER) algorithms are based on shallow features, and the deep networks tend to lose some key features in the expression, such as eyes, nose and mouth. To address the limitations, we present in this paper a novel approach, named CBAM-Global-Efficient Channel Attention-ResNet (C-G-ECA-R). C-G-ECA-R combines a strong attention mechanism and residual network. The strong attention enhances the extraction of important features of expressions by embedding the channel and spatial attention mechanism before and after the residual module. The addition of Global-Efficient Channel Attention (G-ECA) into the residual module strengthens the extraction of key features and reduces the loss of facial information. The extensive experiments have been conducted on two publicly available datasets, Extended Cohn-Kanade and Japanese Female Facial Expression. The results demonstrate that our proposed C-G-ECA-R, especially under ResNet34, has achieved 98.98% and 97.65% accuracy, respectively for the two datasets, that are higher than the state-of-arts.

Keywords Facial expression · Attention mechanism · Residual module

✉ Jing Mu
mujing@xatu.edu.cn

Zhizhe Qian
qzz_922@163.com

Feng Tian
ft47@duke.edu

[1] School of Computer Science and Engineering, Xi'an Technological University, Xi'an, 710021, Shaanxi, China

[2] Division of Natural and Applied Science, Duke Kunshan University, Suzhou, China

1 Introduction

Facial expression is the most intuitive and natural way to convey personal mental activities, thoughts, mental states, and judge emotions. Facial Expression Recognition (FER) has been applied to many fields such as human-computer interaction, mental health assessment, driver fatigue detection, and criminal lie detection. Traditional FER includes 3 steps: image preprocessing, feature extraction, and feature classification. The methods like active shape model (ASM) [16], active appearance model (AAM) [24], appearance feature based local binary pattern (LBP) [32], and Gabor wavelet transform [15] are used for feature extraction, and hidden markov model (HMM) [4], bayesian classification (BN) [26], and support vector machine (SVM) [25] for feature classification. To further increase the accuracy of FER, in recent years, more and more researchers turn to deep learning, taking advantage of its end-to-end learning and direct extraction of features through network training. Wu et al. combined CapsNet [22] and Inception-ResNet-v1 to propose FaceCaps [30] that changes the first layer of the convolutional layer in the original capsule network to Inception-ResNet-v1. FaceCaps extracts the feature of facial expressions and then implements the final expression classification through the primary capsule, digital capsule routing, and fully connected layer. Cao et al. proposed E2-Capsnet network [5] that embeds the AU-attention mechanism to the VGG16 network and combines it with CapsulesNet. Arriaga et al. proposed a real-time convolutional neural network (CNN) for the classification of expressions and achieved good results [1].

To take a step further on feature extraction, especially those key features such as eyes, nose, and mouth, CNNs have been combined with other advanced methods such as attention mechanism, which imitates the human vision mechanism to obtain more useful information. In 2014, Mnih et al. [17] used the attention mechanism on the RNN model for image classification and focused on the key features of the image by assigning different weights to the feature map. In 2018, Roy et al. proposed the squeeze-and-excitation-network (SE-Net) channel attention mechanism [20] to learn the key feature connections between channels, which significantly improved the performance of CNN. Woo et al. proposed a convolutional block attention module (CBAM) attention mechanism [29] that combines channel attention and spatial attention. CBAM improves the network's capture of certain features and key locations. In 2020, Wang et al. proposed a lightweight efficient channel attention network (ECA-Net) [27] that has local cross-channel interaction capabilities to improve the performance of image classification and target detection based on SE-Net. In 2021, Wang et al. [28] used attention branches to extract important local information of facial expressions, and Gera et al. [7] proposed an end to end architecture for FER through a novel spatio-channel attention (SCAN) to capture features of per channel and spatial location.

However, the facial expressions extracted by most FER methods such as [6, 8] are insufficient because they do not pay strong attention to the key features, such as eyes, noses, and mouth. As a result, the deeper network layers such as those in [31, 33] may easily lose the distinctive key feature. [7, 28] used attention mechanism to strengthen the extraction of important features of facial expression, and [30] redesigned a new network FaceCaps to effectively extract expression features under the deep network according to the characteristics of different networks. To address these issues, in this paper we propose a novel framework that combines a strong attention mechanism with a residual network. ResNet is used as the backbone due to its depth feature extraction ability. The strong attention mechanism enhances the extraction of important features of facial expressions by embedding the channel and spatial attention mechanism before and after the residual module, and embeds Global-Efficient Channel Attention (G-ECA) proposed by us combines global and local attention mechanism inside the residual module of ResNet for comprehensively focusing on the important features on channel dimension. To strengthen the attention of facial expressions by different dimensions in different positions of the network, we also integrate the framework with CBAM [29] that consists of the channel and spatial attention mechanism, as shown in Fig. 1. Hence, we dub our framework as CBAM-Global-ECA-ResNet (C-G-ECA-R). Within the framework, we propose the global attention including a one-dimensional convolution with a 16-layer convolution kernel size of 1. Since the size of the facial expression image does not change, the characteristics extracted after the 16-layer one-dimensional convolution are more comprehensive. As a result, it is easy to extract the key characteristics of facial expressions in-depth. Note that the information exchange in the one-dimensional local cross-channel interactive convolution (LC-CI) of ECA, referred to as local attention, is among the partial channels, and the extraction of the information is not sufficient, so it needs to be fused with the global attention. To further explore the fusion strategy of the global and local attention, two variants of G-ECA are proposed in the paper and referred as G-ECA-1 and G-ECA-2. The difference between the two is the extraction process of facial expression by global and local attention mechanisms. Our experiments and ablation study have demonstrated that the proposed network structure has strengthened the attention of facial expression characteristics, reduced information loss during training, and improved the capture capabilities of key features. Consequently, the recognition rate of facial expression has been increased, compared to the state-of-arts. The visualization of the feature maps of facial expressions shows that among various combinations of attention mechanism and ResNet, CBAM-Global-ECA-ResNet34 (C-G-ECA-R34) focuses on the distinguishable key features in the facial expression and offers the best performance.

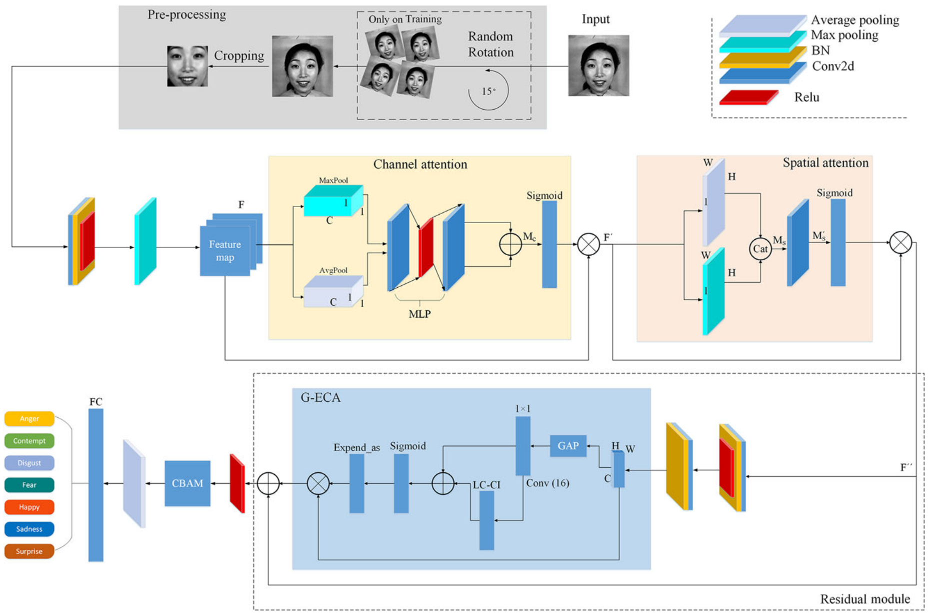


Fig. 1 Structure of C-G-ECA-R. The yellow and pink blocks are the channel and spatial attention in CBAM, respectively. The blue block, G-ECA, is located inside the residual module in the box with dotted lines

2 Methodology

2.1 C-G-ECA-R network structure

As shown in Fig. 1, the gray block is to preprocess input images. The expression image is rotated to improve the training capabilities of the network during the training process, and then the face area in the image is cropped out. The shallow feature map is obtained after the first convolution, batch normalization, activation function, and max-pooling. Afterward, the channel attention mechanism in the yellow block focuses on ’what’ is meaningful from the input image, and the spatial attention mechanism in the pink block to pay attention to ’where’ the key parts such as eyes are. The channel and spatial attention together constitute CBAM that can suppress trivial features, pay attention to the small number of channels and the important information of larger expression feature map from the channel and spatial dimensions, to obtain a new attention feature map. After entering the residual module, the number of channels in the feature map is more and the size is smaller. It passes through the first part of the convolution, batch normalization, and activation function. After the second part of convolution and batch normalization. The feature map enters the blue block and passes through the Global Average Pooling (GAP) layer, where the number of parameters of the model and the occurrence of overfitting are reduced.

Then the feature map is divided into two branches for training, within the channel G-ECA, i.e., the attention mechanism, as shown in the blue block of Fig. 1. In the first branch, the feature map passes through Conv(16) that is a one-dimensional convolution using a 16-layer convolution kernel size of 1, where the feature map is not changed in size, and the information of the feature map is extracted comprehensively and deeply. The global map represented by the global attention is then obtained. In the second branch, the global map passes through LC-CI, i.e., the local attention, that determines how many k adjacent channels are involved in the prediction of attention. Though it does not guarantee that the information exchange between all channels, the complexity of model is reduced. The global and local attention mechanisms are fused to obtain a more comprehensive feature map by adding the corresponding elements of the global and local feature maps. The feature map after the residual module is supplemented with the key feature information from the channel and spatial level through CBAM as an auxiliary before it is input into the classifier for classification.

2.2 Attention mechanism

2.2.1 Channel and spatial attention mechanism (CBAM)

CBAM [29] is proposed to imitate human visuals and used in FER to enhance the feature extraction of facial expressions by focusing on important characteristics of expressions in network training, such as eye, nose, mouth, and texture. CBAM consists of two parts: channel attention and spatial attention, as shown in the yellow and pink blocks in Fig. 1. CBAM is a lightweight module that has the characteristics of fewer parameters and can reduce irrelevant feature information while paying attention to the representation of effective feature information. The channel attention focuses on ’what’ feature of the feature map is more meaningful, while the spatial attention focuses on ’location’ of the information in the feature map, a kind of supplement to the channel attention. The principle of CBAM is to update the weights related to the key feature information so that the network learns the areas that need to be focused on the input expression feature map.

As shown in Fig. 1, the input of the CBAM is the feature map $F \in R^{C\times H\times W}$ of the upper layer. To obtain channel attention feature map, the input feature map passes through the max-pooling and average-pooling, and then through the multi-layer perceptron (MLP). The channel attention map for element-wise summation is $M_c \in R^{C\times 1\times 1}$. It is multiplied by the input feature map F to obtain a new feature map $F' \in R^{C\times H\times W}$, which is described as

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'^{\,C\times H\times W} = M_c(F)^{\,C\times 1\times 1} \otimes F^{\,C\times H\times W} \tag{2}$$

where $M_c{}^{C\times 1\times 1}$ is a one-dimensional channel attention map formed by channel attention, and $\otimes$ represents the multiplication operation of feature map elements.

The new feature map F' is used as the input feature map of spatial attention, passing through two pooling layers, to obtain two two-dimensional feature maps. The spliced feature map is $M_s \in R^{2\times H\times W}$ by the Cat operation, and then passed through a convolution layer to reduce the dimensionality into one channel. The spatial attention map formed is $M_s{}' \in R^{1\times H\times W}$, and then multiplied by the input feature map $F' \in R^{C\times H\times W}$ to obtain the final feature map $F'' \in R^{C\times H\times W}$, which is described as

$$F'' = M_s'(F') \otimes F' \tag{3}$$

$$F''^{\,C\times H\times W} = M_s'(F')^{\,1\times H\times W} \otimes F'^{\,C\times H\times W} \tag{4}$$

where $M_s'(F')^{\,1\times H\times W}$ is a one-dimensional spatial attention map formed by spatial attention.

2.2.2 ECA-Net

The purpose of a lightweight ECA-Net [27] is to improve SE-Net [20], for SE-Net's process of adjusting the dimensionality of the feature map causes the loss of some feature map information. To avoid the loss, ECA-Net uses an LC-CI instead of the squeeze-and-excitation module in SE-Net, which retains the dimension of the feature. The key characteristics of facial expressions are mainly achieved by fast one-dimensional convolution with a convolution kernel size of k, where k represents the coverage of the LC-CI, indicating how many adjacent channels participate in the key feature prediction and expand the weight of key features in the expression. Because ECA-Net enables the information of the feature map to

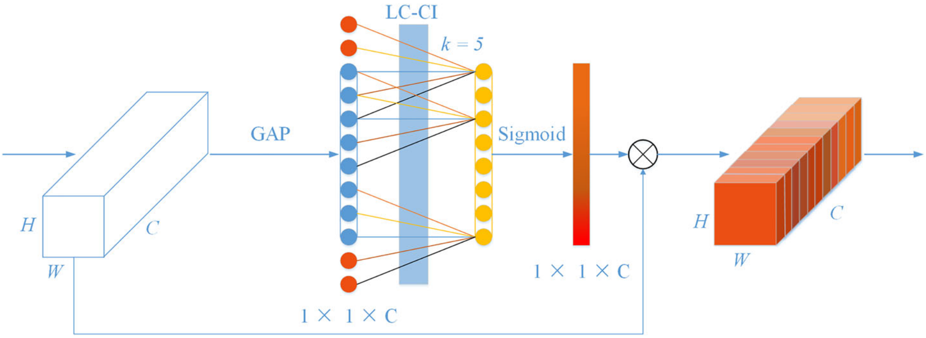


Fig. 2 Architecture of ECA. The blue, red and yellow circles represent one-dimensional expression features, padding elements and the generation features after LC-CI, respectively. The rectangles with gradient red represent the one-dimensional attention feature after Sigmoid

communicate between the partial channels, it avoids the computational complexity caused by crossing all channels and effectively improves network performance. Figure 2 shows the ECA attention mechanism structure.

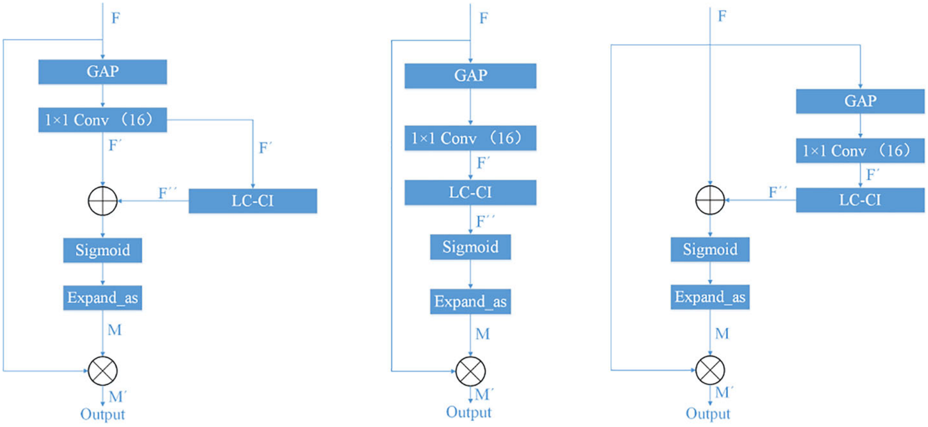
As shown in Fig. 2, the blue circle represents image features after GAP. When the size of the ECA attention convolution kernel is 5, the LC-CI range is also determined, as shown in Fig. 2 with 5 straight lines of different colors. In order not to change the size of the feature map, ECA determines the size of padding by subtracting 1 from k and dividing it by 2. When k=5, then padding=2, as shown in the red circle. The feature map after LC-CI is shown in yellow circles. The characteristics of the image represented by the yellow circles are passed through the Sigmoid function and multiplied by the input feature map to obtain a new feature map that has an ECA attention representation.

2.2.3 Three attention mechanisms based on ECA-Net

Since the information exchange of the feature maps in the one-dimensional LC-CI of ECA-Net is between the partial channels, the extraction of information is not sufficient. Therefore, in this paper, we propose a one-dimensional convolution with a 16-layer convolution kernel size of 1 to be global attention to combine local attention mechanisms, which is conducive to extract the features of expression comprehensively.

To explore the effectiveness of fusion, three novel ECA structures are proposed in this paper: G-ECA, G-ECA-1, and G-ECA-2, as shown in Fig. 3.

Figure 3(a) illustrates the structure of G-ECA that is shown in the blue block of Fig. 1. The symbols F, F' and F'' represents feature maps while M and M'' represent weight matrices. F' and F'' are fused by the addition of their corresponding elements. For instance, when F' (global feature map) focuses on the facial expression and F'' (local feature map) on the eyes, the fusion of F' and F'' will increase the weight of the eyes in the new feature map. The function Expand_as expands the feature map after fusion so that it has the same



(a) G-ECA (b) G-ECA-1 (c) G-ECA-2

Fig. 3 Three different fusion forms of global attention (1 × 1 Conv(16)) and local attention (LC-CI), and the new attention feature map formed after Sigmoid is through Expand_as extends to the same dimension as the input F

dimension as F. G-ECA aims to effectively improve the capture of key information in feature maps. The weight matrix M can be computed as

$$M = Expand\_as(\sigma(C1D_{16}(GAP(F)) \oplus C1D_k(F'))) \tag{5}$$

where $C1D_{16}$ are one-dimensional convolution with a 16-layer convolution kernel size of 1, $C1D_k$ is LC-CI, $\otimes$ is the sum of corresponding elements of the feature map, and σ is the Sigmoid activation function.

Figure 3(b) represents the structure of G-ECA-1. Its purpose is to verify whether adding a global attention representation layer directly to the middle between GAP and LC-CI enhances the attention to facial expression features. The weight matrix M can be computed as

$$M = Expand\_as(\sigma(C1D_k(C1D_{16}(GAP(F))))) \tag{6}$$

Figure 3(c) represents the structure of G-ECA-2. The purpose of G-ECA-2 is to verify whether the weight matrix M formed by the fusion of the F and F'' after global and local attention can enhance the focus on the key features of facial expressions. The weight matrix M can be computed as

$$M = Expand\_as(\sigma(F \oplus C1D_k(C1D_{16}(GAP(F))))) \tag{7}$$

Multiplying M with the input feature map F gives us

$$M' = F \otimes M \tag{8}$$

The weight of LC-CI can be implemented through a fast one-dimensional convolution as

$$\omega = (C1D_k(y)) \tag{9}$$

where k is the convolution kernel size and determines the coverage of the adjacent channel. The one-dimensional convolution of size k is used to generate weights for each feature channel to obtain the correlation between the feature channels. The channel dimension C can be computed as

$$C = 2^{(\gamma * k - b)} \tag{10}$$

where γ and b are set to 2 and 1, respectively, which makes k have a larger interaction range. k is computed as

$$k = \left| \frac{log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{11}$$

where the odd indicates that the nearest odd number is selected, and different channels C will generate a different interaction range k. Note that the (9), (10) and (11) are similar to that presented in [27], where more details can be found.

2.3 Loss function

In this paper, we use the cross-entropy function to measure the difference between the true probability distribution and the predicted probability distribution. The smaller the value of cross-entropy, the better the model predicts. The cross-entropy function avoids the vanishing gradient problem of the network, speeds up the training, and is suitable for facial expression classification. The loss function is defined as

$$L = -\sum_{i=1}^{N} y_i' log(y_i) \tag{12}$$

where N is the number of expression categories, y_i' is the variable (0 or 1). If the predicted expression category is the same as the sample expression category, y_i' is 1, otherwise, it is 0.

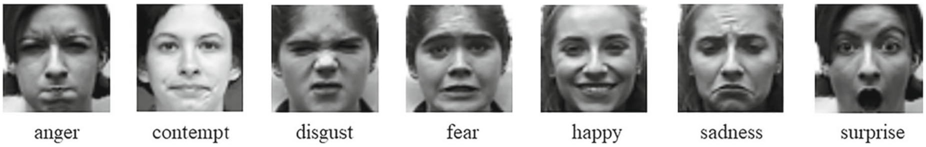


Fig. 4 Sample images in CK+

y_i represents the predicted probability that the observed expression sample belongs to the expression category i. A smaller L indicates that the predicted expression is closer to the real sample expression.

3 Experiment and result analysis

3.1 Environment configuration

The implementation of the model is based on Pytorch, using Python 3.7. All the experiments were carried out using a workstation with 11th Gen Intel(R) Core(TM)i5-1135G7 2.42GHz and 16G RAM.

3.2 Dataset and data enhancement

The experiments choose two datasets publicly available in the FER research field: Extended Cohn-Kanade (CK+), Japanese Female Facial Expressions (JAFFE). CK+ is expanded based on the Cohn-Kanade Dataset, in which a total of 123 subjects participated, ranging in age from 18 to 50 years old, including European and American, non-European, and other races. The dataset contains seven types of expressions: happy (207), sadness (84), fear (75), anger (135), disgust (177), surprise (249), and contempt (54) where the numbers in parentheses denote the number of expressions for a particular type. CK+ is based on the 593 image sequences composed of the subjects, 327 sequences with markers are selected. We selected three frames with peak formation from the labeled sequence and obtained 981 images. JAFFE contains 231 face images, and 7 expressions made by 10 female students selected in Japan, including 6 basic ones: happy (31), sadness (31), fear (32), anger (30), disgust (29), surprise (30), and one neutral expression (30). Sample images are shown in Figs. 4 and 5, respectively.

We divide CK + and JAFFE datasets into three parts: training set, validation set, and test set, with a division ratio of 6:2:2. To improve the generalization ability of the model and reduce the over-fitting problem, first, we randomly rotate images 15 degrees. Then, to ensure that most of the extracted features come from the face area and reduce the interference of irrelevant factors, we crop the face area from the central area of the image, which is more

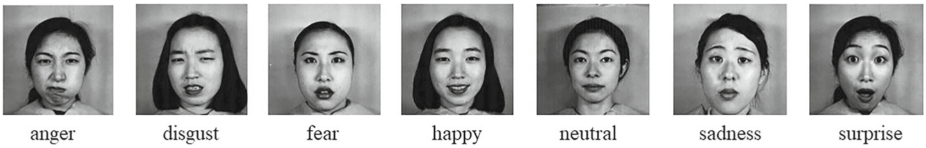


Fig. 5 Sample images in JAFFE

Table 1 Performance comparison with existing methods on CK+

Methods	Accuracy (%)
Lopes et al.(2017) [13]	95.75
Li et al.(2019) [10]	97.38
Sadeghi et al.(2019)[23]	95.11
OAENet(2020) [28]	98.65
Rao et al.(2021) [19]	98.68
SCAN(2021) [7]	97.31
NTF_LRS(2021) [12]	95.77
DCNN(2021) [18]	96.79
C-G-ECA-R34 (Ours)	**98.98**

conducive to the analysis of expression features. Finally, the data is normalized before the training and the Adam optimizer. With the normalization, the activation function have a more reasonable value, the gradient can be spread well and the convergence speed of the model can be accelerated. The initial learning rate is 0.001, and the batch size is 32. In this paper, the training cycles for each experiment are 150.

3.3 Performance comparison and analysis

To verify the performance of C-G-ECA-R proposed in this paper, we compare it with several existing network structures. We choose ResNet34 as the backbone of our network because the experiments later show that C-G-ECA-R34 offers the optimal performance. Tables 1 and 2 present the classification results of various network structures on CK+ and JAFFE, respectively.

Table 1 shows the comparison on CK+.

1) The performance of C-G-ECA-R34 is better than that of conventional methods. Sadeghi et al. [23] utilized the Gabor filter to extract the expression features, but the model requires human intervention, which is not entirely accurate, resulting in a poor recognition on CK+. NTF_LRS constructs tensor models to get the higher-order correlations of the underlying multi-dimensional facial expression data, but ignores the global relationships between adjacent expression samples, failing to achieve the high accuracy of

Table 2 Performance comparison with existing methods on JAFFE

Methods	Accuracy (%)
ARLCP(2019) [21]	94.41
SCNN(2019) [9]	93.02
Li et al.(Without LBP)(2020) [11]	96.53
Avani et al.(2020) [2]	97.48
SCAN(2021) [7]	56.33
HDG(2021) [3]	90.00
HDGG(2021) [3]	91.43
DCNN(2021) [18]	95.63
MFF-CNN(2021) [34]	96.52
C-G-ECA-R34 (Ours)	**97.65**

recognition. On the contrary, our CNN-based model does not require human intervention and achieves more efficient feature extraction through both the shallow and deep parts of the network.

2) C-G-ECA-R34 outperforms the competitive CNN methods. Lopes et al. [13] proposed a new method of CNN combined with preprocessing operations. However, the CNN model is too simple to extract sufficient features. Similar to [13], the simple CNN model used in [10] was unable to achieve optimal recognition rates, despite the new expression cropping strategy proposed. DCNN [18] combines traditional feature extraction with deep learning. However, practical results show that this approach is inferior to deep learning methods, as deep learning is stronger for shallow feature extraction. Compared with the above methods, C-G-ECA-R34 utilizes an attention mechanism to solves the problem of inadequate feature extraction while being able to focus on the key features of expressions, such as eyes and mouth.
3) C-G-ECA-R34 is better than other CNN combined attention methods. Both SCAN and OAENet contain global attention and local attention, and in common with them, C-G-ECA-R34 is also designed for both global and local attention. Unlike them, C-G-ECA-R34 adds effective attention at different positions of the residual module to strengthen the extraction of important features of expression. As a result, C-G-ECA-R34 achieves better recognition.
4) Rao et al. [19] proposed a novel multi-scale graph convolutional network. Based on landmark maps extracted from facial images, a graph segmentation strategy was used to construct a multi-scale GCN structure by exploring homogeneous subgraphs to change the receptive fields. The approach achieves the accuracy of recognition as 98.68% that is close to our C-G-ECA-R34, as shown in Table 1. However, it can extract only 98 facial landmarks, while more subtle facial expressions may be overlooked, as the formation of expressions is a common result of the whole facial landmarks. As CK+ is laboratory data with higher quality, our model is more conducive to capturing the features of expressions clearly through strong attention.

Table 2 represents the comparison on JAFFE.

1) The proposed C-G-ECA-R34 model outperforms conventional methods and CNN methods. Conventional methods, such as ARLCP, effectively encoded significant information of expressions by using histogram. HDG and HDGG used different Histogram of Directional Gradient to distinguish expression features respectively. CNN based methods, such as SCNN, automatically organized several features of facial expressions, DCNN fused expression features through different branches, and a lightweight network MFF-CNN fused multi-feature expressions to better extract features at different levels of abstraction. It is worth noting that despite the high image quality of JAFFE, the approaches ARLCP, HDG and HDGG are sensitive to image noise due to their using histogram. Whereas SCNN, DCNN and MFF-CNN all use feature fusion strategies, it is difficult to find the best features to fuse. Our model is less sensitive to noise by focusing on the important features of expressions and amplifies the corresponding weights. This approach outperforms the feature fusion strategy thus achieves better recognition rates on JAFFE.
2) The proposed method is compared with CNN combined with attention and other competitive strategies. The results show that our proposed model has strong attentional capabilities to enhance the capture of effective features of expressions. The recognition rate of SCAN is high under CK+ in Table 1, but low under JAFFE in Table 2, indicating the limitations of SCAN's recognition capability. Li et al. [11] use attention as part

Table 3 Performances on different layers of the proposed global attention. The dichotomy is used to verify between 1 and 50 layers, and measured by multiple indicators in the table

Model name	Epoch	B_Epoch	Loss	Val_Acc	Test_Acc	Param	GFLOPs	Time	Continue
C-G-ECA_1-R18	150	37	0.001652	97.95	95.93	2.8M	2.1325	21.35	√
C-G-ECA_7-R18	150	77	0.001354	98.97	97.47	2.8M	2.1326	25.12	×
C-G-ECA_13-R18	150	47	0.003690	97.95	97.76	2.8M	2.1328	24.56	√
C-G-ECA_14-R18	150	65	0.002979	97.65	97.72	2.8M	2.1320	24.25	×
C-G-ECA_16-R18	150	**33**	**0.000739**	**98.46**	**97.96**	2.8M	**2.1330**	**22.15**	√
C-G-ECA_17-R18	150	77	0.003380	98.32	97.76	2.8M	2.1329	25.66	×
C-G-ECA_19-R18	150	80	0.006610	98.26	97.55	2.8M	2.1327	25.79	√
C-G-ECA_22-R18	150	76	0.005210	97.27	96.77	2.8M	2.1332	26.55	×
C-G-ECA_25-R18	150	69	0.002993	97.95	96.95	2.8M	2.1325	26.73	√
C-G-ECA_31-R18	150	83	0.001356	97.95	97.86	2.8M	2.1334	26.23	×
C-G-ECA_37-R18	150	67	0.001438	97.95	97.77	2.8M	2.1330	30.41	√
C-G-ECA_43-R18	150	75	0.001532	98.43	97.72	2.8M	2.1335	42.00	×
C-G-ECA_50-R18	150	70	0.016043	98.97	97.86	2.8M	2.1325	44.16	√

of the model, while we focus on different locations of attentiond placement as a way to improve the performance of the model. Avani et al. [2] divide the face into four regions and extract features separately. This sub-regional approach is more rigid and lacks deep exploration of region partitioning.

3.4 Global attention experiment and comparison

In order to explore how many layers of one-dimensional convolution with a convolution kernel size of 1 inside global attention gives the best performance, this experiment verifies the dichotomy between the one-dimensional convolution with the convolution kernel size of 1 from 1 to 50 layers, using C-G-ECA-R18 and CK+. ResNet18 is less complex than ResNet34 or ResNet50, which makes the effect of layers of global attention more obvious. The results will be analyzed according to the Epoch, the best_epoch (B_Epoch), Loss, the accuracy of the verification set (Val_Acc), the accuracy of the test set (Test_Acc), and the number of parameters (Param), GFLOPs, Time(min) and Continue. Here the Continue means whether or not the dichotomy operation is continuous. Table 3 shows the number of one-dimensional convolutions with a convolution kernel size of 1 between 1-50 layers. Figure 6 shows the specific process of the Continue operation.

It can be known from Table 3 that when the number of one-dimensional convolutional layers is 16, C-G-ECA-R18 is optimal compared with other layers. This is reflected in the fact that the best number of Epoch appears earlier, indicating that the network has better training capabilities. Moreover, the loss is the smallest, and the results for both the verification set and test set are good, as shown in Table 3. Although the Param and GFLOPS for different layers are not much different, it can be clearly seen that when the number of layers is 16, the time used for network training is the least. This shows that G-ECA proposed in this paper through one-dimensional convolution with a 16-layer convolution kernel size of 1 has better performance on the C-G-ECA-R18 network than convolutions with other layers in Table 3. Figure 6 shows the complete dichotomy operation process intuitively based

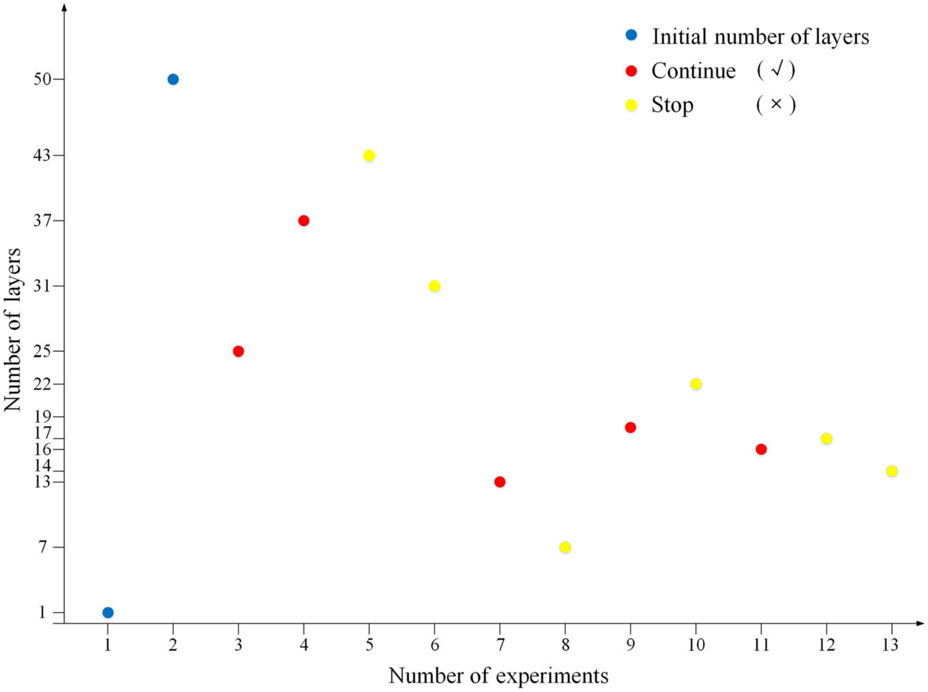


Fig. 6 Process of continuing the dichotomy operation between 1-50 layers. The number of experiments are 11, and the one-dimensional convolution with a convolution kernel size of 1 inside global attention is 16

on the results in Table 3. The blue circle indicates the initial 1 and 50 layers, the red indicates that the dichotomy operation can continue, and the yellow circle indicates that the dichotomy operation is stopped. It can be seen that when the one-dimension convolution with a convolution kernel size of 1 is in 16 layers, the effect is optimal.

3.5 Ablation study

To verify the effectiveness of the proposed strong attention mechanism combined with the residual network structure, we conducted an ablation study with 4 experiments on CK+ and JAFFE and compared the performance of G-ECA and ECA. The network performance was measured by the recognition rate of the verification set (Val_Acc (%)) and the test set (Test_Acc (%)). Using ResNet18 and ResNet34 as the backbone, we compare a few network structures, including (1) +Head_CBAM (H_Attention): the experiment adds CBAM before the residual module; (2) +Tail_CBAM (T_Attention): the experiment adds the CBAM after the residual module; (3) +G-ECA, +G-ECA-1, +G-ECA-2: the experiment put the corresponding attention mechanism inside the residual module.

Experiment 1 Based on the ResNet18, the effects of the various modules on the experimental results are compared, as shown in Table 4.

It can be seen from Table 4 that compared with ResNet18, +H_Attention increases the accuracy of the verification set and the test set by 1.50% and 1.87%, respectively, proving that the addition of attention mechanism improves the attention of the key parts of the expression feature map. The results of +H_Attention + ECA indicate that the ECA

Table 4 Comparison of the accuracy of models with different structures under ResNet18 (CK+)

Model name	Val_Acc (%)	Test_Acc(%)
ResNet18	95.45	94.04
+H_Attention	96.95	95.91
+H_Attention + ECA	97.95	96.95
+H_Attention + ECA + T_Attention	97.44	97.46
+H_Attention + G_ECA + T_Attention (C-G-ECA-R18)	**98.46**	**97.96**

module can improve the attention of facial expressions. Note that while the accuracy of +H_Attention + ECA + T_Attention (C-ECA-R18) for the verification set drops by 0.51%, its accuracy on the test set increases by 0.51%. The results prove that the effect of adding CBAM after the residual network is not large, but it can be used as an auxiliary role to focus on the characteristics of expression. For +H_Attention + G_ECA + T_Attention, we can see that the accuracy for both the verification set and test set has been improved significantly, with an increase of 1.02% and 0.50%, respectively. On the whole, C-G-ECA-R18 network with ResNet18 as the main framework has improved the performance of the CK+ dataset significantly.

Experiment 2 Different from the previous experiment above, we use ResNet34 to compare the performance of various network structures, as shown in Table 5.

It can be seen from Table 5 that as the number of network layers increases, the overall performance of the network also improves. The recognition rate of +H_Attention + ECA under ResNet34 for the verification set is relatively not changed compared to that under ResNet18, and the recognition rate of the test set is increased by 1.01%. The recognition rates of C-ECA-R34 in the verification set and test set are improved by 1.02% and 1.01%, respectively, compared with C-ECA-R18. The recognition rates of the C-G-ECA-R34 in the verification set and test set are increased by 0.51% and 1.02%, respectively, compared with C-G-ECA-R18.

To see how the network performance changes when the network layers deepen, the accuracy curves of C-G-ECA-R18 and C-G-ECA-R34 are shown in Fig. 7.

It can be seen that from the 100^{th} cycle after the training, the accuracy of C-G-ECA-R18 becomes stable and stops increasing; but for C-G-ECA-R34, the recognition rate is still slowly rising. This also proves that C-G-ECA-R can achieve a significant increase in attention to key features even when the number of network layers increases. Through the above experiments, it is clear to know that the network with G-ECA is better than the one with ECA. To further verify the effectiveness of the proposed network structure, we conducted a

Table 5 Comparison of the accuracy of models with different structures under ResNet34 (CK+)

Model name	Val_Acc (%)	Test_Acc(%)
ResNet34	96.53	95.02
+H_Attention	97.23	96.37
+H_Attention + ECA	97.95	97.96
+H_Attention + ECA + T_Attention	98.46	98.47
+H_Attention + G_ECA + T_Attention (C-G-ECA-R34)	**98.97**	**98.98**

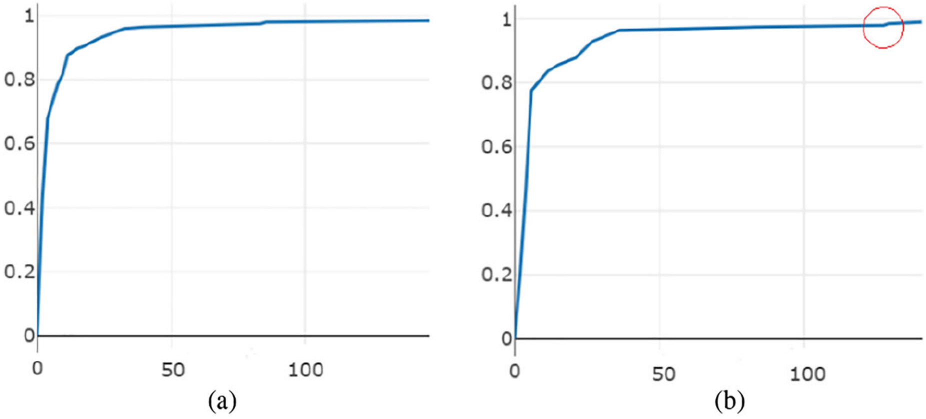


Fig. 7 (a) accuracy of C-G-ECA-R18 under ResNet18. (b) accuracy of C-G-ECA-R34 under ResNet34

similar experiment on C-ECA-R34 and C-G-ECA-R34, but under the JAFFE dataset. The results are given in Table 6, showing that the recognition accuracy of C-G-ECA-R34 is higher than C-ECA-R34.

Experiment 2.1 Performance measurement.

To prove that G-ECA in this paper is better than ECA, we further analyze the confusion matrix of C-ECA-R34 and C-G-ECA-R34 network under CK+ and JAFFE datasets and the corresponding $Precision$, $Recall$, and $Specificity$, ROC curve, and $visualization$ of feature maps. The following equations show the calculation of these values, where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative respectively. $Precision$ is the measure to identify the number of relevant facial expressions among the classified ones. It is also the ratio of the correct predictions to the total predictions.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

The $Recall$ calculates the correctly classified facial expression images over all the facial expression images in the dataset.

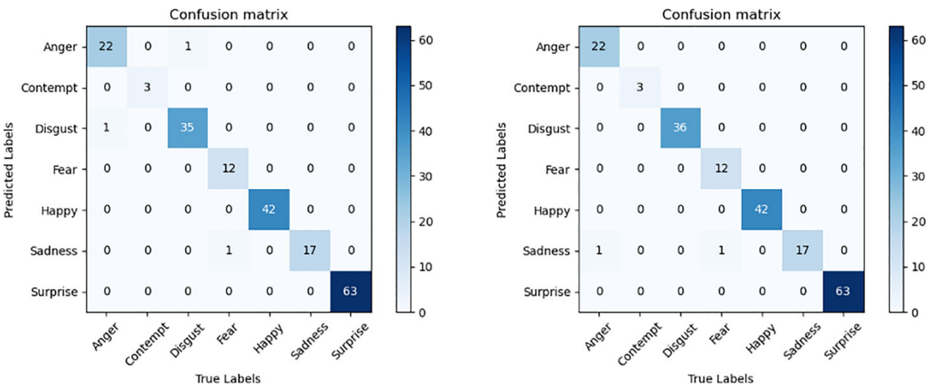
$$Recall = \frac{TP}{TP + FN} \tag{14}$$

The $Specificity$ is also for true expressions, but it represents the proportion of negatives that the facial expressions correctly predict among all real negatives.

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

Table 6 Comparison of the accuracy of models with different structures under ResNet34 (JAFFE)

Model name	Val_Acc (%)	Test_Acc(%)
+H_Attention + ECA + T_Attention (C-ECA-R34)	94.92	94.53
+H_Attention + G_ECA + T_Attention (C-G-ECA-R34)	97.69	**97.65**



(a) C-ECA-R34

(b) C-G-ECA-R34

Fig. 8 Confusion matrix under CK+. Diagonals represent the correct number of different expressions

With the above three evaluation criteria, the confusion matrices for C-ECA-R34 and C-G-ECA-R34 are calculated under CK+, as shown in Fig. 8. Tables 7 and 8 give a comparison of two networks in $Precision$, $Recall$, and $Specificity$.

The $True$ Labels in Fig. 8 represent the real expression, and $Predicted$ Labels represent the predicted expression. Both figures have one false prediction of the real Anger. From the $Precision$ of Tables 7 and 8, the Anger under C-ECA-R34 is 95.6521%, and C-G-ECA-R34 is 100%. The $Precision$ for Disgust is 97.2973% for C-ECA-R34 and 100% for C-G-ECA-R34, respectively. In total, there are three prediction errors in C-ECA-R34 and only two errors in C-G-ECA-R34. It is worth noting that both (a) and (b) consider Fear to be Sadness, indicating that the two types of expressions are more difficult to distinguish. In fact, as shown in Fig. 4, there is some similarity between the two.

From the $Recall$ of Tables 7 and 8, the $Recall$ of C-G-ECA-R34 is 100%, better than 97.2578% under C-ECA-R34 for Disgust. The $Specificity$ has reached 100% for Anger and Disgust. Obviously, we can see that the effect of C-G-ECA-R34 in the CK+ dataset is superior to C-ECA-R34.

For intuitive observation of C-G-ECA-R34 recognition capabilities for different expressions, the performance of the experiment is displayed on the CK+ and JAFFE datasets by

Table 7 Precision, Recall, and Specificity under C-ECA-R34 (CK+)

7 Classes	Precision (%)	Recall (%)	Specificity (%)
Anger	95.6521	95.7569	99.4252
Contempt	100	100	100
Disgust	97.2973	97.2578	99.3788
Fear	100	92.3679	100
Happy	100	100	100
Sadness	94.1176	100	99.4327
Surprise	100	100	100

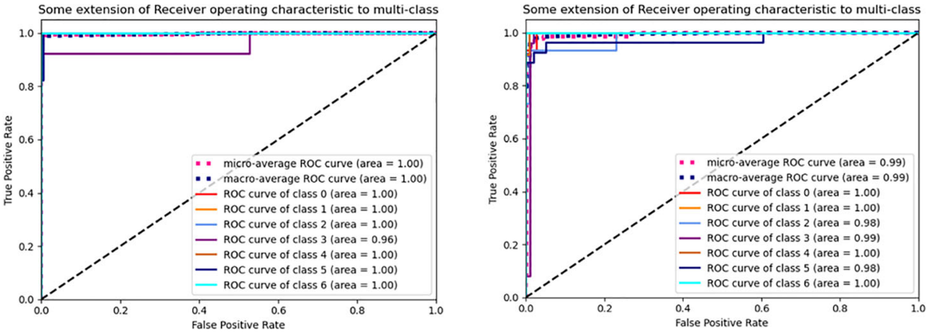
Table 8 Precision, Recall, and Specificity under C-G-ECA-R34 (CK+)

7 Classes	Precision (%)	Recall (%)	Specificity (%)
Anger	100	95.6521	100
Contempt	100	100	100
Disgust	100	100	100
Fear	100	92.3076	100
Happy	100	100	100
Sadness	89.7736	100	98.8888
Surprise	100	100	100

the ROC curve shown in Fig. 9. The continuous line indicates the ability to identify different categories. The interrupted lines are Micro-Average Roc Curve and Macro-Average Roc Curve. The higher the two, the better classification a network achieves.

As shown in Fig. 9, there is a good performance on Micro-Average Roc Curve and Macro-Average Roc Curve. On the CK+ dataset, only one type of expression recognition effect has not reached the highest level. On the JAFFE dataset, only three types of expression recognition have a slight deviation, but the overall recognition rate of C-G-ECA-R34 has reached a very high level. From the confusion matrix, $Precision$, $Recall$, and $Specificity$, and the ROC curve, G-ECA has a significant improvement in the performance of C-G-ECA-R34 compared to ECA, which proves that the G-ECA proposed in this paper has a good effect, and the C-G-ECA-R34 network has good performance in FER.

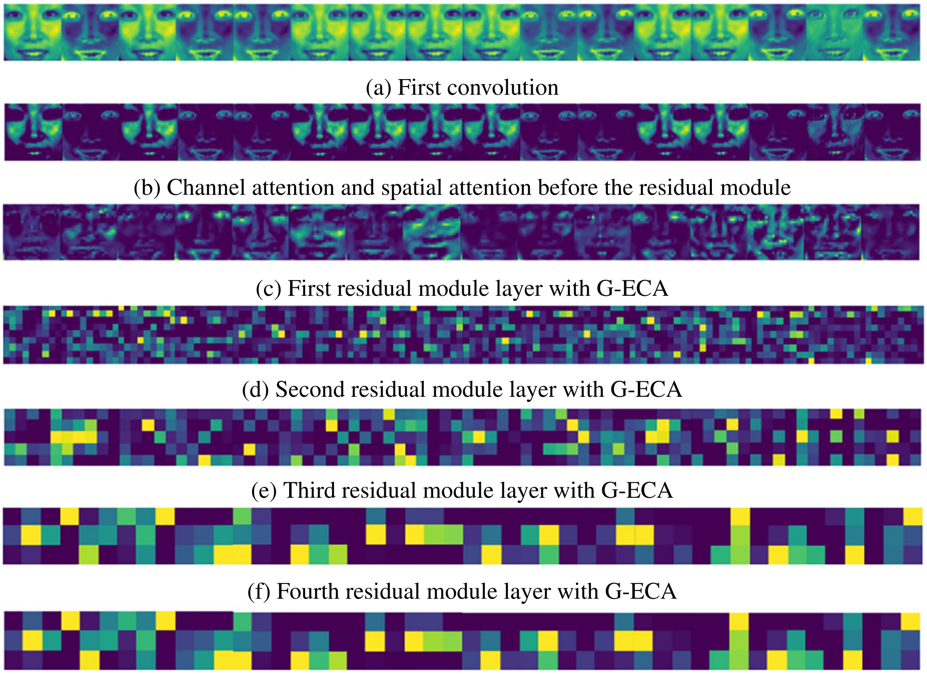
To show key features in the expression that are noted through C-G-ECA-R34, the visualization maps in partial layers during the training process of the CK+ dataset are shown in Fig. 10. The figure shows the visual feature map of the intermediate process of FER based on the C-G-ECA-R34. The yellow and green areas indicate the salient features of facial expressions. As shown in (a)-(c), C-G-ECA-R34 pays good attention to the salient areas of facial expressions, such as the eyes, nose, and mouth areas. It can be seen from (d)-(g) that although the feature map is getting more and more blurred, the yellow and green blocks are still concentrated in the salient area of facial expressions, and other marginal feature ranges can also be learned.



(a) CK+

(b) JAFFE

Fig. 9 The micro-average pays more attention to the expression with more samples, and the macro-average pays attention to the category with less samples. The classes 0-6 stand for Anger, Contempt, Disgust, Fear, Happy, Sadness, and Surprise, respectively



(a) First convolution

(b) Channel attention and spatial attention before the residual module

(c) First residual module layer with G-ECA

(d) Second residual module layer with G-ECA

(e) Third residual module layer with G-ECA

(f) Fourth residual module layer with G-ECA

(g) Channel attention and spatial attention after the residual module

Fig. 10 Visualization of the facial expression characteristics of C-G-ECA-R34

Experiment 3 To verify the network performance of C-G-ECA-R will continue to increase, or stabilize with the increase in the number of network layers, in this experiment, we compare the performance of our proposed networking using ResNet50 and ResNet101 as the backbone. Table 9 gives the accuracy of C-G-ECA-R50 and C-G-ECA-R101 under the verification set and test set of CK+.

It can be seen from Table 9 that when the network deepens, the performance of C-G-ECA-R slightly declines, compared with C-G-ECA-R34 in Table 5, but the recognition rate of the verification set and the test set are relatively close, which proves that the overall performance of the C-G-ECA-R is still stable. As shown in DCANet [14], the effect of the attention mechanism in large network is not as good as that in small network. This observation has been demonstrated from Tables 9 and 5 that C-G-ECA-R achieves the optimization of performance when ResNet34 is used as the backbone.

Table 9 Comparison of network accuracy under ResNet50, ResNet101

Model name	Val_Acc (%)	Test_Acc(%)
+H_Attention + G-ECA + T_Attention (C-G-ECA-R50)	97.95	97.96
+H_Attention + G_ECA + T_Attention (C-G-ECA-R101)	97.75	97.56

Table 10 Performance comparison of G-ECA, G-ECA-1, and G-ECA-2 (CK+)

Model name	Val_Acc (%)	Test_Acc(%)
+H_Attention + G-ECA + T_Attention (C-G-ECA-R34)	**98.97**	**98.98**
+H_Attention + G_ECA-1 + T_Attention (C-G-ECA-1-R34)	97.95	99.49
+H_Attention + G_ECA-2 + T_Attention (C-G-ECA-2-R34)	98.46	98.47

Experiment 4 Based on the above experiments, this attention placement structure of C-G-ECA-R34 is optimal. Therefore, in this experiment, we explore and compare the performance of G-ECA, G-ECA-1, G-ECA-2 under R34. The results are shown in Tables 10 and 11.

It can be seen from Tables 10 and 11 that C-G-ECA-R34 not only has a high recognition rate on the verification set and test set of CK+ and JAFFE, but also the results of the verification set and test set are closer, showing the network performance is stable. For CK+, C-G-ECA-1-R34 is 0.51% higher than the C-G-ECA-R34 in the test set. Moreover, the results show that the verification set accuracy and test set accuracy of C-G-ECA-R34 are closer, proving its stability. C-G-ECA-2-R34 has a slightly lower effect on JAFFE than the other two networks. From the ablation study, it can be concluded that the three network structures combining the strong attention mechanism and the residual network proposed in this paper have a good ability to capture the key information of facial expression features. And among the three network structures, C-G-ECA-R34 achieves the best performance.

4 Conclusion

This paper proposes a C-G-ECA-R network structure for facial expression recognition which can strengthen the capture of key information and improve the recognition rate of facial expressions. C-G-ECA-R first adds channel and spatial attention mechanism before the residual module to extract the key features of expressions, and then adds the G-ECA attention module inside the residual module. The motivation of G-ECA is the weighted fusion of global and local features to enhance the focus on key features of expressions, and it is embedded inside the residual module. CBAM has been added again after the residual module to assist in capturing key expression features. The entire network structure exhibits a strong attentional effect, which is to continuously increase the attention to the salient information in the facial features. Experimental results prove that the C-G-ECA-R improves the recognition rate of facial expressions. Due to the effect of the attention mechanism is decreasing slightly with the increase of the number of network layers, the effect of improving the attention mechanism is challenging in the deeper network. Further ongoing research

Table 11 Performance comparison of G-ECA, G-ECA-1, and G-ECA-2 (JAFFE)

Model name	Val_Acc (%)	Test_Acc(%)
+H_Attention + G-ECA + T_Attention (C-G-ECA-R34)	97.69	**97.65**
+H_Attention + G_ECA-1 + T_Attention (C-G-ECA-1-R34)	97.65	97.63
+H_Attention + G_ECA-2 + T_Attention (C-G-ECA-2-R34)	98.43	96.09

will focus on improving the stability of G-ECA in-depth C-G-ECA-R and other network structures to achieve a better recognition effect on facial expression.

Funding This work is supported by National Natural Science Foundation of china under grant number (No. 62177037) and Education Department of Shaanxi Provincial Government Service Local Special Scientific Research Plan Project under grant number (No. 22JC037).

Data Availability All data can be obtained by contacting the first author.

Declarations

Conflict of Interests The authors declare no conflict of interest.

References

1. Arriaga O, Valdenegro-Toro M, Plöger P (2017) Real-time convolutional neural networks for emotion and gender classification. arXiv:1710.07557
2. Avani V, Shaila S, Vadivel A (2020) Interval graph of facial regions with common intersection salient points for identifying and classifying facial expression. Multimed Tools Appl 80(3):3367–3390
3. Ayeche F, Alti A (2021) HDG And HDGG: an extensible feature extraction descriptor for effective face and facial expressions recognition. Pattern Anal Applic, 1–16
4. Bystroff C, Thorsson V, Baker D (2000) HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. J Mol Biol 301(1):173–190
5. Cao S, Yao Y, An G (2020) E2-capsule neural networks for facial expression recognition using AU-aware attention. IET Image Process 14(11):2417–2424
6. Chen D, Song P (2021) Dual-graph regularized discriminative transfer sparse coding for facial expression recognition. Digital Signal Process, 108
7. Gera D, Balasubramanian S (2021) Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. Pattern Recogn Lett 145:58–66
8. He J, Yu X, Sun B, Yu L (2021) Facial expression and action unit recognition augmented by their dependencies on graph convolutional networks. J Multimodal User Interfaces, 1–12
9. Khaliluzzaman M, Pervin S, Islam M, Hassan M (2019) Automatic facial expression recognition using shallow convolutional neural network. In: IEEE International conference on robotics, automation, artificial-intelligence and internet-of-things (RAAICON). IEEE, pp 98-103
10. Li K, Jin Y, Akram MW, Han R, Chen J (2019) Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis Comput 36(2):391–404
11. Li J, Jin K, Zhou D, Kubota N, Ju Z (2020) Attention mechanism-based CNN for facial expression recognition. Neurocomputing 411:340–350
12. Liu X, Li C, Dai C, Chao H (2021) Nonnegative tensor factorization based on low-rank subspace for facial expression recognition. Mobile Networks and Applications, 1–12
13. Lopes A, Aguiar E, Souza A, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn 61:610–628
14. Ma X, Guo J, Tang S, Qiao Z, Fu S (2021) DCANet: learning connected attentions for convolutional neural networks. In: 2021 IEEE International conference on multimedia and expo (ICME), pp 1-6
15. Mena-Chalco J, Carrer H, Zana Y, Jr R (2008) Identification of protein coding regions using the modified Gabor-Wavelet transform. IEEE/ACM Trans Comput Biol Bioinformatics 5(2):198–207
16. Milborrow S, Nicolls F (2008) Locating facial features with an extended active shape model. In: European conference on computer vision, pp 504-513
17. Mnih V, Heess N, Graves A (2014) Recurrent models of visual attention. Adv Neural Inf Process Syst, 2204–2212
18. Mohan K, Seal A, Krejcar YazidiA (2021) Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. IEEE Trans Instrum Meas 70:1–12
19. Rao T, Li J, Wang X, Sun Y, Chen H (2021) Facial expression recognition with multiscale graph convolutional networks. IEEE MultiMedia 28(2):11–19
20. Roy AG, Nav Ab N, Wachinger C (2018) Concurrent spatial and channel squeeze & excitation in fully convolutional networks. Springer, Cham, pp 421–429

21. Rubel A, Chowdhury A, Kabir M (2019) Facial expression recognition using adaptive robust local complete pattern. In: IEEE International conference on image processing (ICIP), pp 41-45
22. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. Advances in neural information processing systems, 30
23. Sadeghi H, Raie A (2019) Human vision inspired feature extraction for facial expression recognition. Multimed Tools Appl 78(21):30335–30353
24. Smolyanskiy N, Huitema C, Liang L, Anderson S (2014) Real-time 3D face tracking based on active appearance model constrained by depth data. Image Vis Comput 32(11):860–869
25. Suykens J, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300
26. Vinay A, Gupta A, Bharadwaj A, Srinivasan A, Murthy K, Natarajan S (2018) Unconstrained face recognition using bayesian classification. Procedia Comput Sci 143:519–527
27. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q (2020) ECA-Net: Efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 11531-11539
28. Wang Z, Zeng F, Liu S, Zeng B (2021) OAENEt: Oriented attention ensemble for accurate facial expression recognition[J]. Pattern Recogn 112:107694
29. Woo S, Park J, Lee J, Kweon I (2018) CBAM: convolutional block attention module. european conference on computer vision. In: Proceedings of the European conference on computer vision (ECCV), pp 3-19
30. Wu F, Pang C, Zhang B (2021) FaceCaps for facial expression recognition. In: 25th International conference on pattern recognition (ICPR2020)
31. Xie S, Hu H (2019) Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Trans Multimedia 21(1):211–220
32. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29:915–928
33. Zhu X, Ye S, Zhao L, Dai Z (2021) Hybrid attention cascade network for facial expression recognition. Sensors 21(6):2003
34. Zou W, Zhang D, Lee D (2021) A new multi-feature fusion based convolutional neural network for facial expression recognition. Appl Intell, 1–12

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.