J Inf Process Syst, Vol.20, No.4, pp.535~549, August 2024
https://doi.org/10.3745/JIPS.01.0107

ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Framework for Facial Expression Recognition Combining Contextual Information and Attention Mechanism

Jianzeng Chen* and Ningning Chen

Abstract

Facial expressions (FEs) serve as fundamental components for human emotion assessment and human–computer interaction. Traditional convolutional neural networks tend to overlook valuable information during the FE feature extraction, resulting in suboptimal recognition rates. To address this problem, we propose a deep learning framework that incorporates hierarchical feature fusion, contextual data, and an attention mechanism for precise FE recognition. In our approach, we leveraged an enhanced VGGNet16 as the backbone network and introduced an improved group convolutional channel attention (GCCA) module in each block to emphasize the crucial expression features. A partial decoder was added at the end of the backbone network to facilitate the fusion of multilevel features for a comprehensive feature map. A reverse attention mechanism guides the model to refine details layer-by-layer while introducing contextual information and extracting richer expression features. To enhance feature distinguishability, we employed islanding loss in combination with softmax loss, creating a joint loss function. Using two open datasets, our experimental results demonstrated the effectiveness of our framework. Our framework achieved an average accuracy rate of 74.08% on the FER2013 dataset and 98.66% on the CK+ dataset, outperforming advanced methods in both recognition accuracy and stability.

Keywords

Contextual Information, Convolutional Channel Attention, Deep Learning, Facial Expression Recognition, Feature Fusion, Reverse Attention

1. Introduction

Facial expressions (FEs) provide a nonverbal channel for conveying genuine emotions and intentions, serving as a vital means of exchanging emotional information and facilitating interpersonal dynamics [1]. In the field of computer vision, the analysis of human FEs enables the comprehension of human emotions and their integration into a wide array of human–computer interaction systems, spanning service robots, fatigue detection for drivers, and medical services [2]. In social dynamics, complex facial movements and expressions have evolved to convey inner emotions. However, academic circles predominantly delve into the six fundamental emotional categories, as proposed by psychologists Ekman and Friesen, which encompass happiness, anger, sadness, surprise, disgust, and fear [3]. According to [4], in day-to-day human interaction, as much as 55% of the conveyed information is transmitted via FEs. This highlights

Manuscript received May 26, 2023; first revision November 6, 2023; accepted December 9, 2023.

* **Corresponding Author:** Jianzeng Chen (2003083@nut.edu.cn)
School of Computer and Information Engineering, Nanchang Institute of Technology, Nanchang, China (2003083@nut.edu.cn, 2006085@nut.edu.cn)

the considerable research value and significance of facial expression recognition (FER).

Traditional FER methods predominantly rely on shallow learning and manually engineered features. These techniques include principal component analysis (PCA) [5], local binary patterns (LBP) [6], Gabor transformation [7], geometric feature-based extraction [8], and hybrid feature extraction [9]. Nonetheless, these FER approaches are constrained by their reliance on prior knowledge, limited generalization capabilities, and inability to meet the precision and efficiency demands of real-world applications.

With rapid advancements in deep learning (DL) technology, studies on FER using deep neural network models have made significant progress. Convolutional neural networks (CNNs) have gained popularity for image recognition and classification [10]. Researchers have undertaken multifaceted efforts to enhance the accuracy of CNNs for expression recognition. Mollahosseini et al. [11] constructed a 7-layer CNN, initially pretrained on an extensive face dataset, followed by fine tuning with a FE dataset. Their innovative use of the inception layer architecture across multiple datasets for FER yielded superior results compared with traditional methods. However, the limited data in the FE datasets led to overfitting.

Ding et al. [12] introduced the FaceNet2ExpNet approach, employing deep features from the facial network to supervise training of the convolutional layer. Subsequently, they added a randomly initialized fully connected layer and initiated the training from scratch. Ng et al. [13] adapted a pretrained CNN model from the ImageNet dataset, adjusted it using the FER2013 dataset, and fine-tuned the modified model using the EmotiW dataset. They assessed the performance of the model for FER in real-world scenarios; however, the recognition accuracy was not optimal. Verma et al. [14] employed diverse subnetworks to extract rich features and efficiently combined them using an appropriate ensemble technique. This approach comprehensively considered changes in facial features due to significant facial movements and performed well on the CK+ dataset. Liu et al. [15] adopted a strategy involving three parallel multichannel CNNs to learn the global and local features from distinct facial regions. They implemented a joint embedding feature learning strategy to explore identity-invariant and pose-invariant expression representations based on fused regions in the embedding space. However, this method does not achieve precise human facial recognition accuracy in unconstrained environments.

In recent years, researchers have introduced attention mechanism into CNN [16]. By learning and adaptively assigning different weight coefficients to different regions on the feature maps (FMs), the network is capable of obtaining more expressive features, which enhances the efficiency and accuracy of FER. Hu et al. [17] presented a SENet network to obtain the channel dependencies of the features, which significantly improved the performance of the CNN model. Woo et al. [18] introduced the convolutional-block-attention-module (CBAM) concept, in which feature attentional operations were simultaneously performed in the spatial and channel dimensions, and good recognition results were obtained.

In recent years, the rapid development of image dehazing technology has a profound impact on various computer vision domains, including FER [19]. In FER, image quality significantly affects the algorithm performance. In particular, in practical applications such as security monitoring, facial recognition, and emotion analysis, image quality can be compromised by atmospheric conditions and adverse weather, making it challenging to accurately capture FEs [20].

Recent studies have emphasized the importance of image dehazing technology in FER. By applying state-of-the-art image dehazing algorithms, researchers can enhance the image clarity and visibility, improving the accuracy of expression recognition algorithms. This is particularly crucial for capturing expressions in low-light conditions or for conducting real-time facial analyses in outdoor environments. Furthermore, image dehazing technology can assist in reducing noise and enhancing image quality, thereby facilitating the precise capture of facial features and emotions [21]. Consequently, the application of image dehazing technology in FER has become a topic of significant interest, offering new opportunities to enhance the practicality and performance of FER systems. This trend will further drive interdisciplinary research on image dehazing and FER, aiming for a clearer and more accurate FER.

This study delves into the integration of contextual information and multiple attention mechanisms within the VGGNet16 network. In our proposed FER framework, the enhanced VGGNet serves as a backbone network for feature extraction. We introduce a multiscaled feature merging strategy to combine FMs from different levels, thereby enhancing the utilization of lower-level features and achieving precise recognition performance. The main contributions of our study are as follows:

- We employed an improved VGGNet16 as the backbone network for feature extraction. In each backbone block, we implemented an enhanced group convolutional channel attention (GCCA) module to steer the network's focus toward critical areas while suppressing irrelevant ones.
- Five backbone blocks were used to extract multiple features of varying sizes in different layers. The lower-level blocks capture high-resolution edge features with limited semantics. A partial decoder (PD) was introduced at the end of the backbone to aggregate all of the high-level block features and generate a global map. This map guides progressive learning through reverse attention (RA) modules, enabling the network to learn more nuanced expression details.

The remainder of this paper is organized as follows. Section 2 describes the proposed FER framework. Section 3 presents the performance evaluation and comparison of the results obtained for two public datasets. Finally, Section 4 provides a comprehensive summary of the study, highlighting the limitations of the proposed approach and outlining future research directions.

2. Proposed FER Framework

2.1 Network Structure

Fig. 1 illustrates the overall framework of this study. The FE images first undergo processing in the initial two low-level blocks to extract high-resolution, low-level features with limited semantics. To bolster the extraction of boundary features, these features pass through a convolutional layer with a single kernel, thereby enhancing the edge mapping accuracy. The low-level feature C_2 is subsequently directed to the last three high-level blocks within the backbone network. To extract the partial features, a PD module is incorporated at the end of the backbone. The PD module consolidates all of the high-level features from these blocks, producing a comprehensive global map labeled as P_a. Furthermore, an RA module is placed after each high-level block. Each high-level block generates features at different scales and sequentially combines them with features from various blocks originating from the preliminary global map. These outputs serve as inputs to the RA module, enabling the extraction of finer expression details. Finally, multiple features are fused to generate the ultimate FER outcome.

2.2 Backbone Network

In our proposed FER framework, we opted for VGGNet16 as the backbone for feature extraction with certain enhancements to enhance the stability of the FER tasks. A visual representation of the modified backbone is shown in Fig. 2. To streamline the model and safeguard against the loss of fine-grained FE details in high-level features, we omitted the final pooling layer and the fully connected layer in the VGGNet16. In addition, we integrated a GCCA module into each block of the backbone network. This module encourages the network to concentrate on target areas, thereby enhancing the overall model performance.

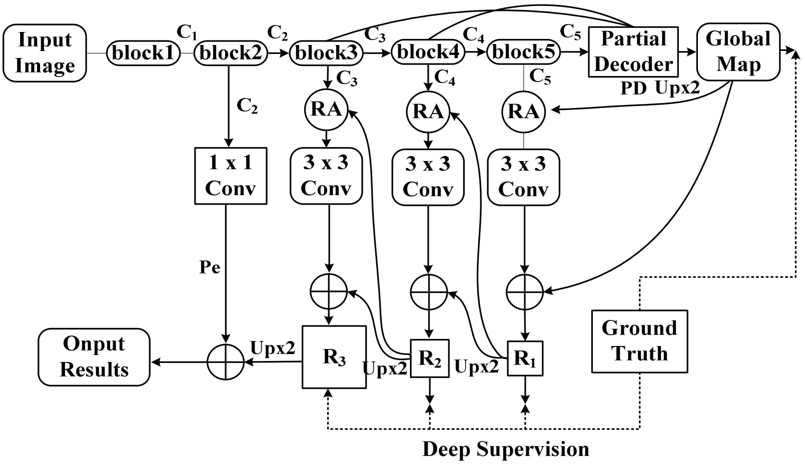


Fig. 1. Structure of the proposed framework. The improved VGGNet16 was used as the backbone network, and feature fusion was employed to extract multiple features (shallow features from low-level blocks in the backbone network and the aggregation of deep internal details from high-level blocks using a PD module). Through the RA mechanism, the currently predicted area was erased from the high-level side to the output features. This guides the entire network to progressively explore the supplementary fine details from top to bottom.

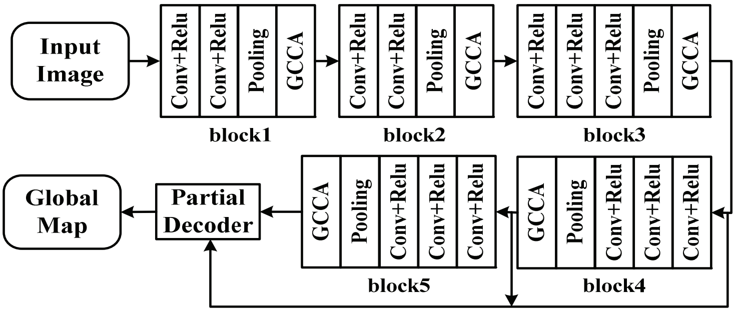


Fig. 2. Structure of the backbone network. To steer the network's focus toward target regions, CBAM was introduced in every block of the backbone network. To empower the network with the capability to learn multiscale features, lateral output components were added after the last four blocks of the backbone network to provide feature information of varying scales.

To equip the network with multiscale feature learning capabilities, we introduced a lateral output component after the last four blocks of the backbone network, offering features of varying scales. As depicted in Fig. 2, the lateral output branch from the low-level block 2 was employed to extract low-level features with limited semantics. Simultaneously, the lateral output branches of high-level blocks 3, 4, and 5 were dedicated to extract high-level features with strong semantic content. The ultimate result was derived by fusing the high-level and low-level features.

2.3 Backbone Network

Within a CNN, the distinct feature channels exhibit varying responses. If each channel is assigned an equal weight, the significance of the individual channels in feature extraction remains inadequately addressed. To maximize the utility of each feature channel, we introduced a channel attention module (i.e., the GCCA module) based on the group convolution concept [22]. The structure of this module is illustrated in Fig. 3.

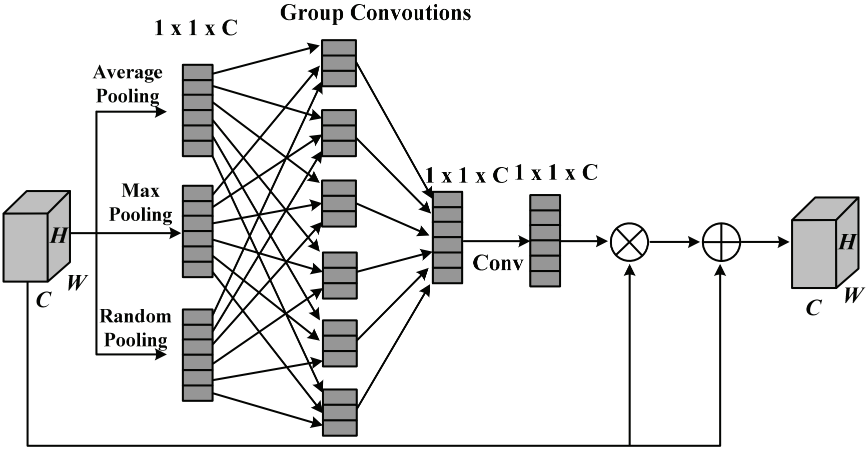


Fig. 3. GCCA module. While the SENet exclusively employs maximum pooling, the proposed module extends it by incorporating average pooling and random pooling. This triad of pooling methods collectively yields a more comprehensive extraction of global features from various channels. Following these three distinct pooling processes, the FM was transformed into three channel descriptors, aligning with the dimensionality of the input FM.

Let input FM be $F \in R^{H\times W\times C}$, which is compressed into three channel descriptors, F_{avg}^{C}, F_{max}^{C}, and F_{rdm}^{C}, with dimensions of 1×1×C after the aforementioned pooling processes. To further learn the correlation between the channels, a group convolution operation was introduced. First, the three channel descriptors were grouped according to different channels, and the global information of the same channel was spliced together to form a new feature vector. Each new feature vector contained three types of global information. Following this, the feature vectors were convoluted by convolutional layers containing 1×1 convolution kernels such that the three types of global information were adaptively fused together, resulting in a channel descriptor with dimensions of 1×1×C. The channel descriptor was subsequently sent to two convolutional layers containing a 1×1 convolution kernel for feature learning. The number of channels in the previous convolutional layer was C/16. The number of channels in the following convolutional layer was C, to learn the weight coefficients of different channels.

$$P(F) = \sigma(W_2(\delta(W_1(\delta(G(F_{avg}; F_{max}; F_{rdm})))))). \tag{1}$$

here, δ and σ are the ReLU and sigmoid functions, G is the group convolution operation, and W_1 and W_2 are the convolution parameters of the first and second convolutional layers, respectively. F_{avg}, F_{max}, and F_{rdm} represent average pooling, max pooling, and random pooling operations, respectively.

The sigmoid function limits the value of each element within the interval [0,1]. If it is directly multiplied by the input FM, the output response of the FM is weakened. Hence, the input FM was weighted using the attention weight coefficients through a point multiplication operation so that the effective features were strengthened and redundant features were restrained. To prevent the weakening of the output response, the input FM was added to the weighted attention FM to strengthen the stability of the model. The ultimate output of the GCCA module is expressed as:

$$F' = M_c(F) \otimes F + F \tag{2}$$

where F' is the output FM by the GCCA module and $M_c(F)$ is the attention weight coefficient.

2.4 Partial Decoder

In the CNN model, high-level features convey semantics, whereas low-level features depict spatial details that are beneficial for refining object boundaries. In contrast to high-level features, low-level features have a relatively smaller impact on the overall performance, and due to their substantial spatial resolution, can result in substantial computational overhead. Hence, a PD module [23] was introduced at the end of the backbone, as illustrated in Fig. 4.

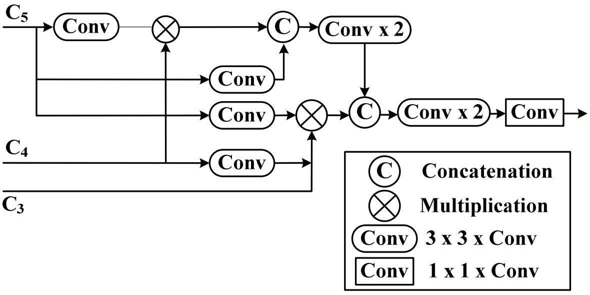


Fig. 4. Structure of the PD module. The PD only incorporates high-level features, discarding the larger resolution features from the low-level layers, which facilitates rapid and precise extraction of target region features.

The PD module exclusively integrates high-level features while discarding lower-resolution shallow features, ensuring swift and accurate extraction of FE features. The process is described as follows. Initially, three sets of high-confidence, high-level features $\{C_i, i = 3,4,5\}$ are extracted from the high-level blocks of the backbone network. Subsequently, these high-level features from the three sets are amalgamated using the partial decoder $pd(\cdot)$. This fusion of features from different levels encourages information from distinct layers to complement one another, culminating in the creation of a preliminary global map $P_d = pd(C_3, C_4, C_5)$. This map serves as a guide for the subsequent progressive learning based on the reverse attention strategy.

2.5 Reverse Attention Module

To meticulously capture detailed expression information from crucial regions, we introduced an RA module within high-level blocks to progressively expand the target area. Commencing from the initial global map generated by the PD module, multiple features of distinct sizes extracted by blocks 5, 4, and 3 serve as inputs and are transmitted to the RA module.

The RA module methodically erases the presently predicted region from the high-level lateral output features, sequentially unveiling the missing details and nuanced features of the essential expression regions that require supplementation from top to bottom. In this approach, the present prediction result is obtained by upscaling information from the deeper network layers. This incremental erasure concept [24] refines the initial rough prediction into a comprehensive and precise prediction outcome.

The reverse attention feature output results from the element-wise multiplication of the high-level output features $\{C_i, i = 3,4,5\}$. The reverse attention weight A_i is expressed mathematically as:

$$R_i = D(C_i \cdot A_i), \tag{3}$$

where $D(\cdot)$ denotes the dot multiplication operation. The RA weight A_i can be obtained by simply subtracting the upsampling prediction of the $(i+1)$-th lateral output from 1, as follows:

$$A_i = 1 - \operatorname{sigmoid}\big(U(C_{i+1})\big), \tag{4}$$

where $U(\cdot)$ denotes the upsampling operation.

2.6 Reverse Attention Module

The loss function primarily quantifies the disparities between the predicted and actual values, and the network training aims to minimize these loss functions. Using the softmax loss function, the neural network output values were mapped within the (0,1) interval, providing probabilities for various classifications. These probabilities were then compared to achieve multi-classification. Although the softmax loss function effectively optimizes interclass spacing, it can misjudge samples of the same FE when there are substantial discrepancies.

To address this issue, we introduced an islanding loss function [25]. By implementing these two functions, the objectives of increasing interclass distances and decreasing intraclass distances were realized. The islanding loss function is an enhancement built upon the center loss. Initially, a cosine distance is computed and 1 is added to extend the range to (0,2), thereby enlarging the distance between different classes. The islanding loss function is mathematically expressed as:

$$L_I = L_C + \lambda_1 L_{I-cos} \tag{5}$$

where L_C represents the center loss function (which is used to optimize the intraclass distance), L_{I-cos} represents the cosine distance of the cluster center, and λ_1 is a hyperparameter indicating the weight ratio in the islanding loss function. L_{I-cos} can be calculated from the following equation:

$$L_{I\text{-cos}} = \sum_{c_j \in N} \sum_{c_k \in N} \left(\frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2} + 1 \right), \tag{6}$$

where *N* represents the set of sample labels, c_k and c_j represent the cluster centers of the *k*-th and *j*-th classes of expressions, respectively, and $\|c_k\|_2$ and $\|c_j\|_2$ represent the Euclidean distances from the cluster center to the origin of the coordinates.

In our proposed framework, the training of the network was optimized by considering the softmax loss and islanding loss functions. Therefore, features of the same class were close to one another, and the distances between dissimilar classes of facial features were increased to achieve better recognition results. The joint loss function can be determined using the following equation:

$$L = L_s + \lambda L_I, \tag{7}$$

where λ denotes the weight ratio of the islanding loss in the joint loss function. Based on the model performance results with different parameter values in the experiments, λ and λ_1 were fixed at 0.005 and 7, respectively

3. Experiment and Analysis

3.1 Experimental Dataset and Data Augmentation

In this experiment, two widely recognized public FER datasets were employed: the FER2013 and CK+ expression datasets.

The CK+ expression dataset [26] encompasses 593 sequences of expression images, spanning from natural to peak expressions, featuring 123 individuals. Of these sequences, 327 have additional expression tags. The dataset encompasses eight fundamental FE classes: anger, contempt, happiness, sadness, surprise, disgust, fear, and neutral. All images portray clear and positive FEs, with annotations meticulously validated by psychologists. For fairness in performance evaluation and comparison, the dataset omits contempt expressions due to their notably small sample size.

The FER2013 dataset [27] comprises 35,887 facial images accompanied by expression labels. This dataset encapsulates seven expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The data collection was conducted in an uncontrolled environment, making it challenging to obtain precise recognition results.

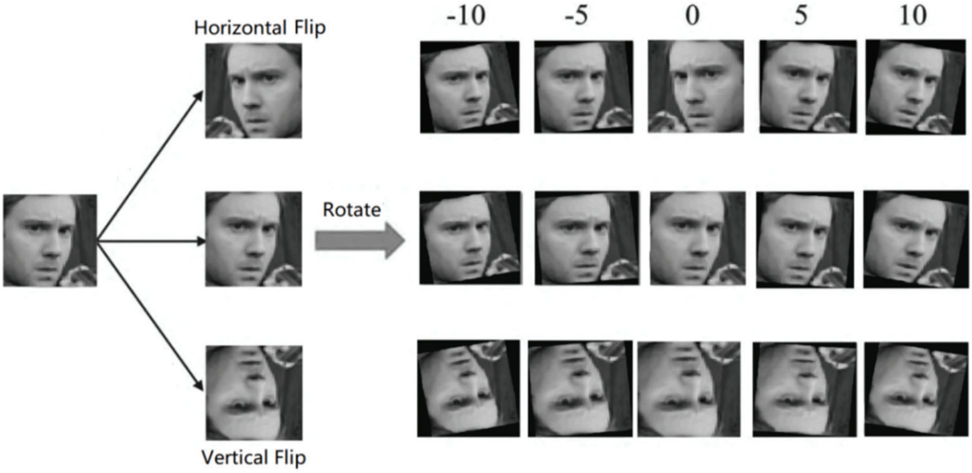


Fig. 5. Data augmentation. The multi-task cascaded convolutional networks (MTC-NN) model was used for facial detection and cropping to obtain nearly background-free facial images. Subsequently, the acquired facial images were scaled and normalized using bilinear interpolation, resulting in saved images with dimensions of 224 × 224 pixels. Subsequently, data augmentation was applied to the used samples, expanding the sample quantity.

Data augmentation was performed on the image samples to enhance the resilience of the model to interference, resulting in an expanded sample pool. Each image underwent flipping and rotation, with a rotation angle of ±10° at 5° intervals. Hence, the number of image samples increased to 15 times the number of image samples of the original dataset. The efficacy of data augmentation is depicted in Fig. 5.

3.2 Performance Evaluation Metrics

Two performance indicators (average accuracy Acc and stability Sta) were used [28] in this experiment. Each method was tested N ($N = 10$) times, and the final average accuracy was determined as follows:

$$Acc = \frac{\sum_{i=1}^{N} Acc_i}{N}, \tag{8}$$

where Acc_i denotes the recognition accuracy of the i-th experiment. Owing to the random initialization of network parameters and random batching of training samples, there is a certain error in each recognition result under the same settings. Therefore, it is fairer and more reliable to use the average value of multiple experiments.

Stability is the mean square error of N experimental results, which is defined as:

$$Sta = \sqrt{\frac{\sum_{i=1}^{N} (Acc_i - Acc)^2}{N}}, \tag{9}$$

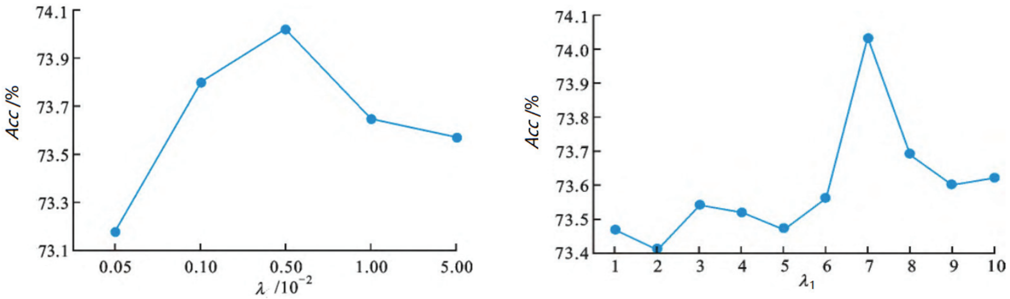
where Sta represents the degree of variation in the experimental results under the same settings.

3.3 Network Training and Parameter Setting

The following hardware were used for the experiment: i5-10400F 2.9 GHz CPU, NVIDIA GTX 1080Ti 8 G GPU, and 16 GB RAM. The following software were used: Ubuntu 18.04.3 64-bit operating system, MATLAB version 2019a, PyTorch GPU 1.4.0 (to establish the training environment), and Python version 3.6.2.

During the experiment, as the training rounds progressed, the learning rate decayed to half of its original value when the loss rate ceased to decrease within three iterations. To mitigate model overfitting, dropout was integrated into the model to randomly deactivate neurons at a set random dropout rate of 0.001. During the model training process, after each iteration of the training set, a validation set was tested, and the loss and accuracy were recorded.

The joint loss function incorporates two weight parameters (λ and λ_1), which require adjustment. Theoretically, the grid technique should be used to compute the optimal weights. It is more appropriate to reduce the constraint weight for intraclass variations within a range of 0.001–0.01. Hence, in the experiment, λ was initially fixed at 0.01, while λ_1 was varied from 1 to 10. The recognition accuracy for the FER2013 dataset is shown in Fig. 6(b). Although the recognition accuracy was not strictly a convex curve, overall, the recognition accuracy was significantly higher when $\lambda_1 = 7$ compared with those for other cases. Consequently, λ_1 was set to 7, while λ was varied at 0.0005, 0.001, 0.005, 0.01, and 0.05. The recognition accuracy for the FER2013 dataset is shown in Fig. 6(a). Fig. 6 shows that the model achieved the highest recognition rate when λ was set at 0.005. Hence, we selected the values 0.005 and 7 for λ and λ_1, respectively.



(a) (b)

Fig. 6. Performance of the model with different weight values: (a) $\lambda_1 = 7$ and (b) $\lambda = 0.01$.

3.4 Ablation Analysis

To assess the influence of various components in the framework on the FER accuracy and stability, we conducted multiple sets of ablation experiments. The aim of these experiments was to evaluate the effectiveness and dependability of different modification strategies within the proposed FER framework. Tests were performed on the FER2013 and CK+ datasets, with consistent experimental parameters across the various test groups. The outcomes are summarized in Table 1.

Table 1. Comparison of the experimental results for various improvement strategies

Model	Multi-feature fusion	GCCA	PD	RA	Joint loss function	FER2013		CK+	
						Acc (%)	*Sta*	*Acc* (%)	*Sta*
1	×	×	×	×	×	69.79	0.38	92.14	0.31
2	√	×	×	×	×	71.53	0.35	93.01	0.27
3	√	√	×	×	×	73.85	0.34	96.72	0.24
4	√	√	√	×	×	73.84	0.26	96.72	0.16
5	√	√	√	√	×	73.97	0.25	97.94	0.13
6	√	√	√	√	√	74.08	0.24	98.66	0.11

"×" denotes the components that were not used, and "√" indicates the components that were incorporated.

Compared with the original VGG16 model (model 1), the different improvement strategies adopted in this study improved the model recognition accuracy and stability to a certain extent. It can be observed from the results that after adding the GCCA module, the effective features were enhanced, the redundant features were suppressed, and the model recognition accuracy was significantly improved. The multi-feature fusion strategy combines the edge features obtained in the low-level blocks as constraints with the high-level semantic features obtained in the high-level blocks, thereby enabling the overall framework to learn more targeted areas and improve capturing of detailed information of subtle facial changes. The effect of the PD module on improving the recognition accuracy was not obvious; however, the module effectively reduced the computational complexity of the overall framework and significantly improved the stability of the model. The subsequent RA module supplemented the missing information and detailed features of the key expression regions to ensure better accuracy while increasing the processing speed. Finally, after applying multiple strategies, we attained an accuracy of 74.08% and 98.66% on the CK+ and FER2013 datasets, respectively.

3.5 Performance Evaluation and Comparison

To further analyze the accuracy for various expression classes, we obtained the confusion matrices of FER results from different sources. The confusion matrices are presented in Figs. 7 and 8.

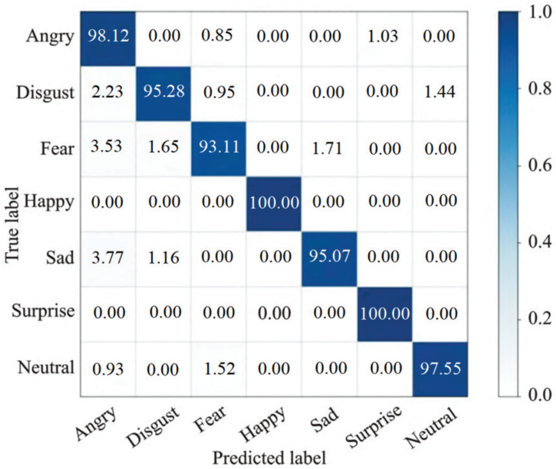


Fig. 7. CK+ dataset confusion matrix: *Acc* (%).

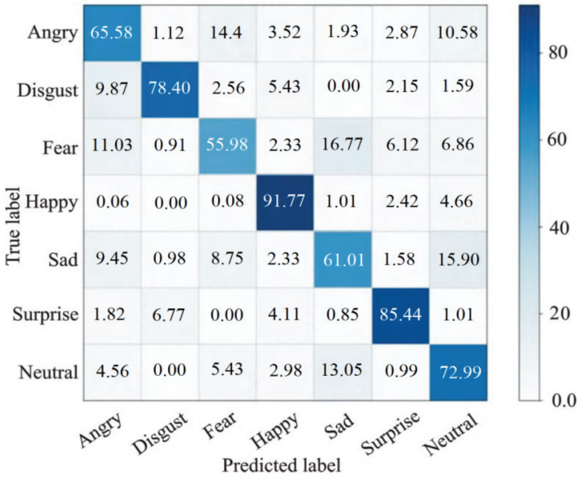


Fig. 8. FER2013 dataset confusion matrix: *Acc* (%).

For the FER2013 dataset, the proposed model had a high recognition accuracy for happy and surprise expressions. However, the model could not accurately identify sad and fear expressions, with a recognition accuracy of only 55.98% and 61.01%, respectively. The distribution of dissimilar classes of samples in the FER2013 dataset was extremely unbalanced. The number of sad images in the training set was only approximately 500, whereas the number of happy images was more than 7,000. Moreover, two classes of expressions (sad and disgust) had similar changes in the mouth or eye areas of the human face, and the distinguishability of the expression features was relatively low, making them prone to recognition errors.

However, for the CK+ dataset, the accuracy for various classes of expressions was significantly higher than that for the FER2013 dataset, and the fluctuation in the recognition accuracy for different classes of expressions was very small. This is because the CK+ dataset is obtained under controlled laboratory conditions, and the facial images are clear without occlusion and environmental factors. In addition, fear, disgust, and sad expressions have a certain degree of similarity, which increases the difficulty of distinguishing between these three classes.

Comparison of the average accuracy and stability of different models tested on the FER2013 and CK+ datasets is presented in Table 2. Our findings indicate that our model achieved the highest accuracy in expression recognition and displayed the best stability across both experimental datasets.

Barman and Dutta [8] and Agarwal and Mukherjee [9] employed traditional FER methods based on manual features. In [8], the authors used an active shape model to extract facial contours and region positions, facilitating the extraction of salient FE features. However, this approach tends to lose key recognition and classification information, leading to relatively poor FER accuracy on both datasets.

In [9], complex non-rigid motion facial components were captured by extracting scale-independent features and tracking pixel motion. Unfortunately, the generalization of this method is rather limited, particularly the performance of the model on the FER2013 dataset, which proved to be unsatisfactory.

Among the deep learning methods, Mollahosseini et al. [11] adopted a fine-tuning strategy after pre-training to achieve better recognition results compared with traditional methods. Nonetheless, network overfitting was a concern, and the feature attention mechanism was not considered. Verma et al. [14] processed image sequences through a visual branch network, introducing jump connections from a low level to a high level to consider underlying features. This significantly improved the model performance, but the method did not account for contextual information or the influence of highly similar expression classes on the recognition accuracy.

Liu et al. [15] proposed a parallel multi-channeled convolutional network to learn effective feature representation through the integration of global and local features, achieving good accuracy and robustness. However, the FER accuracy on the unconstrained environment dataset FER2013 still requires improvement, indicating limited generalizability.

Table 2. Performance comparison of different methods

Method	**FER2013 dataset**		**CK+ dataset**	
	***Acc* (%)**	***Sta***	***Acc* (%)**	***Sta***
Barman and Dutta [8]	48.75	0.31	92.15	0.28
Agarwal and Mukherjee [9]	51.14	0.38	93.47	0.29
Mollahosseini et al. [11]	58.72	0.41	95.36	0.35
Verma et al. [14]	69.44	0.37	96.17	0.34
Liu et al. [15]	72.01	0.35	96.49	0.29
Proposed method	74.08	0.24	98.66	0.11

Our approach achieved both accuracy and stability for both experimental datasets. This success can be attributed to our framework, which uses the improved VGGNet16 as the backbone and incorporates the GCCA module to capture crucial information in the deeper network layers. We fused multiple features, extracted shallow features from the low-level blocks of the backbone, and aggregated the deep details from the high-level blocks using the PD module. Furthermore, the RA mechanism guides the entire network sequentially from top to bottom, allowing the mining of detailed information that requires supplementation. This approach makes full use of contextual information, leading to improved FER accuracy and model stability.

4. Conclusion

In this study, we developed a novel FER framework to address the limitations of traditional FER algorithms, which tend to overlook important features as the network depth increases during feature extraction. This framework is based on multiscale feature fusion, and incorporates an attention mechanism that considers contextual information. The improved VGGNet was employed for FE feature extraction, complemented by a multiscale FM fusion strategy that introduced contextual information, thereby enhancing the recognition accuracy. In addition, we introduced an attention mechanism that improved the channel attention module based on group convolution to extract more expressive FER features. Our results confirmed the high accuracy of our model for FER tasks across various scenarios. However, although the FER2013 dataset can represent uncontrolled non-laboratory environments, it is primarily sourced from the Internet, potentially resulting in limited diversity in image quality and environmental conditions. This implies that the dataset may not comprehensively represent all of the possible real-world scenarios. The experimental dataset includes only seven major FEs, whereas real-world expressions are much more diverse, encompassing a richer array of emotions and emotional expressions. In future research, we plan to further optimize the network structure and explore datasets that closely mimic real-world conditions, thereby enhancing the practical applications of our research.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work was supported by Ministry of Education Industry-University Cooperative Education Fund Project (No. 22087166301756).

References

[1] N. Samadiani, G. Huang, B. Cai, W. Luo, C. H. Chi, Y. Xiang, and J. He, “A review on automatic facial expression recognition systems assisted by multimodal sensor data,” *Sensors*, vol. 19, no. 8, article no. 1863, 2019. https://doi.org/10.3390/s19081863

[2] S. Li and W. Deng, “Deep facial expression recognition: a survey,” *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195-1215, 2022. https://doi.org/10.1109/TAFFC.2020.2981446

[3] A. John, M. C. Abhishek, A. S. Ajayan, S. Sanoop, and V. R. Kumar, "Real-time facial emotion recognition system with improved preprocessing and feature extraction," in *Proceedings of 2020 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2020, pp. 1328-1333. https://doi.org/10.1109/ICSSIT48917.2020.9214207

[4] S. Saeed, A. A. Shah, M. K. Ehsan, M. R. Amirzada, A. Mahmood, and T. Mezgebo, "Automated facial expression recognition framework using deep learning," *Journal of Healthcare Engineering*, vol. 2022, article no. 5707930, 2022. https://doi.org/10.1155/2022/5707930

[5] M. S. Ejaz, M. R. Islam, M. Sifatullah, and A. Sarker, "Implementation of principal component analysis on masked and non-masked face recognition," in *Proceedings of 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019, pp. 1-5. https://doi.org/10.1109/ICASERT.2019.8934543

[6] A. S. Rubel, A. A. Chowdhury, and M. H. Kabir, "Facial expression recognition using adaptive robust local complete pattern," in *Proceedings of 2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019, pp. 41-45. https://doi.org/10.1109/ICIP.2019.8802911

[7] S. Ahmed, M. Frikha, T. D. H. Hussein, and J. Rahebi, "Optimum feature selection with particle swarm optimization to face recognition system using Gabor wavelet transform and deep learning," *BioMed Research International*, vol. 2021, article no. 6621540, 2021. https://doi.org/10.1155/2021/6621540

[8] A. Barman and P. Dutta, "Facial expression recognition using distance and shape signature features," *Pattern Recognition Letters*, vol. 145, pp. 254-261, 2021. https://doi.org/10.1016/j.patrec.2017.06.018

[9] S. Agarwal and D. P. Mukherjee, "Facial expression recognition through adaptive learning of local motion descriptor," *Multimedia Tools and Applications*, vol. 76, pp. 1073-1099, 2017. https://doi.org/10.1007/s11042-015-3103-6

[10] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689-694, 2020. https://doi.org/10.1016/j.procs.2020.07.101

[11] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proceedings of 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, 2016, pp. 1-10. https://doi.org/10.1109/WACV.2016.7477450

[12] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition," in *Proceedings of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, 2017, pp. 118-126. https://doi.org/10.1109/FG.2017.23

[13] H. W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, Seattle, WA, USA, 2015, pp. 443-449. https://doi.org/10.1145/2818346.2830593

[14] M. Verma, H. Kobori, Y. Nakashima, N. Takemura, and H. Nagahara, "Facial expression recognition with skip-connection to leverage low-level features," in *Proceedings of 2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019, pp. 51-55. https://doi.org/10.1109/ICIP.2019.8803396

[15] Y. Liu, W. Dai, F. Fang, Y. Chen, R. Huang, R. Wang, and B. Wan, "Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition," *Information Sciences*, vol. 578, pp. 195-213, 2021. https://doi.org/10.1016/j.ins.2021.07.034

[16] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48-62, 2021. https://doi.org/10.1016/j.neucom.2021.03.091

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132-7141.

[18] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module" in *Computer Vision - ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[19] S. Rathi and S. Bamal, “Deep learning and globally guided image filtering technique based image dehazing and enhancement,” *International Journal of Technical Research & Science,* vol. 2020(Special Issue), pp. 60-68, 2020. https://doi.org/10.30780/specialissue-ICACCG2020/043

[20] T. Jia, J. Li, L. Zhuo, and T. Yu, “Semi-supervised single-image dehazing network via disentangled meta-knowledge,” *IEEE Transactions on Multimedia*, vol. 26, pp. 2634-2647, 2023. https://doi.org/10.1109/TMM.2023.3301273

[21] X. Zhang, J. Li, and Z. Hua, “MFFE: multi-scale feature fusion enhanced net for image dehazing,” *Signal Processing: Image Communication*, vol. 105, article no. 116719, 2022. https://doi.org/10.1016/j.image.2022.116719

[22] T. Zhang, G. J. Qi, B. Xiao, and J. Wang, “Interleaved group convolutions,” in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 4373-4382. https://doi.org/10.1109/ICCV.2017.469

[23] Z. Wu, L. Su, and Q. Huang, “Cascaded partial decoder for fast and accurate salient object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 3907-3916. https://doi.org/10.1109/CVPR.2019.00403

[24] S. Chen, X. Tan, B. Wang, and X. Hu, “Reverse attention for salient object detection,” in *Computer Vision - ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 234-250. https://doi.org/10.1007/978-3-030-01240-3_15

[25] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, “Island loss for learning discriminative features in facial expression recognition,” in *Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, 2018, pp. 302-309. https://doi.org/10.1109/FG.2018.00051

[26] 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 *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, San Francisco, CA, USA, 2010, pp. 94-101. https://doi.org/10.1109/CVPRW.2010.5543262

[27] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, et al., “Challenges in representation learning: a report on three machine learning contests,” in *Neural Information Processing.* Heidelberg, Germany: Springer, 2013, pp. 117-124. https://doi.org/10.1007/978-3-642-42051-1_16

[28] P. Jiang, G. Liu, Q. Wang, and J. Wu, “Accurate and reliable facial expression recognition using advanced softmax loss with fixed weights,” *IEEE Signal Processing Letters*, vol. 27, pp. 725-729, 2020. https://doi.org/10.1109/LSP.2020.2989670

Jianzeng Chen **https://orcid.org/0009-0007-1098-3038**

He is a master's degree and a lecturer at Nanchang Institute of Technology. His research directions include database technology and virtual reality technology.

Ningning Chen **https://orcid.org/0009-0004-3391-7613**

He holds a bachelor's degree and is a lecturer at Nanchang Institute of Technology. His research areas include e-commerce and computer applications.