IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 October 2023, accepted 30 November 2023, date of publication 7 December 2023,
date of current version 21 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340510

RESEARCH ARTICLE

Emotion Recognition in Complex Classroom Scenes Based on Improved Convolutional Block Attention Module Algorithm

LI LI[1] AND DENGFENG YAO[1,2]
[1]Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China
[2]Laboratory of Computational Linguistics, School of Humanities, Tsinghua University, Beijing 100084, China

Corresponding author: Dengfeng Yao (tjtdengfeng@buu.edu.cn)

This work was supported in part by the National Social Science Foundation of China under Grant 21BYY106, in part by the National Natural Science Foundation of China under Grant 62036001, in part by the General Project of the National Language Committee under Grant YB145-25, in part by the Beijing Municipal Natural Science Foundation under Grant 4202028, in part by the Support Plan for Beijing Municipal University Faculty Construction—High-Level Scientific Research and Innovation Team Project under Grant BPHR20220121, in part by the Educational Science Research Project of Beijing Union University (Deepening and Reconstruction of Higher Education Teaching Evaluation in the Intelligent Age) under Grant JK202312, in part by the Jiangsu Province Key Research and Development Program (Industry Prospects and Key Core Technologies) under Grant BE2020047, and in part by the Characteristic-Disciplines Oriented Research Project in Beijing Union University under Grant KYDE40201702.

ABSTRACT This study provides a deep learning-based intelligent recognition technology for student facial expressions in the classroom. This technology realizes the recognition of students' facial expressions and provides a practical approach for classroom assessment and teacher improvement of teaching methods toward achieving smart education. An improved hybrid attention mechanism is designed for the student classroom. The facial expression recognition model addresses issues of instability in the recognition process of student facial expressions in the classroom, the high redundancy of parameters in traditional convolutional neural networks, and the long training time and slow convergence prone to overfitting. In the image modality data, this study proposes a hybrid attention mechanism in the deep neural network model before feature fusion to extract network features with stronger representational capability, enhance the prediction performance of deep neural networks, and improve the interpretability of the model. The improved hybrid attention mechanism is introduced into the deep neural network by modifying the convolutional block attention module with shortcut connections, deepening the network depth of the attention module and enabling it to learn the weight information among feature channels and spatial regions effectively. The proposed student facial expression recognition model achieves an accuracy of 88.71% on the publicly available RAF-DB dataset and an accuracy of 86.14% on the self-collected real classroom teaching video dataset. The proposed technology can be applied in the education field to evaluate student engagement automatically, provide personalized teaching guidance and learning analytics for teachers, and promote the development of intelligent education.

INDEX TERMS Attention mechanism, deep learning, DenseNet, facial expression recognition, classroom student expressions.

I. INTRODUCTION

Facial expressions play a vital role in human communication, conveying diverse emotional states and serving as crucial cues for social interaction. The recognition and understanding of facial expressions have garnered extensive attention in computer vision, finding applications in human-computer interaction, affective computing, and psychological research. Numerous research efforts have been dedicated to developing accurate and robust facial expression recognition algorithms.

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

Early methods for facial expression recognition relied on manually crafted features and traditional machine learning algorithms, such as Support Vector Machines (SVMs) and Hidden Markov Models. However, these approaches often struggled to capture the intricate and subtle variations present in facial expressions, limiting their performance in real-world settings.

With the advent of deep learning, facial expression recognition has witnessed remarkable advancements. Deep neural networks, particularly convolutional neural networks (CNNs), have demonstrated outstanding performance in various computer vision tasks, including facial expression recognition. CNNs excel at automatically learning discriminative features from raw data, obviating the need for manual feature engineering and achieving exceptional accuracy in facial expression analysis.

Researchers have explored various methods and introduced novel techniques to enhance the accuracy and robustness of facial expression recognition further. Attention mechanisms have emerged as a promising avenue for augmenting the discriminative power of CNN in facial expression recognition. Attention mechanisms enable networks to focus on salient facial regions or features, facilitating improved representation learning and capturing fine-grained details.

Several attention-based methods have recently been proposed to address the challenges in facial expression recognition [1], [2]. This study aims to enhance facial expression recognition by introducing a novel attention-based model incorporating spatial and channel attention mechanisms. Building upon prior research, our model introduces new modifications to improve the effectiveness of attention mechanisms in capturing distinctive facial expression features. Experimental evaluations on benchmark datasets validate our proposed approach's exceptional performance and robustness of our proposed approach.

The remainder of this paper is organized as follows: Section II comprehensively reviews prior work on facial expression recognition and attention mechanisms. Section III provides a detailed description of the proposed attention-based model. Section IV presents the experimental setup, results, and analysis. Finally, Section V summarizes the study and discusses future research directions in facial expression recognition.

II. RELATED WORKS

Facial expressions are the primary means through which humans convey emotional information, whereas facial expression recognition algorithms typically identify common facial emotions, such as surprise, fear, happiness, sadness, and anger [3], [4]. With the advancement of research, convolutional neural networks(CNNs) have gained increasing popularity in image processing. In 2012, Krizhevsky et al. proposed the AlexNet network model [5], which achieved great success in the ImageNet visual recognition challenge, sparking the wave of deep learning research. In the 2013 Facial Expression Recognition Challenge (FER2013), Tang [6] applied CNN for expression recognition, utilizing SVM for classification and improving upon the traditional cross-entropy loss function, resulting in a recognition rate of 71.2% on the FER2013 dataset. Furthermore, deep neural networks such as Visual Geometry Group(VGG) [7], Visual Geometry Group Face(VGG-Face) [8], and Google Inception Net(GoogLeNet) [9], have been widely utilized in image processing, demonstrating strong performance and becoming key models for feature extraction. In 2016, He et al. introduced the ResNet model, a residual neural network architecture [10], which addressed the issue of network degradation by incorporating identity connections within residual units, achieving significant advancements in image processing. In 2017, Chollet proposed the Xception model, a lightweight depthwise separable convolutional network, which exhibited faster convergence speed and superior performance in facial expression recognition [11]. Zhao et al. employed a network structure comprising deep belief networks and stacked autoencoders for face classification [12]. Zhang et al. [13] proposed an end-to-end network model based on generative adversarial networks (GAN) that enhances the accuracy of facial expression recognition accuracy by synthesizing facial images and applying discriminative networks. Minaee et al. [14] presented a deep learning method based on attention convolutional networks that allows selective attention to important facial regions. They also utilized visualization techniques to identify key facial areas based on the classifier's output for detecting various emotions.

Current research in facial expression recognition aims to achieve higher recognition rates by designing more complex networks. However, this has led to large model parameter sizes, redundant complexity, and lengthy training times. Furthermore, the obtained results often fail to meet real-time requirements, and facial expression recognition in daily life exhibits significant instability. Dachapally [15] proposed a Vanilla CNN model for expression recognition based on CNN and autoencoders, which somewhat improved the recognition accuracy. However, this network's many parameters increased computational complexity and hindered convergence. Fan et al. [16] introduced a novel multi-region ensemble framework, MRE-CNN, for facial expression recognition, which enhanced the learning capability of CNN by capturing multiple local facial information. This framework improved the recognition rate, but significantly increased the parameter computation of the model. Mollahosseini et al. [17] utilized an inception structure to design a deeper neural network for expression recognition. While the increased network depth improved performance, the substantial parameter computation also had a considerable impact. The Google development team also introduced MobileNet, an efficient lightweight network for mobile devices, which utilizes depthwise separable convolutions instead of standard convolutions to reduce parameter size and improve efficiency. It incorporates width and resolution factors as parameters and has performed well in image-processing tasks [16], [18].

However, its recognition performance is not ideal when applied to expression recognition tasks.In exploring various approaches to facial expression recognition, the study by Barra et al. is of significant importance. Their proposed 'mesh model' demonstrates its efficiency in processing complex data [19].

Although deep learning-based facial expression recognition algorithms have been widely applied, in-depth research on applying these algorithms to smart education in real and complex scenarios remains lacking. When using these deep learning algorithms for feature learning from images or videos, occlusions, lighting conditions, and other factors may affect students' facial expressions. Thus, obtaining reliable facial expression features becomes challenging. Furthermore, while existing expression networks can achieve high recognition accuracy on expression datasets, their large architectures are not readily deployable in practical applications, necessitating the lightweight of facial expression recognition networks.

This study attempts to address the issues above and improve the feature representation capability, predictive performance, and interpretability of deep neural networks by proposing an improved deep neural network model for student facial expression recognition in classroom teaching, utilizing a hybrid attention mechanism. The proposed model aims to achieve facial expression recognition in complex classroom scenarios by effectively learning the facial expression features of students. We have enhanced the convolutional block attention module (CBAM) and introduced a novel attention module that allows the network to focus on important features, suppress irrelevant features, and learn critical feature information. The performance metrics of the model are improved by incorporating this module into the CNN architecture. The effectiveness of the proposed model has been validated through experiments conducted on the RAF-DB facial expression dataset, demonstrating favorable results. When applied in classroom scenarios, compared with conventional CNNs, this method effectively classifies student expressions in complex classroom settings, improving the performance of the student facial expression recognition model in classroom environments.

The main contributions and innovations of this study are summarized as follows:

1) The CBAM attention module has been improved by proposing a novel attention module to enhance the feature extraction capability of the network model. The improved network model deepens the network depth. However, increasing the network model's depth may lead to gradient explosion during training, causing the model to fail to converge. Thus, a skip connection structure is introduced. The attention module's channel number is modified (with different dimension reduction ratios) by utilizing skip connections (shortcut connections) to module CBAM and incorporating the hyperbolic tangent activation function Tanh(); the attention module's channel number is modified (with different dimension reduction ratios). The modified module learns the weight information between various channels in different channel groups, allowing the attention module to learn the weight information effectively on feature channels and spatial regions.

2) A deep attention network model based on skip connections is designed to address the instability in recognizing students' facial expressions in classroom settings and the problems of parameter redundancy, long training time, slow convergence, and overfitting in traditional CNNs. The improved CBAM is integrated into DenseNet and connected to 3×3 convolutional blocks with larger receptive fields to enhance the performance of the network model and strengthen its feature extraction capability on image data, further optimizing the parameter updates of the model. Thus, the network can focus on important features, suppress unnecessary features, and learn crucial feature information, thereby improving the performance metrics of the model. Experimental results comparing the accuracy of different models on the dataset demonstrate that the proposed network model outperforms the baseline model, indicating a significant improvement in the performance of the network model.

3) A student classroom facial expression dataset is constructed to address the scarcity of real-world classroom facial expression datasets. This dataset was created by collecting 90 high-quality classroom teaching videos online. The images in the dataset have a predominant pixel resolution of $1920\times$ 1080, including 16,000 authentic student facial expression images captured during various classroom scenarios.

III. NETWORK FRAMEWORK

The main objective of this section is to detail how improved deep learning techniques, particularly improved CBAM, can be utilized to enhance the model's ability to recognize and interpret facial expressions in complex classroom scenarios. Through careful structural design and algorithmic optimization, our model not only excels in handling the diversity and subtlety of facial expressions, but also significantly improves its computational efficiency and interpretability.

A. ATTENTION MECHANISM

Attention mechanism serves as a primary approach to addressing the issue of information overload. It enables allocating limited computing resources to process more critical information, particularly in situations with limited computational capacity. Researchers have integrated attention mechanisms into various domains of artificial intelligence by rapidly selecting high-value information from a vast pool of data.

In the field of computer vision, attention mechanism has been widely applied in various directions. Different weights can be assigned to features and hidden layer representations in the network by introducing attention mechanism into deep learning networks. During training, these weights can be

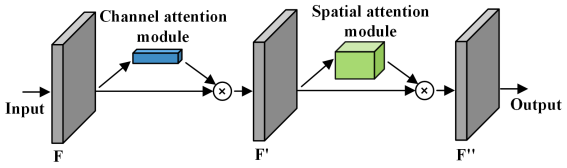

FIGURE 1. CBAM network structure.

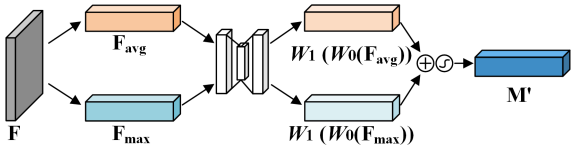

FIGURE 2. Channel attention module.

learned, providing an intuitive and accurate indication of the importance of each feature for the model's learning process. Consequently, attention mechanism effectively interprets the model's prediction performance.

In processing deep neural networks using lightweight attention modules, a novel channel attention network called Squeeze-and-Excitation Networks (SENet) was proposed in [20]. SENet implicitly and adaptively predicts crucial features, exhibiting excellent generalization capabilities on challenging datasets. It has been demonstrated to enhance the ResNet50 deep architecture's performance significantly. Another hybrid attention mechanism, CBAM, was presented in [21]. CBAM consists of two modules: the spatial attention module and the channel attention module. These modules infer the attention maps of the network from spatial and channel dimensions. The network framework of CBAM, as illustrated in Figure 1, comprehensively attends to the feature regions from channel and spatial perspectives.

The structure of the channel attention module is shown in Figure 2. Firstly, two different channel feature longs F_{avg} and F_{max} are generated by global average pooling and maximum pooling operations on the input feature F. Then, these two sets of pooled information are weighted and summed by two fully connected layers, and then this result is activated using the Sigmoid activation function. Finally, the channel attention feature map M' is obtained, which contains information about the degree of attention paid to different channels of the input feature F, and is computed as shown in Equation (1).

$$M'(F) = \sigma\left(W_1\left(W_0\left(F_{\text{avg}}\right)\right) + W_1\left(W_0\left(F_{\max}\right)\right)\right) \quad (1)$$

where σ represents the activation function Sigmoid, W_0 and W_1 are the two shared Multilayer Perceptron weights.

The structure of the spatial attention module is shown in Figure 3. First, the feature map F' undergoes maximum pooling and average pooling operations to capture its spatial information. Next, F' is downscaled to one channel by a 7×7 convolution operation for subsequent computation. Then, the spatial attention feature map M'' is generated using a Sigmoid function, computed as shown in equation (2). where F' is the

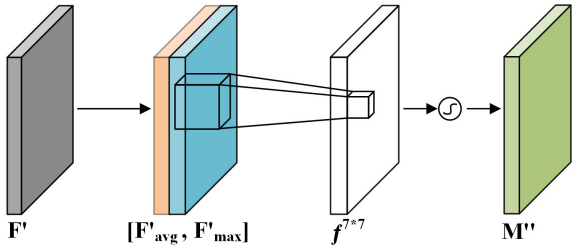

FIGURE 3. Spatial attention module.

product of the original input feature map F and the channel attention feature map M'. Similarly, F'' can be calculated to obtain F', which is calculated in the same manner as F'. By means of the spatial attention module, information about the degree of spatial attention to the input feature map F can be obtained and a corresponding feature map representation can be obtained.

$$M''\left(F'\right) = \sigma\left(f^{7\times7}\left(F'_{\text{avg}}; F'_{\max}\right)\right) \quad (2)$$

where σ represents the activation function Sigmoid, $f^{7\times7}$ denotes a convolution operation with 7×7 convolution kernel size.

Considering the impact of lightweight attention mechanism modules on deep network structures, these modules exhibit strong dependency and adaptability. The same module can yield different effects in different backbone network architectures. Furthermore, embedding the same module in different feature layers of the same backbone network can also influence weight learning in the model. Increasing the depth of the model network enhances its learning capacity but can potentially lead to gradient explosion. The introduction of attention mechanisms in deep learning networks aims to strengthen the representation of critical features.

This study addresses these considerations by deepening the attention module and optimizing the module structure through skip connections. This approach mitigates the risk of gradient explosion and accelerates model convergence. This study also investigates the effects of embedding attention modules into different feature layers of the deep learning backbone network on the predictive structure. Furthermore, this study explores the significant factors influencing predictive performance and aims to enhance the interpretability of the model.

B. IMPROVED DEEP NEURAL NETWORK MODEL WITH MIXED ATTENTION MECHANISM

The current research on facial expression recognition is a complex network designed to achieve a sufficient expression recognition rate, resulting in many network model parameters, redundancy, complexity, extensive training, challenging processing of results to achieve real-time requirements, and unstable face image expression recognition.

Considering that students' facial expressions in classrooms are often occluded, the loss of facial expression information

due to occlusion unavoidably affects the learning of facial expression features. Using a single CNN alone cannot automatically focus on the effective regions of student facial expressions and fails to capture valid facial expression information. This study addresses this issue by proposing an improved deep neural network model for student facial expression recognition in classroom teaching, incorporating a modified hybrid attention mechanism. Specifically, CBAM is enhanced by introducing a novel attention module that enables the network to focus on important features, suppress unnecessary features, and learn crucial feature information. Moreover, this module is integrated into the CNN, enhancing the model's performance metrics and better meeting the real-time facial expression recognition requirements in practical scenarios. The overall architecture of the proposed model is illustrated in Figure 4.Figure 4a illustrates the structure of the DenseNet121 model, while Figure 4b shows the structure of each Bottleneck in the DenseBlock in DenseNet121. Each Bottleneck consists of a 1×1 convolution and a 3×3 convolution serially connected. After this, we embed the improved attention mechanism module onto the 3×3 convolutional layer of Bottleneck in DenseNet121 dense block.

DenseNet [22] connects each layer with other layers in a feed-forwardly.The DenseNet network consists of dense blocks calculated as shown in Equation (1).

$$X_l = H_l\left([X_0, X_1, \ldots, X_{l-1}]\right) \tag{3}$$

TABLE 1. Introduction of DenseNet and introduction of SE, ECA, and CBAM modules to DenseNet comparison on the RAF-DB dataset.

Methods	Accuracy (%)
DenseNet	82.79
SE+DenseNet	83.07
ECA+DenseNet	83.84
CBAM+DenseNet	83.21

where [*] represents the splicing operation; $H_l([*])$ represents the nonlinear transformation of the first layer, such as batch normalization, convolution, or pooling; and X_i denotes the feature output of the layer. Compared with ResNet, DenseNet further optimizes the propagation of information flow. Table 1 shows that incorporating SE, ECA, and CBAM modules individually into DenseNet did not significantly improve performance compared with the original baseline model. This study proposes a novel attention mechanism module that addresses the limitations of the CBAM module to enhance the network's representation capacity and improve the model's predictive performance. The improved module is designed as a parallel structure with the convolutional modules in the backbone network. The input to the backbone network DenseNet is denoted as x, and the convolutional modules are represented as $F1(x)$ in the backbone network. On the right side, the attention module is denoted as $F2(x)$. The attention module consists of the modified Channel Attention Module (CAM) and Spatial Attention Module (SAM), as shown in Figure 4b.

The improved channel attention module consists of three serial structural blocks, mainly including convolution operation $\mathrm{Conv}_1(1 \times 1 \times C)$, batch normalization $\mathrm{BN}_1\left(1 \times 1 \times C'\right)$, rectified linear activation $\mathrm{ReLu}_1\left(1 \times 1 \times C'\right)$, $\mathrm{Conv}_2\left(1 \times 1 \times C'\right)$, batch normalization $\mathrm{BN}_2(1 \times 1 \times C)$, and hyperbolic tangent activation $\mathrm{Tanh}_2(1 \times 1 \times C)$. In the first serial structural block, the dimensionality reduction ratio in the middle is 16, the dimensionality reduction ratio in the second serial structure block is 8, and the dimensionality reduction ratio in the third serial structure block is 4, as shown in Figure 5. Compared with deepening the network width, two methods are utilized: one is to increase the network depth, and the other is to increase the network width; the cost of increasing the network width is often higher than the cost of increasing the depth [23]. Adding skip connections to the deep network can simplify the optimization problem of the deep network, simplify the model structure, reduce the parameters of the network model relatively, and avoid gradient explosion to a certain extent. This skip connection, this skip connection structure is used in the CAM module, and the output of the pooling layer and the output of the first serial structure block are used as the input of the second serial structure block. Therefore, the number of network layers of the attention module can be effectively increased, and more relationships between linear and nonlinear weights can be learned in channels and spaces. The final output of the improved CBAM module can be expressed as $F1(x) + F1(x) \times F2(x)$.

C. SELECTION OF BACKBONE NETWORK

Given the problem of model performance consistency and generalization in unknown data, datasets with different distributions do not perform consistently in the model and different models have different performances for the dataset [24]. As a feature extractor for expression images, the backbone network plays a decisive role in the classification effect. This study selects a suitable backbone network by choosing several common classical backbone networks, the latest backbone networks in the application domain, and alternative backbone networks, including GoogLeNet [9], Xception [11], ResNet [10], and DenseNet [22].

D. EMBEDDING OF ATTENTIONAL MECHANISMS IN THE BACKBONE NETWORK

The attention mechanism can strengthen the feature map of the network, focus on strengthening the contribution of crucial features in network learning, reduce the influence of non-key features in network learning, and then improve the feature representation ability of the network and reduce the noise information of the data, making the deep neural network. The critical features in the expression image data can be automatically obtained, and the network structure

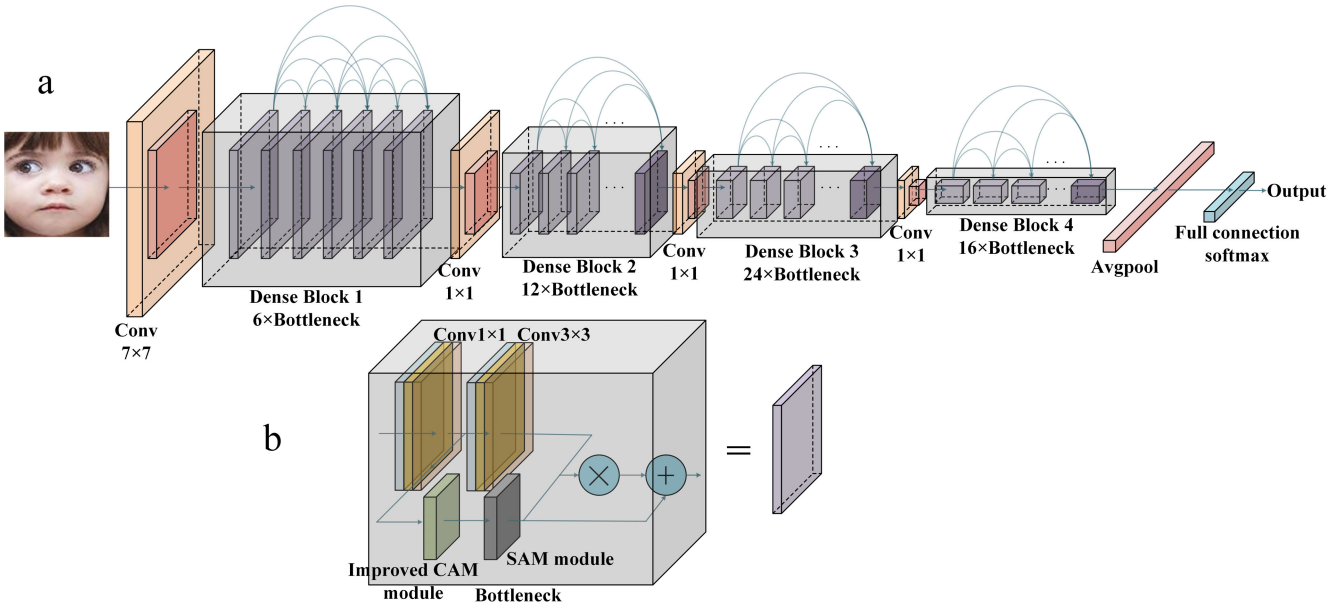

FIGURE 4. Structure of lightweight network model based on attention network.

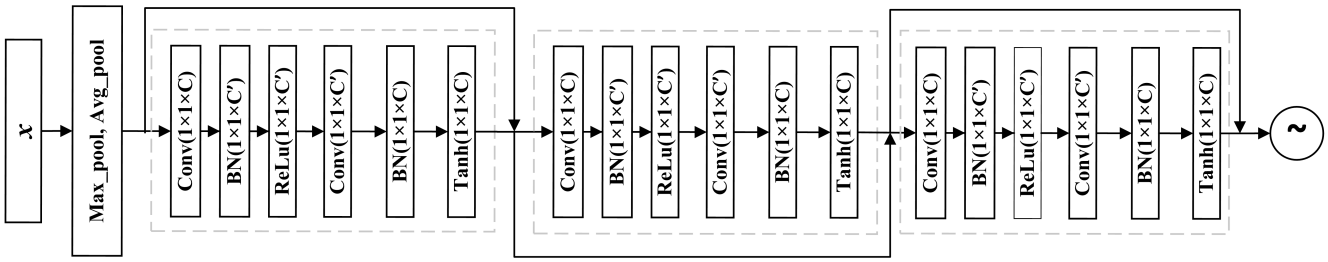

FIGURE 5. Structure of the improved CAM module.

is simplified simultaneously, speeding up the network training and improving the prediction performance of the network model. This study embeds the improved CBAM into the DenseNet backbone network, as shown in Figure 6. When CBAM is embedded into the backbone structure, a skip connection structure is also used. This parallel skip connection structure can ensure that new information can be learned from each module to the greatest extent. Given that large convolution kernel in CNN has a larger receptive field, more parameters need to be updated during the network learning process, and the model is more challenging to learn on this network layer, so the improved attention mechanism module is embedded in DenseNet121 on a 3×3 kernel layer in a dense block.

Recent studies have shown that, in deep networks, the hyperbolic tangent activation function (Tanh) is superior to the nonlinear activation function (ReLU), and the convergence speed is faster [25]. Therefore, the Tanh activation function is used to learn the nonlinear relationship between features in dimensionality reduction and dimensionality enhancement.

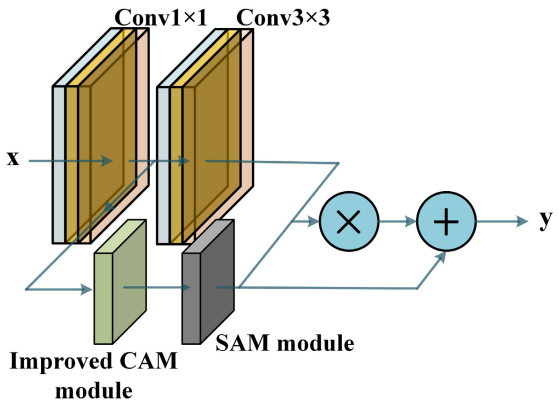

FIGURE 6. Embedding the improved CBAM into the DenseNet121 network.

The image of the Tanh function is a double tangent curve with the center point $(0, 0)$, and the function takes values between -1 and 1. The Tanh function converges faster, its

sensitive region is wider, and the model is easier to converge. The mathematical expression of the Tanh function is shown in Equation (2).

$$\mathrm{Tanh}(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{4}$$

In this study, the loss function of all backbone networks is Cross Entropy Loss, also known as Softmax loss, expressing the output of a multiclassification problem in the form of a probability, the maximum of which is the category to which the corresponding sample belongs. The operation process is shown in Equation (3).

$$L_{\mathrm{crossEntropy}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{h_{y_i}}}{\sum_{j=1}^{N} e^{h_j}}\right) \tag{5}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASETS

This study uses the public facial expression dataset Real-world Affective Faces Database (RAF-DB) [26], [27] and the self-built classroom teaching video dataset to verify the accuracy of the proposed expression recognition and intelligent teaching evaluation methods.

The RAF-DB dataset, containing 29672 irregular face images, has interference from factors such as occlusion, lighting, and interference from different hairstyles, postures, and accessories. The RAF-DB face library contains seven basic and twelve compound expressions to simulate emotion analysis in real-life environments. The dataset has been divided into a training set and a data set, in which the training set has 12271 images and 3,068 images, and the details of the data set are shown in Table 2.

TABLE 2. RAF-DB dataset details.

Dataset RAF-DB	Training data	Test data
Neutral	2524	680
Angry	705	162
Disgust	717	160
Happy	4772	1185
Sad	1982	478
Surprise	1290	329
Fear	281	74
Total	12271	3068

The self-built classroom teaching video dataset is a video encompassing students' expressions and students' behavioral movements, in which 16000 images of students' real expression states are collected from real classroom scenes on the Internet, with seven types of expression labels. The sample dataset is shown in Figure 7, which is mainly used to validate intelligent teaching assessment methods.

B. EXPERIMENTAL SETUP

After completing the pre-processing of the dataset, the dataset is randomly divided into a training set, a validation set and a test set in a 3:1:1 ratio in order to train the images to obtain the network model with optimal results. The training set is used to update the model parameters, the validation set is used to optimize the hyperparameters, and the test set is used for the performance evaluation of external data.

FIGURE 7. Sample example of a self-built classroom teaching dataset.

The optimal training weights on the ImageNet dataset are preloaded onto each backbone network to shorten the training process and enable the model to converge faster. The learning rate in the experiments is 1e-6, the learning rate decay rate is decayed by half every 50 rounds, and the maximum number of training rounds is 150. the batch size is 32. to prevent overfitting, the Dropout is set to 0.5.

The system used is Windows 10, under which the deep learning framework Pytorch is built for model training and Pycharm is used for code experiments. The computer hardware configuration includes an Intel(R) Core(TM) i7-7700K CPU, NVIDIA GeForce RTX 1080Ti GPU, and 32GB RAM.

C. EVALUATION INDEX

The accuracy rate indicates the proportion of correctly identified samples to the total samples, and the calculation is shown in Equation (4). Other performance indicators are added for comprehensive measurement.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where TP is the number of true positive samples,TN is the number of true negative samples,FP is the number of false positive samples, and FP is the number of false negative samples.

D. EXPERIMENTAL RESULTS

This section compares several alternative models experimentally on the RAF-DB dataset, as shown in Figure 8. The Xception model began to converge at approximately 120 iterations, and its classification accuracy was 73.38%. The GoogLeNet model began to converge at approximately 110 iterations, and its classification accuracy was 75.12%. The ResNet model began to converge at approximately 100 iterations, and its classification accuracy is 82.37%. The DenseNet model began to converge at about 90 rounds of iteration, and its classification accuracy rate is 83.76%. The DenseNet model integrated with the CBAM attention module starts to converge at about 70 rounds, and its classification accuracy rate is 84.21%. The improved attention module fusion DenseNet model starts to converge around round 65, and its classification accuracy is 88.71%. Compared with other models, the model proposed in this study has the best classification effect, the network model with skip connection structure is significantly better than other

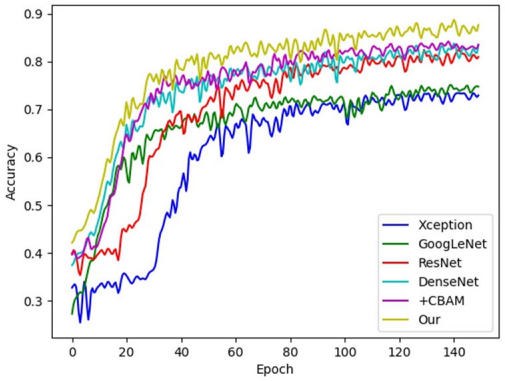

FIGURE 8. Comparison of validation accuracy of different models.

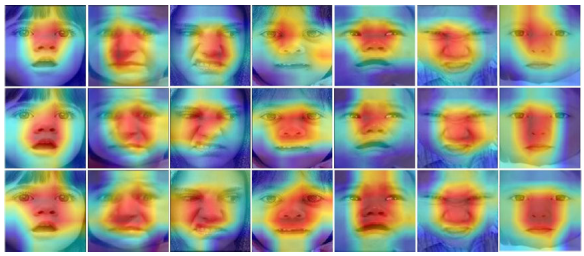

FIGURE 9. Baseline (DenseNet), DenseNet fused with CBAM module, and the proposed method Comparison of facial attention visualization between them. The redder the color in each image, the higher the attention score. Conversely, the greener the color, the smaller the attention score.

classic network models, and the model convergence speed is faster than the baseline CBAM fusion in the DenseNet network model. These experimental results show that the improved attention model can effectively simplify the model structure.

Table 3 compares adding CBAM and the improved attention module to the model and other classic models on the test set, including model accuracy, number of model layers, parameter amount and calculation amount. The experimental results show that in the horizontal comparison, the performance of the backbone network model with the attention mechanism module is generally better than without the attention mechanism module. In the longitudinal comparison, the improved attention module is introduced into the DenseNet backbone network, and the number of model parameters after training is slightly higher than that of the DenseNet model fused with the CBAM module, but its accuracy is higher. The comprehensive evaluation reveals that the DenseNet model incorporating the improved CBAM module has the highest accuracy rate, and the network parameters after adding the new attention mechanism are not much different from the lightweight network GoogLeNet and less than most network models. The structural complexity of the model is lower, and the model converges faster to a certain extent. The amount of calculation is also lower than Resnet and the Resnet network that incorporates the attention mechanism. The DenseNet model that incorporates the improved attention module appears to be better than other classic models.

Based on the data presented in Table 4, the model was comprehensively compared with other existing methods on the RAF-DB dataset. The model achieved an accuracy of 88.71%, demonstrating high accuracy. The experimental results indicate that the model outperformed other methods in the table by 0.35% in terms of accuracy on this dataset, further validating the robustness and reliability of our model, as well as its superior performance in facial expression recognition tasks.

The gradient-weighted class activation mapping (Grad-CAM) can provide visual explanations and generate localization heatmaps for deep networks. Figure 9 shows the localization heatmaps obtained after training the baseline DenseNet, the DenseNet model fused with the CBAM module, and the DenseNet model fused with the improved attention module. The first row represents the localization heatmap of the baseline DenseNet, the second row represents the localization heatmap of the DenseNet model fused with the CBAM module, and the third row represents the localization heatmap of the DenseNet model fused with the improved attention module. These heatmaps illustrate the contribution of pixel regions in the input expression images to the final prediction results during model training. In each image, the higher the attention score, and the redder the color, indicating greater attention. Conversely, the greener the color, the lower the attention score.

Our model's advantages are higher accuracy on the RAF-DB dataset, and excellent performance in practical applications. By taking full advantage of the powerful capabilities of deep learning and innovative algorithm design, our model can accurately capture the subtle but critical features of facial expressions, resulting in improved accuracy and robustness.

Experiments and analyses were conducted on a collected dataset of 16,000 real student facial expression images to test the performance and effectiveness of the proposed classroom teaching evaluation algorithm in real classroom scenarios. The proposed student facial expression recognition model based on the improved hybrid attention mechanism achieved an accuracy of 86.14%. The accuracy curve and confusion matrix of the model on the validation set of real student facial expression images in the classroom are shown in Figure 10. The model demonstrated good recognition performance for ''fear'' and ''happy,'' with accuracies reaching 92%. Even for the category with the lowest recognition performance, ''disgust,'' the model still achieved an accuracy of 81%. The performance gap may be attributed to the visual similarity between facial expression categories and the skewed class distribution in the training dataset. Therefore, additional efforts should be made to avoid class confusion, particularly for more challenging classes such as ''disgust'' and ''sad,'' in order to further improve the performance of our method in the future.

TABLE 3. Comparison of various performance indicators of different models.

Backbone network	Layers	Accuracy	Param (Mb)	FLOPs
Xception	36	0.7337	19.8	2.29G
GoogLeNet	27	0.7512	5.9	0.74G
ResNet	101	0.8237	42.5	7.34G
DenseNet	121	0.8279	6.9	2.89G
CBAM+Xception	36	0.7359	21.4	2.30G
CBAM+GoogLeNet	27	0.7711	8.9	0.74G
CBAM+ResNet	101	0.8434	47.2	7.35G
CBAM+DenseNet	121	0.8321	9.7	2.89G
Improved CBAM+Xception	36	0.7856	23.8	2.30G
Improved CBAM+GoogLeNet	27	0.8052	9.6	0.74G
Improved CBAM+ResNet	101	0.8612	48.5	7.35G
Improved CBAM+DenseNet	121	0.8871	13.5	2.89G

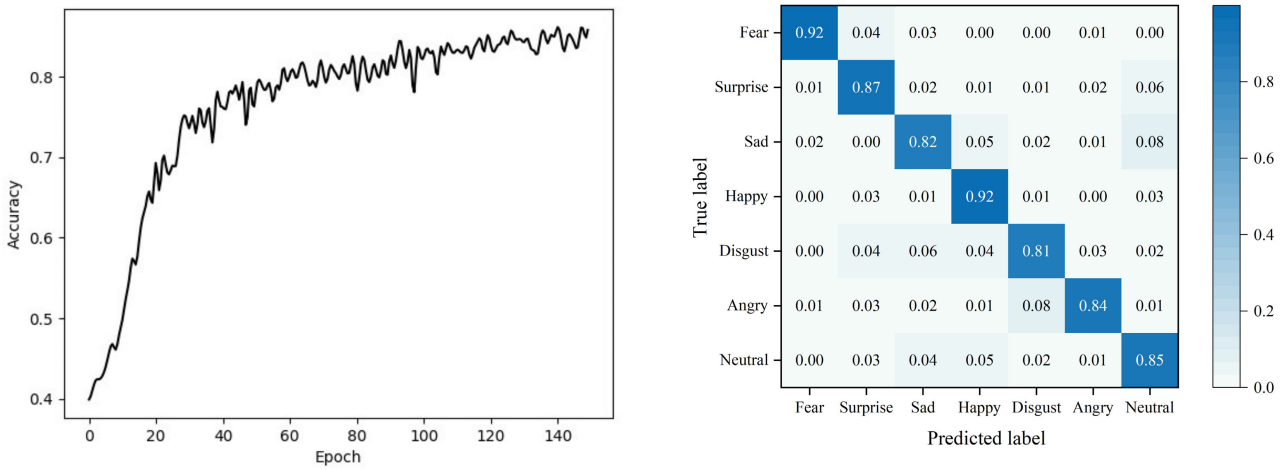

FIGURE 10. Accuracy curve and confusion matrix of the model in recognizing the validation set of classroom expression pictures.

TABLE 4. Performance comparison on the RAF-DB dataset.

Methods	Accuracy (%)
DDA-Loss [28]	86.90
DACL [29]	87.78
IF-GAN [30]	88.33
Eficientrace [31]	88.36
Our	88.71

E. DISCUSSION OF EXPERIMENTAL RESULTS

This section proposes an improved hybrid attention mechanism for student classroom facial expression recognition. Deep neural networks are prone to the issues of gradient explosion and vanishing gradients. An attention module is introduced into the deep network, specifically in the DenseNet architecture, to enhance the network's learning capability on key features. More accurate weight information can be learned by increasing the depth of the attention module. Furthermore, an attention module based on the improved CBAM is proposed, incorporating a skip connection structure. This attention module exhibits structural innovation over the CBAM, enhancing the network's focus on critical regions. The improved network model demonstrates faster convergence compared with other models, effectively simplifying the structure of the deep network model. The model avoids gradient explosion and strengthens its feature representation capability by deepening the attention module. This improvement increases the reliability of the model's classification. Experimental results show that the proposed model performs well on facial expression datasets. It achieves an accuracy of 88.71% on the publicly available RAF-DB dataset and reaches 86.14% accuracy on a self-collected real classroom teaching video dataset.

V. CONCLUSION

This study proposes an improved hybrid attention mechanism for student facial expression recognition in classroom teaching videos. The experimental results demonstrate that the model exhibits high accuracy, robustness, and generalization ability in classroom facial expression recognition, making it suitable for various applications in classroom teaching assessment and beyond. Future research will further integrate this model with other factors of classroom assessment for intelligent classroom evaluation. We have constructed a student classroom facial expression dataset for training and model evaluation to overcome the scarcity of real classroom facial expression datasets. Future research will expand the recognition scenarios to make the model a versatile

framework for classroom facial expression recognition. We will continue to refine the model to adapt to different classroom teaching scenarios, providing accurate and reliable facial expression recognition results as valuable tools and methods in the field of education. Moreover, we will explore additional application domains such as sentiment analysis and learner behavior understanding, further advancing the development of classroom facial expression recognition technology and enabling its broader applications in education.

REFERENCES

[1] L. Yao, S. He, K. Su, and Q. Shao, "Facial expression recognition based on spatial and channel attention mechanisms," *Wireless Pers. Commun.*, vol. 125, no. 2, pp. 1483–1500, Jul. 2022.
[2] H. Ling, J. Wu, J. Huang, J. Chen, and P. Li, "Attention-based convolutional neural network for deep face recognition," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 5595–5616, Mar. 2020.
[3] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[4] P. D. M. Fernandez, F. A. G. Peña, T. I. Ren, and A. Cunha, "FERAtt: Facial expression recognition with attention net," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 837–846.
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[6] Y. Tang, "Deep learning using linear support vector machines," 2013, *arXiv:1306.0239*.
[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[8] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 1–12.
[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[12] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Tech. Rev.*, vol. 32, no. 5, pp. 347–355, 2015.
[13] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.
[14] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, Apr. 2021.
[15] P. Raj Dachapally, "Facial emotion detection using convolutional neural networks and representational autoencoder units," 2017, *arXiv:1706.01509*.
[16] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *Artificial Neural Networks and Machine Learning—ICANN*. Rhodes, Greece: Springer, 2018, pp. 84–94.
[17] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[18] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-MobileNet: An approach of facial expression recognition," *Alexandria Eng. J.*, vol. 61, no. 6, pp. 4435–4444, Jun. 2022.
[19] P. Barra, L. De Maio, and S. Barra, "Emotion recognition by web-shaped model," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11321–11336, Mar. 2023.
[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[23] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proc. Conf. Learn. Theory*, 2016, pp. 907–940.
[24] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger, and N. Jacobs, "Inconsistent performance of deep learning models on mammogram classification," *J. Amer. College Radiol.*, vol. 17, no. 6, pp. 796–803, Jun. 2020.
[25] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Sci.*, vol. 6, no. 12, pp. 310–316, 2017.
[26] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.
[27] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 884–906, Jun. 2019.
[28] A. H. Farzaneh and X. Qi, "Discriminant distribution-agnostic loss for facial expression recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1631–1639.
[29] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2401–2410.
[30] J. Cai, Z. Meng, A. S. Khan, J. O'Reilly, Z. Li, S. Han, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1344–1348.
[31] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3510–3519.

LI LI was born in Hubei, China, in 1997. He received the bachelor's degree in computer science and technology from the College of Computer Engineering, Jingchu University of Technology, in 2019. He is currently pursuing the master's degree with the Robotics Academy, Beijing Union University. His research interest includes computer vision.

DENGFENG YAO was born in 1979. He received the master's degree from Peking University and the Ph.D. degree from Tsinghua University. He is currently a Professor and a Doctoral Supervisor with the Beijing Key Laboratory of Information Service Engineering, Beijing Union University. His research interests include language recognition and computing, and information accessibility.